

Application of QSARs for the design of PET and SPECT imaging agents

Thesis submitted

by

PRIYANKA DE

Doctor of Philosophy (Pharmacy)

Department of Pharmaceutical Technology
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata 700032

2022

Dedicated to Science

**JADAVPUR UNIVERSITY
KOLKATA – 700 032, INDIA**

1. Title of the thesis:

Application of QSARs for the design of PET and SPECT imaging agents

2. Name, Designation & Institution of the Supervisors:

Dr. Kunal Roy

Professor

Drug Theoretics and Cheminformatics Laboratory

Department of Pharmaceutical Technology

Jadavpur University

Kolkata 700 032

West Bengal, India

Dr. Dhananjay Bhattacharyya

Retired Professor

Computational Science Division

Saha Institute of Nuclear Physics

Kolkata 700064

West Bengal, India

3. List of publications:

Papers related to dissertation

Research Articles

8. **De, Priyanka**, and Kunal Roy. "Computational modeling of PET imaging agents for vesicular acetylcholine transporter (VACHT) protein binding affinity: application of 2D-QSAR modeling and molecular docking techniques." *In Silico Pharmacology* 11, no. 1 (2023): 1-14.

7. **De, Priyanka**, and Kunal Roy. "Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness." *European Journal of Medicinal Chemistry Reports* 4 (2022): 100035.

6. **De, Priyanka**, and Kunal Roy. "QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: application of small dataset modeling." *Structural Chemistry* 32, no. 2 (2021): 631-642.

5. **De, Priyanka**, and Kunal Roy. "QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting dopamine receptor." *Theoretical Chemistry Accounts* 139, no. 12 (2020): 1-12.

4. **De, Priyanka**, Joyita Roy, Dhananjay Bhattacharyya, and Kunal Roy. "Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: a QSAR approach." *Structural Chemistry* 31, no. 5 (2020): 1969-1981.

3. **De, Priyanka**, Dhananjay Bhattacharyya, and Kunal Roy. "Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling." *Structural Chemistry* 31, no. 3 (2020): 1043-1055.

2. **De, Priyanka**, Dhananjay Bhattacharyya, and Kunal Roy. "Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease." *Structural Chemistry* 30, no. 6 (2019): 2429-24

1. **De, Priyanka**, Supratik Kar, Pravin Ambure, and Kunal Roy. "Prediction reliability of QSAR models: an overview of various validation tools." *Archives of Toxicology* (2022): 1-17.

Papers not related to dissertation

12. **De, P.**, Kumar, V., Kar, S., Roy, K., & Leszczynski, J. (2022). Repurposing FDA approved drugs as possible anti-SARS-CoV-2 medications using ligand-based computational approaches: sum of ranking difference-based model selection. *Structural Chemistry*, 1-13.

11. Banerjee, A., Chatterjee, M., **De, P.**, & Roy, K. (2022). Quantitative Predictions from Chemical Read-Across and Their Confidence Measures. *Chemometrics and Intelligent Laboratory*, 227.

10. Kumar, V., Kar, S., **De, P.**, Roy, K., & Leszczynski, J. (2022). Identification of potential antivirals against 3CLpro enzyme for the treatment of SARS-CoV-2: A multi-step virtual screening study. *SAR and QSAR in Environmental Research*, 33(5), 357-386.

9. Banerjee, A., **De, P.**, Kumar, V., Kar, S., & Roy, K. (2022). Quick and Efficient Quantitative Predictions of Androgen Receptor Binding Affinity for Screening Endocrine Disruptor Chemicals Using 2D-QSAR and Chemical Read-Across. *Chemosphere*, 309.

8. **De, P.**, Bhayye, S., Kumar, V., & Roy, K. (2022). In silico modeling for quick prediction of inhibitory activity against 3CLpro enzyme in SARS CoV diseases. *Journal of Biomolecular Structure and Dynamics*, 40(3), 1010-1036.

7. Nath, A., **De, P.**, & Roy, K. (2022). QSAR modelling of inhalation toxicity of diverse volatile organic molecules using no observed adverse effect concentration (NOAEC) as the endpoint. *Chemosphere*, 287, 131954.

6. Chatterjee, M., Banerjee, A., **De, P.**, Gajewicz-Skretna, A., & Roy, K. (2022). A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environmental Science: Nano*, 9(1), 189-203.

5. Nath, A., **De, P.**, & Roy, K. (2021). In silico modelling of acute toxicity of 1, 2, 4-triazole antifungal agents towards zebrafish (*Danio rerio*) embryos: Application of the Small Dataset Modeller tool. *Toxicology in Vitro*, 75, 105205.

4. Kumar, V., **De, P.**, Ojha, P. K., Saha, A., & Roy, K. (2020). A multi-layered variable selection strategy for QSAR modeling of butyrylcholinesterase inhibitors. *Current Topics in Medicinal Chemistry*, 20(18), 1601-1627.

3. **De, P.**, & Roy, K. (2018). Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. *SAR and QSAR in Environmental Research*, 29(4), 319-337.

2. **De, P.**, Aher, R. B., & Roy, K. (2018). Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector *Aedes aegypti*: application of ETA indices. *RSC advances*, 8(9), 4662-4670.

1. **De, P.**, Kar, S., Roy, K., & Leszczynski, J. (2018). Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environmental Science: Nano*, 5(11), 2742-2760.

Book chapters not related to dissertation

1. **De, P.,** & Roy, K. (2021). Computational modeling of ACE2-mediated cell entry inhibitors for the development of drugs against coronaviruses. In *In Silico Modeling of Drugs Against Coronaviruses* (pp. 495-539). Humana, New York, NY.

4. List of patents: **Nil**

5. List of Presentation in National/ International:

International conferences

4. **Priyanka De**, Kunal Roy*, (2022) *Computational PBPK modeling of adipose/blood and blood/air partition coefficients: ranking and prioritization of Interbioscreen database compounds using QSPR, QSPPR and read-across approaches*. 23rd EuroQSAR 2022, September 26-30, 2022, Heidelberg, Germany.

3. **Priyanka De**, Kunal Roy* (2021) *A 2D-QSAR approach to explore structural features of nitroaromatics as hypoxic cell radiosensitizers*. International Symposium on Drug Design and Development Research (DDDR- 2021), December 17-18, 2021, GIPER, Kashipur, India. **(Received second prize in Poster Presentation)**

2. **Priyanka De**, Kunal Roy* and Dhananjay Bhattacharyya (2020), *Exploration of nitroimidazoles as radiosensitizers: Application of multilayered feature selection approach in QSAR modeling*. International Conference on Currents Trends in Pharmaceutical and Medical Sciences, February 26-29, 2020, GIPER, Kashipur, India. **(Received third prize in Poster Presentation)**

1. **Priyanka De**, Kunal Roy* and Dhananjay Bhattacharyya (2019) *Exploring molecular features of PET and SPECT imaging agents against amyloid beta plaques: A QSAR and molecular docking approach*. 43rd Indian Biophysical Society Meeting Molecules to Systems, March 15-17, 2019, Indian Biophysical Society Meeting at IISER Kolkata.

National conferences

5. **Priyanka De**. Attended webinar on *8th Indo-US Workshop on Mathematical Chemistry 2022* organised by Sharda University held on September 13-17, 2022.

4. **Priyanka De**. Attended webinar on *Gaining acceptance in next generation PBK modelling approaches for regulatory assessments: Case studies*, OECD Webinar, Organized by European Society of Toxicology In Vitro (ESTIV) held on May 10, 2022.

3. **Priyanka De**. Attended webinar on *Health Informatics Summit* organized by the Department of Computational Biology, Indraprastha Institute of Information Technology Delhi (IIIT-D) and APbians (Bioinformatics Society) held on October 16-19, 2021.

2. **Priyanka De**. Attended webinar on *"Biodiversity and Environmental Protection during Pandemic Outbreak of COVID 19"* organized by International Society of Waste Management Air and Water (ISWMAW) jointly by Chairman Prof. Dr. Sadhan Kumar Ghosh of India and Co-Chairman Dr. Abas Basir of Sri Lanka on 5th June 2020.

1. **Priyanka De** Attended a One-week course on *"Chemometric data processing for express analytical systems in agricultural and environmental studies"* conducted by Prof. D. Kirsanov of St. Petersburg State University, Russia, on 2-7 December 2019, at Dept. of Instrumentation and Electronics Engineering, Jadavpur University.

Workshops

1. **Priyanka De**. Delivered an expert talk on *"QSAR Practical: Hands-on Session on the DTC Lab Tools"* in SERB sponsored workshop on *"Hands-on training on Computer Aided Drug Design and Discovery Tools"* organised by Ganpat University on 11-17 July, 2022.

“Statement of Originality”

I *Priyanka De* registered-on *24th June 2019* do hereby declare that this thesis entitled “*Application of QSARs for the design of PET and SPECT imaging agents*” contains a literature survey and original research work done by the undersigned candidate as part of Doctoral studies. All information in this thesis has been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work. I also declare that I have checked this thesis as per the “Policy on Anti Plagiarism, Jadavpur University, 2019”, and the level of similarity as checked by iThenticate software is **10%***.

Signature of Candidate:

Date:

Certified by Supervisor(s):
(Signature with date, seal)

1.

2.

*Ignoring smaller match of up to 14 words (as per the UGC) and ignoring own publications as the source.

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled “**Application of QSARs for the design of PET and SPECT imaging agents**” submitted by **Smt. Priyanka De**, who got her name registered on 24th June 2018 for the award of Ph.D. (Pharmacy) degree of Jadavpur university is absolutely based upon her own work under the supervision of **Dr. Kunal Roy** and co-supervision of **Dr. Dhananjay Bhattacharyya** and that neither her thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

1. _____
(Signature with Date and Seal)

2. _____
(Signature with Date and Seal)

Acknowledgements

I deem it a pleasure and privilege to work under the guidance of **Dr. Kunal Roy**, Professor, Drug Theoretics & Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata and co-guidance of **Dr. Dhananjay Bhattacharyya**, Retired Professor, Computational Science Division, Saha Institute of Nuclear Physics, Kolkata. I express my deep gratitude and regard to my revered mentors for suggesting the subject of this thesis and rendering me his thoughtful suggestions and rational approaches to this thesis work. I am greatly indebted to Dr. Kunal Roy and Dr. Dhananjay Bhattacharyya for their valuable guidance throughout the work that enabled me to complete the work.

I am thankful to the authorities of Jadavpur University and Saha Institute of Nuclear Physics for providing all the facilities needed for my work. I express deep gratitude to the Head of the Department, Vice-Chancellor, Registrar, Dean, Faculty of Engineering and Technology, Jadavpur University for facilities provided to carry out this work. I also thank the Department of Atomic Energy- Board of Research in Nuclear Sciences (DAE-BRNS) and Indian Council of Medical Research (ICMR), New Delhi for providing the financial support to carry out this work.

With a deep sense of thankfulness and sincerity, I acknowledge the continuous encouragement, perpetual assistance and co-operation from my seniors Probir Kr. Ojha, Pravin Ambure, Rahul Aher, Supratik Kar, Khan Kabiruddin, Vinay Kumar, Pathan Mohsin Khan and Gopala Krishna Jillela. Their constant support and helpful suggestions have tended me to accomplish this work in time. I would like to express my special thanks to my friend Mainak Chatterjee and my juniors Joyita Roy, Arnab Seth, Sapna Kumari Pandey, Aniket Nath, Arkaprava Banerjee, Rahul Paul, Ankur Kumar, Souvik Pore and Shilpayan Ghosh who all have extended their helping hands and friendly cooperation all through my work.

A word of thanks to all those people associated with this work directly or indirectly whose names I have been unable to mention here. Finally, I would like to thank my parents **Mr. Susanta Kumar De** and **Mrs. Kumkum Dey** for all the love and inspirations throughout the entire course of my work. I would like to specially acknowledge my husband **Anirban Roy**, for his constant support and encouragement without which this work would have been incomplete.

PRIYANKA DE

Date:

Place:

Preface

The presented work in this dissertation took nearly a span of four years. The dissertation is based on several *in silico* techniques were employed to study potential PET and SPECT imaging agents targeted against various neurodegenerative diseases and cancer. The main purpose of the dissertation was to utilize various *in silico* tools for identifying and optimizing the potential PET or SPECT candidates against several receptors involved in neurodegenerative diseases or cancer pathogenesis. The main advantage of these *in vivo* molecular imaging is its ability to characterize diseased tissues without invasive biopsies or surgical procedures, and with this information in hand, a more personalized treatment planning regimen can be applied. These imaging has also been used in various aspects of drug development such as understanding drug action and establishing dosage regimens and treatment strategies. In basic terms, PET and SPECT imaging effectively allows the non-invasive visualisation, characterisation and measurement of biological processes at the molecular, cellular, whole organ or body level using specific radionuclide probes.

Observing the scarcity of imaging data, scientific researchers have come up with alternative methods like Computer-aided drug design (CADD). CADD has been extensively explored for facilitating lead discovery and optimization with advantages in terms of both high speed and low cost that finally increases the probability of success in the drug development process. A variety of *in silico* methods have evolved in CADD that have two major application areas, i.e., ligand-based drug design (LBDD) and structure-based drug design (SBDD). In the present study, we have employed both ligand-based (i.e., QSAR and read-across) and structure-based (i.e., molecular docking) drug design techniques as together they become a powerful tool to study potential imaging agents. Further, although several *in silico* techniques were employed, but the major part of the work deals with the development of predictive and statistically robust QSAR models. The QSAR technique plays a vital role in lead optimization step in any drug discovery program, which is significantly utilized to save time, money, and more importantly animal sacrifice. The basic steps involved in developing predictive QSAR models comprise dataset collection, data curation, descriptor calculation, data pre-treatment of calculated descriptors, model development employing various chemometric techniques, model validation, and applicability domain determination. In the present work, QSAR technique was proficiently utilized in understanding the structural features that are favorable for the activity as well as to attain desirable selectivity. Further, the developed QSAR model provided valuable information to design new molecules with improved activity and it is also utilized for predicting the activity of a query or newly designed molecule. Many of the software tools used for model generation and validation are freely available to download from http://teqip.jdvu.ac.in/QSAR_Tools/ and <http://dtclab.webs.com/software-tools>.

The following studies have been performed in this dissertation:

Study 1: Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease

Study 2: Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: A QSAR approach

Study 3: QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting Dopamine receptor

Study 4: Computational modeling of PET imaging agents against vesicular acetylcholine transporter (VACHT) protein binding affinity: Application of 2D-QSAR modeling and molecular docking techniques

Study 5: Exploration of nitroimidazoles as radiosensitizers: Application of multilayered feature selection approach in QSAR modeling

Study 6: QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: Application of Small Dataset Modelling

Study 7: Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness

The work has been presented in this dissertation under the following sections:

Chapter 1 : Introduction

Chapter 2 : Present work

Chapter 3 : Materials and methods

Chapter 4 : Results and discussions

Chapter 5 : Conclusion

References

Appendix : Reprints

The '**Introduction**' part provides the background information on molecular imaging including the history, principle, instrumentation and application of PET and SPECT imaging agents. It also includes detailed information on the computational techniques employed in this work and a brief survey on the *in silico* studies that were performed for the design of PET and SPECT imaging agents. The '**Present Work**' section describes the overall envisaged work of the dissertation. The detailed information related to the descriptors and the methodologies has been provided in the section '**Materials and Methods**', while the results have been thoroughly discussed in the '**Results and Discussions**' section. Finally, '**Conclusion**' has been incorporated followed by '**References**'. The studies thus performed have been published in different *refereed international journals* and also presented in different national and international conferences which have been included under the section '**Reprints**'. However, the work done and presented in this dissertation constitutes a small part of the broad spectrum of envisaged work. Considering the stipulated time limit, only some representative and relevant studies could be performed. Many other interesting aspects arising out of this work could have been investigated in a far more meaningful way, which might be planned in future.

Abbreviations

| | | | |
|-----------|--|--------|--|
| 2D QSAR | Two dimensional QSAR | MLR | Multiple linear regression |
| 3D QSAR | Three dimensional QSAR | MRI | Magnetic resonance imaging |
| 4D QSAR | Four dimensional QSAR | OECD | Organization for economic co-operation and development |
| 5D QSAR | Five dimensional QSAR | PCA | Principal component analysis |
| 6D QSAR | Six dimensional QSAR | PD | Parkinson's disease |
| 7D QSAR | Seven dimensional QSAR | PDB | Protein data bank |
| AD | Alzheimer's disease | PET | Positron emission tomography |
| AD | Applicability domain | PLS | Partial least squares |
| ADMET | Adsorption, distribution, | PRESS | Predicted residual sum of squares |
| ANN | Artificial neural network | PRI | Prediction Reliability Indicator |
| A β | Amyloid-beta | QSAAR | Quantitative structural activity-activity relationship |
| BLI | Bioluminescence imaging | QSAR | Quantitative structure-activity relationship |
| BNN | Bayesian neural network | QSPR | Quantitative structure-property relationship |
| BSS | Best subset selection | QSTR | Quantitative structure-toxicity relationship |
| CADD | Computer aided drug design | RA | Read-across |
| CCC | Concordance correlation coefficient | RDD | Rational drug design |
| CoMFA | Comparative molecular field analysis | REACH | Registration, evaluation, authorisation and restriction of chemicals |
| CoMSIA | Comparative molecular similarity indices | RMSD | Root mean square deviation |
| CT | Computed Tomography | RMSE | Root mean square error |
| DCV | Double cross validation | RMSEP | Root mean square error in prediction |
| DNA | Deoxyribonucleic acid | RSH | Rat striatal homogenate |
| ECHA | European Chemicals Agency | SBDD | Structured-based drug design |
| E-state | Electrotopological state | SD | Standard deviation |
| ETA | Extended topochemical atom | SDEP | Standard deviation of error of prediction |
| EU | European Union | SDF | Structure data format/file |
| FDG | Fluorodeoxyglucose | SEE | Standard error of estimate |
| FDOPA | Fluorodopa | SER | Sensitizer Enhancement Ratio |
| FEOBV | Fluoroethoxybenzovesamicol | SiRMS | Simplex representations of molecular structure |
| FET | Fluoroethyl-tyrosine | SMILES | Simplified Molecular Input Line Entry System |
| FLI | Fluorescence imaging | S-MLR | Stepwise- multiple linear regression |
| FN | False negative | SPECT | Single photon emission computed tomography |
| FP | False positive | SR | Survival Ratio |
| G/PLS | Genetic partial least squares | SVM | structure activity relationship |
| GA | Genetic algorithm | TN | Support vector machine |
| GFA | Genetic function approximation | TP | True negative |
| GQSAR | Group based quantitative | VACHT | True positive |
| HBA | Hydrogen bond acceptor | VIP | Vesicular acetylcholine transporter |
| HBD | Hydrogen bond donor | VS | Variable importance plot |
| HCC | Hepatocellular carcinoma | WHIM | Virtual screening |
| HTS | High throughput screening | | Weighted holistic invariant molecular descriptor |
| ICP | Intelligent Consensus Prediction | | |
| KNIME | Konstanz Information Miner | | |
| kNN | k-nearest neighbor | | |
| LBDD | Ligand-based drug design | | |
| LDA | Linear discriminant analysis | | |
| LFER | Linear free energy relationships | | |
| LMO | Leave-many-out | | |
| LOO | Leave-one-out | | |
| LR | Linear Regression | | |
| MAE | Mean absolute error | | |
| MI | metabolism, excretion and toxicity | | |
| | Molecular imaging | | |

Contents:

| | |
|---|-----------|
| <i>Acknowledgements</i> | i |
| <i>Preface</i> | ii |
| <i>Abbreviations</i> | iv |
| Chapter 1: Introduction | 1 |
| 1.1. Molecular Imaging: A new age disease detection technology | 1 |
| 1.1.1. History of molecular imaging | 1 |
| 1.1.2. Molecular imaging technology | 1 |
| 1.1.3. Radionuclide imaging technology: PET and SPECT imaging | 2 |
| 1.2. Positron Emission Tomography | 4 |
| 1.2.1. PET Principle | 4 |
| 1.2.2. PET radionuclides and their clinical applications | 5 |
| 1.3. Single Photon Emission Computed Tomography | 13 |
| 1.3.1. SPECT Principle | 13 |
| 1.3.2. SPECT radiopharmaceuticals | 14 |
| 1.3.3. Application of SPECT imaging | 14 |
| 1.4. Difference between PET and SPECT imaging | 15 |
| 1.5. Nitroaromatics as Radiosensitizers | 16 |
| 1.6. Quantitative structure-activity relationship (QSAR) analysis | 16 |
| 1.6.1. The formalism | 16 |
| 1.6.2. History of QSAR | 17 |
| 1.6.3. Objectives of QSAR analysis | 19 |
| 1.6.4. Classification of QSAR | 20 |
| 1.6.5. QSAR and OECD Guidelines | 21 |
| 1.6.6. QSAR methodology | 23 |
| 1.6.7. Quantitative Structure Activity-Activity Relationship (QSAAR) modeling | 26 |
| 1.7. Non-QSAR in silico techniques | 30 |
| 1.7.1. Molecular Docking | 30 |
| 1.7.2. Virtual Screening (VS) | 30 |
| 1.7.3. Read-Across | 30 |
| 1.8. Application of QSARs for the design of PET and SPECT imaging agents | 31 |
| Chapter 2: Present Work | 33 |
| 2.1. Datasets employed for the development of different QSAR models | 35 |
| 2.1.1. Dataset I A-C (Study 1) | 36 |
| 2.1.2. Dataset II (Study 2) | 36 |
| 2.1.3. Dataset III (Study 3) | 37 |

| | | |
|---|--|-----------|
| 2.1.4. | Dataset IV (Study 4) | 37 |
| 2.1.5. | Dataset V (Study 5) | 37 |
| 2.1.6. | Dataset VI (Study 6) | 38 |
| 2.1.7. | Dataset VII A-C (Study 7) | 38 |
| Chapter 3: Materials and Methods | | 39 |
| 3.1. | Study 1: Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease | 39 |
| 3.1.1. | The dataset and structure curation | 39 |
| 3.1.2. | Molecular descriptors | 43 |
| 3.1.3. | Dataset splitting | 43 |
| 3.1.4. | Model development | 43 |
| 3.1.5. | Statistical validation metrics | 44 |
| 3.1.6. | Molecular Docking | 44 |
| 3.2. | Study 2: Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: A QSAR approach | 45 |
| 3.2.1. | The dataset | 45 |
| 3.2.2. | Molecular descriptors | 46 |
| 3.2.3. | Dataset Division | 47 |
| 3.2.4. | Variable selection and Model Development | 47 |
| 3.2.5. | Statistical validation metrics | 47 |
| 3.2.6. | Applicability domain (AD) | 47 |
| 3.2.7. | Molecular Docking | 48 |
| 3.3. | Study 3: QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting Dopamine receptor | 48 |
| 3.3.1. | The dataset | 48 |
| 3.3.2. | Molecular descriptors | 50 |
| 3.3.3. | Dataset splitting | 50 |
| 3.3.4. | Variable selection and model development | 50 |
| 3.3.5. | Statistical validation metrics | 51 |
| 3.4. | Study 4: Computational modeling of PET imaging agents against vesicular acetylcholine transporter (VACHT) protein binding affinity: Application of 2D-QSAR modeling and molecular docking techniques | 52 |
| 3.4.1. | The dataset | 52 |
| 3.4.2. | Molecular descriptors | 56 |
| 3.4.3. | Feature selection and model development | 57 |
| 3.4.4. | Machine learning based read across prediction | 57 |
| 3.4.5. | Molecular Docking | 57 |

| | | |
|-----------|--|-----------|
| 3.5. | Study 5: Exploration of nitroimidazoles as radiosensitizers: Application of multilayered feature selection approach in QSAR modeling | 58 |
| 3.5.1. | Descriptor calculation | 60 |
| 3.5.2. | Dataset splitting | 61 |
| 3.5.3. | Variable selection and QSAR model development | 61 |
| 3.5.4. | Statistical validation metrics and domain of applicability | 61 |
| 3.6. | Study 6: QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: Application of Small Dataset Modelling | 62 |
| 3.6.1. | Dataset | 62 |
| 3.6.2. | Molecular descriptors | 63 |
| 3.6.3. | Model development: Application of Small Dataset Modeler | 63 |
| 3.6.4. | Statistical validation metrics | 64 |
| 3.7. | Study 7: Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness | 65 |
| 3.7.1. | The dataset | 65 |
| 3.7.2. | Descriptor calculation | 69 |
| 3.7.3. | Data set splitting and model development | 69 |
| 3.7.4. | Statistical validation metrics | 70 |
| 4. | Results and Discussions | 72 |
| 4.1. | Study 1: Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease | 72 |
| 4.1.1. | Descriptor Interpretation from QSAR models | 72 |
| 4.1.2. | Interpretation of PLS plots | 79 |
| 4.1.3. | Molecular docking | 87 |
| 4.1.4. | QSAR modeling and molecular docking studies for newly designed PET and SPECT imaging agents | 94 |
| 4.2. | Study 2: Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: A QSAR approach | 102 |
| 4.2.1. | Modeling binding affinity of PET tracers towards Adenosine (A _{2A}) receptor | 102 |
| 4.2.2. | Modeling selectivity of PET tracers towards Adenosine (A _{2A}) receptor | 108 |
| 4.2.3. | Intelligent Consensus Predictions | 112 |
| 4.2.4. | Applicability Domain (AD) | 114 |
| 4.2.5. | Comparison with a previously published model | 115 |
| 4.3. | Study 3: QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting Dopamine receptor | 116 |
| 4.3.1. | Modeling binding affinity of PET tracers towards dopamine (D ₂) receptor | 116 |
| 4.3.2. | Mechanistic interpretation | 116 |
| 4.3.3. | Plot Interpretation | 119 |

| | | |
|--------|--|-----|
| 4.4. | Study 4: Computational modeling of PET imaging agents against vesicular acetylcholine transporter (VACHT) protein binding affinity: Application of 2D-QSAR modeling and molecular docking techniques | 123 |
| 4.4.1. | QSAR modeling of binding affinity of PET imaging agents towards VACHT | 123 |
| 4.4.2. | Read-Across based prediction | 126 |
| 4.4.3. | Molecular Docking | 126 |
| 4.5. | Study 5: Exploration of nitroimidazoles as radiosensitizers: Application of multi-layered feature selection approach in QSAR modeling | 129 |
| 4.5.1. | 2D-QSAR model using Dragon descriptors | 129 |
| 4.5.2. | 2D-QSAR model using SiRMS descriptors | 131 |
| 4.5.3. | Applicability Domain Assessment | 134 |
| 4.5.4. | Y-randomization | 136 |
| 4.5.5. | True External Predictions | 136 |
| 4.5.6. | Comparison with the previously published research | 138 |
| 4.6. | Study 6: QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: Application of Small Dataset Modelling | 140 |
| 4.6.1. | Model 1: Modeling Drug Sensitizer Enhancement Ratio (SER) | 141 |
| 4.6.2. | Model 2: Modeling Drug Survival Ratio (logSR) | 141 |
| 4.6.3. | Quantitative Structure Activity-Activity Relationship (QSAAR) models | 142 |
| 4.6.4. | Plot interpretation | 146 |
| 4.6.5. | Applicability Domain assessment | 148 |
| 4.6.6. | Prediction dataset | 150 |
| 4.7. | Study 7: Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness | 152 |
| 4.7.1. | Modeling local nitro datasets | 152 |
| 4.7.2. | Modeling the global nitroaromatics dataset | 156 |
| 4.7.3. | Applicability Domain (AD) assessment | 158 |
| 4.7.4. | Y-randomization test | 160 |
| 4.7.5. | True External Prediction using the global model | 162 |
| | Chapter 5: Conclusions | 173 |
| | References | 177 |
| | Appendix: Reprints | |

CHAPTER 1

INTRODUCTION

Chapter 1: Introduction

1.1. *Molecular Imaging: A new age disease detection technology*

Molecular imaging (MI) is an emerging biomedical research field that integrates cell biology, molecular biology, and diagnostic imaging permitting the visualization, classification, and analysis of biological activities occurring at the cellular and subcellular levels within intact living subjects (Weissleder & Pittet, 2008). This technology allows early disease detection identifying the degree of the disease, choosing disease- and/or patient-specific therapeutic treatment also known as personalized medicine, applying a targeted therapy, and estimating receptor-specific effects of treatment. The purpose of molecular imaging is to link the imaging signal with the molecular events using high-resolution and high-sensitive instruments (Luo et al., 2011). Modern clinical scientists apply molecular imaging technology in studying the basis/cause of the disease from the molecular abnormalities found in the cells. This method of analysis, on the other way, accelerates other important clinical goals of: a) early disease detection b) therapy optimisation for important molecular targets c) forecasting and monitoring response to therapy, and d) disease recurrence monitoring. Radionuclide molecular imaging is one of the earliest and most mature methods of imaging technique which is efficient in detecting any harboring infection. Positron emission tomography (PET) and single photon emission computed tomography (SPECT) imaging were the first molecular imaging modalities used clinically and are used in these generations also due to their advantages of non-invasive localisation, high sensitivity and quantifiability (Anderson & Ferdani, 2009).

1.1.1. *History of molecular imaging*

George Charles de Hevesy, a Hungarian radiochemist, 1920, coined the term *radiotracer* or *radio indicator* and familiarized the tracer principle in the biomedical research field. A true tracer molecule can facilitate the study of homeostatic system and its components without upsetting its function. Later in the late 1920s, two physicians, Blumgart and Weiss, injected solutions of radium-C (^{214}Bi) into the veins of healthy persons and patients with heart disease to study the velocity of the blood. Owing to their revolutionary invention, Hevesy is known as the father of nuclear medicine, while Blumgart is regarded as the father of diagnostic nuclear medicine. Irene Curie and her husband Frederic Joliot's discovery of artificial radioactivity in 1930s and cyclotron discovery by Ernest Lawrence, paved the way for chemists to design radiotracers for the study of specific biochemical processes.

1.1.2. *Molecular imaging technology*

Molecular imaging technology can be subcategorized into different types, viz., a) magnetic resonance imaging (MRI), b) X-ray computed tomography imaging, c) optical imaging (including bioluminescence imaging (BLI) and fluorescence imaging (FLI)), d) radionuclide imaging (involving PET and SPECT) e) ultrasound imaging and, f) multimodality imaging (Chen et al., 2014). The different technical features of MI technologies are summarized in **Table 1.1**. In the present research, we have emphasized on radionuclide imaging including both PET and SPECT imaging methods.

Table 1.1. *In vivo* molecular imaging techniques (mainly non-invasive methods)

| Molecular Imaging modality | Form of energy used | Spatial resolution (nm) | | Acquisition time (s) | Mass of probe necessary (ng) | Sensitivity of detection Mol/l | Depth of penetration (mm) |
|----------------------------|----------------------|-------------------------|--------|----------------------|------------------------------|--------------------------------|---------------------------|
| | | Clinical | Animal | | | | |
| PET | Annihilating photons | 3-8 | 1-3 | 1-300 | 1-100 | 10^{-11} - 10^{-12} | >300 |
| SPECT | γ -photons | 5-12 | 1-4 | 60-2000 | 1-1,000 | 10^{-10} - 10^{-11} | >300 |

| | | | | | | | |
|----------------------------------|---------------------------|---------|-----------|---------|-------------|---------------------|-------|
| Computed Tomography (CT) | X-rays | 0.5-1 | 0.03-0.4 | 1-300 | - | - | >300 |
| Magnetic Resonance Imaging (MRI) | Radio frequency waves | 0.2-1 | 0.025-0.1 | 50-3000 | 10^3-10^6 | $10^{-3}-10^{-5}$ | >300 |
| Bioluminescence imaging (BLI) | Visible to infrared light | - | 3-10 | 10-300 | 10^3-10^6 | $10^{-13}-10^{-16}$ | 1-10 |
| Fluorescence imaging | Visible to infrared light | - | 2-10 | 10-200 | 10^3-10^6 | $10^{-9}-10^{-11}$ | 1-20 |
| Ultrasound | High-frequency waves | 0.1-1.0 | 0.05-0.1 | 0.1-100 | 10^3-10^6 | - | 1-200 |

1.1.3. Radionuclide imaging technology: PET and SPECT imaging

During the last two decades, molecular imaging technologies like positron emission tomography (PET) and single photon emission computed tomography (SPECT) had a significant impact on various disease diagnosis and prognosis. The non-invasive feature of these molecular imaging techniques has largely benefited drug discovery and development procedures. These methods are used to understand drug action and establish dosage regimens and treatment strategies. The inception time of SPECT and PET can be dated back in the 1950s and 1960s. PET imaging was the first of the mainstay modalities to be demonstrated, previously hypothesized in 1950 by Brownell and Sweet. Kuhl and Edwards developed SPECT brain imaging, which they formerly described in 1963 and optimized over multiple restatements, culminating the Mark IV system in 1976.

In these techniques, a radionuclide is synthetically introduced into a biomolecule (a ligand/peptide/antibody/antibody fragment) of possible biological significance and administered to a subject (animal or patient). After the radiotracer is administered to a subject, the consequent uptake of the radiotracer is quantified over time and used to attain evidence about the physiological, cellular, and molecular processes of interest.

1.1.3.1. Radioactive decay and types

Radioactivity is the phenomenon of the spontaneous disintegration of unstable (i.e., radioactive) atomic nuclei due to nuclear instability. In the process of decay, more than one kind of energetic ionizing radiation (particles or electromagnetic radiation) can be emitted. Radiotracers are chemical compounds in which one or more atoms have been replaced by a radioisotope.

Six years after the discovery of radioactivity (1896) by Henri Becquerel of France, physicist Ernest Rutherford and his co-workers found that three different kinds of radiation are emitted in the decay of radioactive substances, which he named alpha, beta, and gamma rays in the sequence of their ability to penetrate matter (**Figure 1.1**). These were recognized as helium nuclei, electrons, and high-energy photons, respectively. Soddy in 1913 coined the term *isotopes* to describe atoms of an element that have different atomic weights, but the same chemical properties. A *nuclear disintegration theory* was proposed which explains that radioactivity is the alteration of one chemical element into another through the release of α or β particles or γ radiation (**Figure 1.1**).

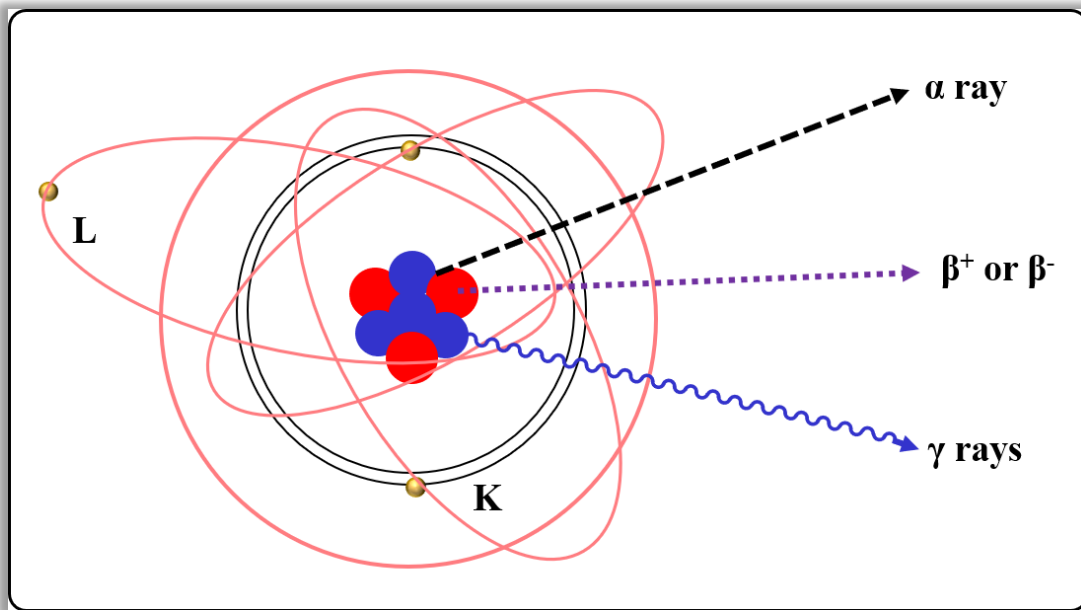


Figure 1.1. Radioactive decay showing α , β and γ emission

Various types of Radioactive Decay

Radioactive decay can be categorized into two main divisions: a) First, which includes a change in the mass number (A) of a radionuclide, and b) Second, where both the parent and the daughter radionuclides have the same mass number (isobaric decay). Four main types are discussed in **Table 1.2**.

a) **Alpha Decay-** A nuclear decay process where an unstable nucleus converts to another element by releasing out a particle composed of two protons and two neutrons. Alpha decay is generally observed in high atomic number elements such as ^{238}U , ^{230}Th , and ^{226}Ra . These radionuclides, which emit α particles may also emit γ photons. Radionuclides with $A > 210$ are large and require to release α particles to reduce their size and become more stable.



Table 1.2. Types of radioactive decay.

| Mode of decay | Cause of instability (of parent nucleus) | Transformation | Example |
|-------------------|--|---|---|
| Alpha decay | Large nucleus | $^A_Z X \rightarrow ^{A-4}_{Z-2} Y + ^4_2 \alpha$ | $^{226}_{88}\text{Ra} \rightarrow ^{222}_{86}\text{Rn} + ^4_2 \alpha$ |
| Beta decay | Neutron rich | $^A_Z X \rightarrow ^A_{Z+1} Y + e^-$ | $^{14}_6\text{C} \rightarrow ^{14}_7\text{N} + e^- + \nu$ |
| Positron emission | Neutron deficient | $^A_Z X \rightarrow ^A_{Z-1} Y + e^+$ | $^{11}_6\text{C} \rightarrow ^{11}_7\text{B} + e^+ + \nu$ |
| Electron capture | Neutron deficient | $^A_Z X + e^- \rightarrow ^A_{Z-1} Y$ | $^{111}_{49}\text{In} + e^- \rightarrow ^{111}_{49}\text{Cd}$ |

b) **Beta decay**- When an atom has either too many protons or too many neutrons in its nucleus, beta decay takes place. Mainly two types of beta decay prevailing: a) **Positive beta decay**- where there is a release of a positively charged beta particle (positron) and a neutrino; b) **Negative beta decay** releases a negatively charged beta particle called an electron and an antineutrino. Negative beta decay is far more common than positive beta decay.

c) **Positron decay**- Radionuclides which lack neutrons are generally unstable and decay to release positive charges either through positron emission or electron capture. These are alternative methods to attain ground state when an unstable nucleus is neutron-deficient and are considered as *inverse beta decay*. Lower atomic numbered elements are more susceptible to positron emission. Positron emission proceeds with the generation of a daughter nucleus having atomic number ($Z-1$), leaving the mass number unaffected. Positron emission is responsible for *annihilation events*, later discussed in **Section 1.2**. ^{11}C , ^{18}F , ^{64}Cu , ^{68}Ga , and ^{124}I , are some important radionuclides used in developing molecular imaging PET that decay by positron emission.

1.2. Positron Emission Tomography

1.2.1. PET Principle

PET imaging technology involves the administration of a radioactive, positron-emitting nuclide, which labels a biomolecule specific to the physiologic process under investigation by PET. PET allows for the three-dimensional mapping of administered positron-emitting radiopharmaceuticals and also enables the study of biological function in both healthy and diseased conditions. Radiopharmaceuticals, labeled with positron-emitting isotopes like ^{11}C and ^{18}F , are administered. The positron-emitting decay process is identified by the alteration of a proton into a neutron along with the emission of a positron (positively charged antiparticle of an electron) and, a neutrino (chargeless particle):



After emission, the positron travels a short distance known as the *positron range* before it annihilates by combining with an electron. During an annihilation event, when a positron unites with an electron nearby, its mass is converted into energy producing two 511 keV γ -rays which travel simultaneously in equal and nearly opposite directions (**Figure 1.2**). This pair of photons is detected by PET scanners, equipped with coincidence γ detectors which hit the detectors almost at the same time. The resolving time between the two coincidence detectors is about 4-5 ns and this time interval is called *coincidence time window*. However, in newly developed PET scanners, the time-of-flight coincidence detectors have a time resolution close to 500 ps (Spanoudaki & Levin, 2010). As soon as the opposite detectors detect the two released photons, within the coincidence window, a coincidence event is logged, and the positron annihilation is expected to have commenced somewhere along the line of response (LOR) connecting the two detectors. In some annihilation events, a detectable coinciding event is not generated either because one of the two γ -rays is absorbed or because it is simply not detected. However, about 97% of the emitting photons are detected. Many such events are summed which help in quantification of line integrals through the isotope distribution. The rationality of this calculation depends on the number of counts collected (Ollinger & Fessler, 1997).

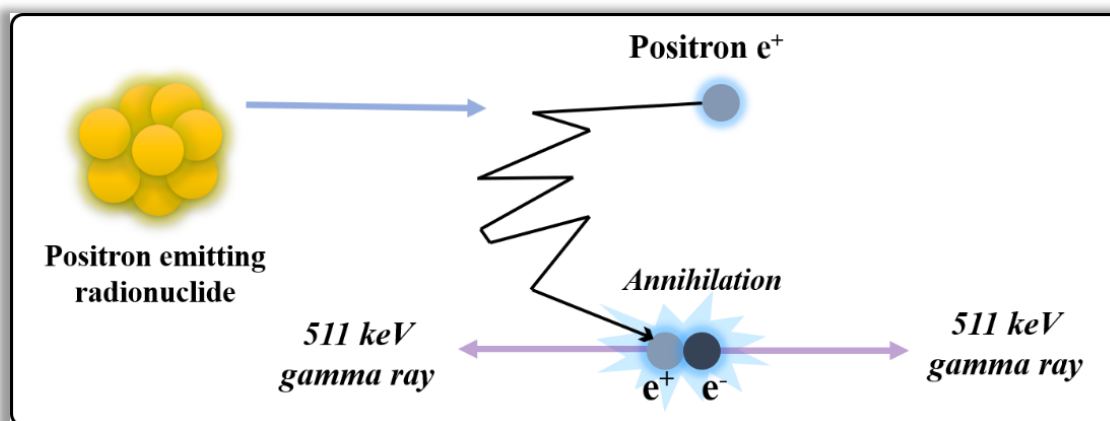


Figure 1.2. The main principle of the PET imaging technique.

Table 1.3. Some commonly used PET radionuclides

| Radionuclide | Half-life ($t_{1/2}$) | E_{\max} (Mev) | β^+ branching fraction |
|---------------------------|-------------------------|------------------|------------------------------|
| ^{11}C Carbon | 20.3 min | 0.96 | 1.00 |
| ^{13}N Nitrogen | 9.97 min | 1.20 | 1.00 |
| ^{15}O Oxygen | 2.1 min | 1.73 | 1.00 |
| ^{18}F Fluorine | 110 min | 0.63 | 0.97 |
| ^{22}Na Sodium | 2.60 y | 0.55 | 0.90 |
| ^{62}Cu Copper | 9.74 min | 2.93 | 0.97 |
| ^{64}Cu Copper | 12.7 h | 0.65 | 0.29 |
| ^{68}Ga Gallium | 67.6 min | 1.89 | 0.89 |
| ^{76}Br Bromine | 16.2 h | Various | 0.56 |
| ^{82}Rb Rubidium | 1.25 min | 2.60, 3.38 | 0.96 |
| ^{124}I Iodine | 4.17 d | 1.53, 2.14 | 0.23 |

A PET study commences with the injection or inhalational administration of a radiopharmaceutical. The scan is initiated after a time lag extending from seconds to minutes to allow for transport and uptake by the organ of interest. Several PET radionuclides with their respective half-life ($t_{1/2}$) are listed in **Table 1.3**.

1.2.2. PET radionuclides and their clinical applications

Majority of the PET radiopharmaceuticals used by medical and clinical researchers are labeled with four common PET radionuclides ^{15}O , ^{13}N , ^{11}C , and ^{18}F . However, metal radionuclides are also used in PET imaging. In this section, we have discussed the clinical applications of different PET radionuclides and their corresponding radiopharmaceuticals.

1.2.2.1. PET Radiopharmaceuticals in Oncology

[^{18}F] labeled compounds- ^{18}F is one of the most commonly used PET radiopharmaceuticals owing to its stable radioisotopic nature. Fluorine is a highly electronegative atom (4.0) when compared with

hydrogen (2.1). Further, carbon-fluorine (C-F) bonds are more stronger and stable *in vivo* than the C–H bonds. Thus, replacing hydrogen with fluorine in the biological system potentiates the half-life of the radiopharmaceutical within the organism. This in turn affects molecules' metabolization, biodistribution, and protein-binding kinetics (Lau et al., 2020). [¹⁸F] fluorodeoxyglucose ([¹⁸F] FDG), the gold standard of PET radiopharmaceuticals is taken up by malignant cells with amplified metabolic and glycolytic rates. [¹⁸F] FDG PET has revolutionized cancer diagnosis because it provides remarkable contrast between the tumor and most normal tissue. Altered glucose metabolism is the central cause of the differential uptake between normal and cancerous cells (Sai et al., 2017). ¹⁸F labeled radiopharmaceuticals have found use in lung cancers, prostate cancer, breast and gynecologic cancers, glioblastoma, hepatocellular carcinoma (HCC), colorectal and pancreatic cancers, solid malignancies, head and neck cancers, and neoplasms (Lau et al., 2020). Radiolabeled amino acid PET radiopharmaceuticals like L-6-[¹⁸F]-fluoro-3,4-dihydroxyphenylalanine ([¹⁸F]-FDOPA), 2-[¹⁸F]-fluoroethyl-tyrosine ([¹⁸F]-FET), 4-fluoroglutamine ([¹⁸F]-FGln), (4S)-4-(3-[¹⁸F]-fluoropropyl)-L-glutamic acid ([¹⁸F]-FSPG), trans-1-amino-3-¹⁸F-fluorocyclobutanecarboxylic acid ([¹⁸F]-FACPC) are widely used in oncology imaging (Qi et al., 2017). Some other examples of F-labeled radionuclides are listed in **Table 1.4**.

[¹¹C] labeled compounds- The short half-life of ¹¹C (~20.4 mins) ensures that the radiopharmaceutical does not involve substantial exposure and facilitates the conduct of multiple studies in a short time interval. [¹¹C]-choline tracers are used in the diagnosis of prostate cancer due to their easy uptake by the malignant cells during cell proliferation. [¹¹C]-acetate is widely used in urological malignancies, renal cell carcinoma, and bladder cancer (Grassi et al., 2012). [¹¹C]-erlotinib, a small molecule radiotracer, can detect lung carcinomas and colorectal cancer in PET scans (Lau et al., 2020).

[¹²⁴I]-Labeled Compounds- These radiotracer molecules play a dual role in cancer diagnosis, i.e., they serve both as an imaging agent as well as provide therapeutic benefit. The therapeutic property of ¹²⁴I radiopharmaceutical is due to the long half-life (~4.18 days) and physical properties of the positron-emitting isotope of iodine. These radionuclides can also be used in mAb development for the potential cure of thyroid and parathyroid cancer (Ranger & Haubner, 2020; Samnick et al., 2018; Wright & Lapi, 2013). ¹²⁴I-tagged small molecules are tested for various targets: ¹²⁴I-dRFIB, ¹²⁴I-IUdR, and ¹²⁴I-CDK4/6 inhibitors of cell proliferation; ¹²⁴I-MIBG for adrenergic activity; ¹²⁴I-hypericin targeting protein kinase C; ¹²⁴I-IAZA and ¹²⁴I-IAZG as hypoxia agents; and ¹²⁴I-FIAU against herpes virus thymidine kinase (Cascini et al., 2014). A few other application includes: ¹²⁴I-IPPM compounds targeting opioid receptors, ¹²⁴I-IPQA participates in EGFR kinase activity, and ¹²⁴I-labeled-6-anilinoquinazoline derivatives irreversibly bind to EGFR. **Table 1.4** lists some important radionuclides and their corresponding radiopharmaceuticals used in cancer diagnosis.

Table 1.4. PET Radiopharmaceuticals used in oncology

| Radiotracer | Disease | Molecular target | Function | Properties |
|--------------------------------------|--|---|---|--|
| [methyl- ¹¹ C] methionine | Urinary, gynecological, liver, and lung cancer | L-type amino acid transporter system and Na ⁺ dependent system | imaging the rate of protein synthesis | the short half-life of [¹¹ C] restricts the availability for PET scanning; [¹¹ C]MET has been also widely used in various brain tumors |
| [¹¹ C]CO | wide applications in clinical | a variety of chemotypes | The production of a wide range of drug- | it requires the presence of transition metals (e.g., Pd) |

| | | | | |
|-----------------------------------|---|--|---|--|
| | research | (amides, ketones, acids, esters, and ureas) | like molecules and radioligands | as reagents; poor solubility of in organic solvents and high dilution in inert gas |
| [¹¹ C]acetate | prostate cancer, hepatocellular carcinoma, lung cancer, nasopharyngeal carcinoma, renal cell carcinoma, bladder carcinoma, and brain tumors | all over the body | tracer for cytoplasmic lipid synthesis (increased in tumors); measurements of myocardial oxygen consumption | acetate is recruited by cells to convert into acetyl- CoA by acetyl-CoA synthetase; rapidly picked up by cells; originally employed in cardiology; salt vector |
| [¹¹ C]erlotinib | Cell lung carcinoma, colorectal cancer | epidermal growth factor receptor | tracing specific binding for activating mutations of the EGFR kinase | small molecule vector; has a structure identical to the clinically used drug |
| [¹¹ C]choline | prostate cancer | Phospholipid synthesis | tumor imaging; diagnostic agent | salt vector; as the proliferation of cancer cells gets higher, tumor cells exhibit an increased rate of the radiotracer's uptake |
| [¹⁸ F]F-choline | prostate cancer | Phospholipid synthesis | primary staging, biochemical recurrence | salt vector; greater accuracy when compared to [¹⁸ F]FDG |
| [¹⁸ F]FDOPA | glioma, neuroendocrine tumors, prostate cancer | amino acid transport; a multiple-target molecule | image a large variety of neuroendocrine tumors and pancreatic beta cell hyperplasia | Amino acid vector; good modality for detection of persistent and residual medullary thyroid cancer |
| [¹⁸ F]FDG | neoplasm | glucose metabolism | tracer used for detection, staging and management of many types of cancer | [¹⁸ F]-FDG accumulates in poorly proliferating and hypoxic cancer cells |
| [¹⁸ F]afatinib | lung carcinoma, colorectal cancer | epidermal growth factor receptor | detection of EGFR-positive tumors | small molecule |
| [¹²⁴ I]I-codrituzumab | hepatocarcinoma | glypican 3 | detects tumor localization in most patients with HCC | antibody vector |
| [¹²⁴ I]I-girentuximab | renal cell carcinoma | Carbonic anhydrase 9 | Discriminates between clear-cell RCC (ccRCC) and non-ccRCC | antibody vector |
| [⁶⁴ Cu]-DOTA- | breast, lung, | a promising | diagnostic/imaging; | first in-human use in 2013 |

| | | | | |
|-----------------------------------|---|---|--|--|
| AE105 | colorectal, prostate, and bladder cancer | uPAR- PET ligand in several preclinical validation studies; peptide antagonists AE105 | prognostic in cancer invasion and metastasis | |
| [⁶⁸ Ga]citrate | Prosthetic joint/bone infections | N.A. | diagnosis of bone infection | ⁶⁸ Ga-citrate has additional advantages over ⁶⁷ Ga for the analysis of bone infections |
| [⁸⁹ Zr]Zr-bevacizumab | Solid malignancies; particularly for malignant breast lesions | Vascular endothelial growth factor receptor | Early detection; VEGF-A overexpression | antibody vectors |

1.2.2.2. PET Radiopharmaceuticals in Neurology

PET imaging has aided in investigations of the underlying pathophysiology of different neurological conditions. It has been employed to investigate metabolism, receptor binding, and alterations in regional blood flow. One of the major applications is in the favor of elucidating complex neurological disorders such as Parkinson's disease (PD), Huntington's disease (HD), multiple sclerosis (MS), and dementias or Alzheimer's disease (AD). Some of the important PET radiopharmaceuticals used in neurodegenerative diseases are enlisted in **Table 1.5**. The molecular sensitivity in the central nervous system (CNS) allows the PET radiotracers for the quantification of target-ligand interactions with good selectivity in humans giving information about disease pathology. Widely accepted radiopharmaceuticals for brain imaging involve [¹⁸F]-FDOPA tracers for dopamine synthesis in PD and schizophrenia, [¹⁸F]-FDG analogs for imaging the glucose metabolism alterations and translocator proteins detection in AD and/or PD (Chételat et al., 2020; Minoshima et al., 2021). Additionally, [¹¹C]-PIB compounds are used for tracking the amyloid β plaque accumulation in AD (Blazhenets et al., 2021).

A PET tracer should have the potential to cross the blood-brain barrier (BBB) while the tracer's selectivity ultimately impacts its usefulness and applicability. Therefore, they should follow essential criteria: a) molecular weight should be less than 500 kDa; b) lipophilic coefficient between 1 and 5; and c) topological polar surface area should be below 90 Å² (Minoshima et al., 2016; S. Y. Yap et al., 2021). Benzothiazole and benzoxazole derivatives like [¹⁸F]-flutemetamol, [¹⁸F]-florbetapir, and [¹⁸F]-florbetaben are used for the detection of pathological amyloid depositions within the brain tissue. Tracers like [18F]-AV-1451 and [18F]-THK help in the detection of aggregation rates of tau proteins (Harada et al., 2018). Adenosine 2A receptors (A_{2A}) are GPCRS targeted by CNS neurotransmitters and highly targeted in multiple neurological disorders. In this case, most reliable tracers developed for targeting A_{2A} are [¹¹C]-TMSX and [¹¹C]-SCH442416.

Table 1.5. PET Radiopharmaceuticals used in neurology

| Radiotracer | Molecular target/ Disease | Function | Properties |
|-----------------------|------------------------------|---|--|
| [¹¹ C]MET | brain gliomas and metastases | differentiation of tumor regrowth; delineation of | the tracer's stability in its final formulation is not well documented in the literature |

| | | | |
|---|---|---|---|
| N-[¹¹ C]-methyl- flumazenil | neuronal damages, epilepsy, stroke-induced penumbral, infarction, and AD | gliomas. binds to the benzodiazepine sites of GABA _A receptors | excellent kinetic properties for image quantification |
| [¹¹ C]raclopride | psychiatric, PD, addiction, attention-deficit hyperactivity disorder, schizophrenia | tracer for dopamine function in the striatal cortex | most widely used PET radiotracer for measuring DA changes in dopamine rates at the synaptic level |
| [¹¹ C]UCB-J | targeting synaptic vesicle proteins SV2A; primary interest in epilepsy, or diseases associated with synaptic loss | imaging SV2A expression in synaptic vesicles | leading SV2A tracer; good selectivity, fast kinetics |
| [¹¹ C]MK-3168 | pain, addiction, and Tourette syndrome | fatty acid amide hydrolase associated receptors | slow kinetics and rapid metabolism in humans |
| [¹¹ C]Martinostat | schizophrenia; cerebellum uptake; | CNS quantification of HDACs | high brain uptake |
| [¹¹ C]PS13 | dysfunctions of enzymes within the CNS | imaging COX-1 | limited data available; low plasma free fraction, suitable kinetic profile |
| [¹⁸ F]FT | brain tumor; does not serve as a substrate to protein synthesis; glioma ¹⁸ F-FET uptake is not significantly influenced by changes in the BBB permeability | good diagnostic performance; highly specific for glioma | evaluation of its applicability in non-clinical research is still lacking; overcomes known limitations of [¹⁸ F] FDG: increased uptake in the inflammatory environment and elevated background signal in normal brain |
| [¹⁸ F]AV-1451 | tau proteins (tauopathies), AD | AD assessment; distinguishes between disease stages | high selectivity over amyloid; fast kinetics |
| [¹⁸ F]MK-6240 | AD | imaging neurofibrillary tangles | low bindings in healthy controls; strong correlation with cognitive AD scores |
| [¹⁸ F]PM-PBB3 | AD, cerebral accumulations of tau deposits | tau imaging | lower binding in basal ganglia and thalamus when compared to [¹¹ C]-PBB3 |
| [¹⁸ F]FEOBV | targeting cholinergic system; Cerebellar grey matter, and striatum | Acetylcholine transporters and cholinergic synapses; studying degenerative conditions | improved signal-to-noise over previous VACht tracers; but slow kinetics |

| | | | |
|-----------------------------------|--|--|--|
| [¹⁸ F]UCB-H | targeting synaptic vesicle proteins SV2A; target has no reference region | binds specifically to the synaptic vesicle glycoprotein 2A | lower sensitivity compared to [¹¹ C]-UCB-J |
| [¹⁸ F]FIMX | PD, addiction, epilepsy, neuropathic pain, and depression | Targeting metabotropic glutamate receptors mGluR1 | very good in vivo block response; fast kinetics |
| [¹⁸ F]BCPP-EF | Targeting mitochondrial complexes MC1 | Quantitative imaging of MC-1 activity in the living brain | high brain uptake, suitable kinetics, large dynamic range in vitro |
| [⁶⁴ Cu]Cu-SARTATE | Neuroendocrine tumors | Targeting somatostatin receptor 2 | peptide vectors |
| [⁶⁸ Ga]DOTA-TOC | | | |
| [⁶⁸ Ga]Ga-DOTA-NOC | | | |
| [⁶⁸ Ga]Ga NODAGA-JR11 | | | |
| [⁶⁸ Ga]Ga-DOTA-TATE | | | |

1.2.2.3. PET Radiopharmaceuticals for Cardiovascular Diseases

Nuclear cardiology has extended its spectrum over the past few years and is applied in the detection of heart function, circulation of blood, non-invasive imaging of myocardial viability, and cardiac inflammation. [¹³N]-ammonia is a promising tracer for the myocardial uptake assessment (Li et al., 2014) or the myocardial blood flow measurement (Nesterov et al., 2014). ¹³N and ¹⁵O labeled inorganic radiopharmaceuticals have been widely used for cardiac perfusion imaging. Small organic tracer molecules like ¹¹C-epinephrine and ¹⁸F-fluorodopamine were useful to image presynaptic sympathetic nervous system of the heart. Cationic radiotracer like [⁸²Rb]-chloride targets Na⁺/K⁺ ATPase cotransporters and helps in myocardial perfusion imaging. A wide range of PET radiopharmaceuticals used for cardiovascular disease imaging is given in Table 1.6.

Table 1.6. PET Radiopharmaceuticals used in cardiology

| Radiotracer | Disease | Molecular target | Function | Properties |
|------------------------------------|--|---|--|---|
| [⁸² Rb]chloride | Cardiac conditions | cardiac tissue | diagnostic; monitoring the cardiac flow | low delivered radiation exposure for a rest/stress test |
| [¹⁵ O]H ₂ O | Myocardial perfusion, cerebral and tumor perfusion | N.A. | tracer for quantitative measurement of cerebral blood flow | the short half-life of ¹⁵ O results in challenges in clinical use |
| [¹³ N]NH ₃ | cardiovascular events, PC and encephalopathy | myocardial tissue, liver, kidneys and brain | imaging agent for assessing regional blood flow in tissues; for elucidation of NH ₃ metabolism in patients with | ammonia ¹³ N enters the myocardium through the coronary arteries; well-validated radiotracer for clinical management; it is also used in PC due to the up-regulation of NH ₃ during glutamine |

| | | | | |
|--------------------------------------|---|--|---|--|
| | | | hepatic encephalopathy; potentially a tumor | synthesis in tumors |
| | | | imaging agent | |
| [¹⁵ O]CO | cardiovascular events | myocardial tissue | myocardial function | the most common tracers used for non-invasively measuring oxygen consumption and blood volume |
| [¹⁸ F]flurpiridaz | | mitochondrial complex I | | novel PET tracer |
| [¹⁸ F]FBnTP | | mitochondrial membrane | | Rapid myocardial uptake and retention and high myocardium/liver, myocardium/blood and myocardium/lung contrast in animal studies; few human studies reported to date |
| [¹⁸ F]FTPP | Myocardial perfusion | mitochondrial membrane | diagnostic/imaging | |
| [¹⁸ F]FDHR | | mitochondrial complex I | | |
| [¹⁸ F]FDM | atherosclerosis | Mannose receptors | Progressive inflammation in atherosclerotic plaques | it is (mannose) an isomer of glucose that is taken up by macrophages through glucose transporters |
| [¹⁸ F]macroflor | atherosclerosis | macrophage-targeted polyglucose nanoparticle | immunoimaging; nanoparticle uptake | noninvasive assessment of the immune system in atherosclerosis |
| [⁶⁴ Cu]DOTA-ECL1i | Lung inflammation | Chemokine receptor type 2 (CCR2) | detection of CCR2- directed inflammation | sensitive and specific detection of CCR2+ cells |
| [⁶⁴ Cu]DOTA-DAPTA-comb | initiation and progression of atherosclerosis | Chemokine receptor CCR5 | specific imaging of CCR5 | nanomedicinal approach toward cardiovascular diseases |
| [⁶⁸ Ga]DOTATAT E/DOTANOC | Inflammatory conditions related to plaques | N.A. | Functional imaging of plaques | increased uptakes in coronary arteries and large arteries; comparable diagnostic accuracy |

1.2.2.4. PET Radiopharmaceuticals for Bacteria Imaging

Based on different bacterial strains (Gram-positive, Gram-negative, Gram-positive and negative, and others), Auletta *et al.* classified the research studies focusing on bacterial PET imaging. Gram-positive bacteria imaging produced better imaging results compared to other strains used during the studies (Auletta *et al.*, 2019). ¹²⁴I-labeled FIAU revealed better results in animal models than in humans (Zhang *et al.*, 2016). [¹⁸F]-FDG-6-P agents can distinguish between infection and sterile inflammation and are highly expressed in several bacteria (Mills *et al.*, 2015). **Table 1.7** enumerates some PET radiopharmaceuticals used in bacterial imaging.

Table 1.7. PET Radiopharmaceuticals used in bacteria imaging

| Radiotracer | Type of bacteria | Properties |
|---|---|--|
| *[¹⁸ F]FHM (maltohexose) | <i>S. aureus</i> | better than FDG in differentiating non-infection inflammation from infection |
| *[¹⁸ F]FDS (sorbitol) | <i>K. pneumoniae</i> | better than FDG to detect lung infection from inflammation |
| *[¹⁸ F]maltotriose | N.A. | imaging bacterial infections in animals;future applications in clinics |
| *[¹⁸ F]FDS | <i>E. coli</i> , <i>Enterobacteriaceae</i> | diagnosis and monitoring therapy;diagnostic for infections |
| *[¹⁸ F]FDG-6-P | <i>S.aureus</i> | potential to differentiate infection from inflammation |
| *[¹⁸ F]isonicotinic acid | <i>M.tuberculosis</i> | non-invasive approach to localize infectious foci;tested only on mice |
| *[¹⁸ F]FIAU | <i>E. coli</i> , <i>P.aeruginosa</i> | engineered pathogens for evaluating experimental therapeutics |
| *[¹⁸ F]PABA | <i>S.aureus</i> | non-invasive tool for detecting/localizing/monitoring infections |
| *[⁶⁸ Ga]TAFC | <i>A.fumigatus</i> | very promising for the detection of infections with high sensitivity |
| *[⁶⁸ Ga]FOXE | | |
| [⁶⁴ Cu]ProT(prothrombin) | <i>S.aureus</i> | non-invasive detection with an analog of ProT |
| [⁶⁴ Cu]JF5 mAb | <i>A.fumigatus</i> | localized aspergillus infection |
| [¹⁸ F]maltose | <i>E. coli</i> | identifying drug resistance; promising for bacterial infection imaging |
| [¹⁸ F]trimethoprim | <i>E. coli</i> , <i>P.aeruginosa</i> <i>S.aureus</i> | infection imaging |
| [⁶⁸ Ga]UBI-29-41 | <i>S.aureus</i> | non-toxic, identify infectious foci in humans; correlated with the degree of infection needs further studies; |
| [⁶⁸ Ga]UBI-31-38 | <i>S.aureus</i> | good localization of infection site; promising results in humans |
| [⁶⁸ Ga]TBIA101 (depsidomycin derivative) | <i>M. tuberculosis</i> <i>S.aureus</i> | imaging inflammation but not necessarily infection; non-specific |
| [¹²⁴ I]FIAU (fialuridine) | <i>S.aureus</i> | well tolerated but of limited value for the detection of prosthetic joint infection; low image quality/specificity |
| [¹¹ C]PABA (para-aminobenzoic acid) | <i>E.coli</i> | imaging living bacteria in humans |

1.2.2.5. PET Radiopharmaceuticals for Inflammation/Infection

Inflammation or infection is linked with a variety of diseases directly or indirectly. Therefore, molecular imaging of inflammation in various conditions (like stroke, Alzheimer's disease, atherosclerosis, autoimmune diseases, and even malignant conditions) provides ample information relating to disease diagnosis or prognosis. B cells are one of the main therapeutic targets essential for

controlling immunological responses. BTK is a cytoplasmic tyrosine kinase expressed by B cells and are being studied for the treatment of B-cell malignancies. Radiolabelled BTK inhibitors are important in the monitoring and treatment of B-cell-mediated diseases. [^{11}C]- ibrutinib a potential PET imager used for inflammation imaging presented >98% radiochemical purity and 19.89 to 20.15 min half-life (Donnelly et al., 2022).

^{11}C or ^{18}F -labeled isoquinoline carboxamide derivatives are used in PET imaging of peripheral tissue translocator proteins which are expressed during inflammation (Hatori et al., 2012). PET imaging has shown promising results in atherosclerosis detection, lung lesion absorption, neuroendocrine tumor imaging, etc. [^{18}F]-FDG is a good agent for biopsy, as it can detect the most active infection sites. Also, it helps in treatment monitoring and regulation and has a great impact on large vessel vasculitis imaging (Ankrah et al., 2019; Douglas et al., 2019).

1.3. Single Photon Emission Computed Tomography

Single-photon emission computed tomography (SPECT) is a nuclear imaging modality used frequently in diagnostic medicine. It gives a three-dimensional nuclear image with combined knowledge obtained from scintigraphy with that of computed tomography. This allows a three-dimensional display offering better detail, contrast, and spatial information. SPECT imaging uses radionuclides that directly emit **gamma (γ) rays** such as technetium-99m ($^{99\text{m}}\text{Tc}$) and iodine-123 (^{123}I). Generally, the half-lives of SPECT radiotracers are longer than those used in PET imaging (Table 1.8). This makes them more accessible for imaging and longer radiosynthesis times make them more viable.

1.3.1. SPECT Principle

SPECT machines combine an array of gamma cameras (ranging from one to four cameras) that rotate around the patient (Lee et al., 2000). Radionuclide distribution within tissues can be determined spatially using specially designed gamma cameras rotating around the patient. The use of multiple gamma cameras increases detector efficiency and spatial resolution. Three-dimensional images are then constructed from the projection data obtained from the cameras (Lee et al., 2000; Van Paesschen et al., 2003). **Figure 1.3** shows the basic principle of how SPECT imaging works.

Table 1.8: Commonly used radionuclides for SPECT imaging

| Nuclide | Half-life/h | Principal photon emission energies/MeV | Type of emission |
|--------------------------|-------------|--|---------------------|
| ^{123}I | 13.2 | 0.16 | Electron capture |
| $^{99\text{m}}\text{Tc}$ | 6 | 0.14 | Isomeric transition |
| ^{111}In | 67.9 | 0.17/0.25 | Electron capture |
| ^{67}Ga | 78.3 | 0.09/0.19/0.30 | Electron capture |
| ^{201}Tl | 73.1 | 0.17 | Electron capture |

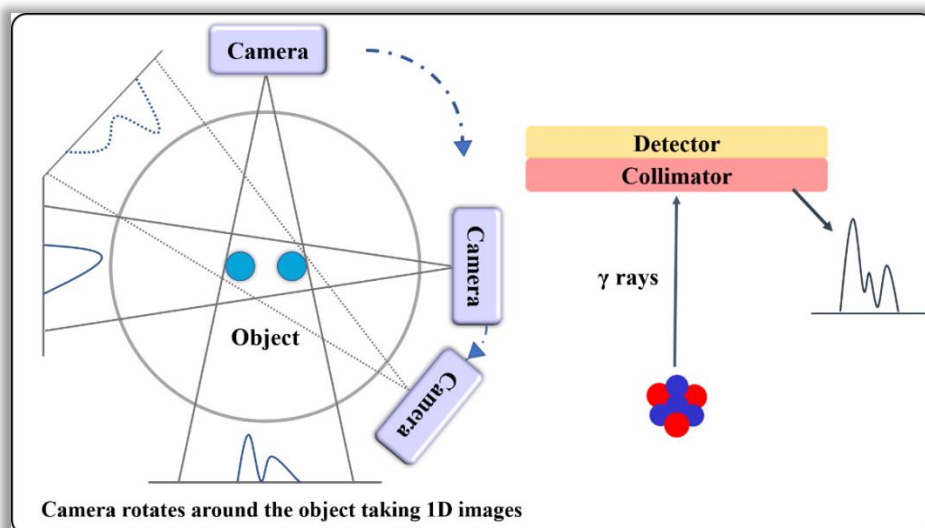


Figure 1.3. Principles of SPECT imaging

1.3.2. SPECT radiopharmaceuticals

SPECT radiopharmaceuticals are used to diagnose neurodegenerative diseases, cancer, and infections by emitting gamma (γ) radiation. The longer half-life of these chemicals has the advantage of enabling SPECT imaging studies to be conducted over longer periods. **Figure 1.4** shows some commonly used technetium (^{99m}Tc) labeled SPECT radiopharmaceuticals. The important features required for a good SPECT radiotracer are:

- Easy availability
- Carrier free
- Non-toxic
- Free from α and β particles emission (with little emission)
- Biological half-life not greater than the time of study
- Suitable energy range
- Chemically reactive to form coordinate covalent bonds with the compound which is to be labeled.

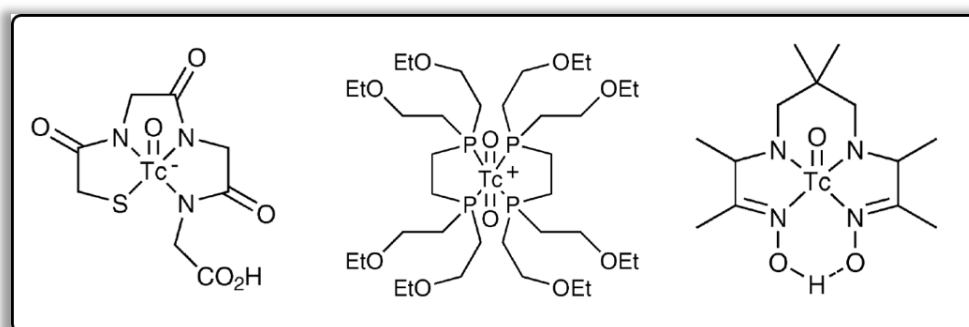


Figure 1.4. Examples of ^{99m}Tc radiopharmaceuticals.

1.3.3. Application of SPECT imaging

During SPECT imaging, the clinician must have prior knowledge about: a) the goal of the scan b) the associated risk to the patient, and c) the expense of the isotope. SPECT imaging not only depends on the selection of the correct radioisotope but the main achievement lies beneath the fact that the isotope

has to be successfully bound to a biologically active ligand, which will interact with the body tissues to deliver the isotope to the desired location. SPECT imaging depends on more than just the selection of the correct radioisotope. SPECT scan has found immense applicability in the medical fields mainly in cardiovascular diseases, brain disorders, and cancer. SPECT scans are also indicated for non-cardiac and non-neurological conditions such as osteomyelitis, spondylolysis, parathyroid disease, pulmonary embolism, and abscess localization.

1.3.3.1. SPECT imaging in Oncology- SPECT imaging has played a wholesome role in clinical oncology in identifying various tumors overexpressing particular receptors. Receptors such as somatostatin receptors (breast, brain, and small cell lung cancer tumors), prostate-specific membrane antigen (prostate cancer), gastrin-releasing peptide receptor (prostate, breast, pancreas, small cell lung cancer, and colorectal tumors), melanocortin receptor (melanomas), and integrin $\alpha_v\beta_3$ receptor (brain, lung, ovary, breast, and skin cancer) were targeted (Rezazadeh & Sadeghzadeh, 2019).

1.3.3.2. SPECT Radiopharmaceuticals for Cardiovascular Events- Tc labeled radiopharmaceuticals have shown good results in the diagnosis risk assessment of coronary artery disease (CAD). Currently, in use, three SPECT imaging agents [^{201}Tl]-Cl, [$^{99\text{m}}\text{Tc}$ (I)]-sestamibi, and [$^{99\text{m}}\text{Tc}$ (V)]-tetrofosmin have revealed promising results (Watson & Glover, 2010). The applicability of these diagnostic agents relies based on their good pharmacokinetic properties (half-life, high first-pass extraction, linear relation between uptake and blood flow, and rapid clearance).

1.3.3.3. SPECT Radiopharmaceuticals in Neurological Disorders- Commonly used SPECT imaging agents for neurological disorders (mainly for Parkinson's disease) are: [$^{99\text{m}}\text{Tc}$ (V)]-HMPAO, [^{123}I]-ioflupane and [$^{99\text{m}}\text{Tc}$ (I)]-TRODAT-1 (Adak et al., 2012; Valotassiou et al., 2018). 123I-based imidazopyridine compounds are used for amyloid beta imaging (Chen et al., 2015).

1.4. Difference between PET and SPECT imaging

Both PET and SPECT imaging agents are widely used in the medical field for the diagnosis of different disease pathologies. However, both have unique features which make them better from each other. Table 1.9. enlists some advantages, disadvantages, and clinical use of both PET and SPECT imaging agents explaining how unique they are from each other.

Table 1.9. Comparison between PET and SPECT imaging agents

| Method | Advantages | Disadvantages | Clinical Use | In vivo animal use |
|--------|---|--|---|--|
| PET | -high sensitivity -3D acquisition -good resolution within a physical limit | -isotopes are of short half-life -isotopes produced in cyclotrons -expensive process -higher tissue dose required | -[^{18}F]FDG is a routine imaging agent in the diagnosis of cancer -special application in neurology and cardiology | Currently evolving -microPET -high-density avalanche chamber |
| SPECT | -resolution limited by technology (submillimeter) -low sensitivities -can differentiate | -2D planar images -semiquantitative data only | -readily available tracer -a wide range of clinically tested tracers available | -pinhole collimator -dedicated cameras evolving |

between isotopes
with different
radiation energies

1.5. Nitroaromatics as Radiosensitizers

Radiosensitizers are promising chemicals or pharmaceuticals that are intended to enhance injury to tumor tissue by accelerating DNA damage and producing free radicals. They usually tend to augment the lethal effects of radiation. Based on the mechanism of DNA damage and repair, G E Adams, a pioneer in the field of radiation therapy, categorized radiosensitizers into five classes: (1) suppressor of intracellular thiols or other endogenous radioprotective substances; (2) formation of cytotoxic substances by radiolysis of the radiosensitizer; (3) inhibitors of repair of biomolecules; (4) thymine analogs that can incorporate into DNA; and (5) oxygen mimics that have electrophilic activity (Gong et al., 2021). Oxygen is regarded as the best radiosensitizer by far; however, metabolic consumption of oxygen limits its diffusion into hypoxic tumor cells. Hypoxia has a principal role in cancer progression operating angiogenesis, vasculogenesis, activation of a glycolytic shift in metabolism, invasion enhancement, and metastasis (De & Roy, 2021). Nitroheterocyclic compounds like nitroimidazoles, nitrofurans and nitrothiophenes are recognized **oxygen-mimetic agents** in which electron-rich reactive nitro group reacts with DNA, after which the DNA and nitro group adduct causes DNA strand breakage and subsequent cell death. The mechanism for DNA damage by aromatic nitro compounds is shown in **Figure 1.5**. The bioreduction ability of nitroaromatics allows the generation of free radicals in intracellular environments with a low oxygen concentration; a typical circumstance occurring in solid tumors encompassing areas of hypoxia resulting from inadequate blood supply (Chin Chung et al., 2011; Wardman, 2007). Thus, these compounds are potential targets for the detection of hypoxic cells in cancer patients.

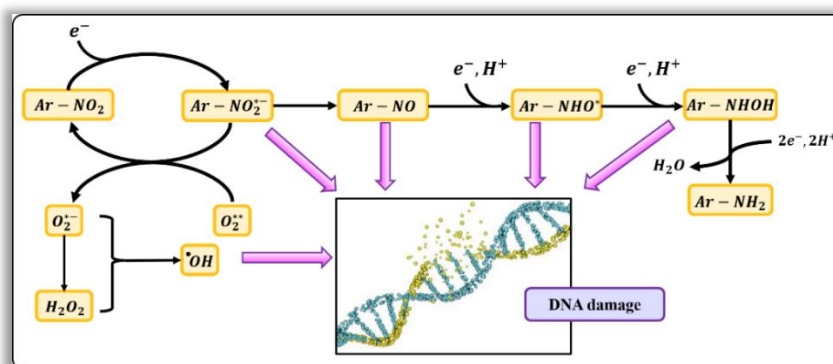


Figure 1.5. Mechanism of nitroaromatic radiosensitizers in DNA damage

1.6. Quantitative structure-activity relationship (QSAR) analysis

1.6.1. The formalism

Drug discovery has been recently oriented towards the modeling and design of new molecules to discover potent molecules having improved therapeutic activity and less toxicity. *In silico* approaches play a vital role in this practice of rational drug discovery. The ideology of QSAR analysis assumes that the molecules available in nature contain information for their physical, chemical, or biological (including toxicological) activity which can be suitably described in terms of mathematical correlation. Eq. 1.3 describes the response elicited by chemicals to be a mathematical function of chemical features.

$$\text{Chemical Response} = f(\text{Chemical attributes}) = f(\text{Structure, Property}) \quad (1.3)$$

Here, response refers to any type of physicochemical property, activity, or toxicity shown by the chemicals, while the chemical attributes correspond to the quantitative information obtainable from the chemicals using suitable theoretical/ experimental techniques. The nomenclature of the technique is usually done depending on the nature of the response. Accordingly, three broad types can be identified namely QSAR, QSPR, and QSTR denoting the response to be a biological activity (e.g., anti-cancer, anti-malarial, anti-diabetic, anti-tubercular, anti-cholinergic, anti-bacterial, etc.), physicochemical property (melting point, boiling point, molar refractivity, lipophilicity, viscosity, aqueous solubility etc.) as well as toxicity (non-systemic such as ecotoxicity/environmental toxicity as well as systemic such as hepatotoxicity, cardiotoxicity, pulmonary toxicity, nephrotoxicity, etc.) respectively. However, we shall use the term QSAR to denote QSAR/QSPR/QSTR analyses in a broad sense. It may be noted that various physicochemical properties of chemicals can also be used as a chemical feature as shown on the right side of Eq. 1.3.

1.6.2. History of QSAR

The history of the correlation of chemical features can be traced back to the nineteenth century with the theory proposed by Mendeleev who used the ‘rule of eight’ (Tute, 1990) for identifying similar chemicals. In 1868, the first mathematical notion in QSAR was reported by Crum-Brown and Fraser (Brown & Fraser, 1868) who expressed the possible mathematical correlation between the biological activity of various alkaloids with their molecular constitution using the following equation 1.4.

$$\Phi = f(C) \quad (1.4)$$

Here the physiological action Φ of a chemical in a biological system is shown as the function (f) of its constitution C . Thus, an alteration in the chemical constitution, ΔC , would be reflected by an alteration in biological activity $\Delta \Phi$. It may be noted that ‘chemical constitution’ at that time was not a vividly defined principle, but rather an effort to express elemental composition which provided the conceptual platform for the modern predictor variables/descriptors.

Körner (Körner, 1874) theorized the change in color of (physicochemical property) of disubstituted benzenes be correlated with their differing chemical structure. Following that, in the year 1884, Mills (Mills, 2009) found the melting point and boiling point of compounds to be correlated with chemical composition. About ten years later, cytotoxicity and aqueous solubility of diverse organic compounds were reported to be inversely correlated by Richet (Richet, 1893). Meyer (Meyer, 1899) and Overton (Overton, 1899), independently suggested that the narcotic (depressant) action of a group of organic compounds is correlated with their olive oil/water partition coefficients.

Hammett, in 1935, provided a revolutionary contribution here by establishing a relationship between the chemical reactivity and structural features of benzene derivatives using rate constant and electronic constant terminologies. The famous Hammett constant is described by the following equation (Eq. 1.5 and Eq. 1.6) where k_X and k_H are the rate constants and K_X and K_H are the equilibrium constants of substituted and unsubstituted benzenes, σ_X is the Hammett constant providing electronic information and ρ is a constant.

$$\log \frac{K_H}{K_X} = \rho \sigma_X \quad (1.5)$$

$$\log \frac{k_H}{k_X} = \rho \sigma_X \quad (1.6)$$

Hammett ‘ σ ’ was defined by ionization constant terms allowing its correlation with a linear free energy-based formalism using the Gibbs equation, i.e., $\Delta G^0 = -RT \ln K$ (where, ΔG^0 is the Gibbs free energy change at the standard state, R is the ideal gas constant, T is the temperature in Kelvin) and thus Hammett’s equation is considered the beginning of the linear free energy relationship (LFER) concept. By employing the LFER technique, Taft introduced steric feature E_s and allowed a separate assessment of polar, steric, and resonance effects by performing acid- and base-catalyzed hydrolysis of aliphatic esters. Swain and Lupton (1968) provided further information on the resonance and polar effects. The notable observation of Hammett and Taft paved the next foothold observations by Hansch and Fujita. Corwin Hansch is considered the ‘Father of QSAR’ (Martin & Stouch, 2011) for his notable contribution in the QSAR paradigm. By using the Hammett constant and a hydrophobicity measure, Hansch and Muir (Hansch et al., 1962) performed structure-activity relationship analysis using plant growth regulators. Using the octanol/water system, a new hydrophobic measure ‘ π ’ was introduced to represent the partition coefficient of the substituent using the contribution of the whole molecule (see equation 1.7).

$$\pi_X = \log P_X - \log P_H \quad (1.7)$$

Here, the subtraction of the logarithmic partition coefficient of the derivative (P_X) and parent (P_H) molecule gives the substituent hydrophobicity π_X . This was followed by combining Hammett’s electronic constant (σ) and the hydrophobicity measure (π) into a single equation by Fujita and Hansch. Fujita identified that the free energy variables, e.g., π , $\log P$, σ , etc. can be combined into a single equation to represent biological activity by implementing a logarithmic transformation of the concentration term to keep the LFER formalism. Because of the presence of linear free-energy related variables, Hansch analysis is also termed as the ‘extra thermodynamic approach’. Eq. 1.8 is an LFER model showing biological activity to be composed of hydrophobic and electronic factors.

$$\log \left(\frac{1}{C} \right) = k_1 \pi + k_2 \sigma + k_3 \quad (1.8)$$

This model was later modified by Hansch who considered that drug molecules undergo a ‘random walk’ to reach the target receptor, and he proposed a parabolic relationship between $\log(1/C)$ and $\log P$ and the following equations were obtained likewise.

$$\log \left(\frac{1}{C} \right) = k_1 (\log P) - k_2 (\log P)^2 + k_3 \sigma + k_4 \quad (1.9)$$

$$\log \left(\frac{1}{C} \right) = k_1 \pi - k_2 \pi^2 + k_3 \sigma + k_4 \quad (1.10)$$

Another LFER equation was also proposed by Hansch containing the hydrophobicity constant, Hammett’s electronic constant, and Taft steric parameter (see equation 1.11).

$$\log \left(\frac{1}{C} \right) = k_1 \pi + k_2 \sigma + k_3 E_s + k_4 \quad (1.11)$$

Later in 1976, Kubinyi (Kubinyi, 1976) provided a bilinear model by modifying the parabolic relationship (equation 1.12) of Hansch. The parameters a , b , and c can be calculated by linear multiple regression analysis, while the non-linear term β must be derived by a stepwise iteration process or Taylor series iteration.

$$\log \left(\frac{1}{C} \right) = a \log P - b \log(\beta P + 1) + c \quad (1.12)$$

Along with Hansch's approach, other notable methodologies also came into light in the 1960s. The famous Free-Wilson approach (Free & Wilson, 1964) mathematically correlates the biological activity of a congeneric series of chemicals with the common contribution of the parent moiety plus the contribution of each structural substituent. Eq. 1.13 presents the Free-Wilson model where BA is the biological activity assumed to comprise the average parent moiety contribution μ and the contribution of each substituent a_i . Here, x_i is a Boolean variable denoting the presence (=1) or absence (=0) of a specific structural fragment.

$$BA = \sum a_i x_i + \mu \quad (1.13)$$

Fujita and Ban (1971) considered a log transformed response and provided a modified version of this equation to overcome its shortcomings (Eq. 1.14).

$$\log BA = \sum G_i x_i + \mu \quad (1.14)$$

Here, μ is the contribution of the parent moiety, G_i denotes the contribution of chemical fragments and x_i represents the presence/ absence of G_i .

In 1973, C. Hansch and S. Unger defined good practice in QSAR, comprising five criteria that must be considered before selecting a "best equation". The five criteria were as follows, i) selection of independent variables considering the widest possible number of variables; ii) justification of the choice of independent variables by statistical procedures; iii) the principle of parsimony, i.e., all things being equal, one accepts the simplest model; iv) a number of terms: one should have at least five to six data points per variable to avoid chance correlations; v) try to find a qualitative model of physicochemical or biochemical significance, etc.

1.6.3. Objectives of QSAR analysis

The principal objectives and significance of QSAR analysis are (Cronin et al., 2003):

- a. Prediction of new analogs of a compound with better property
- b. Exploring and a better understanding of the mode of action
- c. Lead optimization with decreased activity
- d. Reduction of wet laboratory experimentation
- e. Cost, time, and manpower reduction by evolving more effective compounds using a scientifically less exhaustive approach.
- f. To develop expert systems for predicting various toxicity endpoints of potential drug candidates.
- g. Identifying pollution prevention measures.
- h. Identifying scientific data gaps.

The key objective of QSAR analysis is the development of a mathematical relationship between the response and chemical features of a series of compounds. Hence, mathematically for a series of compounds, this is about developing a correlation between a Y and several X variables where Y is the dependent variable remaining on the left side of the equation and the X variables are independent entities. The response or the Y variable is also termed as 'endpoint' while the X features are predictor variables, and a QSAR relationship can be simply stated as follows in Equation 1.15:

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_n x_n \quad (1.15)$$

In Eq. 1.15 Y is the response variable (independent) being modeled for a set of n number of compounds with the predictor variables x_1, x_2, \dots, x_n etc. possessing coefficient values a_1, a_2, \dots, a_n respectively, and a_0 is the constant term. With the values of Y and x variables known, the coefficients (a_1, a_2, \dots, a_n) along with the constant (a_0) can be easily solved by giving an explicit mathematical equation correlating biological activity or toxicity or physicochemical properties of compounds with their chemical features.

1.6.4. Classification of QSAR

1.6.6.1. Based on dimensionality

QSAR models can be classified into different classes based on the descriptor complexity. They are (i) 0D-QSAR, (ii) 1D-QSAR, (iii) 2D-QSAR, (iv) 3D-QSAR, (v) 4D-QSAR, (vi) 5D-QSAR, and (vii) 6D-QSAR and (viii) 7D-QSAR. Table 1.8 and Figure 1.6 give examples of descriptors based on their dimensions.

Table 1.8. Classification of QSAR technique based on its dimension

| Dimension | QSAR Method |
|-----------|---|
| 0D-QSAR | Descriptors involving molecular formulas, like molecular weight etc. |
| 1D-QSAR | Physicochemical properties of molecular structure, such as lipophilicity, solubility etc. |
| 2D-QSAR | Structural patterns, i.e., the topology of the molecules (without 3D representation) |
| 3D-QSAR | Activity is correlated with three-dimensional structure of the ligands. |
| 4D-QSAR | Ligands are represented as an ensemble of configurations. |
| 5D-QSAR | As 4D-QSAR + explicit representation of different induced-fit models. |
| 6D-QSAR | As 5D-QSAR + simultaneous consideration of different solvation models. |
| 7D-QSAR | Such analysis comprises of real receptor or target-based receptor model data |

1.6.6.2. Based on chemometric methods

Based on the type of chemometric methods used, QSAR methods are classified as linear and non-linear. Linear methods include stepwise multiple linear regression (S-MLR), principal component analysis (PCA), partial least-squares (PLS), and genetic function approximation (GFA). Although, both PLS and GFA techniques can also be employed in developing non-linear models. Further, improvements in the chemometric field have also created several methods of building predictive models, including non-linear regression and algorithmic techniques like support vector machine (SVM), artificial neural networks (ANN), k-nearest neighbors (kNN), and Bayesian neural nets (BNN) (Roy & Mitra, 2011), etc.

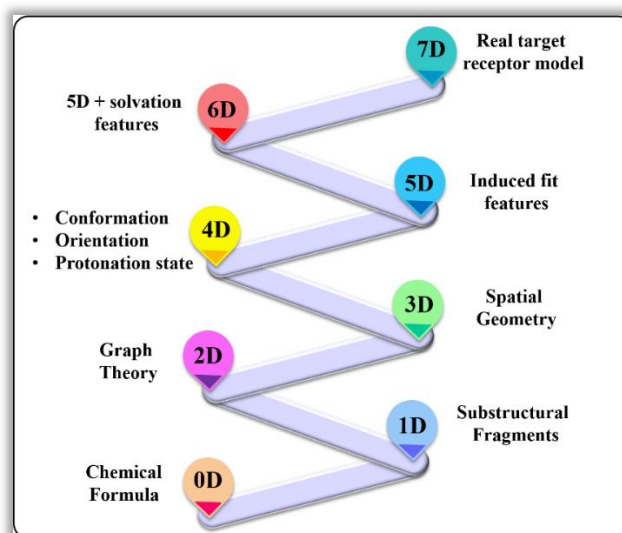


Figure 1.6. Classification of descriptors based on dimensionality.

1.6.6.3. Based on the number of the dependent variables

Depending on the number of dependent variables, QSAR can also be classified as single-target QSAR, and multi-target QSAR (see Figure 1.5). Studies have indicated that multi-target QSAR studies are highly beneficial in the case of complex diseases like AD, PD, and cancer.

1.6.5. QSAR and OECD Guidelines

QSAR/QSTR is an interdisciplinary study of chemistry, biology, and statistics. The predictions for the essential structural requirements needed for obtaining a molecule with optimized activity/toxicity provide a good platform for the synthesis of a relatively lesser number of chemicals with improved activity/ toxicity/property of interest. To achieve the aforementioned objectives, it is necessary to follow some guidelines adopted by the Organization for Economic Co-operation and Development (OECD) <http://www.oecd.org/dataoecd/33/37/37849783.pdf>. The guidelines suggest that a valid QSAR/QSPR/QSTR should have (Dearden et al., 2010; Gramatica, 2007):

- i) **Principle 1:** a defined end point;
- ii) **Principle 2:** an unambiguous algorithm;
- iii) **Principle 3:** a defined domain of applicability;
- iv) **Principle 4:** appropriate measures of goodness of fit, robustness, and predictivity;
- v) **Principle 5:** a mechanistic interpretation, if possible

A brief overview of the aforementioned OECD principles is explained below:

i) OECD Principle 1: A defined endpoint- This principle commands transparency to be maintained during the selection of endpoint data for modeling. It is expected that the QSAR models are to be developed using homogeneous datasets comprising single protocol/assay-generated response data. A QSAR scientist should take utmost care while verifying the experimental protocols, quality of the data, concentration unit, etc. thoroughly during compiling activity/ property/ toxicity data from varying sources. Another critical consideration to be maintained is the mechanism/ mode of action for all the chemicals used should be common. This implies that the compounds used for developing a model must work via the same mode/ mechanism of action. A defined endpoint has a noteworthy impact on developing QSAR models and the principal features can be summarized as:

- a) A well-defined endpoint must portray the variation of chemical structures within a dataset.
- b) Detailed information on the employed test protocols mentioning the factors able to cause variation, uncertainty, and possible deviation from standardized test guidelines.
- c) Differences in the protocols implemented must not make a marked difference in values for a given endpoint.
- d) Differences made within a protocol e.g., involving media, reagents, etc. must not be irrational.
- e) The chemical domain of the test protocol must encase the domain defined by the model.
- f) The endpoint measured using a given test protocol and the endpoint being modeled must be the same concerning specific assessment of chemical hazard.

ii) OECD Principle 2: An unambiguous algorithm- This principle states an unambiguous methodology to be used while developing predictive QSAR models. This comprises the methodology employed during data pre-treatment, dataset division, and the selection of features. Hence, this rule focuses to bring transparency in model building rendering it not only reproducible to others but also making it explanatory in achieving the endpoint estimates. OECD recognizes the following essential elements for maintaining methodological transparency during (Q)SAR model development.

- a) A dataset of chemical compounds along with their endpoint and descriptor values i.e., the QSAR data matrix.
- b) Clear depiction of the descriptor computation steps as well as their measurement.
- c) Description of the training and test sets along with a definite justification for the removal of outlier observations if any.
- d) Description on the mathematical models portraying the relationship between endpoint and descriptor, and the extracted chemical information thereof.
- e) Statistical parameters for judging the reliability of the prediction.

iii) Principle 3: A defined domain of applicability- The third principle of OECD portrays the importance of the chemical/ response domain of applicability. Any QSAR model developed using a set of chemicals possesses a distinct theoretical space providing a reliable predictive result within that domain. Netzeva et al. (2005) have defined the applicability domain of QSAR models as follows: *“The applicability domain (AD) of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability.”* The domain of applicability of a model using the training set molecules checks whether the prediction of test set molecules is trustworthy or not. The AD of a model depends on three major characteristics a) structural information, b) physicochemical feature, and c) response space. Because of the possible involvement of multiple mechanistic basis in various regulatory endpoints, (Q)SAR models can be developed on specific chemical classes those act via the same mechanism of action. Some of the methods for defining an AD include a range of individual descriptors, distance-based methods such as Euclidean distance, Mahalanobis distance, Manhattan distance, distance to model in the X-space approach etc. Another distance-based formalism involving Hotelling’s test and the associated leverage statistics can also be used. A warning value of the leverage (h^*) can be computed using the formula $3p'/n$ where n represents number of training set chemicals and p' is the number of descriptors plus one. The detection of an outlier can be performed using the confidence limits of the AD defining techniques.

Another approach could be the fragment-based technique where the test set/ query molecule can be split into structural fragments and checking can be done to verify whether the fragments are presented by the corresponding training set domain or not. The following points as identified by the OECD should be followed with respect to AD of a QSAR model.

- a) Defining confidence limits characterizing an AD.
- b) AD for structural alerts and fragment-based QSAR techniques.
- c) Assessment of the strength, limitations, and applicability of AD methods
- d) Implementation of tools that allows AD determination along with other statistical operations as an integrated operation.

iv) OECD Principle 4: Appropriate measures of goodness-of-fit, robustness, and predictivity-

The fourth OECD principle provides knowledge on the statistical verdict of stability and predictivity of a model. The internal model performance by fitness and robustness measure using a training set, and external predictivity using test set is measured. This provides a suitable balance between the extreme conditions namely overfitting and underfitting of the model-based prediction. Division of the dataset into training and test sets is one of the principal strategies to determine the internal stability and external predictivity of a developed QSAR model. In simpler terms, all such validation exercise is meant to verify the closeness of a prediction/ estimation by a model with respect to its experimental observation.

v) OECD Principle 5: A mechanistic interpretation, if possible- The fifth OECD principle attempts in aiding a good mechanistic basis for the response being modeled. Definite information on the mechanism of action of chemicals towards a process can guide the design and development of only desired analogs. Molecular descriptors play a crucial role in proving mechanistic information towards a modeled endpoint. Hence, various types of experimental as well as theoretical descriptors containing sufficient chemical diagnostic potential should be of interest for developing predictive QSAR models. Various expert systems on QSAR modeling here help the user to gather chemical knowledge towards a given process. The expert systems are usually characterized by their induced statistical rules and their ability to provide expert knowledge. A quick overview at the OECD guidelines for the development and validation of QSAR models is shown in **Figure 1.4**.

1.6.6. QSAR methodology

Predictive QSAR model development comprises several steps namely i) data preparation, ii) data analysis, iii) data validation, and iv) data interpretation where the 'data' refers to the response and predictor variables. It should be noted that mathematics plays only an abstract platform here to provide the quantitative correlation and hence the preparation and analysis of the chemical data must be operated carefully to avoid the loss of any essential chemical attribute. A brief discussion on the steps of QSAR modeling is presented below.

1.6.6.1. Data preparation: This component comprises the preparation of the QSAR data matrix composed of response and predictor variables.

- The biological/ toxicological/ physicochemical response is at first subjected to conversion into a molar unit followed by logarithmic transformation maintaining data uniformity.
- The next task involves drawing the chemical structures using suitable chemical drawing software applications such as ChemDraw, MarvinSketch, ChemSketch, etc. followed by saving them

in a suitable format e.g., MDL molfile (.mol). The drawn structures are usually subjected to a cleaning operation as implemented in the software to remove any error left due to bond length, angle, etc. The chemical structures can also be collected from public databases such as **PubChem** (<https://pubchem.ncbi.nlm.nih.gov/search/index.html>), **ZINC** (<https://zinc.docking.org/>), **ChemSpider** (<http://www.chemspider.com/>), etc. The desired stereochemical configuration needs to be checked while collecting chemical structures from public databases.

- Depending upon the purpose of modeling, the chemical structures drawn/ collected might be subjected to conformational analysis and energy minimization operation.
- The final files are submitted to descriptor computing software (Dragon, Padel or AlvaDesc) for the generation of theoretical predictor variables. At the initial stage, the computed descriptors can be subjected to a variance and correlation check to remove redundant chemical features and attributes with constant or near-constant values throughout the dataset. This is commonly known as data pretreatment. Furthermore, the user can employ different software applications for the generation of a separate class of descriptors all of which can be pooled together in a single data matrix along with the experimental variables (if any) and then subjected to data pretreatment operation.
- At this point, the user has a descriptor matrix containing many variables and a single column of the response (usually) which needs to be clubbed into a spreadsheet to form the final QSAR data matrix containing a column denoting a serial number of the compounds, a column for the response (activity/ property/ toxicity), and the descriptors both obtained from experimental and theoretical operations. An additional column denoting the name of the chemicals can be added for quick identification of any compound.

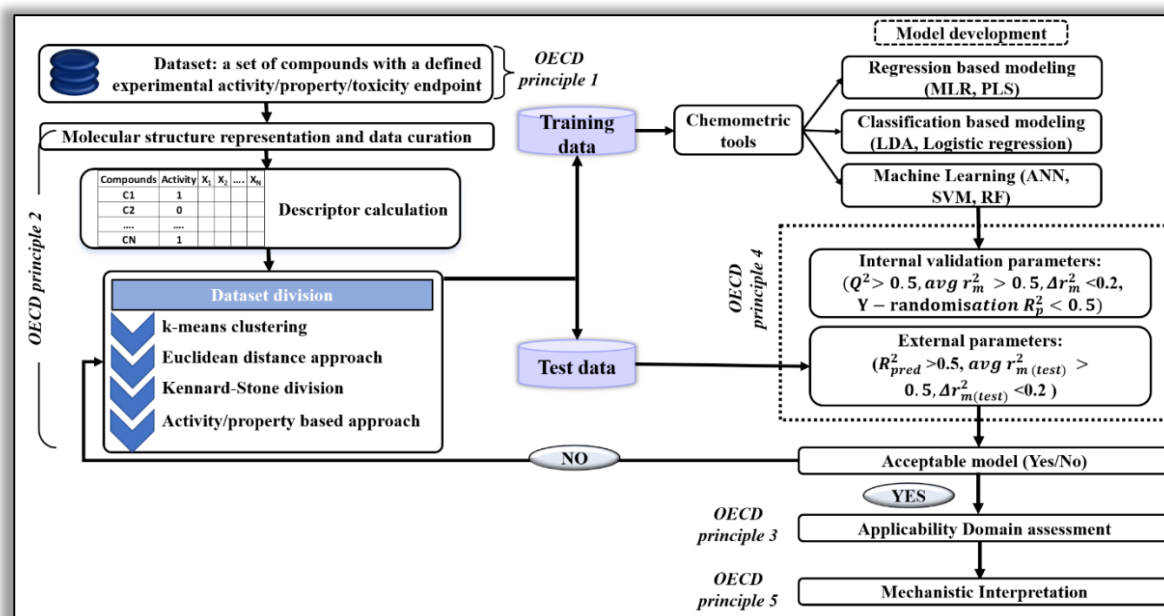


Figure 1.7. QSAR methodology based on OECD guidelines.

1.6.6.2. Data analysis: This component consists of dataset division, feature selection and model development.

- The performance of a predictive model is ascertained by dividing the whole dataset into a **training set** and a **test set** based on chemical similarity. The training set is employed for model development (i.e., the equation), while the test set (not used during model development) is used to judge the external predictivity of the model. The internal predictivity, i.e., the predictive performance of the model on the training set is also judged. Usually, a higher number of compounds are allotted to

the training set compared to the test set. The total dataset is divided such that the test set compounds lie within the chemical space of the training set, i.e., the training set becomes representative of the test set. The methods employed for the dataset division may involve a) treatment of the predictor variables, b) treatment of the response variable, and c) random selection. The first approach i.e., the predictor variable-based division first attempts to assign the divided compounds into separate groups or classes based on their chemical similarity determinable from a suitable operation on the descriptor matrix. This is followed by the selection of a user-defined fraction of compounds into training and test sets from each such obtained group. Some of the techniques to divide chemicals into groups are k-means clustering, Kennard Stone algorithm, sphere exclusion principle, principal component analysis (PCA) based selection, Kohonen's self-organizing map, statistical molecular design, Extrapolation-oriented test set selection, etc. In the response variable-based division approach, the compounds are assumed to be diverse based on their biological/physicochemical/ toxicological response values. Here, the whole data matrix is first sorted using the response column followed by the selection of a predefined fraction of compounds into training/ test set from different zones maintaining a pattern e.g., every fourth compound, etc. In the random division approach, compounds are randomly classified into training and test sets following a user-defined fraction. Sometimes a combination of response variable-based and predictor variable-based approaches may also be employed e.g., compounds may be assigned into different structurally similar groups using any of the above-mentioned techniques followed by the selection of compounds into training/ test set using the sorted response formalism separately from each group.

- **Selection of features** refers to the identification of the important predictor variables suitable for developing a correlation with the response variable. Many software applications are capable of generating hundreds or thousands of different molecular descriptors. Typically, only some of them are significantly correlated with the activity. Thus, appropriate feature selection tools must be used. Furthermore, many of the descriptors are intercorrelated. This has negative effects on several aspects of QSAR analysis. Some statistical methods require that the number of compounds is significantly greater than the number of descriptors. Various chemometric tools are employed for the selection of the potential variables with respect to an endpoint data from the whole descriptor matrix. The selected variables are then subjected to suitable statistical operations leading to the development of the final model. Some of the feature selection tools employed in chemometric modeling studies include stepwise variable selection, genetic algorithm, best subset selection, variable subset selection, factor analysis, etc.

- The **model development** step dictates that the selected best features are to be combined in a single equation employing an explicit formalism. Multiple linear regression (MLR), partial least squares (PLS), etc. are the algorithm employed for the development of quantitative regression-based equations while linear discriminant analysis (LDA) produces a classification model. All these techniques are preceded with a feature selection step and the techniques are known as stepwise-MLR, GFA-MLR, G/PLS (genetic PLS), PLS-DA (PLS followed by discriminant analysis), etc.

1.6.6.3. Model validation: Following the development of predictive models, the next essential task becomes the determination of its statistical reliability. The objective of QSAR analysis is not model development only but also to apply it for the prediction of response of untested/ new chemicals, it is necessary to ascertain its stability as well as predictivity. Various statistical metrics are computed to judge the model's fitness (R^2 , R_{adj}^2 , etc.), internal stability (Q_{LOO}^2 , $r_{m(LOO)}^2$) as well as external predictivity (R_{pred}^2 , $r_{m(Test)}^2$) and values above the threshold limits identify model acceptability. It may be noted that by 'internal stability' we aim to portray the stability of prediction determined using

the training set compounds only, i.e., compounds used for developing the model, while external predictivity refers to the judgment on test set prediction. Some additional metrics can also be employed to judge the overall predictivity e.g., $r_m^2(\text{overall})$. For the validation of discriminant model parameters such as sensitivity, specificity, accuracy, precision, F-value, receiver operating characteristic (ROC) analysis, etc. can be employed. Table 1.9 and 1.10 enlists major regression and classification validation metrics respectively.

1.6.6.4. Model interpretation: Once a QSAR model has been developed and has been considered acceptable from the values of the metrics, the final important part remains with the mechanistic interpretability of the modeled features. Establishing a suitable basis between the chemistry of the compounds and biological/ toxicological action or physicochemical property helps in understanding the mechanism of action involved. Accordingly, by combining the experimental results and observation from the model, one can explicitly explain each step of the process of behavioral manifestation of chemicals. Such knowledge is useful in designing and developing potent analogs.

The methodological workflow of QSAR modeling studies with reference to the OECD-recommended principles has been schematically presented in **Figure 1.7**.

1.6.7. Quantitative Structure Activity-Activity Relationship (QSAAR) modeling

Quantitative structure activity-activity relationship (QSAAR) models are mathematical expressions correlating two biological endpoints, with the aim to extrapolate any one explicit activity endpoint when the experimental data is not available. This advanced technique can overcome the additional cost of manifold experimental procedures. One endpoint acts as a predictor variable and offers to predict the other endpoint (De & Roy, 2021; Gajewicz-Skretna et al., 2021). QSAAR is also applicable when same endpoint is present for different species leading to an interspecies model. Here, the QSAAR or (quantitative activity–activity relationship) QAAR models predict an endpoint (which is a dependent variable) for a specific species employing the same endpoint (response in the form of activity, property, or toxicity) for another species along with selected structural and physicochemical features as a predictor or explanatory or independent variables (descriptors) (Kar et al., 2016; De et al., 2018). Thus, extrapolating data from one endpoint to another helps filling the data gaps without wasting time, money, and animal study maintaining the 3R's approach intended to a replacement, reduction, and refinement of animals.

Table 1.9. Validation metrics for regression modeling (Roy & Mitra, 2011).

| Parameters | Equation | Description |
|--|--|---|
| Determination coefficient (R^2) | $R^2 = 1 - \frac{\sum(Y_{obs} - Y_{calc})^2}{\sum(Y_{obs} - \overline{Y_{training}})^2}$ | Metric to check the goodness-of-fit of a regression model. It measures the variation of observed data with the predicted ones. The maximum possible value for R^2 is 1, which defines a perfect correlation. Y_{obs} denotes the observed response values for the training set, and Y_{calc} denotes the calculated response values for the training set of compounds. $\overline{Y_{training}}$ is the mean observed response of the training set compounds. |
| Explained variance or adjusted R^2 (R_{adj}^2) | $R_{adj}^2 = \frac{\{(n - 1) X R^2\} - p}{n - p - 1}$ | Modified version of the determination coefficient. The R_{adj}^2 parameter incorporates the information of the number of samples and the independent variables used in the model. n is the number of training set compounds and p is the number of predictor variables. |
| Leave-one-out cross-validation (Q_{LOO}^2) | $Q_{LOO}^2 = 1 - \frac{\sum(Y_{obs(training)} - Y_{pred(training)})^2}{\sum(Y_{obs(training)} - \overline{Y_{training}})^2}$ | Cross-validated R^2 (Q^2) is checked for internal validation. $Y_{obs(training)}$ is the observed response, and $Y_{pred(training)}$ is the predicted response of the training set molecules based on the leave-one-out (LOO) technique |
| Predictive R^2 or R_{pred}^2 or Q_{ext}^2 ($F1$) | $Q_{ext(F1)}^2 = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \overline{Y_{training}})^2}$ | This metric employed for judging external predictivity. It is a measure of correlation between the observed and predicted data of test set. $Y_{obs(test)}$ is the observed response, and $Y_{pred(test)}$ is the predicted response of the test set molecules. $\overline{Y_{training}}$ denotes the mean observed response of the training set. |
| Q_{ext}^2 ($F2$) | $Q_{ext(F2)}^2 = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \overline{Y_{test}})^2}$ | It helps in the judgment of predictivity of a model using the test set ($\overline{Y_{test}}$). |
| Q_{ext}^2 ($F3$) | $Q_{ext(F3)}^2 = 1 - \frac{[\sum(Y_{obs(test)} - Y_{pred(test)})^2]/n_{test}}{[\sum(Y_{obs(train)} - \overline{Y_{training}})^2]/n_{train}}$ | $Q_{ext(F3)}^2$ is measured to determine external predictivity employing both training and test set features. $Y_{obs(test)}$ is the observed response, and $Y_{pred(test)}$ is the predicted response of the test set molecules. $Y_{obs(training)}$ is the observed response and $\overline{Y_{training}}$ denotes the mean observed response of the training set molecules. The threshold for $Q_{ext(F3)}^2$ is 0.5. |

Concordance correlation coefficient (CCC)

$$CCC = \frac{p_c}{\frac{2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})}}$$

Root mean square error in predictions ($RMSE_p$)

$$RMSE_p = \sqrt{\frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{n_{test}}}$$

r_m^2 metrics

$$\bar{r}_m^2 = \frac{r_m^2 + r'_m{}^2}{2} \quad \text{and} \quad \Delta r_m^2 = |r_m^2 - r'_m{}^2|$$

$$\text{where } r_m^2 = r^2 X(1 - \sqrt{r^2 - r_0^2})$$

$$r'_m{}^2 = r^2 X(1 - \sqrt{r^2 - r_0^2})$$

Predicted residual sum of squares ($PRESS$)

$$PRESS = \sum (Y_{obs} - Y_{pred})^2$$

Standard deviation of error of prediction ($SDEP$)

$$SDEP = \sqrt{\frac{PRESS}{n}}$$

Mean absolute error (MAE)

$$MAE = \frac{1}{n} X \sum |Y_{obs} - Y_{pred}|$$

The concordance correlation coefficient (CCC) measures both precision and accuracy detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. 'n' denotes the number of compounds, and x_i and y_i signify the mean of observed and predicted values, respectively.

It gives a measure of model external validation. A lower value of this parameter is desirable for good external predictivity.

r^2 is the squared correlation coefficient value between observed and predicted response values, and r_0^2 and $r'_0{}^2$ are the respective squared correlation coefficients when the regression line is passed through the origin by interchanging the axes. For the acceptable prediction, the value of all Δr_m^2 metrics should preferably be lower than 0.2 provided that the value of \bar{r}_m^2 is more than 0.5.

Sum of squared differences between experimental and predicted data. Y_{obs} and Y_{pred} correspond to the observed and LOO predicted values.

The value of standard deviation of error of prediction ($SDEP$) is calculated from $PRESS$. N refers to the number of observations.

This is also known as average absolute error (AAE) and is considered a better index of errors in the context of predictive modeling studies.

Table 1.10. Validation metrics for classification modeling (De et al., 2022).

| Sl No. | Classification metric | Equation |
|--------|---------------------------------------|--|
| 1 | Sensitivity | $\text{Sensitivity} = \frac{TP}{TP + FN}$ |
| 2 | Specificity | $\text{Specificity} = \frac{TN}{TN + FP}$ |
| 3 | Precision | $\text{Precision} = \frac{TP}{TP + FP}$ |
| 4 | Accuracy | $\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$ |
| 5 | F-measure | $F - \text{measure}(\%) = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Sensitivity}}}$ |
| 6 | G-means | $G - \text{means} = \sqrt{\text{Specificity} \times \text{Sensitivity}}$ |
| 7 | Cohen's Kappa (κ) | $P_r(a) = \frac{(TP + TN)}{(TP + FP + TN + FN)}$ $P_r(e) = \frac{\{(TP + FP) \times (TP + FN)\} + \{(TN + FP) \times (TN + FN)\}}{(TP + FN + FP + TN)^2}$ $\text{Cohen's } K = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$ |
| 8 | Mathews correlation coefficient (MCC) | $MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$ |

1.7. Non-QSAR *in silico* techniques

1.7.1. Molecular Docking

Molecular docking is a “structure-based drug design” (SBDD) computational method used for the investigation of the behavior of small molecules at the binding site of the target protein and predict the ligand-receptor complex structures (Meng et al., 2011). Docking plays an important role in the prediction of a stable protein-ligand complex orientation and the ligand-receptor interaction details. This helps in the proper investigation and interpretation of the vital mechanism of the biologically active compounds. Molecular docking is broadly classified into three sub-categories namely, (i) protein-small molecule docking, (ii) protein-protein docking, and (iii) protein-nucleic acid docking. The docking is enacted out using two steps: 1) prediction of a stable conformation and orientation of ligand, 2) evaluation of binding affinity and binding orientation of the ligand at the active binding site (Dar & Mir, 2017). Molecular docking is carried out in basic four steps 1) Receptor/Target selection and its preparation, 2) Small molecule/Ligand Preparation, 3) Molecular docking and 4) Docking analysis. Basic tools and online servers available for molecular docking are AutoDock Vina (<http://vina.scripps.edu/>), Schrodinger software (<https://www.schrodinger.com/platform>), Molegro virtual docker software 6.0 (MVD) (<https://molegrovirtualdocker.weebly.com/>), Biovia Discovery Studio (<https://www.3dsbiovia.com/>), “Achilles” Blind Docking Server (<https://bio-hpc.ucam.edu/achilles/>) and FlexX (<https://www.biosolveit.de/FlexX/>). Molecular docking finds a vast application in the drug discovery process such as in lead optimization, hit identification, drug-DNA interaction studies, chemical mechanism studies, structure-activity studies, combinatorial library design and assistance in X-ray crystallography in substrate binding (Agarwal & Mehrotra, 2016; Meng et al., 2011).

1.7.2. Virtual Screening (VS)

The discovery of innovative leads with potential interaction with specific targets is of central importance to early-stage drug discovery, conventionally achieved by wet-lab high-throughput screening (HTS). However, the high cost and low hit rate associated with HTS have stimulated the development of computational alternatives and the broad application of cheaper and faster screening *in silico* (Clark, 2008; Ripphausen et al., 2010). To meet the demands of the economically driven pressure of industry computational chemistry (CADD, molecular docking, etc) combined with virtual screening have come up as the newest and fastest method in order to develop new chemical entities. Structure-based approaches (QSAR, molecular docking) have a mounting number of success rates and are arguably the most widely applied one in practice (Clark, 2008).

1.7.3. Read-Across

In the European Union (EU), the European Chemicals Agency (ECHA) has defined read-across (RA) for chemicals in general (Oomen et al., 2015) as a technique for predicting endpoint information for one substance (target substance) by using data for the same endpoint from another substance or other substances (source substances). The RA works on the principle based on structural similarity, i.e., following an assumption that similar structures should exhibit similar physicochemical, biological, and toxicological properties. RA acts like local QSAR models, wherein, chemically or biologically similar compounds are used for the weighted prediction of target test compounds (Chatterjee et al., 2022a). The following four schemes have been proposed in read-across data gap filling: (i) one-to-one, (ii) one-to-many, (iii) many-to-one, and (iv) many-to-many.

1.8. Application of QSARs for the design of PET and SPECT imaging agents

Quantitative structure–activity relationships (QSAR) studies, as progressive tools in modeling and prediction of many physiochemical properties, allow cost savings by reducing the laboratory resources needed and the time required to investigate and design new compounds by desired properties (Roy et al., 2015b). QSAR techniques aim to develop consistent relationships between any property or activity and physicochemical properties for a series of compounds so that these “rules” can be used to evaluate new chemical entities (Roy et al., 2015a; Roy, 2015). Different QSAR/QSPR methodologies have been utilized to model the complex formation of different metal ions with organic ligands and to design PET and SPECT imaging agents.

Relatively few attempts have been made to apply the QSAR/QSPR techniques in modeling and designing imaging agents with applications in cancer and neurodegenerative diseases. A few studies are briefly explained below:

- Kovac et al. (Kovac et al., 2010) have reported 3D-QSAR studies for vesamicol and benzovesamicol derivatives as PET radioligands for the vesicular acetylcholine transporter which can be used for quantitative visualization of early degeneration of cholinergic neurons. Linear Genetic Function Approximation (GFA) model and a 3D QSAR model confirmed the spatial impact on affinity for VACHT via steric descriptors and the Van der Waals coefficient.
- Hocke et al. (Hocke et al., 2008) have reported computer-assisted prediction of D3 selectivities of new fluoroalkoxy-substituted receptor ligands employing 3D-QSAR analysis. 3D-QSAR models were able to predict subtype selectivities of dopaminergic test compounds. Receptor binding experiments confirmed the computer-assisted molecular design revealing subnanomolar D3 affinities and excellent selectivity profiles.
- Long and Liu (Long & Liu, 2010) have reported QSAR models for predicting the radiosensitization effectiveness of nitroimidazole compounds by combining heuristic method (HM) and projection pursuit regression (PPR), for descriptor selection and correlation modeling.
- Yang et al. (Yang et al., 2015) have reported 3D-QSAR studies for structurally identical ^{18}F - and ^{125}I -labeled benzyloxybenzene derivatives which could be used for PET/SPECT Imaging of β -Amyloid Plaques. Molecular docking and 3D-QSAR models predicted excellent binding to A β fibers.
- Salahinejad and Mirshojaei (Salahinejad & Mirshojaei, 2016) have established molecular modeling methods for predicting the liver and kidney uptakes of Tc-99m labeled quinolone antibiotics. Three-dimensional quantitative-activity relationships (3D-QSAR) models were developed using comparative molecular field analysis and grid-independent descriptors procedures.
- Salihinejad (Salahinejad, 2015) also reported 3D and 2D QSPR to model the complexation formation of bifunctional coupling agents with $^{64}\text{Cu}(\text{II})$ and $^{67/68}\text{Ga}(\text{III})$ radiometal ions. The information obtained could be very useful to design the most efficient ligands and find new matching chelators to radiometals for radiopharmaceutical applications.
- Ambure and Roy (Ambure & Roy, 2015) used a congeneric series of 44 imaging agents, including 17 PET and 27 SPECT imaging agents to understand the structural features required for having essential binding affinity against A β plaques. 2D-quantitative structure-activity relationship (2D-QSAR) and group-based QSAR (G-QSAR) models have been developed using genetic function approximation (GFA) and validated using various statistical metrics.

The present research aims at the development of predictive chemometric model for PET and SPECT imaging agents with application in cancer, neurodegenerative diseases like Alzheimer’s and Parkinson’s disease, vesicular acetylcholine transporters, and imaging of Dopamine receptors. The work also aspires to find multifunctional imaging agents, i.e., agents that bind to the receptors and

also provide therapeutic benefit. Furthermore, it was strived to obtain predictive models for nitroaromatics to study their radiosensitization properties. The structure and biological activity data was collected from the literature as provided in the **Present Work** section. The collected data was first subjected to data curation workflow as available in <https://sites.google.com/site/dtclabdc/> . Molecular descriptors were calculated using various descriptor tools as discussed in the **Methods and Materials** section. For QSAR model development, attempt was made to develop more straightforward and interpretable models like multiple linear regression, partial least squares, genetic function approximation derived models etc. Rigorous procedures of model validation involving cross-validation, Y-scrambling, external validation and tests for applicability domain etc. was performed using strategies available in <https://dtclab.webs.com/software-tools> in order to select the best predictive models. All model development and validation were done based on the OECD recommended strategies. A proper mechanistic interpretation and conclusion is provided in the **Results and Discussions** section.

CHAPTER 2

PRESENT WORK

Chapter 2: Present Work

Molecular imaging technologies involving radionuclides such as positron emission tomography (PET) and single photon emission computed tomography (SPECT) have a significant impact on many aspects of healthcare (Pimlott & Sutherland, 2010). For example, these techniques are used for detecting diseases in early stages (screening), identifying extent of disease, selecting disease-and patient-specific treatment (personalized medicine), applying a directed or targeted therapy, and measuring molecular-specific effects of treatment (Pysz et al., 2010). These imaging agents have been used for diagnostic imaging in different disease conditions including *cancer* and *neurodegenerative diseases*. The main advantage of in vivo molecular imaging is its ability to characterize diseased tissues without invasive biopsies or surgical procedures, and with this information in hand, a more personalized treatment planning regimen can be applied. Molecular imaging has also been used in various aspects of drug development such as understanding drug action and establishing dosage regimens and treatment strategies. Molecular imaging has the potential to improve therapeutic monitoring by, for example, measuring the direct effect of a drug at an earlier time point before overt morphological-anatomical changes become visible on imaging (Pysz et al., 2010). In basic terms, molecular imaging effectively allows the non-invasive visualisation, characterisation and measurement of biological processes at the molecular, cellular, whole organ or body level using specific imaging probes. These molecular imaging probes (sometimes called tracers due to the subpharmacological amounts administered) provide an analytical signal which is detected by a particular method resulting in either a two- or three-dimensional image (Pimlott & Sutherland, 2010). Of the non-invasive imaging technologies available, PET and SPECT are the most sensitive techniques for imaging function in vivo.

Imaging agents (Ametamey et al., 2008; Meikle et al., 2005) used for PET are radiolabelled with radionuclides that decay by the emission of a positively charged particle called the positron. A significant number of PET nuclides exist for the incorporation into biomolecules. The isotopes generally selected for PET imaging have half-lives comparable to the half-life of the process being imaged. While most of these in theory can be used for PET imaging, it is mainly ^{11}C or ^{18}F labelled molecular probes which are employed. The main advantage of PET imaging over SPECT is that the radiolabelled imaging agent is essentially indistinguishable from its nonradioactive counterpart. Carbon is the main constituent of naturally occurring compounds and thus, replacement of carbon-12 with carbon-11 produces only a negligible isotope effect. Fluorine is not normally found in biomolecules, but the substitution of a hydrogen atom or a hydroxyl group by a fluorine atom is a commonly applied bioisosteric replacement. Fluorine and hydrogen are similar in size (van der Waal's radii of hydrogen and fluorine are 1.20 and 1.35 Å, respectively) and thus, this replacement induces only a slight steric perturbation. Fluorine is considerably more electronegative than hydrogen but this change in the electronic properties of the molecule can quite often be advantageous producing molecular probes with improved potency. SPECT imaging (Ametamey et al., 2008; Meikle et al., 2005) uses radionuclides that directly emit γ -rays, such as iodine-123 (^{123}I) and technetium-99m ($^{99\text{m}}\text{Tc}$), and are generally of a lower energy than those used for PET. PET systems are generally more sensitive than SPECT systems which in turn translates into a higher resolution for PET compared with SPECT. However, the use of longer-lived radionuclide and the relatively lower costs of gamma cameras make SPECT imaging much more widely available for clinical use than PET scanners.

Generally, SPECT isotopes have a considerably longer half-life than PET isotopes which makes SPECT tracers more available for imaging and longer radiosynthesis times make them more practical. In contrast, the shorter half-life of the PET isotopes limits the availability of PET imaging, requiring

an on-site cyclotron, and makes radiolabelling protocols more challenging. The longer half-life of SPECT radionuclides also has the advantage of enabling SPECT imaging studies to be conducted over longer time periods, whereas imaging studies using the shorter half-life PET radionuclides may require more complicated modelling. New imaging agent development for complex diseases like Alzheimer's disease (AD), Parkinson's disease (PD), cancer etc. is becoming more and more painstaking, expensive time-consuming. In such cases, *in silico* or computational technique serves as an efficient tool for identifying/screening and optimizing the lead molecules to overcome the tedious and expensive procedure of synthesis and analysis of at least thousands of possible bioactive molecules. In recent years, computer-aided drug design (CADD) has been extensively explored for facilitating lead discovery and optimization with advantages in terms of both high speed and low cost that finally increases the probability of success in the drug development process. A variety of *in silico* methods have evolved in CADD that have two major application areas, i.e., ligand-based drug design and structure-based drug design. Structure-based drug design techniques like molecular docking rely on three-dimensional (3D) knowledge of the target protein (enzyme or receptor) structure and its active/binding site to investigate various interactions as well as binding energy. On the other hand, ligand-based drug design techniques like QSAR and ligand-based pharmacophore modeling rely on knowledge of ligands that interact with the target of interest and are usually very helpful approaches when the structure of the target is not known. Structure-based and ligand-based drug design techniques together become a powerful tool to study potential ligands for one or more targets.

In the present thesis work, several *in silico* techniques were employed to study potential PET and SPECT imaging agents targeted against various neurodegenerative diseases and cancer. The main purpose of the dissertation was to utilize various *in silico* tools for identifying and optimizing the potential PET or SPECT candidates against several receptors involved in neurodegenerative diseases or cancer pathogenesis. We have also developed a number of predictive QSAR models studying radiosensitization effectiveness of various nitroaromatic compounds to understand their role in hypoxic cancer cells. Although we have employed several *in silico* techniques such as QSAR, molecular docking, virtual screening etc., but the major part of the work deals with the development of predictive and statistically robust QSAR models. Quantitative structure–property/activity relationships (QSPR/QSAR) studies, as progressive tools in modeling and prediction of many physicochemical properties, allow cost savings by reducing the laboratory resources needed and the time required to investigate and design new compounds by desired properties (Roy et al., 2015b). The aim of QSAR techniques is to develop consistent relationships between any property or activity and physicochemical properties for a series of compounds so that these “rules” can be used to evaluate new chemical entities (Roy et al., 2015a; K Roy, 2015). The concept of QSAR has been applied to modeling imaging agents and metallic radiopharmaceuticals only to a limited extent. There is enough scope of further application of QSAR theories in this area for refinement of the existing models leading to development of new models with enhanced robustness and predictivity which can be used to design new imaging agents and radiopharmaceuticals having potential applications in imaging analysis for diseases like cancer and neurodegeneration.

The present work is further categorised into two different sections, where the first section (Section I) involves development of predictive chemometric models for PET and SPECT imaging agents with application in neurodegenerative diseases (like AD and PD), vesicular acetylcholine transporters and imaging of dopamine receptors. This part also includes development of predictive models for multifunctional imaging agents. The second section (Section II) involves development of predictive models for radiosensitization effectiveness of various nitroaromatic compounds to understand their role in hypoxia.

2.1. Datasets employed for the development of different QSAR models

For performing the requisite *in silico* studies, several datasets were collected from various reliable sources as mentioned in **Table 2.1**.

Table 2.1. Datasets employed in the present work.

| Datasets | Target/Endpoint | No. of compounds | Class of compounds | References |
|----------|--|------------------|-----------------------------------|--|
| I-A | Binding affinity towards amyloid beta in Alzheimer's disease | 38 | PET imaging agents | (Cohen et al., 2012; Herholz & Ebmeier, 2011; H. F. Kung et al., 2010; Mathis et al., 2003; Ono et al., 2006; Schilling et al., 2016; Zhu et al., 2014) |
| | | 73 | SPECT imaging agents | (Alagille et al., 2011; Fuchigami et al., 2015; M. P. Kung et al., 2002; Mathis et al., 2003; Maya et al., 2009, 2016; Ono et al., 2013; Pan et al., 2013; Qu et al., 2007; Yang et al., 2013) |
| I-C | Binding affinity towards tau protein in Alzheimer's disease | 31 | Both PET and SPECT imaging agents | (Declercq et al., 2016; Hashimoto et al., 2015; Matsumura et al., 2011; Okamura et al., 2005, 2013; Ono et al., 2011; Pan et al., 2013; Tago et al., 2014, 2016) |
| II | Binding affinity and selectivity data towards A _{2A} adenosine receptors for diagnosis of Parkinson's disease | 35 | Xanthine ligand-based PET tracers | (Tamiji et al., 2018) |
| III | Binding affinity towards Dopamine (D2) receptor for diagnosis of Parkinson's disease | 34 | PET tracers | (Baldessarini et al., 1991; Chumpradit et al., 1993; Gao, Ram, et al., 1990; Murphy et al., 1990; Sipos et al., 2008; Søndergaard et al., 2005; Tóth et al., 2006; Vasdev et al., 2006) |
| IV | Binding affinity towards vesicular acetylcholine transporter (VACHT) | 19 | PET imaging agents | (Kovac et al. 2010; Tu et al. 2015, 2009) |
| V | Radiosensitization effectiveness expressed as C _{1.6} (pC _{1.6}) | 84 | Nitroimidazoles | (Long & Liu, 2010) |
| VI | Sensitizer Enhancement Ratio (SER) and Survival Ratio (SR) | 21 | Nitroimidazole sulfonamides | (Bonnet et al., 2018) |
| VII-A | Radiosensitization effectiveness (pC _{1.6}) | 18 | Nitrofurans | (Naylor et al., 1990) |
| VII-B | | 11 | Nitrothiophenes | (Threadgill et al., 1991) |
| VII-C | | 84 | Nitroimidazoles | (Long & Liu, 2010) |

2.1.1. Dataset I A-C (Study 1)

The experimental binding affinity (K_i) data for 38 PET (Cohen et al., 2012; Herholz & Ebmeier, 2011; H. F. Kung et al., 2010; Mathis et al., 2003; Ono et al., 2006; Schilling et al., 2016; Zhu et al., 2014) and 73 SPECT imaging agents (Alagille et al., 2011; Fuchigami et al., 2015; M. P. Kung et al., 2002; Mathis et al., 2003; Maya et al., 2009, 2016; Ono et al., 2013; Pan et al., 2013; Qu et al., 2007; Yang et al., 2013) against beta amyloid ($A\beta$) plaques and 31 (25 PET compounds and 6 SPECT compounds) imaging agents (Declercq et al., 2016; Hashimoto et al., 2015; Matsumura et al., 2011; Okamura et al., 2005, 2013; Ono et al., 2011; Pan et al., 2013; Tago et al., 2014, 2016) against tau protein were obtained from different literatures. Due to limited number of data available for tau protein, the PET and SPECT data were combined to form a single dataset. In the present study, the binding affinity values for both the PET and SPECT dataset compounds expressed as K_i (nM) were converted to negative logarithm of K_i (pK_i) values. Following the strict Organization for Economic Co-operation and Development (OECD) guidelines, significant descriptors were selected from the large initial pool of descriptors using multilayered variable selection strategy using the double cross validation (DCV) method followed by the best subset selection (BSS) method prior to the development of the final PLS models. The developed models showed significant statistical performance and reliability. Molecular docking studies have been performed to understand the molecular interactions between the ligand and receptor, and the results are then correlated with the structural features obtained from the QSAR models. Furthermore, we have also designed some imaging agents based on the information provided by the models developed and some of them are predicted to be similar to or more active than the most active imaging agents present in the original dataset

2.1.2. Dataset II (Study 2)

The experimental binding affinity and selectivity data of 35 xanthine ligand-based PET tracers were taken from a previously published literature (Tamiji et al., 2018) and applied for QSAR modeling to determine the essential structural features needed for binding affinity and explore the structural requirements necessary to be present in the antagonists for selectivity towards A_{2A} adenosine receptors. The experimental values of selectivity and binding affinity (K_i) ranged from 0.1–20 nM and 7.84–16,500 nM respectively. The experimental values were converted into negative logarithm scale during modeling and were used as independent values. No compounds with binding affinity data were removed during modeling but some compounds (mentioned in Materials and Methods Section) with no experimental selectivity values were eliminated during modeling. Here, the binding affinity and selectivity were separately used as endpoints or independent variables in modeling. The division of the dataset into training and test sets was done using a random method, while the feature selection for the binding affinity was done using Genetic Algorithm (GA). The best model with five descriptors was obtained using the spline option in the GA run. QSAR models with four descriptors were also developed for $A_{2A}R$ selectivity, where significant descriptors were selected from the large pool of descriptors using stepwise regression method followed by Best Subset Selection (BSS) method. Furthermore, to improve the quality of the external predictions, we used the “Intelligent Consensus Predictor” tool (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/). Both the models showed robustness in terms of statistical parameters. Molecular docking studies have been carried out to understand the molecular interactions between the ligand and receptor, and the results are then correlated with the structural features obtained from the QSAR models. Furthermore, the information derived from the newly found descriptors gives an insight for the development of new candidate PET tracers for the use in PD.

2.1.3. Dataset III (Study 3)

Dopamine (D2) receptor binding affinity (K_i) data of 34 PET imaging agents was taken from different literatures as mentioned in Table 2.1. The experimental binding affinity for all the compounds was measured using the same assay protocol, i.e., rat striatal homogenate (RSH) assay method. This data was applied in the development of a 2D-QSAR model to determine the essential structural features required for good binding to the D2 receptor. The binding affinity (K_i) values for the PET imaging agents were converted to their negative logarithm (pK_i) form and then used for modeling. The present study explores quantitative structure—activity relationship analysis of 34 PET imaging agents targeted toward dopamine D2 receptor. The dataset division into training and test sets was done using Euclidean distance division method, while the feature selection was done by double cross-validation-genetic algorithm method. Finally, a five-descriptor partial least squares regression model was derived after carrying out the best subset selection applied on the significant descriptors. The developed model showed robustness in terms of statistical parameters. Finally, the structural information derived from the model descriptors gives an insight for the development of new candidate D2-PET imaging for the use in PD.

2.1.4. Dataset IV (Study 4)

In this study, 2D quantitative structure-activity relationship (2D-QSAR) models for 19 positron emission tomography (PET) imaging agents targeted against presynaptic vesicular acetylcholine transporter (VAcHT) were developed. VAcHT assists in the transport of ACh into the presynaptic storage vesicles and it becomes one of main targets for the diagnosis of various neurodegenerative diseases. For our present work, the binding affinity (K_i) values of 19 PET imaging agents acting against vesicular acetyl choline transporter was procured from different previously published literature (Kovac et al., 2010; Tu et al., 2009, 2015). The aim was to understand the important structural features of the PET imaging agents required for their binding with VAcHT. This was done by feature selection using Genetic Algorithm followed by the Best Subset Selection method and developing a Partial Least Squares- based 2D QSAR model using the best feature combination. The developed QSAR model showed significant statistical performance and reliability. Using the features selected in the 2D-QSAR analysis, we have also performed similarity-based chemical read-across predictions and obtained encouraging external validation statistics. Further, molecular docking analysis was also performed to understand the molecular interactions occurring between the PET imaging agents and the VAcHT receptor. The molecular docking results were correlated with the QSAR features for better understanding of the molecular interactions. This research serves to fulfil the experimental data gap, highlighting the applicability of computational methods in the PET imaging agents' binding affinity prediction.

2.1.5. Dataset V (Study 5)

Nitroimidazoles and related analogues are efficient radiation sensitivity enhancers, and they particularly work on hypoxic tumor cells. A data of 86 nitroimidazoles possessing radiosensitizing properties are used for two-dimensional QSAR (2D-QSAR) study (Long & Liu, 2010). Radiosensitization capacities of the compounds can be understood by radiosensitization effectiveness, expressed as $C_{1.6}$, which can be represented as the corresponding concentration of a given compound when its sensitization enhancement ratio (SER) accomplishes 1.6. A higher value of $C_{1.6}$ indicates lower bioactivity of radiosensitization effectiveness. For analysis purpose, the source literature had converted the endpoint $C_{1.6}$ to its negative logarithmic scale ($pC_{1.6}$, where $pC_{1.6} = -\log(C_{1.6})$). Two compounds (one radical and one salt) were removed and the final dataset of 84 compounds is used for model development. In the current study, we have developed two partial least squares (PLS) regression-based two-dimensional quantitative structure-activity relationship (2D-QSAR) models using the novel class of 84 nitroimidazole compounds to understand their radiosensitization

effectiveness (pC1.6). Feature selection was done by genetic algorithm along with stepwise regression, while model validation was performed using various stringent validation criteria following the strict rules of OECD guidelines of QSAR validation. The variables included in the models were obtained from Dragon (version 7.0) and simplex representation of molecular structures (SiRMS) (version 4.1.2.270) software. The developed models were robust, externally predictive, and useful tools to predict the radiosensitization effectiveness of nitroimidazole compounds. True external prediction was carried out using a group of six nitroimidazole derivatives and the model reliability was checked using the Prediction Reliability Indicator tool (<http://dtclab.webs.com/software-tools>). Furthermore, the developed models will give an insight for development of new radiosensitizers with enhanced radiation sensitivity.

2.1.6. Dataset VI (Study 6)

In vitro radiosensitization data of selected compounds involving sensitizer enhancement ratio (drug SER) and survival ratio (drug SR) was obtained from a previously published research work (Bonnet et al., 2018). A dataset of 21 compounds was selected for 2D-QSAR modeling. The present study explores the features essential to show radiosensitization properties by nitroimidazole sulphonamide derivatives using QSAR and quantitative structure activity-activity relationship (QSAAR) modelling (Lessigiarska et al., 2006). Two dimensional (2D) descriptors obtained from Dragon and SiRMS software were utilised during the development of well validated models. A small dataset of nitroimidazole sulfonamides is used for modelling in the current study where splitting of dataset into training and test sets would cause loss of chemical information leading to unreliable models. The models were developed using the small dataset modeler software (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/), and model validation was performed using various stringent validation criteria. The developed models are robust, predictive, and should be useful tools to predict the radiosensitization of nitroimidazole sulfonamides. Furthermore, we have used the “prediction reliability indicator” tool to check the predictive ability of the developed models using 14 external nitroimidazole sulfonamide derivatives.

2.1.7. Dataset VII A-C (Study 7)

The radiosensitization effectiveness (pC1.6) data for three nitroaromatics datasets (nitrofurans, nitrothiophenes and nitroimidazoles) were obtained from the previously published literature (Long & Liu, 2010; Naylor et al., 1990; Threadgill et al., 1991). The datasets comprised 18 nitrofurans analogues, 11 nitrothiophenes and 84 nitroimidazole derivatives in the composite set. ‘C1.6’ is a term used to explain the radiosensitization capacities; this is the molar concentration of the compound required to give a sensitizer enhancement ratio (SER) of 1.6. Thus, lower value for C1.6 will give greater sensitizing efficiency. For an efficient analysis, the C1.6 values were converted into their negative logarithmic scale (pC1.6). The work comprises two parts: (i) local modeling using individual datasets; and (ii) global modeling by clubbing the three datasets. The two-dimensional descriptors were calculated using Dragon (version 7.0) software. The developed models were obtained using various feature selection techniques applied in “Small Dataset Modeling” and “Double Cross Validation” tools available from <https://dtclab.webs.com/software-tools>. Finally, the models were validated using stringent metrics following the Organisation for Economic Co-operation and Development (OECD) guidelines. The developed models are robust, predictive, and are useful tools to predict the radiosensitization of newly developed nitroaromatics. Furthermore, the global model was used to predict two external sets comprising 10 and 47 compounds, and the prediction ability was validated using the “Prediction Reliability Indicator” tool.

CHAPTER 3

MATERIALS AND METHODS

Chapter 3: Materials and Methods

The present dissertation aims at implementing a transparent methodological framework for the development of predictive QSAR models for PET and SPECT imaging agents targeted against various neurodegenerative and oncological diseases. We have endeavoured to maintain explicitness for computation of the descriptors, thinning of the variable matrix, selection of potential features as well as judgment of robustness and predictivity of the models. In this section, we have described here the details of datasets comprising the structures along with their **binding affinity** or **radiosensitivity** data and methodologies employed to carry out the *in silico* studies namely, QSAR and virtual screening. The section has been divided in the following parts:

- Details of datasets consisting chemical structures along with their activity or toxicity data.
- General description of methods implemented for developing QSAR models.
- Study wise specific description of methodologies utilized in each study.

3.1. Study 1: Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease

3.1.1. The dataset and structure curation

The experimental binding affinity (K_i) data for 38 PET (Cohen et al., 2012; Herholz & Ebmeier, 2011; H. F. Kung et al., 2010; Mathis et al., 2003; Ono et al., 2006; Schilling et al., 2016; Zhu et al., 2014) and 73 SPECT imaging agents (Alagille et al., 2011; Fuchigami et al., 2015; M. P. Kung et al., 2002; Mathis et al., 2003; Maya et al., 2009, 2016; Ono et al., 2013; Pan et al., 2013; Qu et al., 2007; Yang et al., 2013) against beta amyloid ($A\beta$) plaques and 31 (25 PET compounds and 6 SPECT compounds) imaging agents (Declercq et al., 2016; Hashimoto et al., 2015; Matsumura et al., 2011; Okamura et al., 2005, 2013; Ono et al., 2011; Pan et al., 2013; Tago et al., 2014, 2016) against tau protein were obtained from different literatures. Due to limited number of data available for tau protein, the PET and SPECT data were combined to form a single dataset. In the present study, the binding affinity values for both the PET and SPECT dataset compounds expressed as K_i (nM) were converted to negative logarithm of K_i (pK_i) values. All the structures for both the datasets were drawn in MarvinSketch software version 15.12.7.0 (<https://www.chemaxon.com>) with proper aromatization and hydrogen bond addition. The data set is composed of various classes of heterogeneous molecular structures as given in the **Table 3.1** along with their pK_i values.

Table 3.1. PET and SPECT imaging datasets against amyloid beta and tau protein.

| PET imaging agents against $A\beta$ Plaques | | | |
|---|-------------|--|--------|
| Serial No. | Compound ID | SMILES Structure | pK_i |
| 1 | A-P-1 | <chem>c1cc(ccc1c1nc2c(s1)cc(cc2)O)NC</chem> | 5.071 |
| 2 | A-P-2 | <chem>c1(cc(c(cc1)NC)F)c1nc2c(s1)cc(cc2)O</chem> | 5.155 |
| 3 | A-P-3 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCF)NC</chem> | 4.638 |
| 4 | A-P-4 | <chem>c1(ccc(cc1)NC)/C=C/c1ccc(nc1)OCCOCCOCCF</chem> | 4.593 |
| 5 | A-P-7 | <chem>c1cccc2c1sc(n2)c1ccc(cc1)NC</chem> | 3.959 |
| 6 | A-P-8 | <chem>c1cc(cc2c1cc(o2)c1c(nc(cc1)NC)F)O</chem> | 4.638 |
| 7 | A-P-21 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCF)NC</chem> | 4.538 |
| 8 | A-P-22 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCF)NC</chem> | 4.174 |
| 9 | A-P-23 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCF)NC</chem> | 4.357 |
| 10 | A-P-24 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCF)NC</chem> | 4.222 |

| | | | |
|----|--------|--|-------|
| 11 | A-P-25 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCOCCOCCF)NC</chem> | 4.125 |
| 12 | A-P-26 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCOCCOCCF)NC</chem> | 4.086 |
| 13 | A-P-27 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCOCCOCCOCCF)NC</chem> | 4.046 |
| 14 | A-P-28 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCF)NC</chem> | 3.733 |
| 15 | A-P-29 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCF)NC</chem> | 3.914 |
| 16 | A-P-30 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCF)NC</chem> | 3.201 |
| 17 | A-P-31 | <chem>c1c(ccc(c1)/C=C/c1ccc(cc1)O)NC</chem> | 4.620 |
| 18 | A-P-43 | <chem>c1c2c(ccc1)nc(s2)c1ccc(cc1)N</chem> | 3.432 |
| 19 | A-P-44 | <chem>c1c2c(ccc1)nc(s2)c1ccc(cc1)N(C)C</chem> | 4.398 |
| 20 | A-P-45 | <chem>c1c2c(ccc1C)nc(s2)c1ccc(cc1)N</chem> | 4.022 |
| 21 | A-P-46 | <chem>c1c2c(ccc1C)nc(s2)c1ccc(cc1)NC</chem> | 4.000 |
| 22 | A-P-47 | <chem>c1c2c(ccc1OC)nc(s2)c1ccc(cc1)N</chem> | 4.155 |
| 23 | A-P-48 | <chem>c1c2c(ccc1OC)nc(s2)c1ccc(cc1)NC</chem> | 4.310 |
| 24 | A-P-49 | <chem>c1c2c(ccc1OC)nc(s2)c1ccc(cc1)N(C)C</chem> | 4.721 |
| 25 | A-P-50 | <chem>c1c2c(ccc1O)nc(s2)c1ccc(cc1)N</chem> | 3.337 |
| 26 | A-P-51 | <chem>c1c2c(ccc1O)nc(s2)c1ccc(cc1)N(C)C</chem> | 4.357 |
| 27 | A-P-52 | <chem>c1c2c(ccc1C#N)nc(s2)c1ccc(cc1)N</chem> | 3.194 |
| 28 | A-P-53 | <chem>c1c2c(ccc1C#N)nc(s2)c1ccc(cc1)NC</chem> | 4.066 |
| 29 | A-P-54 | <chem>c1c2c(ccc1C#N)nc(s2)c1ccc(cc1)N(C)C</chem> | 3.959 |
| 30 | A-P-55 | <chem>c1c2c(ccc1Br)nc(s2)c1ccc(cc1)N</chem> | 4.143 |
| 31 | A-P-56 | <chem>c1c2c(ccc1Br)nc(s2)c1ccc(cc1)NC</chem> | 4.770 |
| 32 | A-P-57 | <chem>c1c2c(ccc1Br)nc(s2)c1ccc(cc1)N(C)C</chem> | 4.538 |
| 33 | A-P-58 | <chem>c1cc(cc2c1oc(c2)c1ccc(cc1)N)OC</chem> | 4.638 |
| 34 | A-P-59 | <chem>c1cc(cc2c1oc(c2)c1ccc(cc1)N)O</chem> | 3.939 |
| 35 | A-P-60 | <chem>c1cc(cc2c1oc(c2)c1ccc(cc1)NC)OC</chem> | 4.886 |
| 36 | A-P-61 | <chem>c1cc(cc2c1oc(c2)c1ccc(cc1)NC)O</chem> | 5.155 |
| 37 | A-P-62 | <chem>c1cc(cc2c1oc(c2)c1ccc(cc1)N(C)C)OC</chem> | 3.921 |
| 38 | A-P-63 | <chem>c1cc(cc2c1oc(c2)c1ccc(cc1)N(C)C)O</chem> | 4.553 |

SPECT imaging agents against A β Plaques

| Serial No. | Compound ID | SMILES Structure | pKi |
|------------|-------------|--|-------|
| 1 | A-S-2 | <chem>c1c(ccc(c1)N(C)C)/C=C/C(=O)/C=C/c1ccc(cc1)OCCC[N]12[Re]3(=O)(SCC1)N(CC2)CCS3</chem> | 3.607 |
| 2 | A-S-3 | <chem>c1c(ccc(c1)N(C)C)/C=C/C(=O)/C=C/c1ccc(cc1)OCCCC[N]12[Re]3(=O)(SCC1)N(CC2)CCS3</chem> | 3.866 |
| 3 | A-S-4 | <chem>c1cc(ccc1N(C)C)/C=C/C(=O)/C=C/c1ccc(cc1)OCCC[N]12CCS[Re]31(=O)N(C(=O)C2)CCS3</chem> | 2.918 |
| 4 | A-S-5 | <chem>c1cc(ccc1N(C)C)/C=C/C(=O)/C=C/c1ccc(cc1)OCCCC[N]12CCS[Re]31(=O)N(C(=O)C2)CCS3</chem> | 3.228 |
| 5 | A-S-6 | <chem>c1(ccc2n(c1)cc(n2)c1ccc(nc1)N1CCN=N1)I</chem> | 4.738 |
| 6 | A-S-7 | <chem>c1(ccc(cc1)C#Cc1nc(c(c1)Br)OCCOCCOCCF)N(C)C</chem> | 3.951 |
| 7 | A-S-8 | <chem>c1(ccc(cc1)C#Cc1nc(c(c1)Br)OCCO)N(C)C</chem> | 4.174 |
| 8 | A-S-9 | <chem>c1(ccc(cc1)C#Cc1nc(c(c1)Br)OCCOCCOCCF)NC</chem> | 3.883 |
| 9 | A-S-10 | <chem>c1(ccc(cc1)C#Cc1nc(c(c1)Br)OCCO)NC</chem> | 4.796 |
| 10 | A-S-11 | <chem>c1c(ccc(c1)N(C)C)C#Cc1nc(c(c1)I)OCCOCCOCCF</chem> | 3.775 |
| 11 | A-S-12 | <chem>c1(ccc(cc1)C#Cc1nc(c(c1)I)OCCO)N(C)C</chem> | 4.036 |
| 12 | A-S-13 | <chem>c1(ccc(cc1)C#Cc1nc(c(c1)I)OCCO)NC</chem> | 3.903 |

| | | | |
|----|--------|---|-------|
| 13 | A-S-14 | <chem>c1c2c(ccc1I)oc(c2)c1cnc(cc1)N</chem> | 3.987 |
| 14 | A-S-15 | <chem>c1c2c(ccc1I)oc(c2)c1cnc(cc1)NC</chem> | 4.532 |
| 15 | A-S-16 | <chem>c1c2c(ccc1I)oc(c2)c1cnc(cc1)N(C)C</chem> | 4.627 |
| 16 | A-S-17 | <chem>c1c2c(cc(c1)I)c(=O)cc(o2)/C=C/c1ccc(cc1)OC</chem> | 3.682 |
| 17 | A-S-18 | <chem>c1c2c(cc(c1)I)c(=O)cc(o2)/C=C/c1cc(c(cc1)OC)OC</chem> | 3.575 |
| 18 | A-S-19 | <chem>c1c2c(cc(c1)I)c(=O)cc(o2)/C=C/c1ccc(cc1)O</chem> | 3.456 |
| 19 | A-S-20 | <chem>c1c2c(cc(c1)I)c(=O)cc(o2)/C=C/c1ccc(cc1)OCCO</chem> | 2.656 |
| 20 | A-S-21 | <chem>c1c2c(cc(c1)I)c(=O)cc(o2)/C=C/c1ccc(cc1)N</chem> | 3.231 |
| 21 | A-S-22 | <chem>c1c2c(cc(c1)I)c(=O)cc(o2)/C=C/c1ccc(cc1)NC</chem> | 3.770 |
| 22 | A-S-33 | <chem>c1c2c(cc(c1)I)C(=O)/C(=C/c1ccc(cc1)O)/O2</chem> | 4.893 |
| 23 | A-S-34 | <chem>c1c2c(cc(c1)I)C(=O)/C(=C/c1ccc(cc1)OCCO)/O2</chem> | 4.979 |
| 24 | A-S-35 | <chem>c1c2c(cc(c1)I)C(=O)/C(=C/c1ccc(cc1)OCCOCCO)/O2</chem> | 4.474 |
| 25 | A-S-36 | <chem>c1c2c(cc(c1)I)C(=O)/C(=C/c1ccc(cc1)OCCOCCOCCO)/O2</chem> | 4.592 |
| 26 | A-S-37 | <chem>c1c2c(cc(c1)OC)sc1n2cc(n1)c1ccc(cc1)N</chem> | 3.526 |
| 27 | A-S-38 | <chem>c1c2c(cc(c1)OC)sc1n2cc(n1)c1ccc(cc1)NC</chem> | 4.215 |
| 28 | A-S-39 | <chem>c1c2c(cc(c1)OC)sc1n2cc(n1)c1ccc(cc1)N(C)C</chem> | 3.232 |
| 29 | A-S-40 | <chem>c1c2c(cc(c1)F)sc1n2cc(n1)c1ccc(cc1)N</chem> | 2.876 |
| 30 | A-S-41 | <chem>c1c2c(cc(c1)F)sc1n2cc(n1)c1ccc(cc1)NC</chem> | 3.419 |
| 31 | A-S-42 | <chem>c1c2c(cc(c1)F)sc1n2cc(n1)c1ccc(cc1)N(C)C</chem> | 3.368 |
| 32 | A-S-43 | <chem>c1c2c(cc(c1)Br)sc1n2cc(n1)c1ccc(cc1)N</chem> | 3.541 |
| 33 | A-S-44 | <chem>c1c2c(cc(c1)Br)sc1n2cc(n1)c1ccc(cc1)NC</chem> | 3.462 |
| 34 | A-S-45 | <chem>c1c2c(cc(c1)Br)sc1n2cc(n1)c1ccc(cc1)N(C)C</chem> | 3.363 |
| 35 | A-S-46 | <chem>c1c2c(cc(c1)OC)sc1n2cc(n1)c1ccc(cc1)I</chem> | 3.963 |
| 36 | A-S-47 | <chem>c1c2c(cc(c1)F)sc1n2cc(n1)c1ccc(cc1)I</chem> | 3.378 |
| 37 | A-S-48 | <chem>c1c2c(cc(c1)Br)sc1n2cc(n1)c1ccc(cc1)I</chem> | 3.676 |
| 38 | A-S-49 | <chem>c1c2c(cc(c1)C)sc1n2cc(n1)c1ccc(cc1)I</chem> | 3.752 |
| 39 | A-S-50 | <chem>c1c2c(cc(c1)OC)sc1n2cc(n1)c1ccc(cc1)Br</chem> | 4.027 |
| 40 | A-S-51 | <chem>c1c2c(cc(c1)C)sc1n2cc(n1)c1ccc(cc1)Br</chem> | 3.585 |
| 41 | A-S-52 | <chem>c1c(cc(c(c1)OC)C(=O)O)/C=C/c1ccc(c(c1)I)/C=C/c1cc(c(cc1)OC)C(=O)O</chem> | 5.770 |
| 42 | A-S-53 | <chem>c1c(cc(c(c1)O)C(=O)O)/C=C/c1ccc(c(c1)I)/C=C/c1cc(c(cc1)O)C(=O)O</chem> | 6.000 |
| 43 | A-S-54 | <chem>c1c(cc(c(c1)O)C(=O)O)/C=C/c1ccc(c(c1)Br)/C=C/c1cc(c(cc1)O)C(=O)O</chem> | 5.959 |
| 44 | A-S-55 | <chem>c1c(ccc(c1)OC)/C=C/c1ccc(c(c1)Br)/C=C/c1ccc(cc1)OC</chem> | 2.179 |
| 45 | A-S-56 | <chem>c1c(ccc(c1)O)/C=C/c1ccc(c(c1)Br)/C=C/c1ccc(cc1)O</chem> | 4.658 |
| 46 | A-S-57 | <chem>c1c(ccc(c1)O)/C=C/c1ccc(c(c1)I)/C=C/c1ccc(cc1)O</chem> | 4.699 |
| 47 | A-S-61 | <chem>c1c2c(ccc1OCCC[N]13CCS[Re]41(=O)N(C(=O)C3)CCS4)nc(s2)c1ccc(cc1)N(C)C</chem> | 4.000 |
| 48 | A-S-62 | <chem>c1c2c(ccc1N1CCN3[Re]41(=O)[N@](CC3)(CCS4)C)nc(s2)c1cccc1</chem> | 2.255 |
| 49 | A-S-63 | <chem>c1c2c(ccc1N1CCN3[Re]1(=O)SCC[N](CC3)(C)C)nc(s2)c1ccc(cc1)F</chem> | 2.210 |
| 50 | A-S-64 | <chem>c1c2c(ccc1N1CCN3[Re]1(=O)(SCCN(CC3)C)C)nc(s2)c1ccc(cc1)OC</chem> | 3.071 |
| 51 | A-S-65 | <chem>c1c2c(ccc1N1CCN3[Re]41(=O)[N](CC3)(CCS4)C)nc(s2)c1ccc(cc1)N(C)C</chem> | 3.523 |
| 52 | A-S-66 | <chem>c1c2c(ccc1)nc(s2)c1ccc(cc1)N1CCN2[Re]31(=O)[N](CC2)(CCS3)C</chem> | 2.423 |
| 53 | A-S-67 | <chem>c1c2c(ccc1F)nc(s2)c1ccc(cc1)N1CCN2[Re]31(=O)[N](CC2)(CCS3)C</chem> | 2.928 |
| 54 | A-S-68 | <chem>c1c2c(ccc1O)nc(s2)c1ccc(cc1)N1CCN2[Re]31(=O)[N](CC2)(CCS3)C</chem> | 3.053 |
| 55 | A-S-69 | <chem>c1c2c(ccc1OC)nc(s2)c1ccc(cc1)N1CCN2[Re]31(=O)[N](CC2)(CCS3)C</chem> | 3.060 |
| 56 | A-S-70 | <chem>c1c2c(ccc1)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(N2)SCC[N]3(CC1)C</chem> | 3.046 |
| 57 | A-S-71 | <chem>c1c2c(ccc1F)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(N2)SCC[N]3(CC1)C</chem> | 2.947 |
| 58 | A-S-72 | <chem>c1c2c(ccc1OC)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(N2)SCC[N]3(CC1)C</chem> | 3.215 |
| 59 | A-S-73 | <chem>c1c2c(ccc1)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(O2)SCC[NH]3CC1</chem> | 2.963 |

| | | | |
|----|--------|---|-------|
| 60 | A-S-74 | <chem>c1c2c(ccc1F)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(O2)SCC[NH]3CC1</chem> | 3.194 |
| 61 | A-S-75 | <chem>c1c2c(ccc1OC)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(O2)SCC[NH]3CC1</chem> | 3.523 |
| 62 | A-S-76 | <chem>c1c2c(ccc1)nc(s2)c1cc2c(cc1)O[Re]13(=O)N2CC[NH]1CCS3</chem> | 2.553 |
| 63 | A-S-77 | <chem>c1c2c(ccc1F)nc(s2)c1cc2c(cc1)O[Re]13(=O)N2CC[NH]1CCS3</chem> | 2.646 |
| 64 | A-S-78 | <chem>c1c2c(ccc1OC)nc(s2)c1cc2c(cc1)O[Re]13(=O)N2CC[NH]1CCS3</chem> | 2.854 |
| 65 | A-S-79 | <chem>c1c2c(ccc1)nc(s2)c1cc2c(cc1)S[Re]13(=O)N2CC[NH]1CCS3</chem> | 2.578 |
| 66 | A-S-80 | <chem>c1c2c(ccc1F)nc(s2)c1cc2c(cc1)S[Re]13(=O)N2CC[NH]1CCS3</chem> | 3.032 |
| 67 | A-S-81 | <chem>c1c2c(ccc1OC)nc(s2)c1cc2c(cc1)S[Re]13(=O)N2CC[NH]1CCS3</chem> | 2.879 |
| 68 | A-S-82 | <chem>c1c2c(ccc1)nc(s2)c1ccc2c(c1)S[Re]13(=O)N2CC[NH]1CCS3</chem> | 3.420 |
| 69 | A-S-83 | <chem>c1c2c(ccc1F)nc(s2)c1ccc2c(c1)S[Re]13(=O)N2CC[NH]1CCS3</chem> | 3.509 |
| 70 | A-S-84 | <chem>c1c2c(ccc1OC)nc(s2)c1ccc2c(c1)S[Re]13(=O)N2CC[NH]1CCS3</chem> | 3.367 |
| 71 | A-S-85 | <chem>c1c2c(ccc1)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(N2CC[S]3C)SCC1</chem> | 2.699 |
| 72 | A-S-86 | <chem>c1c2c(ccc1F)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(N2CC[S]3C)SCC1</chem> | 2.830 |
| 73 | A-S-87 | <chem>1c2c(ccc1OC)nc(s2)c1cc2c(cc1)N1[Re]3(=O)(N2CC[S]3C)SCC1</chem> | 2.750 |

PET and SPECT Imaging Agents against Tau protein

| Serial No. | Compound ID | SMILES Structure | pKi |
|------------|-------------|---|-------|
| 1 | T-P-1 | <chem>c1c2c(ccc1)nc([nH]2)/C=C/c1ccc(cc1)N(CC)CC</chem> | 3.921 |
| 2 | T-P-2 | <chem>c1c2c(ccc1OCCF)nc(o2)/C=C/c1ccc(cc1)NC</chem> | 3.194 |
| 3 | T-P-3 | <chem>c1c(enc(c1)F)c1cc2c(cc1)c1c([nH]2)ccnc1</chem> | 3.959 |
| 4 | T-P-4 | <chem>C1N(CCC(C1)CCF)c1nc2n(cc1)c1c(n2)cccc1</chem> | 3.027 |
| 5 | T-P-5 | <chem>c12c(ccc(c1)OCC(CF)O)nc(cc2)c1ccc(cc1)N(C)C</chem> | 3.108 |
| 6 | T-P-6 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N)OCC(CF)O</chem> | 2.444 |
| 7 | T-P-7 | <chem>c1(cc2c(cc1)nc(cc2)c1ccc(cc1)NC)OCC(CF)O</chem> | 2.979 |
| 8 | T-P-8 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N)OCCF</chem> | 2.227 |
| 9 | T-P-9 | <chem>c1c(ccc2c1ccc(c2)C(=C(C#N)C#N)C)N(CCF)C</chem> | 1.580 |
| 10 | T-P-10 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N(C)C)OCCF</chem> | 2.290 |
| 11 | T-P-11 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N(C)C)OCCCF</chem> | 2.059 |
| 12 | T-P-12 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N(C)C)OCC(CF)O</chem> | 2.996 |
| 13 | T-P-13 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)NCC)OCC(CF)O</chem> | 2.553 |
| 14 | T-P-14 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)NC)OCC(CF)O</chem> | 2.536 |
| 15 | T-P-15 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N(C)C)OCC(CF)O</chem> | 2.757 |
| 16 | T-P-16 | <chem>c1cc(cc2c1nc(cc2)c1ccc(en1)NC)OCC(CF)O</chem> | 2.077 |
| 17 | T-P-17 | <chem>c1cc(cc2c1nc(cc2)c1ccc(en1)N(C)C)OCC(CF)O</chem> | 2.585 |
| 18 | T-P-18 | <chem>c1c(ccc2c1nc(cc2)c1ccc(en1)N(C)C)OCC(CF)O</chem> | 1.848 |
| 19 | T-P-19 | <chem>c1cc(cc2c1nc(cc2)c1ccc(en1)N(CC)CC)OC(CF)CO</chem> | 2.007 |
| 20 | T-P-20 | <chem>c1c(ccc2c1sc(n2)/C=C/C=C/c1enc(cc1)NC)O</chem> | 3.593 |
| 21 | T-P-21 | <chem>c1c(ccc2c1nc(cc2)c1ccc(cc1)NC)O</chem> | 2.684 |
| 22 | T-P-22 | <chem>c1c(ccc2c1nc(cc2)c1enc(cc1)NC)O</chem> | 1.957 |
| 23 | T-P-23 | <chem>c1cc(cc2c1nc(cc2)c1ccc(cc1)N)O</chem> | 2.442 |
| 24 | T-P-24 | <chem>c1cc(cc2c1nc(cc2)c1enc(cc1)NC)O</chem> | 2.517 |
| 25 | T-S-25 | <chem>C1(=S)S/C(=C\c2oc(cc2)c2cc(ccc2)I)/C(=O)N1CCC(=O)OCC</chem> | 1.311 |
| 26 | T-S-26 | <chem>C1(=S)N/C(=C\c2oc(cc2)c2cc(ccc2)I)/C(=O)N1CCC(=O)OCC</chem> | 1.810 |
| 27 | T-S-27 | <chem>C1(=S)N/C(=C\c2oc(cc2)c2cc(ccc2)I)/C(=O)N1CCc1c[nH]cn1</chem> | 2.194 |
| 28 | T-S-28 | <chem>c1c(ccc2c1sc(n2)/N=N/c1ccc(cc1)N)I</chem> | 3.138 |
| 29 | T-S-29 | <chem>c1c(ccc2c1sc(n2)/N=N/c1ccc(cc1)NC)I</chem> | 3.863 |
| 30 | T-S-30 | <chem>c1c(ccc2c1sc(n2)/N=N/c1ccc(cc1)N(C)C)I</chem> | 4.319 |
| 31 | T-P-31 | <chem>c1(ccc(cc1)CCc1ccc(cc1)NC)O</chem> | 2.085 |

3.1.2. Molecular descriptors

The molecular descriptor is the result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number. QSAR models were developed with a selected class of molecular descriptors (2-dimensional) comprising E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom centred fragments and molecular property descriptors, calculated using Dragon 7 software (Dragon version 7, Kodesrl, Milan, Italy, 2016; software available at <http://www.taletе.mi.it/index.htm>). Intercorrelated descriptors (inter-correlation values larger than 0.9) were removed from the descriptor pool to reduce the size of the descriptor matrix. Finally, a pool of 335 descriptors was obtained for PET imaging agents and a pool of 529 descriptors was obtained for SPECT imaging agents targeted against A β protein. For the tau dataset a reduced pool of 263 descriptors from 418 descriptors was employed for model development. A descriptor pool of 633 descriptors was obtained for the A β dataset. In order to reduce the redundant and incompetent data, inter-correlated descriptors (correlation value larger than 0.9) were removed from the descriptor pool, and finally, 539 descriptors were taken for modeling. For the tau dataset a reduced pool of 263 descriptors from 418 descriptors was employed for model development.

3.1.3. Dataset splitting

The main objective in QSAR study is to obtain a well validated QSAR model which is possible with proper division or splitting of the dataset into training and test set. Ideally, the division must be executed in such a way so that points representing both training and test set are well distributed within the whole descriptor space occupied by the entire dataset. Rational data division helps in providing an unbiased external validation with uniform distribution of compounds into training and test sets (Golbraikh et al., 2003). One of the extensively used methods is the Euclidean Distance based division (Golmohammadi et al., 2012), which was used for division of the A β imaging dataset (for both PET and SPECT datasets) into training (~75%) and a test set (~25%). The combined PET and SPECT dataset targeted against the tau protein was divided into a training set (~70%) and a test set (~30%) based on *k*-Medoids division method (Park & Jun, 2009). The *k*-medoids algorithm is a local heuristic method that runs just like *k*-means clustering when updating the medoids. This method tends to select *k* most middle objects as initial medoids. The algorithm involves calculation of the distance matrix once and uses it for finding new medoids at every iterative step.

3.1.4. Model development

A critical evaluation procedure was carried out in order to have the best model with good statistical significance for both internal and external validation metrics. During the development of models for individual subsets, i.e., for PET and SPECT imaging agents targeted against A β , we have used Stepwise Multiple Linear Regression (S-MLR) (Khan & Roy, 2018; Pope & Webster, 2012) method implemented in Double Cross Validation (DCV) tool (version 1.2) (Roy & Ambure, 2016) and finally Partial Least Squares (PLS) regression (Khan & Roy, 2018; Wold et al., 2001) was used to develop the models. In case of tau dataset, a descriptor pool 26 descriptors were selected using Genetic Algorithm (GA) (Khan & Roy, 2018) modeling implemented in Double Cross Validation (DCV) tool (version 1.2). Then the final model was generated using PLS regression method using descriptors selected from Best Subset Selection (BSS).

In both the cases, the Double Cross Validation (DCV) method helped in the generation of the most statistically significant and robust models. DCV aids in the generation and selection of models to produce a better predictive model. DCV is a method where the training set compounds are further divided into 'n' calibration and validation sets, can result in diverse compositions of the modeling set,

thus removing any bias in descriptor selection. Additionally, a model with the lowest prediction errors in the validation set is selected; thus, providing an optimum solution in terms of predictivity in most cases. The tool comprises two nested cross-validation loops: the internal and external cross-validation loops. In the external loop, the compounds in the dataset are divided into training set compounds and test set compounds. The training set compounds goes to the internal loop for the purpose of model development and model selection, and the test set is used exclusively for checking model predictivity. In the internal loop, the training set is further split into calibration and validation sets repetitively by employing the k-fold cross-validation technique (in this study, $k = 10$) (Baumann & Baumann, 2014) and producing k iterations to construct calibration and validation sets. At the end, the best models were selected based on various validation metrics.

3.1.5. Statistical validation metrics

The current study utilizes multiple approaches for assessment of model quality for measurement of the fitness, stability, robustness and predictivity of the developed models. The validation was done using both internal and external validation metrics (Roy et al., 2015b). The fitting potential of the model is established by the determination coefficient (R^2) whereas internal validation dealing with the predictive ability of the model based on training set compounds is usually established by a cross-validated squared correlation coefficient, Q_{LOO}^2 (leave-one-out or LOO). However, Q^2 is not the ultimate quality measuring metric to determine the performance of the model for a new set of compounds. Thus, for new external compounds (or test compounds), various external validation metrics are used such as Q_{F1}^2 and Q_{F2}^2 (Chirico & Gramatica, 2012; Roberto Todeschini et al., 2016). Additionally, r_m^2 metrics (Roy et al., 2012), root mean square error (RMSE), and mean absolute error (MAE) are also calculated (Roy et al., 2016). The applicability domain (AD) (discussed later) (Gadaleta et al., 2016) was performed according to the DModX (distance to model in the X-space) approach using SIMCA-P software.

3.1.6. Molecular Docking

In the present work, we have implemented molecular docking analysis to understand the intermolecular interactions occurring between the PET and SPECT imaging agents and protein beta amyloid and tau proteins separately. The protein structures in the present case are retrieved from the Protein Data Bank with PDB ID: 2LMN (Paravastu et al., 2008) for A β protein and PDB ID: 6FAU (Andrei et al., 2018) for tau protein. Docking was performed in CDOCKER module of receptor-ligand interaction implemented in BIOVIA Discovery Studio 2018 (<http://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/requirements/technical-requirements-410.html>) (Wu et al., 2003). The structure of the beta amyloid protein does not contain any bound ligand; therefore, the active site was defined in the BIOVIA Discovery Studio platform in receptor ligand interaction section using the option “define site from receptor cavities” before docking. A total of 11 active sites were generated by the software, however, we have selected site 1 ($x: 51.610$, $y: 30.947$, and $z: 70.698$) because in the other sites, either the ligands were not able to dock or were docked outside/away from the docking site. In case of tau, the X-ray crystal structure of the protein consists of two chains A and C and four bound ligands (two peptide residues, Ace-Arg-Thr-Pro-Sep-Leu-Pro-Gly in chain A, Thr-Pro-Sep-Leu-Pro-Gly in chain C and two instances of D3W ((2- $\{R\}$)-2-[($\sim\{R\}$)-(2-methoxyphenyl)-phenyl-methyl]pyrrolidine) one in each chain. Due to structural similarity between the chain structures, we have used only one chain (chain A) for our docking purpose. Before docking the target ligands, the protein was prepared by removing the duplicate amino acid conformers, addition of hydrogen, and generation of docking site. The active site ($x: 17.355$, $y: -8.685$, $z: -12.366$) was defined in the BIOVIA Discovery Studio platform from the ligand binding domain of the bound peptide residue and D3W by selecting them

and generating site “from current selection” program in receptor-ligand interaction section of the software. The bound ligands were then removed for new molecule docking purpose.

The target ligands (imaging agents) were subjected to ligand preparation to obtain a series of ligand conformers in both cases using the small molecules module in Discovery Studio. Each of these conformers was used in the CDOCKER module involving CHARMM interaction energy for molecular docking (Wu et al., 2003). The ligand poses were ranked using the CDOCKER interaction energy parameters (kJ/ mol), and the top scoring (most negative, thus favorable to binding) poses are kept. The best pose obtained was further analyzed by considering intermolecular polar and non-polar interactions.

3.2. Study 2: Chemometric modeling of PET imaging agents for diagnosis of Parkinson’s disease: A QSAR approach

3.2.1. The dataset

The experimental binding affinity and selectivity data of 35 xanthine ligand-based PET tracers were taken from a previously published literature (Tamiji et al., 2018) and applied for QSAR modeling to determine the essential structural features needed for binding affinity and explore the structural requirements necessary to be present in the antagonists for selectivity towards A_{2A} adenosine receptors. The experimental values of selectivity and binding affinity (K_i) ranged from 0.1–20 nM and 7.84–16,500 nM respectively and the details are provided in **Table 3.3**. The experimental values were converted into negative logarithm scale during modeling and were used as independent values. No compounds with binding affinity data were removed during modeling but some compounds (**14**, **32**, **33**, and **34**) with no experimental selectivity values were eliminated during modeling. Here, the binding affinity and selectivity were separately used as endpoints or independent variables in modeling. The compounds for both the dataset were represented in MarvinSketch software version 15.12.7.0 (<https://www.chemaxon.com>) with proper aromatization and addition of hydrogen bond as necessary.

Table 3.3. Structures and experimental A_{2A}R binding affinity [pA_{2A}R(BA)] and A_{2A}R selectivity [log A_{2A}R(Sel)] values.

| Compound No. | SMILES structures | pK _i (A _{2A} R) | log A _{2A} R(Sel) |
|--------------|--|-------------------------------------|----------------------------|
| 1 | <chem>O=C1C=CN2[C@H](NC(=C2n1)C1CC(CCC1)CN1CCN(CC1)C1CCC(C1)OCCOC)N</chem> | -0.447 | 2.779 |
| 2 | <chem>O=C1C=CN2C(NC(C2n1)C1CC(CCC1)CN1CCN(CC1)C1C(C(C1)OCCOC)F)N</chem> | -0.431 | 2.808 |
| 3 | <chem>O=C1C=CN2N(N1)C(NC(C2)C1CCCC(C1)N1CCN(CC1)C1CCC(C1)OCCOC)N</chem> | 0.000 | 3.025 |
| 4 | <chem>N1C(NC(N2NC(N12)C1OCCC1)N)N1C[C@@H]2N(CC1)C[C@H](C2)COc1cc(CCC1)F</chem> | 0.699 | 4.217 |
| 5 | <chem>C1(NC2C(NC(N2N1C)CCCC)N)[C@@H]1N=NC=N1</chem> | -0.820 | 1.076 |
| 6 | <chem>C1(NC2C(NC(N2N1C)CCCC)N)[C@@H]1N=NC=N1</chem> | -0.519 | 0.894 |
| 7 | <chem>C1(NC2C(NC(N2N1C)CCc1cccc1)N)[C@@H]1N=NC=N1</chem> | -0.672 | 1.231 |
| 8 | <chem>O1CCCC1C1NN2C(NC3C(C2N1)NcN3CCN1CCN(CC1)C1CCC(C1)OC)N</chem> | 1.000 | 3.229 |
| 9 | <chem>O1CCCC1C1NN2C(NC3C(C2N1)NcN3CCN1CCN(CC1)C1CCC(C1)OCCOC)N</chem> | 0.046 | 2.825 |
| 10 | <chem>O1CCCC1C1NN2C(NC3C(C2N1)NcN3CCN1CCC(CC1)C1CCC(C1)OCCOC)N</chem> | 0.155 | 2.715 |

| | | | |
|----|--|--------|-------|
| | OC)N | | |
| 11 | n1c(nc(n2nc(nc12)c1occc1)N)N(C)CCN1CCN(CC1)c1c(cc(cc1)F)F | -0.602 | 2.312 |
| 12 | n1c(nc(n2nc(nc12)c1occc1)N)N1CCN(CC1)Cc1c(c(ccc1Cl)F)F | -0.699 | 2.000 |
| 13 | n1c(nc(n2nc(nc12)c1occc1)N)N1CCN(CC1)Cc1c(cc(cc1F)F)F | -0.477 | 2.636 |
| 14 | n1c(nc(c2c1n(nc2)Cc1cc(ccc1)OC)c1occc1)N | -0.301 | - |
| 15 | n1c(nc(n2nc(nc12)c1occc1)N)NC[C@H]1N(CCC1)Cc1c(c(ccc1)F)F | -0.699 | 2.398 |
| 16 | n1c(nc(n2nc(nc12)c1occc1)N)NC[C@H]1N(CCC1)Cc1c(cc(c(c1)F)F)F | -0.903 | 2.398 |
| 17 | n1c(nc(n2nc(nc12)c1occc1)N)NC[C@H]1N(CCC1)Cc1c(cc(cc1F)F)F | -0.301 | 2.903 |
| 18 | n1c(nc(n2nc(nc12)c1occc1)N)NC[C@H]1N(CCC1)Cc1c(c(ccc1F)F)Cl | -0.602 | 2.398 |
| 19 | c1(ccc(cc1)OO)C(=O)Nc1sc2c(ncc(c2n1)OC)N1CCOCC1 | -0.477 | 2.653 |
| 20 | n1nc(nc1c1cc(ccc1)OC)Cc1cc(c(cc1)C)C | -1.301 | 1.839 |
| 21 | o1c(ccc1)c1nn2c(n1)c1c(nc2N)n(nc1)CCN1CCN(CC1)c1cccc1F | 0.222 | 2.951 |
| 22 | o1c(ccc1)c1nn2c(n1)c1c(nc2N)n(nc1)CCN1CCN(CC1)c1c(cc(cc1)F)F | 0.222 | 3.204 |
| 23 | o1c(ccc1)c1nn2c(n1)c1c(nc2N)n(nc1)CCN1CCN(CC1)c1ccc(c(c1)F)F)F | 0.222 | 3.176 |
| 24 | o1c(ccc1)c1nn2c(n1)c1c(nc2N)n(nc1)CCN1CCN(CC1)c1cccc(c1)OCCOC | -0.041 | 3.127 |
| 25 | o1c(ccc1)c1nn2c(n1)c1c(nc2N)n(nc1)CCN1CCN(CC1)c1c(ccc(c1)OCCOC)F | 0.398 | 3.240 |
| 26 | o1c(ccc1)c1nn2c(n1)c1c(nc2N)n(nc1)CCN1CCN(CC1)c1c(ccc(c1)OCCOC)F | 0.222 | 3.064 |
| 27 | o1c(ccc1)c1nn2c(n1)cc(nc2N)N(C)CCc1ccc(cc1)OC | -0.255 | 2.675 |
| 28 | c1(nc(cc2n1nc(n2)c1occc1)OCCN1CCN(CC1)c1ccc(cc1)OCCOC)N | -0.447 | 2.607 |
| 29 | C1(=N[C@@H](Cc2n1nc(n2)c1occc1)SCCN1CCN(CC1)c1ccc(c1)OCCOC)N | -0.176 | 2.985 |
| 30 | c1(nc(cc2n1nc(n2)c1occc1)N(CCN1CCN(CC1)c1ccc(cc1)OCCOC)C)N | 0.000 | 3.199 |
| 31 | c1(nc(cc2n1nc(n2)c1occc1)N(CCN1CCN(CC1)c1cc(cc(c1)F)F)C)N | -0.398 | 2.841 |
| 32 | c1(ccc(cc1)CO)C(=O)Nc1sc2c(n1)c(ccc2c1cccc1)OC | 0.301 | - |
| 33 | s1c(ccc1C)C(=O)Nc1sc2c(n1)c(ccc2c1cc(ccc1)N)OC | 0.097 | - |
| 34 | o1c(ccc1C)C(=O)Nc1sc2c(n1)c(ccc2c1cc(ccc1)N)OC | 0.000 | - |
| 35 | C(=O)(N1[C@@H](CCC1)COC)c1cc(c2n(c1)nc(n2)c1oc(cc1)Br)N | -1.204 | 1.818 |

3.2.2. Molecular descriptors

In the present study, QSAR models were developed using a selected class of two-dimensional molecular descriptors involving E-state indices, connectivity, constitutional, functional, 2D atom

pairs, ring, atom centered fragments, molecular property descriptors and Extended topochemical atom (ETA) indices. The ETA descriptors were calculated using PaDel-Descriptor software (C. W. Yap, 2011) whereas the non-ETA descriptors were calculated using Dragon 7 software. Intercorrelated ($|r| > .95$), constant (variance < 0.0001), and other incompetent and redundant data was removed using an in-house software available at <http://dtclab.webs.com/software-tools> before model development.

3.2.3. Dataset Division

Dataset division is a crucial part of QSAR modeling in order to develop a properly validated and robust model. Rational data division ensures an unbiased external validation along with uniform data distribution (Golbraikh et al., 2003). The division of the dataset into training set (~70%) and test set (~30 %) was performed employing random dataset division method (Golbraikh & Tropsha, 2000) for both binding affinity and selectivity end points. The training set was used for model development and test set was used for model validation.

3.2.4. Variable selection and Model Development

Prior to the model development, variable selection strategies such as Genetic Algorithm (GA) (Devillers, 1996; Khan & Roy, 2018) and stepwise regression (Khan & Roy, 2018; Pope & Webster, 2012) were applied for the binding affinity and selectivity, respectively, to extract the important and influential descriptors and created a reduced pool of descriptors. After obtaining the important descriptors, model development was done. The best model with five descriptors was obtained using the spline option in the GA run on Discovery Studio version 4.1 for the binding affinity. On the other hand, for A_{2A}R selectivity, four models with four descriptors were selected from the Best Subset Selection (BSS) method based on MAE criteria (Roy et al., 2016). Further to improve the quality of the external prediction via “intelligent” selection of multiple models, the “Intelligent consensus predictor” tool (Roy et al., 2018) was applied (DTC Lab QSAR Tools http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab). The methodology of the present work is given in **Figure 3.1**.

3.2.5. Statistical validation metrics

The statistical quality of the models developed in the present study was rigorously examined using multiple approaches to check the robustness and predictivity of the developed models. All the models were validated both externally and internally. Various parameters like determination coefficient R^2 , explained variance R^2_a , variance ratio (F), and standard error of estimate (s) were computed. Internal predictivity parameters such as predicted residual sum of squares (PRESS) and leave-one-out cross-validated correlation coefficient (Q^2_{LOO}) were also calculated along with external predictivity parameters like R^2_{pred} or Q^2_{F1} , Q^2_{F2} and concordance correlation coefficient (CCC) (Roy & Mitra, 2011). It has been reported that consensus models are better in performance in comparison to an individual model (Roy & Mitra, 2011). Therefore, “Intelligent Consensus Predictions (ICP)” were applied using multiple models to see whether the quality of predictions can be increased through an intelligent selection.

3.2.6. Applicability domain (AD)

Applicability domain (AD) (Gadaleta et al., 2016) is a theoretical region in the chemical space developed based on modeled descriptors and modeled response of the training set, where the developed model could make predictions basing on some logical reliability. Here, we have checked AD using standardization approach using the tool developed in our laboratory.

3.2.7. Molecular Docking

Molecular docking analysis has been implemented in the present work that helps in understanding the intermolecular interactions taking place between the PET tracer antagonists and the A_{2A} receptor. The protein structure for adenosine A_{2A} receptor is retrieved from the protein data bank with PDB ID: 3UZA (Congreve et al., 2012). The X-ray crystal structure of the protein consists of a bound ligand T4G commonly known as 6-(2,6-Dimethylpyridin-4-yl)-5-phenyl-1,2,4-Triazin-3-amine (Formula: C₁₆H₁₅N₅). Before docking the target PET tracers, protein preparation was done by cleaning the protein for any missing residues, explicit hydrogen addition and generation of the docking site. The generation of active docking site was done in the BIOVIA Discovery Studio platform from the ligand binding domain of the bound ligand T4G by the selection of the ligand and generating the site “from current selection” program in receptor-ligand interaction module of the software. After the generation of the active ligand binding domain (x: 47.473, y: 25.697, and z: 28.736), the bound ligand was removed for new molecule docking. For ligand preparation, the PET tracers were put through small molecule module in Discovery Studio platform where a series of ligand conformers were generated. Each of these generated conformers was then used in the CDOCKER module energy for molecular docking involving CHARMM interaction (Wu et al., 2003). The CDOCKER interaction energy parameter (kcal/mol) was checked for all the receptor ligand complexes, and the top scoring (most negative, thus favourable to binding) poses were kept.

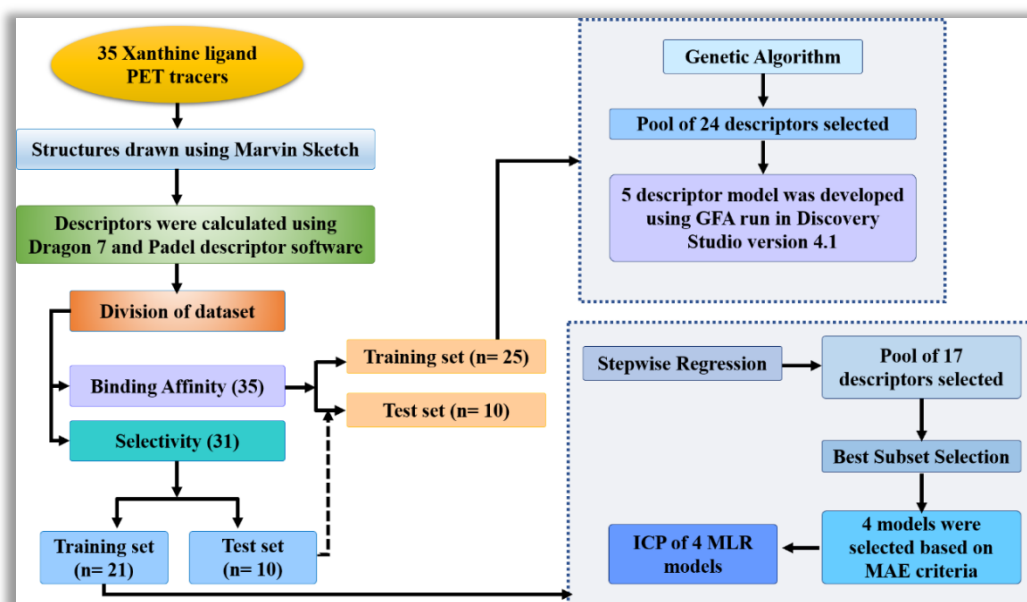


Figure 3.1. The methodology of present QSAR modeling.

3.3. Study 3: QSAR modeling of PET imaging agents for the diagnosis of Parkinson’s disease targeting Dopamine receptor

In the present study, a QSAR model with two-dimensional (2D) molecular descriptors was developed to explore the correlations of the molecular structure of a series of PET tracers against the binding affinity of dopamine (D₂) receptor.

3.3.1. The dataset

Dopamine (D₂) receptor binding affinity (K_i) data of 34 PET imaging agents was taken from different literatures as mentioned in **Table 3.3**. The experimental binding affinity for all the compounds was measured using the same assay protocol, i.e., rat striatal homogenate (RSH) assay method. This data was applied in the development of a 2D-QSAR model to determine the essential structural features

required for good binding to the D2 receptor. The binding affinity (K_i) values for the PET imaging agents were converted to their negative logarithm (pK_i) form and then used for modeling. The compounds were represented using the MarvinSketch software (available from <https://chemaxon.com/marvin>) with proper aromatization and addition of hydrogen bond as necessary.

Table 3.2. Dataset compounds with their observed binding affinity (in pK_i).

| Compound No. | Structure | pK_i | Reference |
|--------------|--|--------|-----------------------------------|
| 1 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3ccc1)C</chem> | 2.321 | (Sipos et al., 2008) |
| 2 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)O)C</chem> | 4.420 | (Gao, Baldessarini, et al., 1990) |
| 3* | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)Cl)C</chem> | 2.652 | (Gao, Baldessarini, et al., 1990) |
| 4 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)Br)C</chem> | 2.752 | (Gao, Baldessarini, et al., 1990) |
| 5 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)SC)C</chem> | 2.262 | (Tóth et al., 2006) |
| 6 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)C)C</chem> | 2.684 | (Sipos et al., 2008) |
| 7* | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)OC)C</chem> | 3.951 | (Gao, Baldessarini, et al., 1990) |
| 8 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)c1cccc1)C</chem> | 2.932 | (Sipos et al., 2008) |
| 9 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)c1ccc(cc1)O)C</chem> | 3.401 | (Sipos et al., 2008) |
| 10 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)c1ccc(cc1)C)C</chem> | 1.460 | (Søndergaard et al., 2005) |
| 11 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)c1ccc(cc1)F)C</chem> | 1.839 | (Søndergaard et al., 2005) |
| 12 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3ccc1)CC</chem> | 4.658 | (Gao, Ram, et al., 1990) |
| 13 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3ccc1)CCC</chem> | 4.097 | (Gao, Ram, et al., 1990) |
| 14 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)OC)CCC</chem> | 4.770 | (Gao, Baldessarini, et al., 1990) |
| 15 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)Br)CCC</chem> | 4.824 | (Baldessarini et al., 1991) |
| 16 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)N)CCC</chem> | 4.036 | (Baldessarini et al., 1991) |
| 17 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)O)CCC</chem> | 5.276 | (Gao, Baldessarini, et al., 1990) |
| 18 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)F)CCC</chem> | 5.921 | (Baldessarini et al., 1991) |
| 19* | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)SC)CCC</chem> | 3.428 | (Tóth et al., 2006) |
| 20 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)SCC)CCC</chem> | 3.108 | (Tóth et al., 2006) |
| 21 | <chem>c1c2c(c(c(c1)O)O)c1c3C(C2)N(CCC3cc(c1)SCCC)CCC</chem> | 2.807 | (Tóth et al., 2006) |
| 22* | <chem>c1(cccc(c1)[C@H]1CN(CCC1)CCC)O</chem> | 3.114 | (Vasdev et al., 2006) |
| 23 | <chem>c1c(c(c(c(c1)OC)C(=O)NC[C@@H]1CCCN1CC)I)S(=O)(=O)N</chem> | 3.824 | (Chumpradit et al., 1993) |

| | | | |
|-----|--|-------|---------------------------|
| 24 | <chem>c1c(c(c(c1Cl)OC)C(=O)NC[C@@H]1CCCN1CC)O)Cl</chem> | 3.959 | (Chumpradit et al., 1993) |
| 25* | <chem>c1c(c(c(c1Cl)OC)C(=O)NC[C@@H]1CCCN1CC)O)CC</chem> | 5.046 | (Chumpradit et al., 1993) |
| 26* | <chem>c1c(c(c(c1)OC)C(=O)NC[C@@H]1CCCN1CC)O)I</chem> | 4.367 | (Chumpradit et al., 1993) |
| 27 | <chem>c1c(cc(c(c1)OC)C(=O)NC[C@@H]1CCCN1CC)I</chem> | 3.523 | (Chumpradit et al., 1993) |
| 28* | <chem>c1c(cc(c(c1OC)OC)C(=O)NC[C@@H]1CCCN1CC)I</chem> | 5.602 | (Chumpradit et al., 1993) |
| 29 | <chem>c1c(c(c(c1OC)OC)C(=O)NC[C@@H]1CCCN1CC)O)I</chem> | 5.721 | (Chumpradit et al., 1993) |
| 30 | <chem>c1c(cc(c2c1CCO2)C(=O)NC[C@@H]1CCCN1Cc1ccc(cc1)F)I</chem> | 4.975 | (Chumpradit et al., 1993) |
| 31 | <chem>c1c(cc(c(c1OC)OC)C(=O)NC[C@@H]1CCCN1Cc1ccc(cc1)F)Br</chem> | 4.833 | (Chumpradit et al., 1993) |
| 32 | <chem>c1c(cc(c(c1OC)OC)C(=O)NC[C@@H]1CCCN1CC)CCCF</chem> | 5.699 | (Chumpradit et al., 1993) |
| 33 | <chem>c1ccc(c(c1OC)OCCF)C(=O)NC[C@@H]1CCCN1Cc1ccc(c1)F</chem> | 2.886 | (Chumpradit et al., 1993) |
| 34 | <chem>1cc(c(c1)OC)C(=O)NC[C@H]1N(CCC1)CC)O</chem> | 2.507 | (Murphy et al., 1990) |

Note: Compounds marked with '*' are test set compounds

3.3.2. Molecular descriptors

QSAR models were developed using a selected class of two-dimensional molecular descriptors. The descriptors were E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments and molecular property descriptors. These descriptors were calculated using Dragon 7 descriptor calculator. A total of 403 Dragon descriptors were calculated. Before the development of the QSAR model, the data was curated (Tropsha, 2010) by removing intercorrelated ($|r| > 0.95$), constant (variance < 0.0001), and other noisy and redundant data by using data pretreatment software developed in our laboratory and available from <http://dtclab.webs.com/software-tools>. After data pretreatment, the number of descriptors was reduced to 179.

3.3.3. Dataset splitting

Splitting of the dataset into training and test sets is a vital step in QSAR modeling and enables the development of a robust and well validated model. Data division must be done in such a way that the points representing both training and test set are well scattered within the whole descriptor space defined by the entire dataset. The training set is used for model development and the test set for model validation. The division of the dataset was executed by one of the most extensively used methods, Euclidean Distance division method, where the Euclidean distances for all of the compounds in the dataset are calculated and the compounds are then sorted, based on the Euclidean distance (Golmohammadi et al., 2012).

3.3.4. Variable selection and model development

The main aim of the present study has been to develop a well validated QSAR model to understand the binding of PET imaging agents towards dopamine (D2) receptor for the diagnosis of Parkinson's

disease. Critical selection of statistically significant descriptors ensures improvement in the quality of the model. Prior to development of the QSAR model, a number of significant descriptors were extracted using Double Cross Validation-Genetic Algorithm (DCV-GA) approach applied on the training set compounds (Devillers, 1996; Khan & Roy, 2018; Roy & Ambure, 2016). Finally, a Partial Least Squares (PLS) regression model was generated using descriptors selected from the best subset selection (BSS).

Double cross validation (DCV) is an attractive statistical design which combines both model generation and model assessment with the aim to produce better models (Roy & Ambure, 2016; Wold et al., 2001). Sometimes the fixed composition of a training set can lead to biased descriptor selection. DCV method helps in better descriptor selection by dividing the training set into 'n' calibration and validation sets. This results in diverse compositions of the modeling set, thus removing any bias in descriptor selection. DCV technique consists of two nested cross-validation loops commonly known as internal and external cross-validation loops. In the external loop, the data objects are split randomly into disjoint subsets known as training set compounds and test set compounds. The training set compounds are involved in the internal loop for the purpose of model development and model selection, and the test set is used solely for the intention of checking model predictivity. Further, in the internal loop, the training set compounds are repetitively split into calibration (construction) and validation sets by employing the k-fold cross validation technique (here, k=10) and producing k iterations to construct calibration and validation sets. The calibration objects are used to derive different models by altering the tuning parameter(s) of the model (i.e., the descriptors) whereas the validation objects are used to guess the models' error. The model with the lowest cross validated error is selected. The test compounds in the outer loop are employed to assess the predictive performance of the selected model.

In the current study, descriptor selection in DCV platform was done using Genetic Algorithm (GA) approach. GA is a model optimisation approach with an algorithm inspired by the theory of evolution (Devillers, 1996). GA has five basic steps: (i) coding of variables; (ii) initiation of population; (iii) evaluation of the response; (iv) reproduction; and (v) mutation. Steps (iii) to (v) alternate until a termination criterion is reached. The criterion can be based on a lack of improvement in the response or simply on a maximum number of generations or on the total time allowed for the elaboration.

3.3.5. Statistical validation metrics

Validation of the robustness and predictive ability of the developed models is a very crucial step in a QSAR study. A meticulous examination of the statistical quality of the developed model has been done to judge the robustness in terms of reliability and predictivity measures using various internal and external validation parameters. For determining the quality of the developed model, statistical parameters like determination coefficient R^2 and explained variance R_a^2 were calculated. Other parameters including internal predictivity parameters such as predicted residual sum of squares (PRESS) and leave-one-out cross-validated correlation coefficient (Q^2_{LOO}) were also calculated along with external predictivity parameters like R^2_{pred} or Q^2_{F1} , Q^2_{F2} and concordance correlation coefficient (CCC) (Roy & Mitra, 2011). Further, r_m^2 metrics (i.e., \bar{r}_m^2 and Δr_m^2) were also calculated for both training and test set compounds (Ojha et al., 2011). Validation using mean absolute error (MAE) based criteria for both external and internal validation was done (Roy et al., 2016). The Q^2_{ext} based criteria do not always interpret the correct prediction quality because of the impact of the response range as well as the distribution of the values of the response in both the training and test set compounds; so, MAE was calculated to check the average error (Roy et al., 2016). **Figure 3.2** shows the flowchart of the present work methodology.

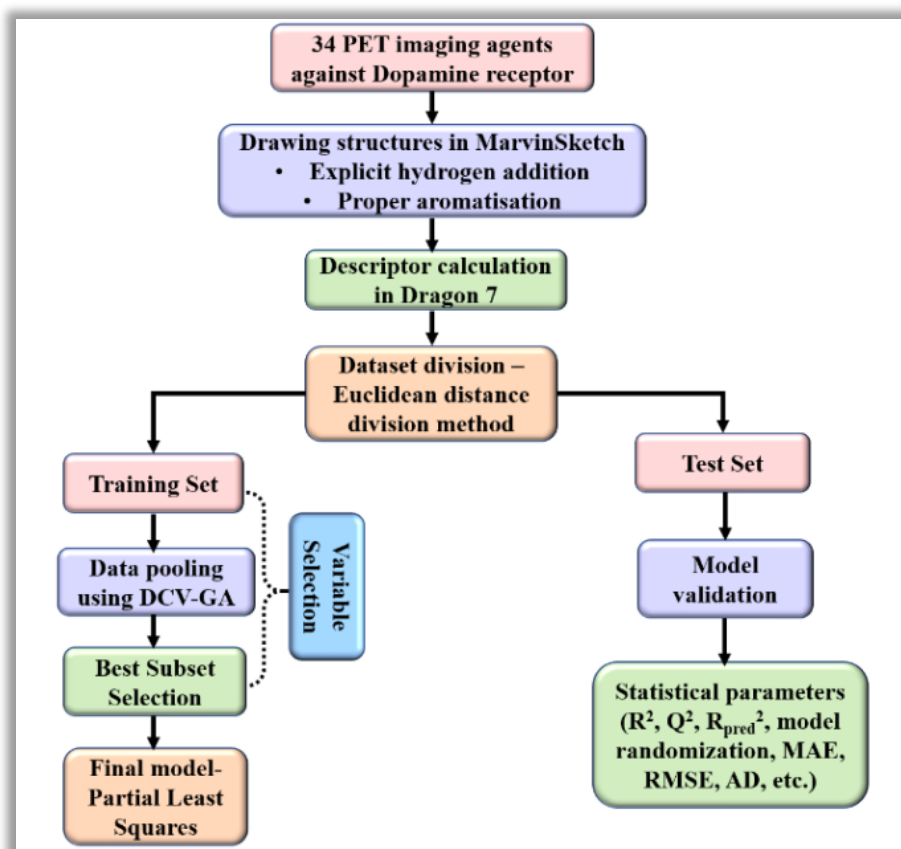


Figure 3.2. Flowchart of the present work methodology.

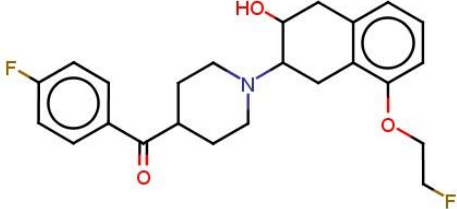
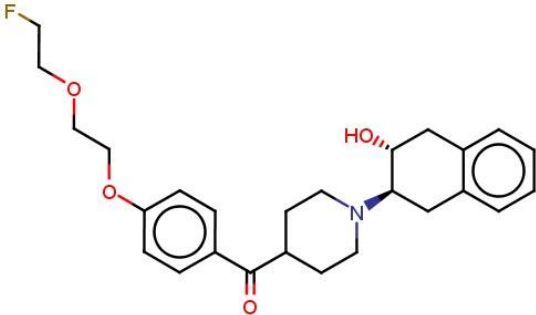
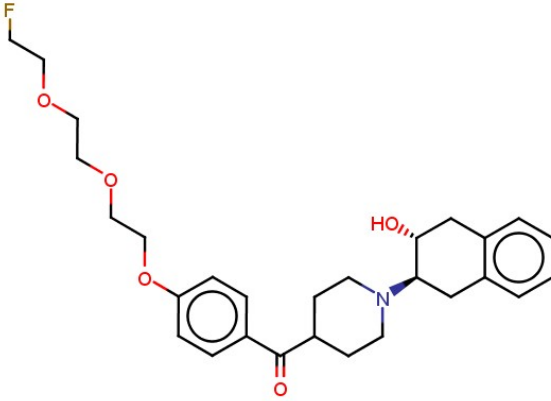
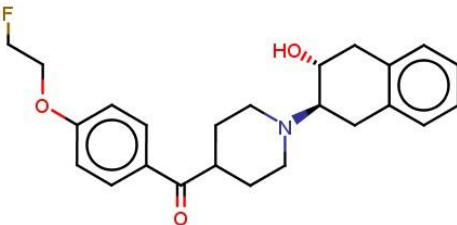
3.4. Study 4: Computational modeling of PET imaging agents against vesicular acetylcholine transporter (VAcHT) protein binding affinity: Application of 2D-QSAR modeling and molecular docking techniques

In the present study, 2D quantitative structure-activity relationship (2D-QSAR) models were developed for 19 positron emission tomography (PET) imaging agents targeted against presynaptic vesicular acetylcholine transporter (VAcHT). VAcHT assists in the transport of ACh into the presynaptic storage vesicles and it becomes one of main targets for the diagnosis of various neurodegenerative diseases.

3.4.1. The dataset

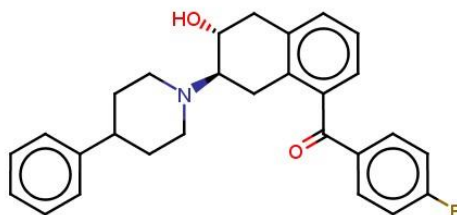
According to the OECD principle, dataset selection with a defined endpoint is first essential step while developing a QSAR model. For our present work, the binding affinity (K_i) values of 19 PET imaging agents acting against vesicular acetyl choline transporter was procured from different previously published literature (Kovac et al., 2010; Tu et al., 2009, 2015). The binding affinity data which was expressed as K_i were converted to its negative logarithmic form (pK_i). The structures obtained from different sources were then represented in MarvinSketch version 15.12.7.0 software with proper explicit hydrogen addition and aromatisation. The 19 PET imaging agents used for the present study is given in **Table 3.4**.

Table 3.4. PET radiotracers target vesicular acetylcholine transporters (VACHT)

| Compound ID | Structure | pKi |
|-------------|--|-------|
| 1 |  | 2.239 |
| 2 |  | 3.060 |
| 3 |  | 2.921 |
| 5 |  | 2.770 |

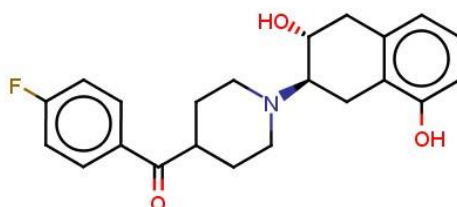
6

2.569



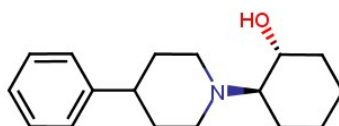
7

2.337



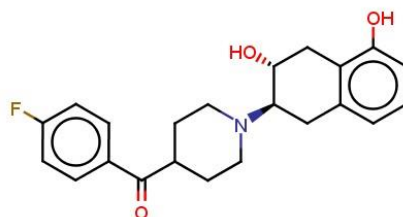
9

2.261



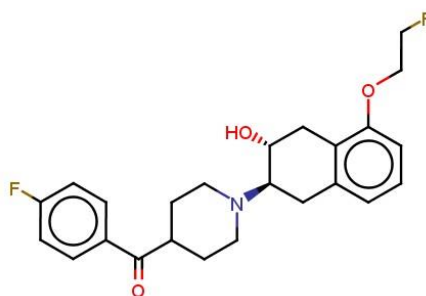
10

1.252



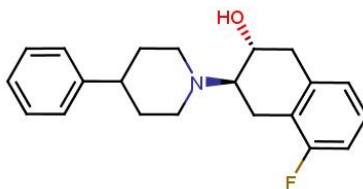
11

1.032



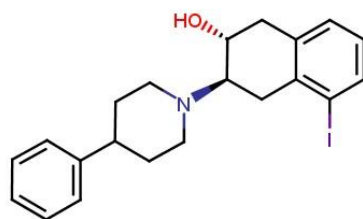
12

2.185



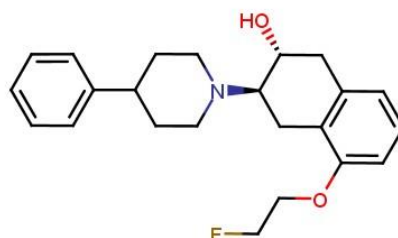
15

1.801



16

1.476



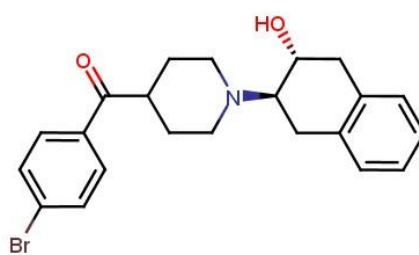
20

3.658



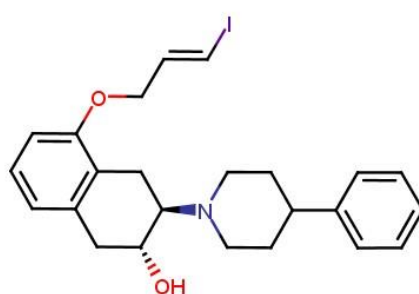
21

3.602



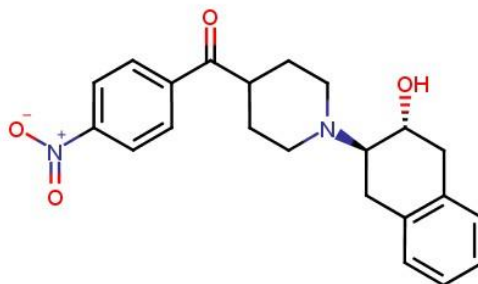
22

3.347



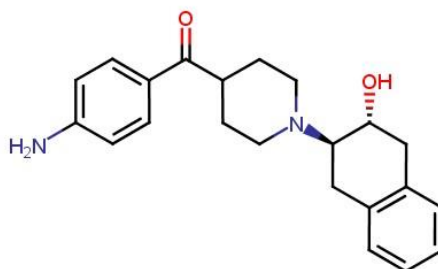
23

3.319



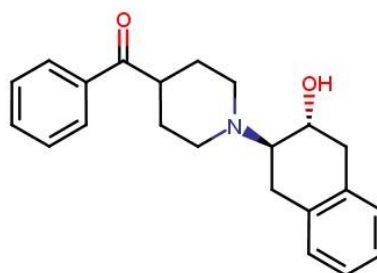
25

2.770



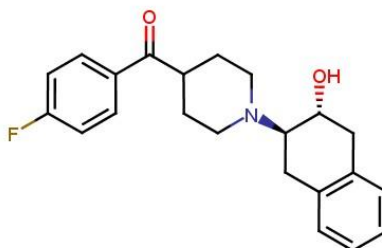
27

2.367



29

0.967



3.4.2. Molecular descriptors

The molecule descriptor is a fundamental component of QSAR and other in-silico models since it formally represents a molecule's structure numerically. Descriptors provide a mathematically meaningful relationships between the molecular structure and biological activities, physico-chemical and toxicological properties of chemicals (Mauri et al., 2017). Descriptors can be classified into different categories depending on the process of calculation or scheme of experimental determination or concept of the origin. For the ease of interpretation, the present work involved the use of eight main types of two-dimensional (2D) descriptors, viz., E-state indices, extended topochemical atom (ETA), connectivity, constitutional, functional, 2D atom pairs, ring, atom centered fragments and molecular property descriptors. The descriptors were calculated using alvaDesc descriptor calculator

(Alvascience, alvaDesc version 2.0.6, 2021, <https://www.alvascience.com>). With the intention to minimize the redundant and incompetent data, inter-correlated descriptors (correlation greater than 0.95) were removed from the original descriptor pool. This resulted in a final pool of 188 descriptors which was used as input variables for QSAR modeling.

3.4.3. Feature selection and model development

In general, a QSAR model development involves a training set and a test for model development and validation purposes respectively. However, owing to the small number of compounds in our dataset, we did not perform the general method of data division. It is natural that all the descriptors calculated through AlvaDesc will not be able to describe the binding properties of the PET imaging agents. Therefore, to further reduce the data pool, we have applied Genetic Algorithm (Sukumar et al., 2014) feature selection method to choose essential features required for binding. Further, we have executed the Best Subset Selection (available at <http://dtclab.webs.com/software-tools>) on the reduced pool of 12 descriptors obtained from the GA. Finally, the acquired pool of descriptors was applied to develop the final model using the partial least squares (PLS) regression (Wold et al., 2001).

3.4.4. Machine learning based read across prediction

In the current work, we have employed a machine learning based Read Across prediction which relies on similarity approaches. The predictions were made using the tool Quantitative Read Across v4.0 developed by Chatterjee *et al.* (Chatterjee *et al.*, 2022b) available at <https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>. The main similarity approaches involved in this tool are Euclidean distance-based similarity, Gaussian kernel function, and Laplacian kernel function-based similarity estimation. For this method we have divided the dataset into training and test sets. The prediction scheme starts with initial optimisation of hyperparameters (sigma and gamma values; distance and similarity threshold) which requires division of the training into sub-training and sub-test sets into different combinations. This step is followed by selection of the best setting of hyperparameters which is then applied to the original training and test sets.

3.4.5. Molecular Docking

In this study, the molecular docking study was performed using the most and least active compounds from the initial dataset to identify the interaction pattern with the target. Owing to the unavailability of any protein structure for VACHT in protein data bank, we have retrieved the predicted protein structure from the AlphaFold Protein Structure Database (Available from <https://alphafold.ebi.ac.uk/entry/Q16572>) with the UniProt: Q16572, Source organism: Homo sapiens (Human), and AlphaFold id: AF-Q16572-F1-model_v2. We have then validated the reliability of the predicted structure using the Ramachandran plot server embedded in Biovia Discovery Studio 4.1 which represents the good quality of the model (see **Figure 3.3**). In this study, multiple active sites at the surface of the protein were predicted using the Biovia discovery studio 4.1 client platform from the “define and edit binding site” using the module “generate active site from receptor cavities”, and the ligand was docked into each site to identify the favorable binding site (identified most favorable active site coordinate x: 16.478, y: 6.38307, Z: -15.9527, the radius of the sphere: 26). Initially, a total of sixteen binding sites were identified where the standard compound “vesamicol” was docked. Ligand preparation was performed using selected high and low active compounds by running them through the Discovery Studio platform's ‘small-molecule module’, where several ligand conformers were formed. Each of these generated conformers was subsequently employed in the CDOCKER module for molecular docking using a CHARMM-based molecular dynamic scheme (Wu et al., 2003). The CDOCKER interaction energy parameter (kcal/mol) was examined for all receptor-ligand complexes, and the highest-scoring (more negative; hence favorable to binding) poses with only non-covalent interactions (ionic bonds, hydrophobic interactions, hydrogen bonds, etc.) were kept for future investigation.

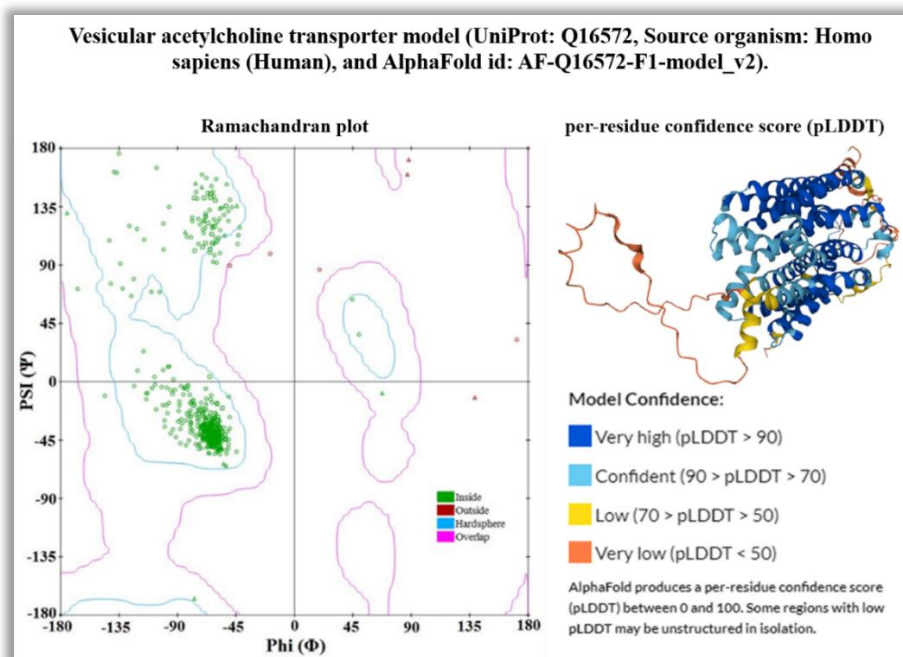


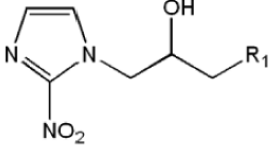
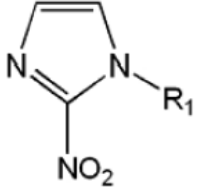
Figure 3.3. Ramachandran plot and per-residue confidence score (pLDDT) for Vesicular acetylcholine transporter model (UniProt: Q16572, Source organism: Homo sapiens (Human), and AlphaFold id: AF-Q16572-F1-model_v2). Ramachandran plot shows 435 residues (97.098%) residues in most favoured region, 10 (2.232%) residues reside in the preferable region and only 3 (0.670%) residues in the unfavourable region. The predicted structures contain atomic coordinates and per-residue confidence estimates on a scale from 0 to 100, with higher scores corresponding to higher confidence. This confidence measure is called pLDDT.

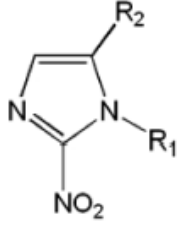
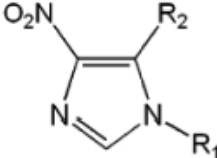
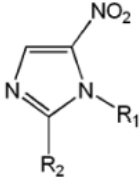
3.5. Study 5: Exploration of nitroimidazoles as radiosensitizers: Application of multilayered feature selection approach in QSAR modeling

A data of 86 nitroimidazoles (Table 3.5) possessing radiosensitizing properties are used for two-dimensional QSAR (2D-QSAR) study (Long & Liu, 2010). Radiosensitization capacities of the compounds can be understood by radiosensitization effectiveness, expressed as $C_{1.6}$, which can be represented as the corresponding concentration of a given compound when its sensitization enhancement ratio (SER) accomplishes 1.6. A higher value of $C_{1.6}$ indicates lower bioactivity of radiosensitization effectiveness. For analysis purpose, the source literature had converted the endpoint $C_{1.6}$ to its negative logarithmic scale ($pC_{1.6}$, where $pC_{1.6} = -\log(C_{1.6})$). Two compounds (one radical and one salt) were removed and the final dataset of 84 compounds is used for model development. The structures of the compounds were drawn in MarvinSketch software (version 14.10.27) [with proper aromatization and hydrogen bond addition and saved as MDL .mol, a recommended format for further descriptor calculation.

Table 3.5. Nitroimidazole dataset and their respective observed radiosensitization effectiveness values ($pC_{1.6}$).

| Compound Number | Structure | R ₁ | R ₂ | pC _{1.6} |
|-----------------|-----------|--------------------------------------|----------------|-------------------|
| 1 [#] | | -OCH ₃ | - | 3.52 |
| 2 | | -OH | - | 3.00 |
| 3 [#] | | -OCH(CH ₃) ₂ | - | 3.00 |
| 4 | | -OCH ₂ CH=CH ₂ | - | 3.59 |

| | | | | | |
|-----------------|---|---|--|-----------------------------------|------|
| 5 [#] | | -F | - | 3.54 | |
| 6 | | -NHC(CH ₃) ₃ | - | 3.60 | |
| 7 | | -NHC ₆ H ₄ OCH ₃ (p) | - | 2.89 | |
| 8 [#] | | -NHCH ₂ C ₆ H ₅ | - | 4.22 | |
| 9 | | -NHCH ₂ C ₆ H ₄ OCH ₃ (p) | - | 3.89 | |
| 10 | | -NHCH ₂ C ₆ H ₄ OCH ₃ (o) | - | 4.10 | |
| 11 | | -NHCH ₂ C ₆ H ₄ OCH ₃ (m) | - | 4.22 | |
| 12 [#] | | -NHCH(Cyclo-C ₆ H ₁₁) | - | 4.00 | |
| 13*(P-1) | | -NH-2,2,6,6-tetramethylpiperidine-O· | - | 4.05 | |
| 14 |  | -N(CH ₃) ₂ | - | 3.82 | |
| 15 | | -N(CH ₂ CH ₃) ₂ | - | 3.89 | |
| 16 | | -N(cyclo-C ₆ H ₁₁) ₂ | - | 3.92 | |
| 17 [#] | | -aziridine | - | 3.12 | |
| 18 | | -pyrrolidine | - | 3.11 | |
| 19 | | -pyrrolidin-3-ol | - | 2.96 | |
| 20 | | -pyrrolidin-2-ylmethanol | - | 3.89 | |
| 21 [#] | | -piperidine | - | 4.00 | |
| 22 | | -piperidin-3-ol | - | 4.10 | |
| 23 | | -piperidin-4-ol | - | 3.80 | |
| 24 | | -piperidin-3-ylmethanol | - | 3.74 | |
| 25 [#] | | -piperidin-2-ylmethanol | - | 3.48 | |
| 26 | | -morpholine | - | 3.40 | |
| 27 | | -piperazine-CH ₃ | - | 3.70 | |
| 28 | | -piperazine-CH ₂ -CH ₂ -OH | - | 3.40 | |
| 29 [#] | | -azepane | - | 4.00 | |
| 30 | | | -CH ₂ OH | - | 3.52 |
| 31 | | | -CH ₂ CH ₂ OCH ₃ | - | 3.30 |
| 32 | | | -CH ₂ CH ₂ OCH ₂ CH ₃ | - | 3.30 |
| 33 | | | -CH ₂ CH ₂ OC ₆ H ₅ | - | 3.60 |
| 34 | | | -CH ₂ COCH ₃ | - | 3.40 |
| 35 [#] | | | -CH ₂ COOCH ₃ | - | 3.77 |
| 36 | | | -CH ₂ CONHCH ₂ CH ₂ OH | - | 3.52 |
| 37 [#] | | | -CH ₂ CH ₂ SO ₂ CH ₃ | - | 3.70 |
| 38 | | | -CH ₂ CH ₂ SO ₂ CH ₂ CH ₃ | - | 3.52 |
| 39 | | | -CH ₂ CH ₂ SO ₂ C ₆ H ₅ | - | 3.85 |
| 40 [#] | | | -(CH ₂) ₂ -N(CH(CH ₃) ₂) ₂ | - | 4.15 |
| 41 | |  | -(CH ₂) ₂ -2-pyridyl | - | 3.30 |
| 42 | -(CH ₂) ₂ -pyrrolidino | | - | 3.52 | |
| 43 | -(CH ₂) ₄ -pyrrolidino | | - | 4.12 | |
| 44 [#] | -(CH ₂) ₈ -pyrrolidino | | - | 4.22 | |
| 45 | -(CH ₂) ₂ -piperidino | | - | 4.12 | |
| 46 | -(CH ₂) ₃ -piperidino | | - | 4.19 | |
| 47 | -(CH ₂) ₄ -piperidino | | - | 4.28 | |
| 48 [#] | -(CH ₂) ₆ -piperidino | | - | 4.40 | |
| 49 | -(CH ₂) ₈ -piperidino | | - | 3.92 | |
| 50 | -(CH ₂) ₃ -morpholino | | - | 3.40 | |
| 51 | -(CH ₂) ₄ -morpholino | | - | 4.15 | |
| 52 [#] | -(CH ₂) ₅ -morpholino | | - | 4.30 | |
| 53 | -(CH ₂) ₆ -morpholino | | - | 4.22 | |
| 54 | -(CH ₂) ₈ -morpholino | | - | 4.10 | |
| 55 | -(CH ₂) ₁₁ -morpholino | | - | 3.92 | |
| 56*(P-2) | | | -(CH ₂) ₄ -4-methylmorpholine-I | - | 3.00 |
| 57 | | | -CH ₃ | -CH ₃ | 3.74 |
| 58 | | | -CH ₃ | -CH ₃ =CH ₂ | 2.77 |

| | | | | | |
|-----------------|--|---|---|------------------|------|
| 59 |  | -CH ₃ | -CH(CH ₃) ₂ | 3.26 | |
| 60 [#] | | -CH ₃ | -CH ₂ OH | 3.00 | |
| 61 | | -CH ₃ | -C(CH ₃) ₂ OH | 3.22 | |
| 62 | | -CH ₃ | -CH(OH)CH ₂ OH | 4.70 | |
| 63 | | -CH ₃ | -CHO | 4.00 | |
| 64 [#] | | -CH ₃ | -COOCH ₃ | 3.82 | |
| 65 | | -CH ₃ | -CH=N(O)CH ₃ | 3.92 | |
| 66 | | -CH ₃ | -CH=N-NH ₂ | 3.60 | |
| 67 [#] | | -CH ₃ | -CH=N-piperidino | 3.70 | |
| 68 | | -CH ₃ | -CH=N-piperizino | 3.00 | |
| 69 | | -CH ₂ CH ₂ OH | -CH ₃ | 3.46 | |
| 70 | | -CH ₂ COOCH ₂ CH ₃ | -CH ₂ CH ₃ | 1.85 | |
| 71 | |  | -CH ₂ CH ₂ -morpholino | H | 1.80 |
| 72 | | | -CH ₂ CH(OH)CH ₂ OCH ₃ | -CH ₃ | 2.17 |
| 73 [#] | H | | -S-CH ₂ -COO-C ₂ H ₅ | 2.77 | |
| 74 | H | | -S-1'-(3-aminopurine) | 2.43 | |
| 75 | H | | -SO ₂ -NH ₂ | 2.77 | |
| 76 [#] | H | | -SO ₂ -N(CH ₃) ₂ | 2.74 | |
| 77 | H | | -SO ₂ -NH-phenyl | 2.40 | |
| 78 | H | | -SO ₂ -NH-CH ₂ -morpholine | 2.59 | |
| 79 [#] | H | | -SO ₂ -O-phenyl | 2.59 | |
| 80 | H | | -SO ₂ -O-4-chlorophenyl | 2.60 | |
| 81 |  | | -CH ₃ | -CH ₃ | 2.40 |
| 82 | | | -CH ₂ CH ₂ OH | -CH ₃ | 2.85 |
| 83 | | -CH ₂ CH(OH)CH ₂ OH | -CH ₃ | 2.70 | |
| 84 | | -CH ₂ CH(OH)CH ₂ Cl | -CH ₃ | 2.62 | |
| 85 | | -CH ₂ CH ₂ -morpholino | H | 2.57 | |
| 86 | | -CH ₂ CH ₂ SO ₂ CH ₂ CH ₃ | -CH ₃ | 3.52 | |

[#] Test set compounds.

* Two compounds are present in the prediction set.

3.5.1. Descriptor calculation

For developing the first 2D-QSAR model, a pool of 270 descriptors was calculated using Dragon version 7 (available at <http://www.taletе.mi.it/index.htm>.) software. This model was developed using specific classes of descriptors including E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom centered fragments and molecular property descriptors. Additionally, SiRMS descriptors were calculated using SiRMS (Version 4.1.2.270) (Kuz'min et al., 2005) tool. Simplex representations of molecular structure (SiRMS) descriptors symbolize a class of diverse molecular features developed from 1D to 4D molecular structures. These are tetratomic fragments of different simplex descriptors having predefined chirality, composition and symmetry (Kuz'min et al., 2005). SiRMS descriptors consider both connected and unconnected fragments and also take into account not only the nature of atoms but also their different chemical and physical properties like charge, lipophilicity, electronegativity, atomic refraction, donor/acceptor of hydrogen in the potential Hbond, etc. In our study, we have used 2D SiRMS descriptors only in order to avoid conformational complexity and energy minimization requirements for higher dimensional descriptors and to derive reproducible models. The constant (variance<0.0001), intercorrelated ($|r| > 0.95$) and other incompetent data were removed using an in-house software available at <http://dtclab.webs.com/software-tools> before model development.

3.5.2. Dataset splitting

A well validated QSAR model is the main objective of any QSAR study which can be obtained through proper division of the dataset into training (used for model development) and test (used for model validation) sets. An unbiased external validation with uniform distribution of compounds into training and test sets can be obtained through rational dataset division (Golbraikh et al., 2003). For 2D-QSAR modeling, the whole dataset utilized for modeling was divided into training (75%) and test (25%) sets using modified k-Medoids (Modified k-medoid GUI 1.3) (Park & Jun, 2009) method of dataset division.

3.5.3. Variable selection and QSAR model development

Development of well-validated QSAR models was the main aim of the present study in order to understand the radiosensitization effectiveness of the dataset compounds. Critical evaluation process helped in the selection of statistically significant models. In this study, we have built two QSAR models; a 2D-QSAR model to deduce a relationship between the molecular properties of the nitroimidazoles and their radiosensitization properties. For the model with Dragon descriptors, a pool of 32 descriptors were selected using Genetic Algorithm (GA) (Devillers, 1996; Khan & Roy, 2018) modeling implemented in double cross validation (DCV) (Roy & Ambure, 2016) tool (version 1.2). Then, the final model was generated using Partial Least Squares (PLS) regression (Khan & Roy, 2018; Wold et al., 2001) method using descriptors selected from best subset selection (BSS). In case of SiRMS, the number of descriptors generated was large, i.e., about more than ten thousand. Handling of this large data is very complicated and so we have applied stepwise regression on the large pool of SiRMS descriptors to find out the essential descriptors contributing to the radiosensitization properties of the dataset. After descriptor thinning, the obtained pool of 300 descriptors was further subjected to multilayered stepwise regression to obtain a manageable number of descriptors and run best subset selection for development of five descriptors models. From the developed models obtained after best subset selection, we have selected one model based on different validation parameters for the test set. Finally, we have run a partial least squares regression (PLS) using SIMCA-P software (available at www.umetrics.com) and developed a PLS model.

3.5.4. Statistical validation metrics and domain of applicability

The statistical quality of the derived models was rigorously checked to judge the robustness in terms of reliability and predictivity measures using various internal and external validation parameters. In the present work we have computed various statistical parameters like determination coefficient R^2 , explained variance R_a^2 , variance ratio (F), and standard error of estimate (s). Since these quality parameters are not sufficient to assess the predictive ability of the model, additional metrics were computed that could properly validate the predictions. For internal predictions, leave-one-out cross-validation ($Q_{(LOO)}^2$) was reported, and for external predictions, parameters like R_{pred}^2 or Q_{F1}^2, Q_{F2}^2 and concordance correlation coefficient (CCC), were calculated (Roy & Mitra, 2011). We have also calculated r_m^2 metrics (i.e., $\overline{r_m^2}$ and Δr_m^2) for both training and test set compounds (Kunal Roy & Mitra, 2011). We have also validated the models using mean absolute error (MAE) based criteria for both external and internal validation (Roy et al., 2016). This was done since the Q_{ext}^2 based criteria do not always offer the correct indication of the prediction quality because of the influence of the response range as well as the distribution of the values of response in both the training and test set compounds (Roy et al., 2016). The Applicability Domain (AD) gives a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable (Gadaleta et al., 2016). AD assessment for both the models was performed using DModX (distance to model in the X-space) approach at 99% confidence level.

3.6. Study 6: QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: Application of Small Dataset Modelling

The present study explores the features essential to show radiosensitization properties by nitroimidazole sulphonamide derivatives using QSAR and quantitative structure activity-activity relationship (QSAAR) modelling (Lessigiarska et al., 2006). Two-dimensional (2D) descriptors obtained from Dragon and SiRMS software were utilised during the development of well validated models. A small dataset of nitroimidazole sulfonamides is used for modelling in the current study where splitting of dataset into training and test sets would cause loss of chemical information leading to unreliable models.

3.6.1. Dataset

In vitro radiosensitization data of selected compounds involving sensitizer enhancement ratio (drug SER) and survival ratio (drug SR) was obtained from a previously published research work (Bonnet et al., 2018). A dataset of 21 compounds given in **Table 4.6** was selected for 2D-QSAR modeling. Sensitizer Enhancement Ratio (**SER**) can be defined as the ratio of radiation dose for 1% survival without or with the drug in a condition where HCT116 cells (human colorectal carcinoma cell line) were exposed to the drug at 6–29 Gy radiation for 1 hour. Survival Ratio can be explained using the following expression: “SR= (cell survival with radiation)/(cell survival with drug and with radiation) interpolated from the radiation dose response curves at 15 Gy”. During modeling, the drug SER values were used as provided in the original article but drug SR values were converted into their logarithmic form (**logSR**) for analysis. The compounds were drawn in MarvinSketch software (version 14.10.27) (available at <https://chemaxon.com/marvin>) with hydrogen bond addition and proper aromatization and saved as MDL.mol, a suggested format for further descriptor calculation.

Table 4.6: Dataset of 21 compounds used for modeling.

| Serial Number | Compound Number | Structure (SMILES) | Drug SER | Log Drug SR |
|---------------|-----------------|---|----------|-------------|
| 1 | 1 | <chem>c1(n(ccn1)CC(COC)O)[N+](=O)[O-]</chem> | 1.4 | 0.833 |
| 2 | 2 | <chem>c1(n(ccn1)CC(=O)NCCO)[N+](=O)[O-]</chem> | 1.339 | 0.663 |
| 3 | 4 | <chem>c1n(c(cn1)[N+](=O)[O-])CCN1CCOCC1</chem> | 1.8 | 1.652 |
| 4 | 6 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCOC)[N+](=O)[O-]</chem> | 1.2 | 0.462 |
| 5 | 7 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCO)[N+](=O)[O-]</chem> | 1.11 | 0.255 |
| 6 | 8 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCN1CCOCC1)[N+](=O)[O-]</chem> | 1.28 | 0.591 |
| 7 | 12 | <chem>c1(n(ccn1)CS(=O)(=O)NN1CCOCC1)[N+](=O)[O-]</chem> | 1.11 | 0.301 |
| 8 | 14 | <chem>c1(n(ccn1)CCS(=O)(=O)NCCCO)[N+](=O)[O-]</chem> | 1.27 | 0.623 |
| 9 | 15 | <chem>c1(n(ccn1)CCS(=O)(=O)NCCCN1CCOCC1)[N+](=O)[O-]</chem> | 1.357 | 0.699 |
| 10 | 16 | <chem>c1n(cc(n1)[N+](=O)[O-])CS(=O)(=O)NCCCOC</chem> | 1.105 | 0.114 |
| 11 | 19 | <chem>c1n(c(cn1)[N+](=O)[O-])CS(=O)(=O)NCCCO</chem> | 1.81 | 2.057 |
| 12 | 21 | <chem>c1n(c(cn1)[N+](=O)[O-])CS(=O)(=O)NCCCN1CCOCC1</chem> | 1.43 | 0.914 |
| 13 | 22 | <chem>c1n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC</chem> | 1.56 | 1.415 |
| 14 | 24 | <chem>c1n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCO</chem> | 1.81 | 2.212 |
| 15 | 26 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC)C</chem> | 1.34 | 0.681 |
| 16 | 28 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCO)C</chem> | 1.176 | 0.208 |
| 17 | 30 | <chem>c1(n(c(cn1)[N+](=O)[O-</chem> | 1.68 | 1.447 |

| | | | | |
|----|----|---|------|-------|
| | |)CCS(=O)(=O)NCCCCN1CCOCC1)C | | |
| 18 | 31 | c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN(C)C)C | 1.57 | 1.173 |
| 19 | 34 | c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN1CCCC1)C | 1.54 | 1.134 |
| 20 | 35 | c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN1CCCC1)C | 1.71 | 1.380 |
| 21 | 38 | c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NN1CCC(CC1)N(C)C)C | 1.67 | 1.398 |

3.6.2. Molecular descriptors

The molecular descriptor is the “final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment” (Consonni & Todeschini, 2010). A selected class of 356 2D molecular descriptors was calculated from Dragon version 7 (software available at <http://www.taletе.mi.it/index.htm>.) software. These comprised E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom centered fragments and molecular property descriptors. Intercorrelated ($|r| > 0.95$) and constant (variance < 0.0001) variables and other incompetent data were removed using a software available at <http://dtclab.webs.com/software-tools> prior to model development. This resulted in 224 Dragon descriptors which were used for modeling. Further, SiRMS descriptors were calculated using SiRMS (version 4.1.2.270) (Kuz'min et al., 2005) tool and used along with Dragon descriptors during modeling. Simplex representations of molecular structure (SiRMS) descriptors are a class of molecular descriptors developed from 1D to 4D molecular structures involving tetratomic fragments of different simplex descriptors having predefined chirality, composition, and symmetry (Kuz'min et al., 2005).

3.6.3. Model development: Application of Small Dataset Modeler

Before development of a QSAR model, the dataset is generally divided into a training set (calibration) and a test set (validation). Further, a double cross validation method (Roy & Ambure, 2016) of model development involves two nested cross-validation loops: internal (inner) and external (outer) cross-validation loops as elaborately discussed in **Section 3.1.4**. However, the present study deals with a small dataset containing a limited number of data points (21 compounds), and splitting of this dataset into training and test sets is not desirable. Small dataset modeling (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/) involves the DCV method of modeling for small datasets without dividing the dataset into training and test sets (Ambure et al., 2019). Here the ‘modeling set’ in the inner loop is not generated. However, deriving all possible combinations (k) of the validation set (containing n compounds) and the calibration set (containing n – r compounds) is followed. The tool has an option for the user to define the number of compounds to be kept in the validation set (r) depending on which the calibration and validation sets are defined. Calibration set compounds are used for the generation of Genetic Algorithm-Multiple Linear Regression (GA-MLR) (Devillers, 1996) models, and the validation sets are utilized for model prediction purpose. A number of internal and external validation metrics are calculated in the exhaustive double cross-validation technique for all the selected models. Additionally, the software also derives Partial Least Squares (PLS) (Wold et al., 2001) regression models corresponding to each MLR model. Further, the selection of best/top model can be done in any of the five following methods mentioned:

- i) Model (MLR/PLS) with the lowest mean absolute error or MAE (95%) in the validation set is selected.

- ii) Model (MLR/PLS) with the lowest MAE (95%) in the modeling set is selected.
 - iii) Model (MLR/PLS) with the highest $Q_{\text{Leave-many-out}}^2$ (modeling set).
 - iv) Application of consensus modeling by using top ranking models selected based on the MAE (95%) values in the respective validation sets. Two types of consensus approaches include: a) simple arithmetic average of predictions from all the selected top models. b) weighted average of predictions by assigning appropriate weights to the selected top models based on the mean absolute error obtained from leave-one-out cross-validation, $\text{MAE}_{\text{cv}(95\%)}$.
 - v) A pool of unique descriptors from the top 3 models with lowest MAE (95%) of the validation set is used. These descriptors are used for further model development purpose. In case of MLR, Best Subset Selection (BSS) method is used which finds the best combinations of descriptors out of all the possible combinations of unique descriptors present in the selected models. In case of PLS models, the models are formed by all descriptors selected in the top models through a PLS run.
- The approach proposed in small dataset modeler (**Figure 3.4**) thus ensures the division of small dataset internally within the DCV algorithm without the actual need of a test set. Thus, there is no requirement of the dataset division. The small dataset modeling approach combines data curation, exhaustive double cross validation, and optimal model approaches including consensus predictions for model development, particularly for small datasets.

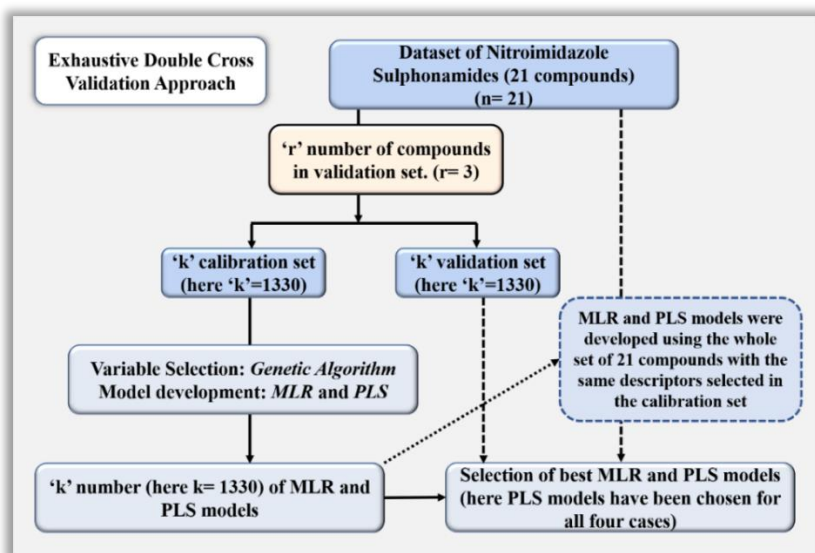


Figure 3.4. The approach adopted to develop QSAR models for small-sized dataset using Small Dataset Modeler

3.6.4. Statistical validation metrics

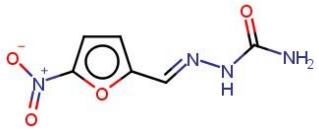
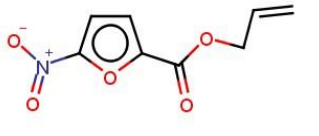
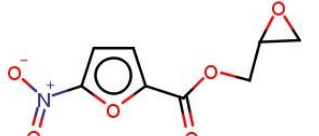
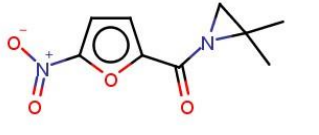
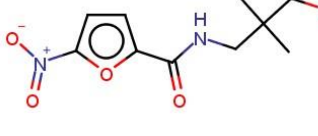
A rigorous analysis using multiple approaches of assessment of the model quality for measurement of the fitness, stability, robustness, and predictivity of the developed models was carried out. In the present work we have computed various statistical parameters like determination coefficient (R^2) and leave one out squared correlation coefficient (Q_{LOO}^2) for internal validation. We have also calculated the leave-many-out squared correlation coefficient ($Q_{LMO(20\%)}^2$) for the final PLS models (Roy et al., 2015a). Further, r_m^2 metrics (Ojha et al., 2011), root mean square error (RMSE), and mean absolute error (MAE) were also calculated (Roy et al., 2016).

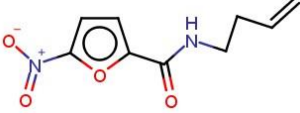
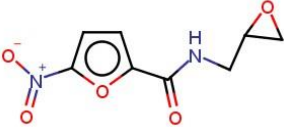
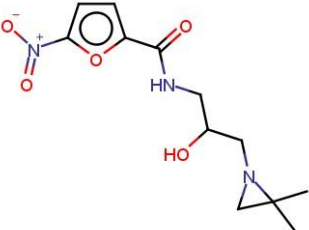
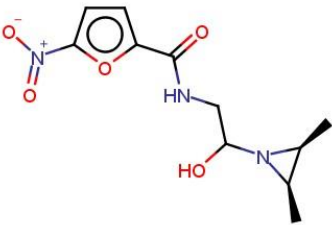
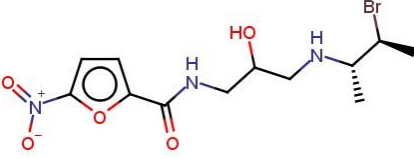
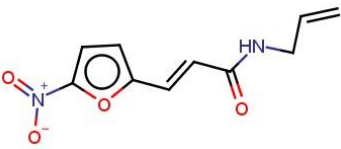
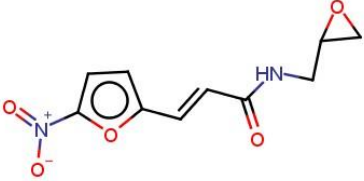
3.7. Study 7: Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness

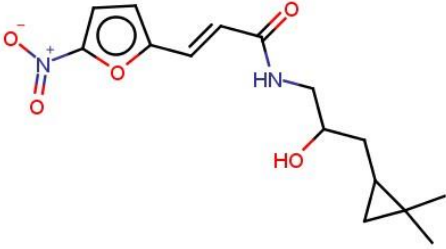
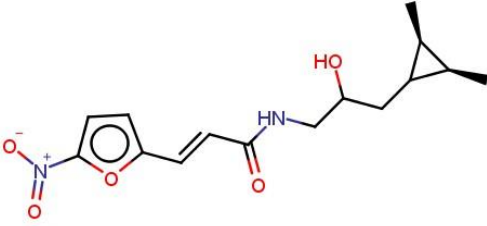
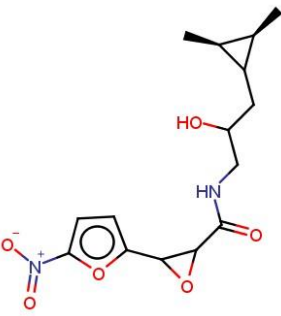
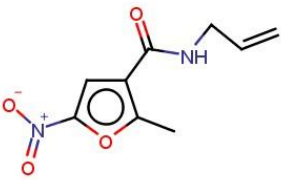
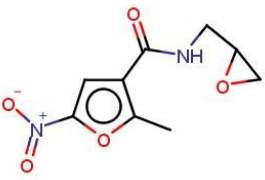
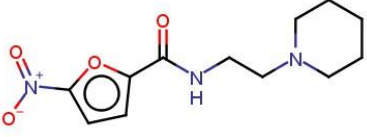
3.7.1. The dataset

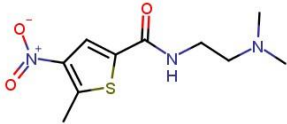
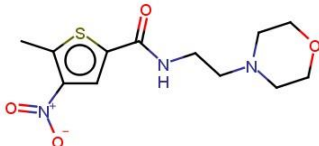
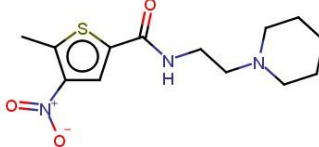
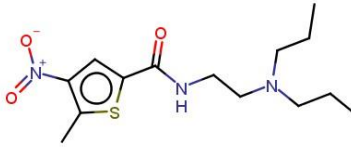
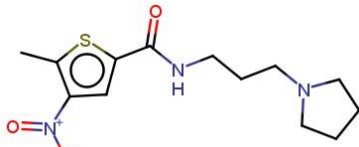
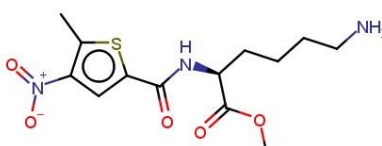
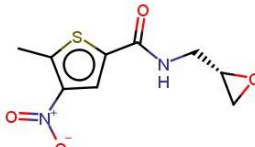
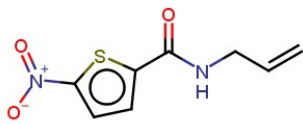
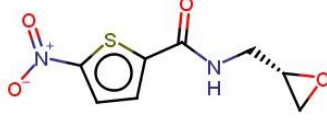
The radiosensitization effectiveness ($pC_{1.6}$) data for three nitroaromatics datasets (nitrofurans, nitrothiophenes and nitroimidazoles) were obtained from the previously published literature (Long & Liu, 2010; Naylor et al., 1990; Threadgill et al., 1991). The datasets comprised 18 nitrofurans, 11 nitrothiophenes and 84 nitroimidazole derivatives in the composite set. ' $C_{1.6}$ ' is a term used to explain the radiosensitization capacities; this is the molar concentration of the compound required to give a sensitizer enhancement ratio (SER) of 1.6. Thus, lower value for $C_{1.6}$ will give greater sensitizing efficiency. For an efficient analysis, the $C_{1.6}$ values were converted into their negative logarithmic scale ($pC_{1.6}$). The structures in the datasets were drawn in MarvinSketch software (version 14.10.27) (software available at <https://chemaxon.com/marvin>) with proper aromatization and hydrogen bond addition and saved as MDL.mol format.

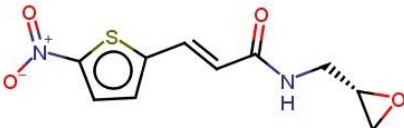
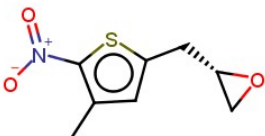
Table 4.7. Experimental radiosensitization effectiveness ($pC_{1.6}$) data for three nitroaromatics datasets (nitrofurans, nitrothiophenes and nitroimidazoles)

| Nitrofuran Dataset | | |
|--------------------|---|------------|
| Compound ID | Structure | $pC_{1.6}$ |
| NF-1 |  | 1.301 |
| NF-2 |  | 1.301 |
| NF-3 |  | 1.602 |
| NF-4 |  | 1.000 |
| NF-5 |  | 1.301 |

| | | |
|-------|--|-------|
| NF-6 |  | 1.301 |
| NF-7 |  | 1.699 |
| NF-8 |  | 1.398 |
| NF-9 |  | 1.301 |
| NF-10 |  | 1.523 |
| NF-11 |  | 1.523 |
| NF-12 |  | 2.097 |

| | | |
|-------------------------------|---|--------------|
| NF-13 |  | 2.097 |
| NF-14 |  | 1.456 |
| NF-15 |  | 1.602 |
| NF-16 |  | 1.125 |
| NF-17 |  | 1.398 |
| NF-18 |  | 2.000 |
| Nitrothiophene Dataset | | |
| Compound ID | Structure | pC1.6 |

| | | |
|------|---|-------|
| NS-1 |  | 1.000 |
| NS-2 |  | 0.000 |
| NS-3 |  | 1.155 |
| NS-4 |  | 1.155 |
| NS-5 |  | 1.523 |
| NS-6 |  | 1.301 |
| NS-7 |  | 0.699 |
| NS-8 |  | 1.097 |
| NS-9 |  | 1.301 |

| | | |
|---|---|-------|
| NS-10 |  | 1.301 |
| NS-11 |  | 1.222 |
| Nitroimidazole dataset (Same as given in Study 5: Table 3.5) | | |

3.7.2. Descriptor calculation

Before a QSAR model is developed, the structural information is converted into numerical values known as descriptors (Todeschini & Consonni, 2008). The three curated datasets were used for the calculation of descriptors using Dragon version 7 software. Specific classes of descriptors were used for model development including: connectivity, constitutional, topological, E-state indices, functional, 2D atom pairs, 2D autocorrelation, ring, atom-centered fragments and molecular property descriptors. Descriptors were pre-treated to reduce redundant and noisy data; constant (variance < 0.0001) and intercorrelated ($|r| > 0.95$) variables were removed using an in-house software available at <http://dtclab.webs.com/software-tools> before model development.

3.7.3. Data set splitting and model development

Rational splitting of a dataset into training and test sets is a crucial step before a QSAR model development leading to the establishment of the models' predictive power. However, a general problem faced by in silico researchers during the development of ideal QSAR models is the non-availability of sufficient data suitable for data set splitting. Datasets with 25-50 datapoints or even less are difficult to divide into training and test sets and there is less chance of getting robust and predictive models. Ambure et al. (Ambure et al., 2019) proposed a method for small datasets which does not require the step of data set division. "Small dataset modelling" as proposed by these authors involves the Double Cross Validation (DCV) method (Baumann & Baumann, 2014; Roy & Ambure, 2016). In this method, the entire dataset of n compounds is taken under consideration. The process involves the generation of all possible combinations (k) of the validation set (each containing r compounds) and the calibration set (containing $n - r$ compounds). Here, the user is allowed to set the 'r' value, i.e., the number of compounds to be retained in the validation set and depending on that all probable combinations of calibration and validation sets are generated. The models are generated using Multiple Linear Regression (MLR) (Aiken et al., 2003) method using Genetic Algorithm (GA) method of feature selection. In this scheme of exhaustive DCV, several important validation metrics are calculated for all the elected models. The selection of the best models is dependent on a set of criteria discussed in the source literature (Ambure et al., 2019). In the current study, the number of data points for nitrofurans and nitrothiophenes is relatively very small (18 and 11 respectively) for dataset division. Hence, we have utilised the "small dataset modelling" technique for efficient model development for these datasets. For the nitrofurans dataset, we have chosen the best Multiple Linear Regression (MLR) (Aiken et al., 2003) model developed using the MLR plus Validation 1.3 tool available from <https://dtclab.webs.com/software-tools>. However, for the nitrothiophene dataset, the descriptors of the best MLR model were subjected to Partial Least Squares (PLS) regression [34]

using the Partial Least Squares tool (available from <https://dtclab.webs.com/software-tools>). Note that PLS is a robust and generalized version of MLR which converts the original sets of descriptors into new latent variables which are lower in number in comparison to the descriptors appearing in corresponding MLR model (Wold et al., 2001). PLS can handle numerous and noisy variables and do not suffer from the inter-correlation problem.

In case of the nitroimidazole dataset with 84 datapoints, the Genetic Algorithm Multiple Linear Regression (GA-MLR) (Devillers, 1996) method was applied for the feature selection on the whole dataset. A pool of ten descriptors (features) was selected after this process which were further subjected to the Best Subset Selection (BSS) method which finds the best combinations of descriptors out of all the possible combinations of unique descriptors present in the selected models. The best descriptor combination obtained in this process were further subjected to PLS regression using the Partial Least Squares tool (available from <https://dtclab.webs.com/software-tools>) to obtain better quality model. In this work, we have not divided the nitroimidazole dataset though it has sufficient amount of data points because division of the data set was earlier performed by our group in a previous work (De et al., 2020). Thus, we have developed three local models from undivided data sets: nitrofurans model, nitrothiophene model and nitroimidazole model. These data sets were further clubbed to form a global dataset which was then modelled.

During modeling the global dataset, the compounds were split into training and test sets using Kennard-Stone method (Saporo et al., 2012) in Dataset Division GUI 1.2 software tool available from <https://dtclab.webs.com/software-tools>. The dataset was divided into training and test sets in 7:3 ratio. Here, Genetic Algorithm method was used in the Double Cross Validation tool for variable selection. A pool of 16 descriptors was selected and the final model was generated using PLS regression method using the Partial Least Squares tool (available from <https://dtclab.webs.com/software-tools>) using descriptors selected from best subset selection (BSS). Figure 3.4 shows the flowchart of present work methodology showing local and global modeling.

3.7.4. Statistical validation metrics

During the course of the present work, we have performed rigorous analysis using multiple approaches of assessment of the model quality for measurement of the stability, robustness, fitness, and predictivity of the developed models. We have computed various statistical metrics like determination coefficient (R^2), adjusted determination coefficient (R_{adj}^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) for internal validation (Roy, 2007). We have also computed the leave-many-out squared-correlation coefficient ($Q_{LMO(20\%)}^2$) (Roy et al., 2015a). For external validation, in case of the global model, parameters like R_{pred}^2 or Q_{F1}^2 , Q_{F2}^2 and concordance correlation coefficient (CCC) were calculated (Kunal Roy & Mitra, 2011). Furthermore, we have also calculated the r_m^2 metrics (both Δr_m^2 and $\overline{r_m^2}$) (Ojha et al., 2011) and validated the models using root mean squared error (RMSE) and mean absolute error (MAE) based criteria (Roy et al., 2016).

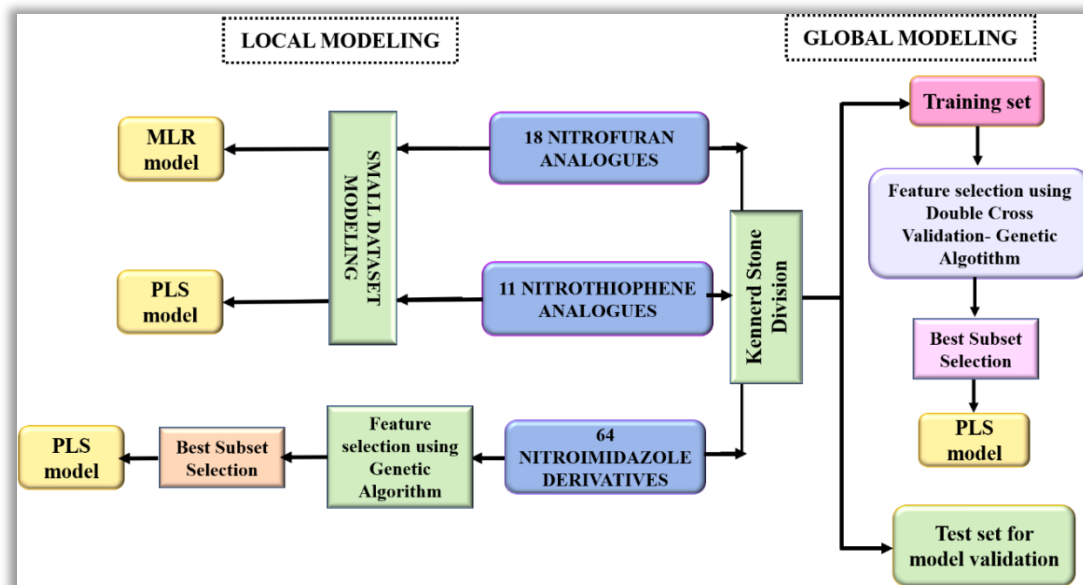


Figure 3.5. The methodology of the present work involving local and global QSAR models.

CHAPTER 4

RESULTS AND DISCUSSIONS

4. Results and Discussions

4.1. Study 1: Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease

In the present study, PET and SPECT imaging agent datasets for both A β plaques and tau fibrils were modeled for their binding affinity using the PLS regression method. For A β dataset, the models for the individual PET and SPECT datasets were developed using PLS regression method after Stepwise Multiple Linear Regression (S-MLR) method. In case of the tau dataset, the final descriptors for the PLS model were obtained from Best Subset Selection (BSS) which was carried out on a pool of descriptors obtained from Double Cross Validation-Genetic Algorithm (DCV-GA) method of model development. The developed models are statistically robust and predictive to be used for data gap filling as suggested by the obtained values of the different validation metrics as given later.

4.1.1. Descriptor Interpretation from QSAR models

4.1.1.1. Modeling of PET imaging agents against A β plaques

PLS model 1 having 4 latent variables (LV) shown in Table 1 gives acceptable values of the determination coefficient R^2 (0.766) and cross-validated determination coefficient ($Q_{LOO}^2=0.600$). The predictivity of the model was analyzed by predictive r^2 (or $r_{pred}^2 = 0.534$) or Q_{F1}^2 which shows acceptable predictivity for the test set compounds. The experimental and predicted pKi values for model 1 are given in the Supplementary Materials. The scatter plot of observed versus predicted pKi values are given in **Fig. 4.1(a)**.

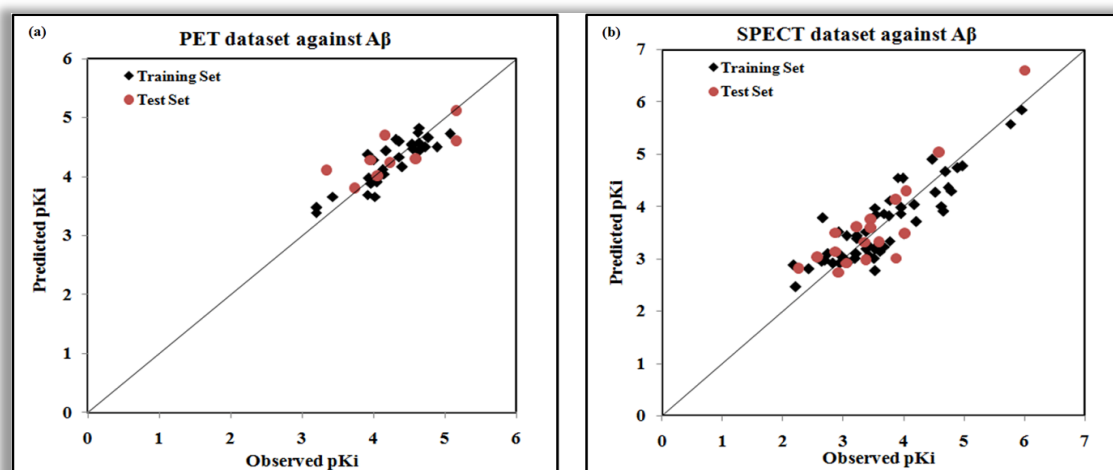


Figure 4.1. Observed vs predicted scatter plot for PET (a) and SPECT (b) imaging agents against beta amyloid

The descriptor **TPSA(Tot)** (a molecular property related descriptor) representing the topological polar surface area using N, O, S, P polar contributions shows a negative correlation to the binding affinity of PET imaging agents. The TPSA descriptor shows the importance of interaction of the O-, N-, S- and P- centered fragments towards beta amyloid plaques (**Figure 4.2**). For example, compounds like

A-P-30 (TPSA(Tot)=122.79), **A-P-29** (TPSA(Tot)=104.33) and **A-P-52** (TPSA(Tot)=90.94) having more number of O-, N-, S- and P- centered fragments have low pKi values (3.20, 3.91 and 3.19 respectively). On the other hand, compounds like **A-P-63** (TPSA(Tot)=36.61), **A-P-31** (TPSA(Tot)=32.26) and **A-P-21** (TPSA(Tot)=30.49) having lower number of aforementioned fragments have high pKi values (4.55, 4.62 and 4.54 respectively). From this observation, we can conclude that hydrophobicity enhances the binding of PET imaging agents to amyloid plaques.

The descriptor **T(O..S)**, a 2D atom pair descriptor, denotes the sum of topological distances between oxygen and sulfur. This descriptor has a positive contribution to the binding affinity of the imaging agents, thus with an increase in the total sum of topological distances between oxygen and sulfur atoms, the binding affinity will increase and vice versa (**Figure 4.2**). In compounds like **A-P-1**, **A-P-51**, **A-P-48** and **A-P-49**, the high values for T(O..S) (T(O..S) = 4) contribute to higher pKi values (5.07, 4.36, 4.31 and 4.72 respectively) whereas in compounds like **A-P-43**, **A-P-30** and **A-P-52** the descriptor value is low (T(O..S) = 0 for all) resulting in low pKi values (3.43, 3.20 and 3.19 respectively).

The descriptor **B10[C-C]**, another 2D atom pair descriptor, denotes the presence or absence of C-C at topological distance 10. The positive regression coefficient of this parameter suggested that presence of such fragment at the topological distance 10 enhances the binding affinity (**Figure 4.2**) as shown in compounds like **A-P-1**, **A-P-51**, **A-P-48** and **A-P-49**. On the other hand, compounds like **A-P-52**, **A-P-43** and **A-P-59** show poor binding affinity due to the absence of such fragments. Here, size (the distance between C and C atoms at 10 reflects the size of the molecules) plays an important role for the binding affinity.

The descriptor **nArX** (functional group count descriptor) represents the number of halogen (X) on the aromatic ring contributing positively towards the binding affinity of the PET imaging agents (**Figure 4.2**). In compounds like **A-P-8**, **A-P-56** and **A-P-57**, the presence of one halogen on the aromatic ring contributes for the high binding affinity (pKi = 4.64, 4.77 and 4.56 respectively) whereas in compounds like **A-P-43**, **A-P-52** and **A-P-59** the absence of halogen group on the aromatic ring reduces the pKi value (3.43, 3.19 and 3.94 respectively).

The descriptor **nHDon** (functional group count descriptor) denotes the number of donor atoms for H-bonds (N and O). The descriptor shows a positive contribution towards binding affinity (pKi) as shown in compounds like **A-P-1** (**Fig.2**), **A-P-8**, **A-P-31** and **A-P-58** all having two hydrogen bond donor sites and hence have higher pKi values (5.07, 4.64, 4.62 and 4.64 respectively). On the other hand, in compounds like **A-P-30** (nHDon = 1) and **A-P-62** (nHDon = 0) (**Figure 4.2**), the pKi values are low (3.20 and 3.92 respectively).

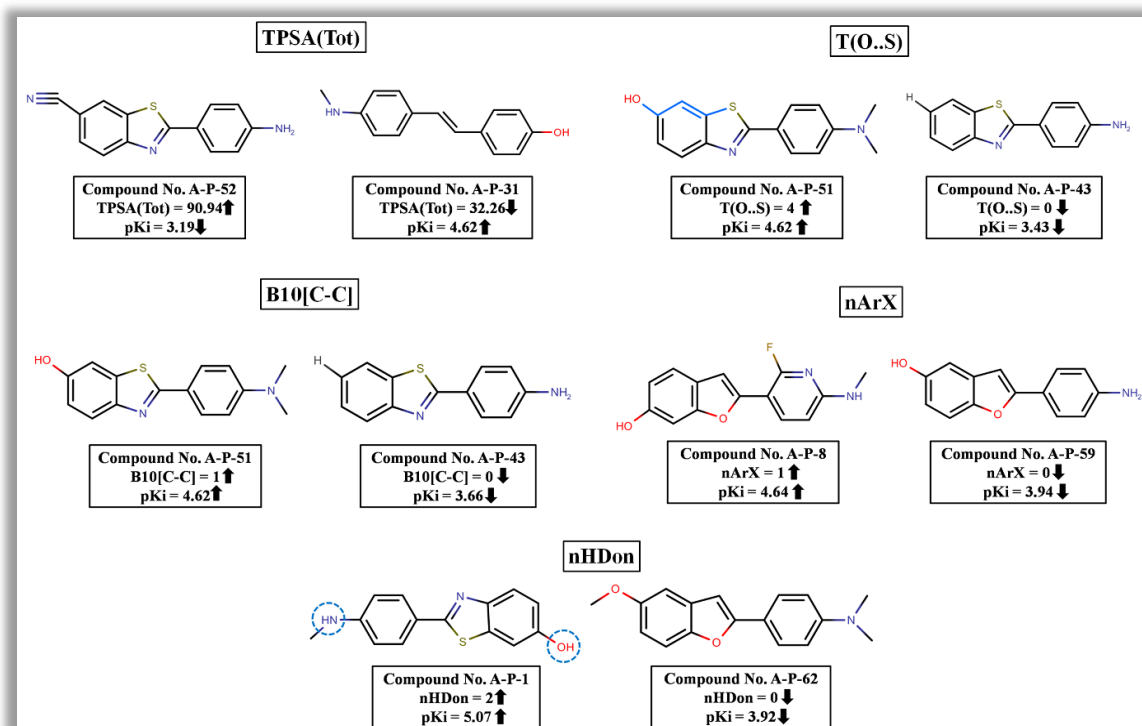


Figure 4.2. Descriptor contribution on binding affinity with respect to model 1 (PET dataset against beta amyloid)

4.1.1.2. Modeling SPECT imaging agents against A β plaques

PLS model 2 with 3 LVs (in **Table 1**) could explain 77.1% of the variance (adjusted determination coefficient). The leave one out (LOO) cross-validated determination coefficient ($Q^2 = 0.758$) above the critical value of greater than 0.5 suggests the statistical reliability of the model. The experimental and predicted pKi values for model 2 are given in the Supplementary Materials. The scatter plot of observed versus predicted pKi values are given in **Fig. 4.1(b)**.

The descriptor **SAacc**, a molecular property type descriptor, denotes the surface area of acceptor atoms from P_VSA-like descriptors. It shows a positive contribution to the binding affinity of SPECT imaging agents as shown in **Fig. 4.3**. The positive regression coefficient indicates that with an increase in the descriptor value, the binding affinity will increase as seen in compounds like **A-S-54**, **A-S-5** and **A-S-35** and vice versa as seen in compounds like **A-S-73**, **A-S-76** and **A-S-85**. Thus, the presence of hydrogen bond donor atoms is beneficial for good binding to beta amyloid plaques.

The descriptor **F05[C-C]** (a 2D atom pair descriptor), depicts the frequency of C-C at the topological distance 5, and it has a negative contribution towards the binding affinity pKi. This indicates that with an increase in the descriptor value (which is an indicator of size and shape), the pKi value will decrease and vice versa as shown in **Fig. 4.3**. In compounds like **A-S-5**, **A-S-65**, **A-S-64** and **A-S-63** the values for the descriptors are high as **A-S-26**, **A-S-25**, **A-S-24** and **A-S-23** respectively thus making the pKi values low (3.23, 3.52, 3.07 and 2.21 respectively) whereas in compounds like **A-S-6**, **A-S-15** and **A-S-16** (having low F05[C-C] values) the pKi values are high (4.74, 4.53 and 4.63).

Table 1: QSAR models for PET and SPECT imaging agents

| Model No. | Target | Dataset | Model | Latent Variables (LVs) |
|-----------|---------------------------|------------------------|--|------------------------|
| 1 | Amyloid beta (A β) | PET | $pKi = 3.987 - 0.012 \times \mathbf{TPSA}(\mathbf{Tot}) + 0.112 \times \mathbf{T}(\mathbf{O..S}) + 0.685 \times \mathbf{B10}[\mathbf{C} - \mathbf{C}] + 0.382$ $\times \mathbf{nArX} + 0.224 \times \mathbf{nHDon}$ $N_{train} = 29, R^2 = 0.766, R_{adj}^2 = 0.727, Q_{loo}^2 = 0.600, MAE(train) = 0.236, RMSE_c = 0.219$ $N_{test} = 9, Q_{F1}^2 = 0.534, Q_{F2}^2 = 0.534, MAE(test) = 0.296, RMSE_p = 0.393$ | 4 |
| 2 | Amyloid beta (A β) | SPECT | $pKi = 3.536 + 0.015 \times \mathbf{SAacc} - 0.042 \times \mathbf{F05}[\mathbf{C} - \mathbf{C}] - 0.284 \times \mathbf{F09}[\mathbf{C} - \mathbf{F}] - 0.911$ $\times \mathbf{nR10} + 0.252 \times \mathbf{F03}[\mathbf{C} - \mathbf{I}]$ $N_{train} = 55, R^2 = 0.771, R_{adj}^2 = 0.758, Q_{loo}^2 = 0.700, MAE(train) = 0.367, RMSE_c = 0.394$ $N_{test} = 18, Q_{F1}^2 = 0.739, Q_{F2}^2 = 0.736, MAE(test) = 0.369, RMSE_p = 0.421$ | 3 |
| 3 | Tau | PET and SPECT combined | $pKi = 0.157 + 0.0185 \times \mathbf{D/Dtr09} + 0.139 \times \mathbf{SaaCH} - 0.176 \times \mathbf{SssCH2} - 0.467$ $\times \mathbf{B08}[\mathbf{N} - \mathbf{F}]$ $N_{train} = 22, R^2 = 0.910, R_{adj}^2 = 0.889, Q_{loo}^2 = 0.839, MAE(train) = 0.229, RMSE_c = 0.198, \overline{r_{m(loo)}^2} = 0.781, \Delta r_{m(loo)}^2 = 0.032$ $N_{test} = 9, Q_{F1}^2 = 0.865, Q_{F2}^2 = 0.850, MAE(test) = 0.275, RMSE_p = 0.325,$ $\overline{r_{m(test)}^2} = 0.768, \Delta r_{m(test)}^2 = 0.114$ | 3 |

The descriptor **F09[C-F]** (a 2D atom pair descriptor), denotes the frequency of C-F at the topological distance 9 and shows a negative correlation with the binding affinity. This descriptor indicates both presence of a fluorine atom and size of the compound. A higher occurrence of C-F at topological distance 9 will decrease the binding affinity as observed in compounds **A-S-41** ($pK_i = 3.42$), **A-S-47** ($pK_i = 3.38$) and **A-S-63** ($pK_i = 2.21$) whereas in compounds like **A-S-52**, **A-S-54**, **A-S-34** and **A-S-33** with the absence of such groups the binding affinity is high (5.77, 5.96, 4.98 and 4.89 respectively) (shown in **Fig.4.3**).

The descriptor **nR10**, a ring descriptor, indicates the number of 10 membered rings present in the compounds (here **4H-1-benzopyran** ring), and the descriptor provides a negative contribution to the binding affinity. Compounds like **A-S-17**, **A-S-18** and **A-S-20** each containing one 10 membered ring has a low binding affinity value ($pK_i = 3.68$, 3.58 and 2.66 respectively) whereas compounds like **A-S-52**, **A-S-54** and **A-S-34**, the absence of any 10 membered ring contributes to higher values of the binding affinity (shown in **Fig.4.3**).

The descriptor **F03[C-I]**, (a 2D atom pair descriptor), represents the frequency of C-I at the topological distance 3 has a positive contribution towards the binding affinity. Thus, with an increase in the value for this descriptor, the pK_i value will increase as seen in compounds **A-S-52**, **A-S-33** and **A-S-34** (5.77, 4.89 and 4.98 respectively) whereas with a decrease in the value of **F03[C-I]**, the pK_i value will also decrease as seen in **A-S-76**, **A-S-78** and **A-S-85** (2.55, 2.85 and 2.70 respectively) (**Fig.4.3**).

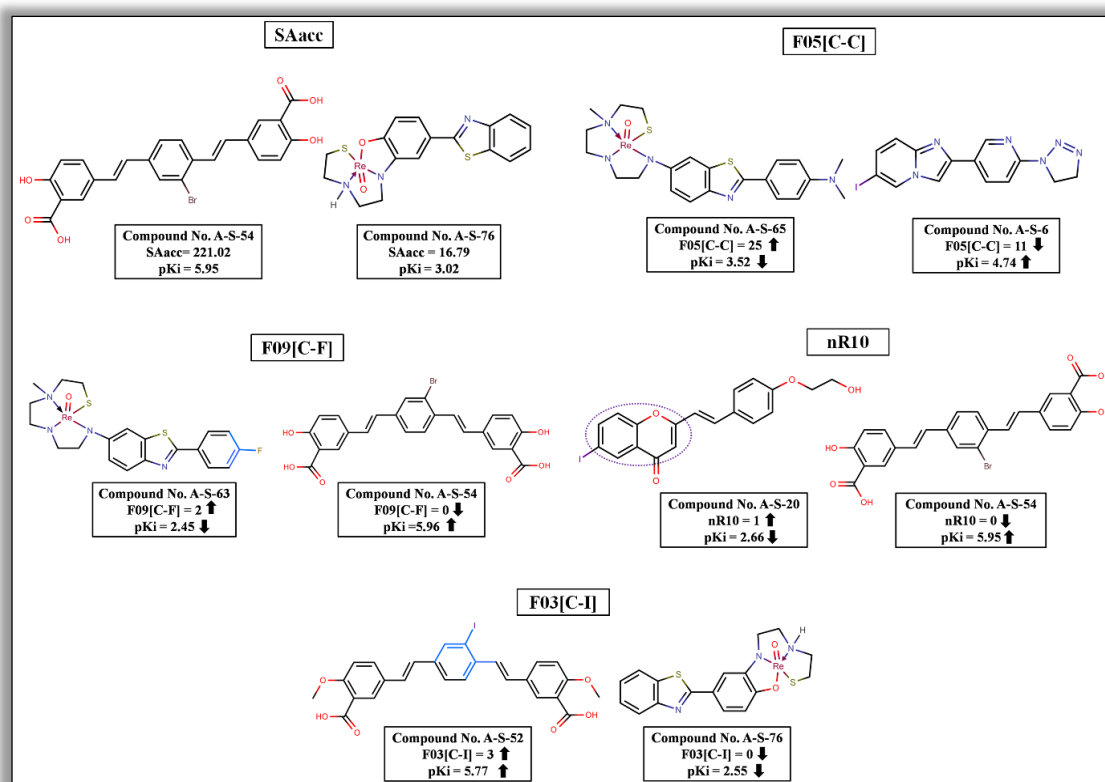


Figure 4.3. Descriptor contribution on binding affinity with respect to model 2 (SPECT dataset against beta amyloid)

4.1.1.3. Modeling PET and SPECT imaging agents against tau protein

PLS model 3 with 3 latent variables (LVs) evolved as the best model, and it could show good statistical robustness and predictivity. Acceptable values for determination coefficient R^2 (0.910) and cross-validated determination coefficient ($Q_{LOO}^2=0.899$) were obtained. The predictivity of the model was analyzed by $predictive r^2$ (or $r_{pred}^2 = 0.865$) or Q_{F1}^2 which shows good predictivity for the test set compounds. The scatter plot of observed versus predicted pKi values are given in **Fig. 4.4**.

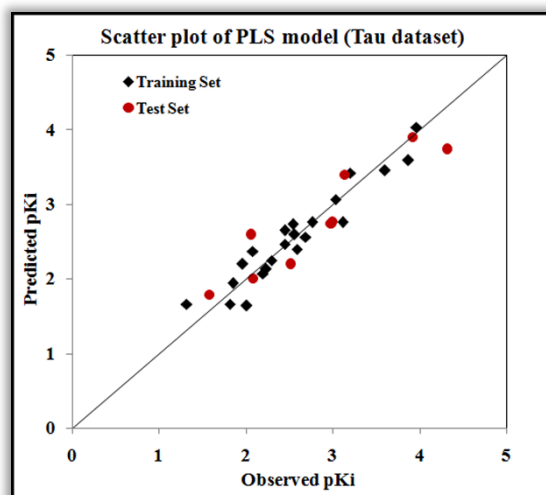


Figure 4.4. Observed vs predicted scatter plot for PET and SPECT imaging agents against tau protein

The descriptor with the highest contribution, **D/Dtr09** (i.e., distance/detour ring of order 9) is a ring descriptor which is based on operations made on distance or detour matrix D/Δ . The detour matrix is square symmetric matrix that contains the ratios of the lengths of the shortest to the longest path between any pair of vertices. The term D/Δ is calculated by:

$$D/\Delta = \sum_{i=1}^A \sum_{j=1}^A (D/\Delta)_{ij}$$

Here, Δ is the detour distance (Benfenati, 2011; Chartrand et al., 1993). This descriptor shows a positive contribution which indicates its positive influence on the binding affinity of the imaging agents as observed in compounds like **T-P-3** ($D/Dtr09=112.603$), **T-P-1** ($D/Dtr09=96.788$) and **T-P-30** ($D/Dtr09=87.895$) having higher binding affinities (3.96, 3.92 and 4.32 respectively). On the other hand, compounds like **T-S-25** ($pKi=1.31$), **T-P-9** ($pKi=1.58$) and **T-S-26** ($pKi=1.81$) have low pKi values corresponding to low values for the descriptor ($D/Dtr09=0$ for all three compounds). **Figure 5** shows how the descriptor $D/Dtr09$ contributes towards the binding affinity of the imaging agents.

The next important descriptor **SaaCH**, an E-state descriptor, denotes the sum of E-state of atom type aaCH where aaCH represents -CH groups in benzene nucleus. The descriptor shows a positive contribution to the binding affinity suggesting that the presence of such groups would increase the binding affinity as seen in compounds like **T-P-5** ($SaaCH=17.67$, $pKi=3.11$) and **T-P-1** ($SaaCH=16.71$, $pKi=3.92$), while in compounds like **T-S-25** ($SaaCH=11.67$, $pKi=1.31$) and **T-S-26** ($SaaCH=11.59$, $pKi=1.81$), the occurrence of such fragment is less resulting in less binding affinity (**Figure 5**).

The third descriptor **SssCH2** is also an E-state descriptor, signifying the sum of E-state of atom type ssCH2 (-CH₂-), which has a negative regression coefficient. This indicates that with an increasing descriptor value, the binding affinity will decrease as seen in compounds **T-P-19** (SssCH₂=0.78, pKi=2.01) and **T-S-25** (SssCH₂=0.65, pKi=1.31) (in **Fig. 5**). The opposite occurs when the descriptor value is less, i.e., the binding affinity becomes higher as observed in compounds like **T-P-2** (SssCH₂=-0.48, pKi=3.19) and **T-P-5** (SssCH₂=-0.88, pKi=3.11) (in **Figure 5**).

The least important descriptor is **B08[N-F]**, a 2D atom pair descriptor, which denotes the presence or absence of N-F at the topological distance 8. The negative regression coefficient of this parameter suggests that presence of such fragment at the topological distance 8 is detrimental to the binding affinity as shown in compounds like **T-P-8** (pKi=2.23) and **T-P-18** (pKi=1.85). On the other hand, compounds like **T-S-28**, **T-S-29** and **T-S-30** show good binding affinity due to the absence of such fragments. **Figure 5** shows the contribution of B08[N-F] descriptor.

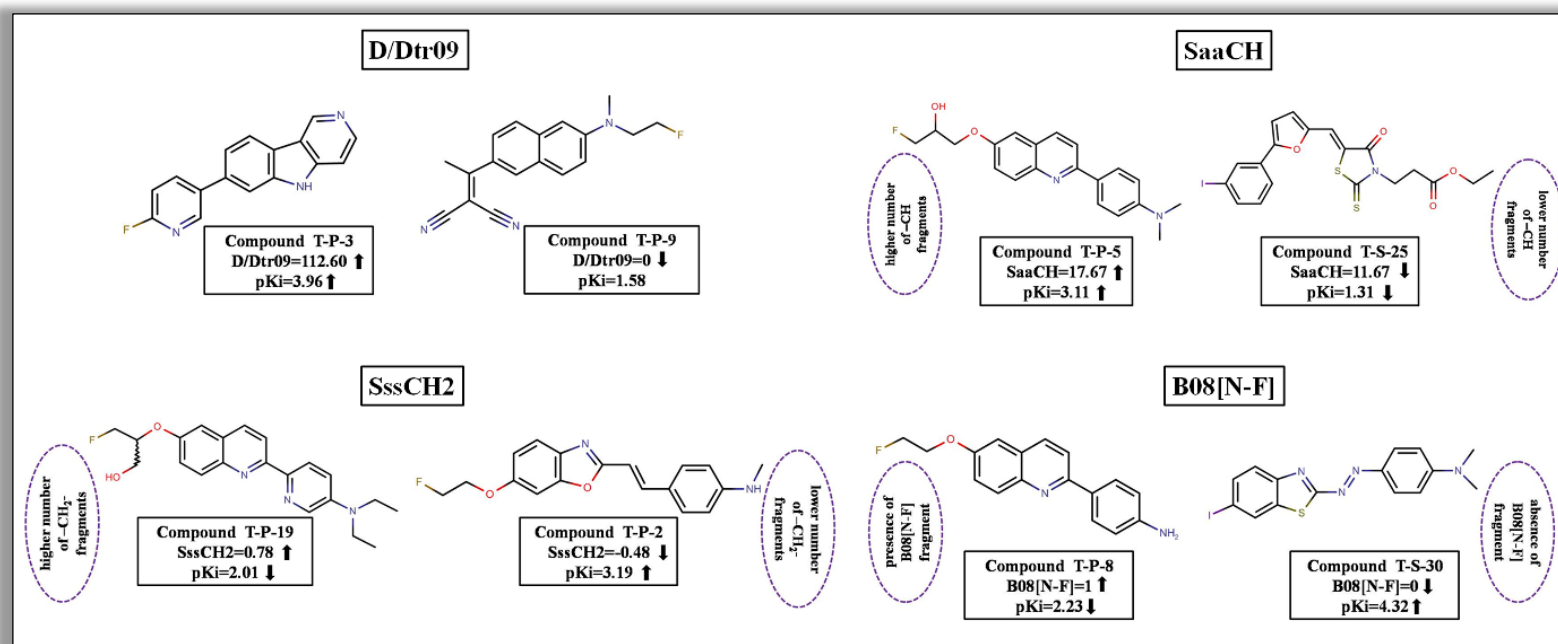


Figure 4.5. Descriptor contribution on binding affinity with respect to model 3 (PET and SPECT in tau protein)

4.1.2. Interpretation of PLS plots

4.1.2.1. Variable importance plot

The variable importance plot (VIP) (Akarachantachote et al., 2014) signifies the order of contribution of each descriptor. The most and least important descriptors can be identified using this plot. A variable with VIP score >1 indicates the descriptor's higher statistical significance as compared to the one with a lower VIP value. The descriptors from higher to lower contribution for all the three models are given in Fig. 4.6.

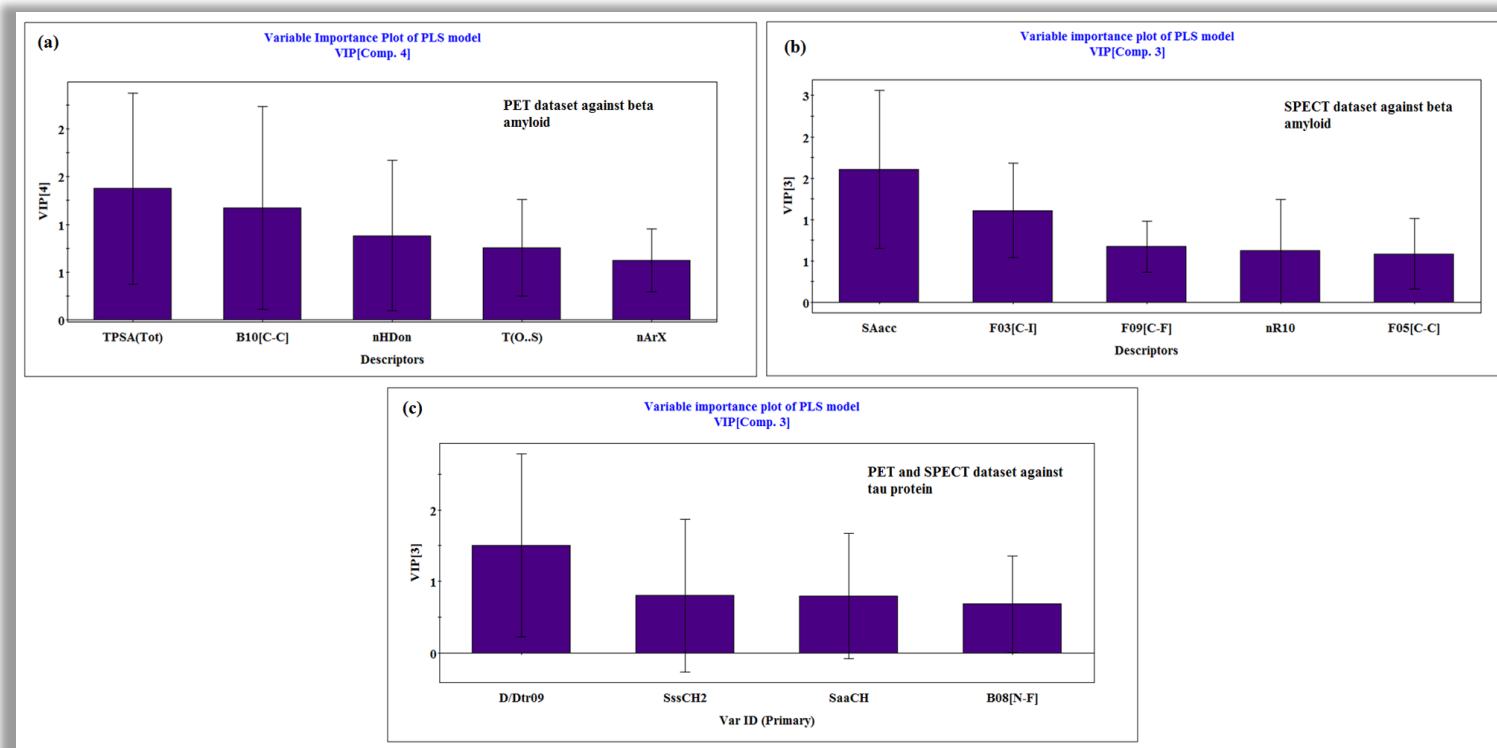


Figure 4.6. Variable importance plot (VIP) of the three PLS models

4.1.2.2. Regression coefficient plot

The regression coefficient plot (Wold et al., 2001) (Fig. 4.7) gives information about the positive or negative contribution of descriptors towards the activity of the compounds. In case of model 1 for the PET dataset against A β fibrils, the descriptors like T(O..S), B10[C-C], nArX and nHDon having a positive regression coefficient signify that with an increase in the descriptor value the binding affinity increases, whereas descriptor having negative coefficients like TPSA(Tot) decrease the binding affinity with their increasing numerical values. In case of model 2 for the SPECT dataset against A β fibrils, SAacc and F03[C-I] descriptors have positive contributions (positive regression coefficients) whereas other three descriptors (F05[C-C], F09[C-F and nR10) have negative regression coefficients. For model 3, i.e., in case of tau protein dataset, it was found that descriptors D/Dtr09 and SaaCH have a positive regression coefficient and other two descriptors like SssCH2 and B08[N-F] have negative coefficients thereby decreasing pK_i values significantly.

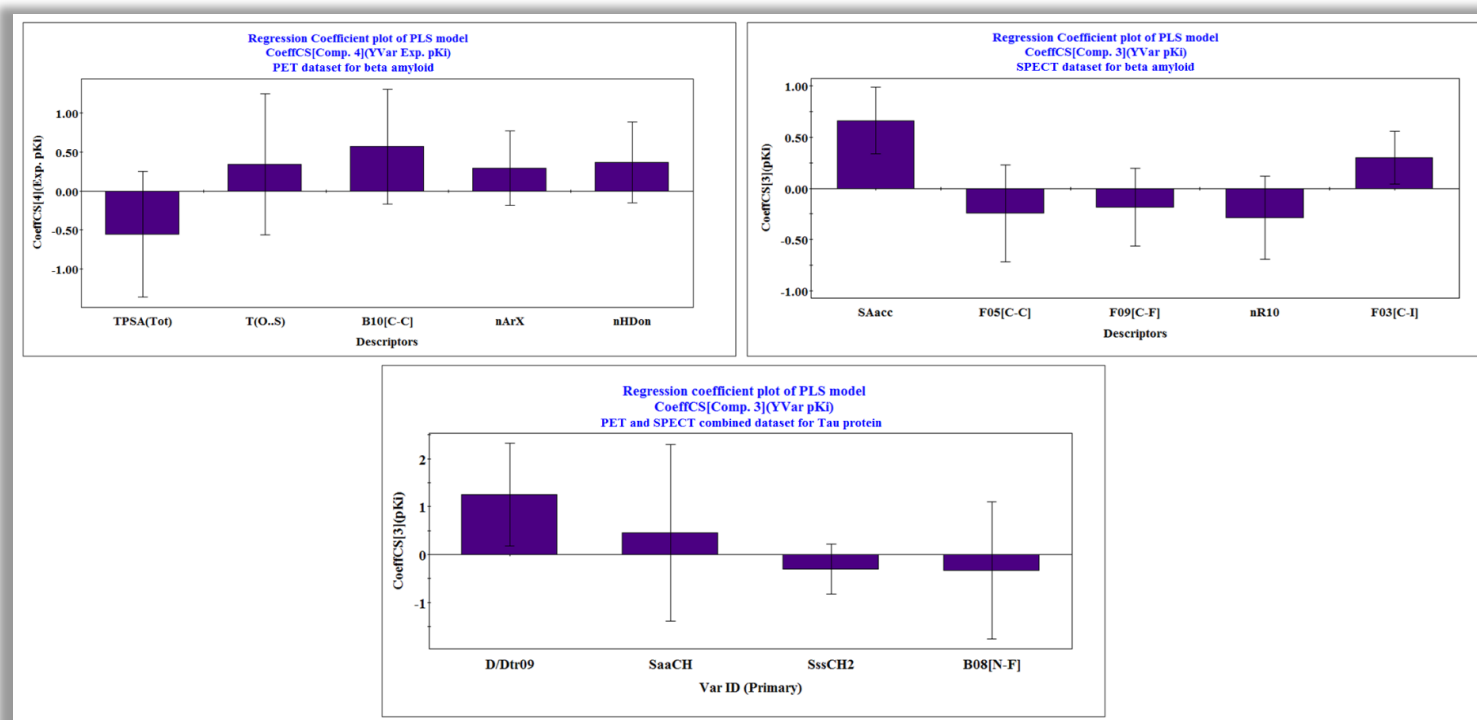


Figure 4.7. Regression coefficient of the three PLS models

4.1.2.3. Score plot

The distribution of the compounds in the latent variable space as defined by the scores is expressed in a score plot (**Figure 4.8**) (Jackson, 2005). From the plot, one can conclude that compounds that are situated near each other have similar characteristics or properties, whereas compounds which are far from each other have dissimilar properties with respect to their binding affinity. Compounds which are outside the ellipse in the plot are outliers. The Score plots for the derived models are shown in **Figure 4.8**.

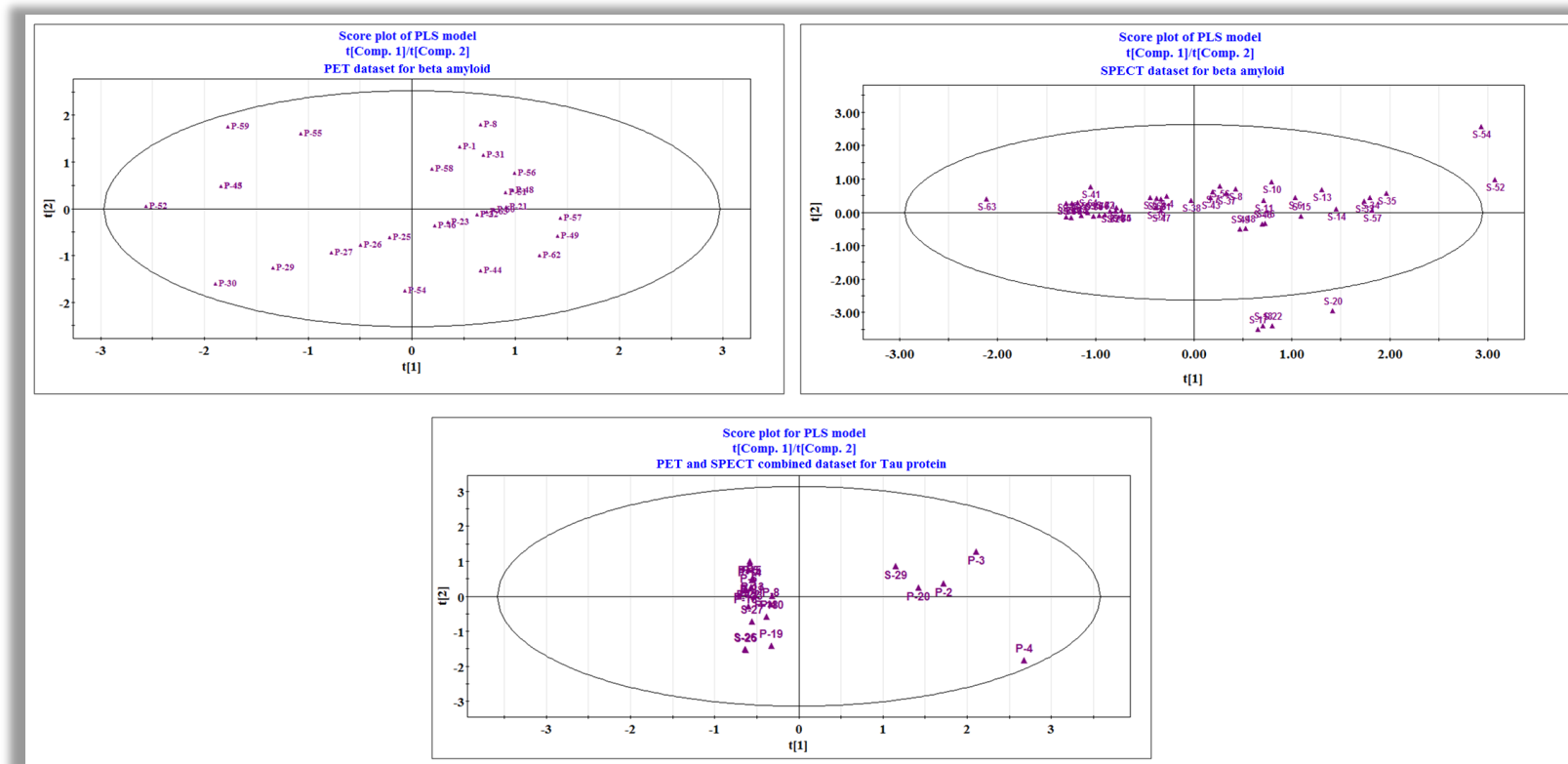


Figure 4.8. Score plot of the PLS models

4.1.2.4. Loading plot

The relationship between the X-variables and Y-variables can be understood by the loading plot (Fig. 4.9) (Wold et al., 2001). The loading plot was developed using the first two PLS components in all the three cases. The influence of different variables on the model can be understood from the loading plot. Descriptors that are grouped together have similar meanings and similar effects on the response. Descriptors with different meaning are situated at a considerable distance from each other. Any descriptor situated far from the plot origin is considered to have a greater impact on the response.

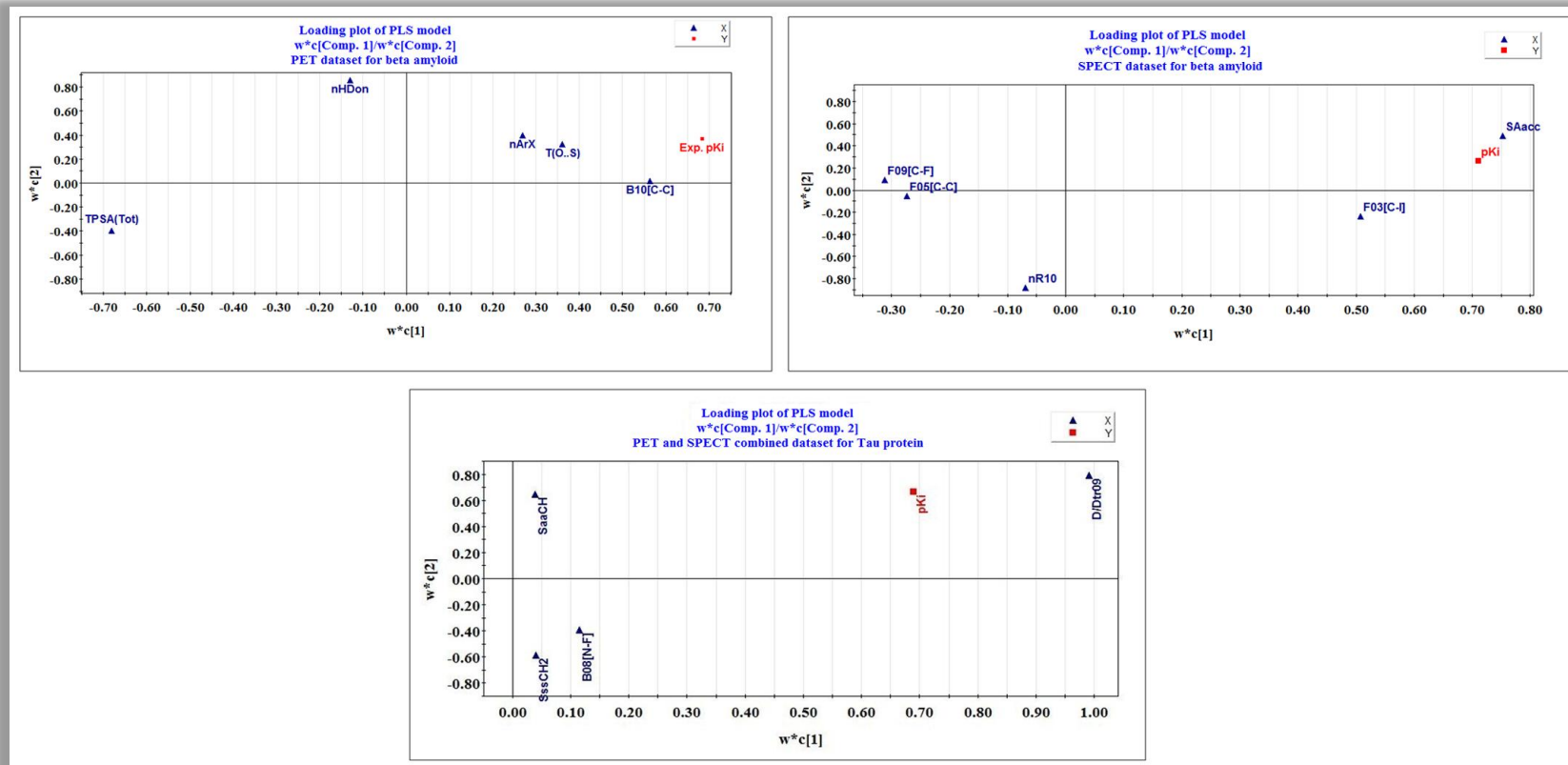


Figure 4.9. Loading plot of the three PLS models

4.1.2.5. Applicability Domain

The applicability domain (AD) provides a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable (Gadaleta et al., 2016). The AD assessment of the proposed model for the imaging agents were performed according to the DModX (distance to model in the X-space) approach using SIMCA-P software. **Figures 4.10, 4.11 and 4.12** show the AD plots of the three models.

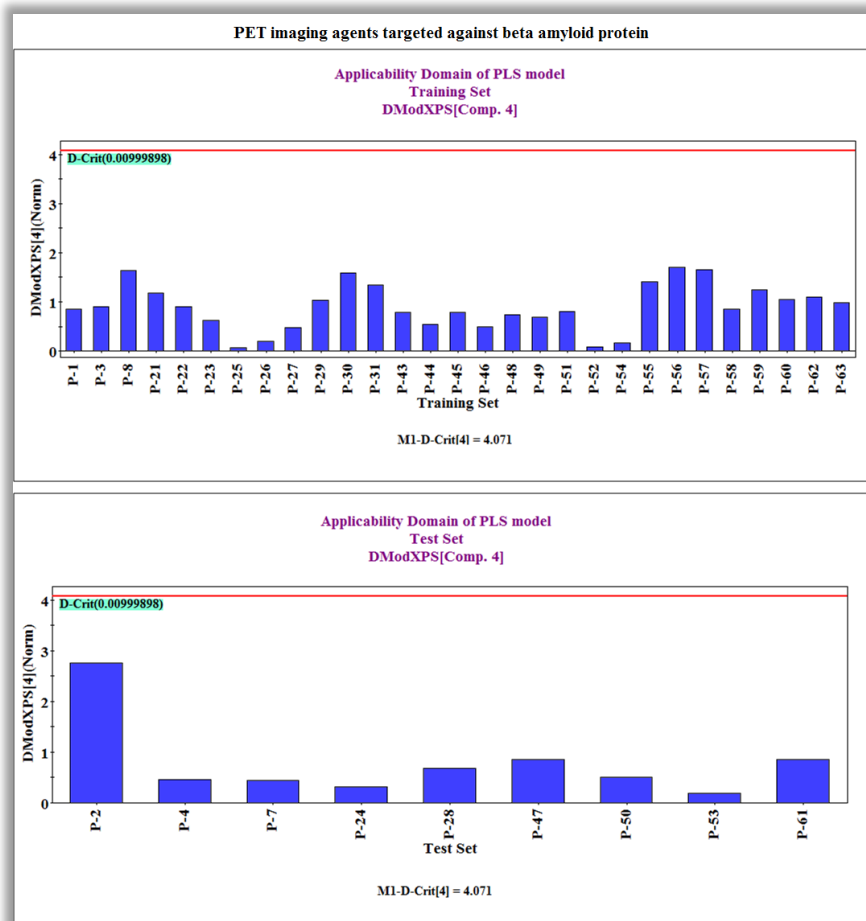


Figure 4.10. DModX Applicability Domain plot of Model 1

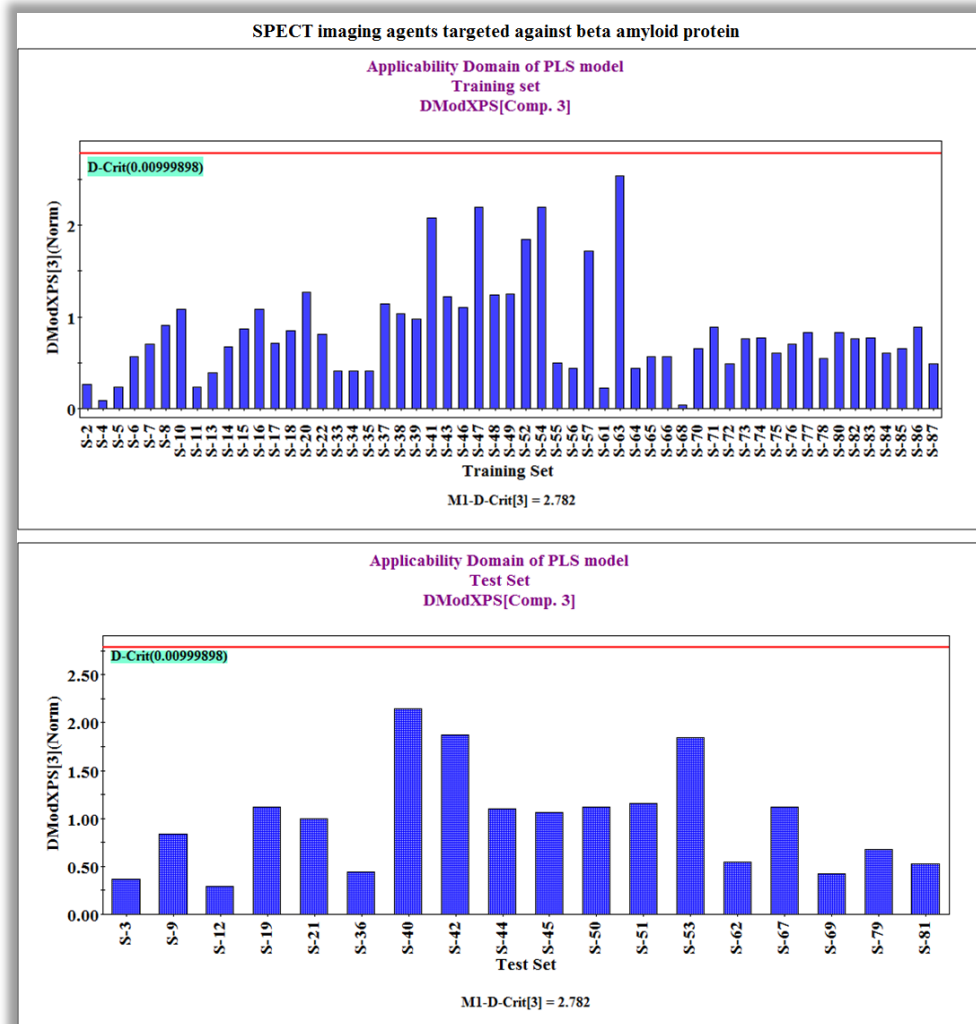


Figure 4.11. DModX Applicability Domain plot of Model 2

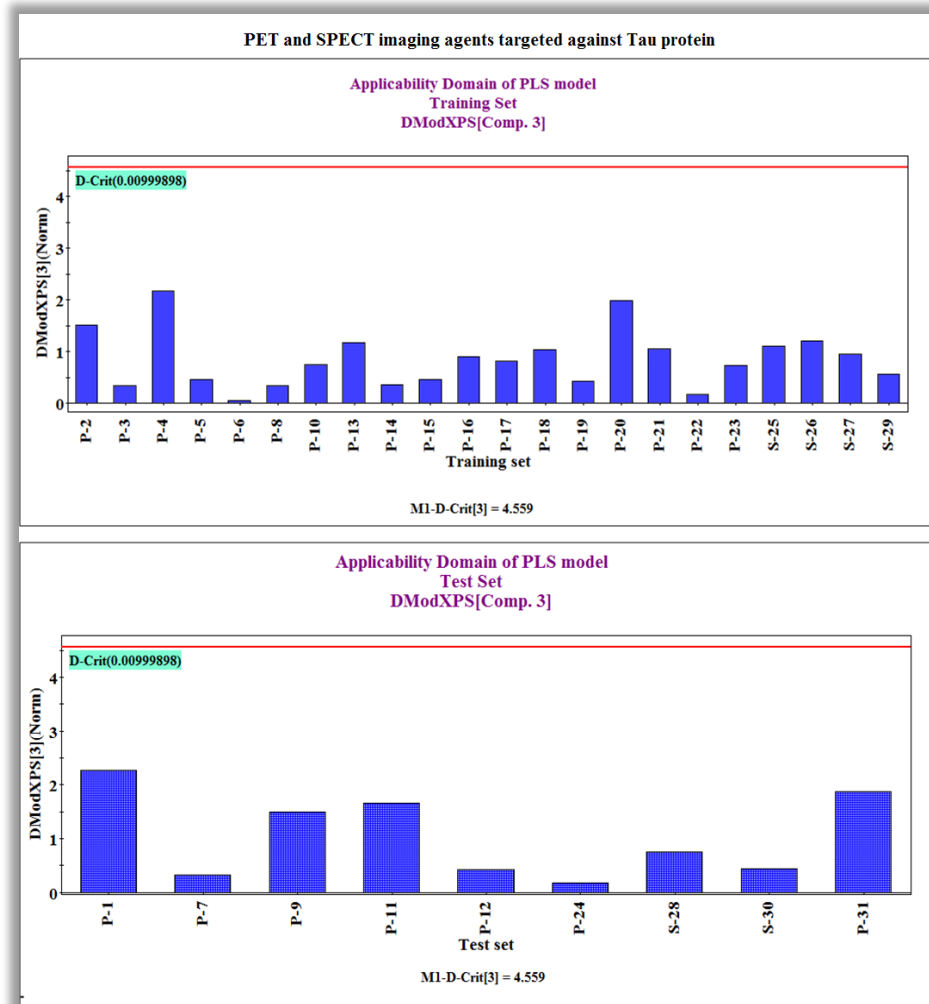


Figure 4.12. DModX Applicability Domain plot of Model 3

4.1.2.6. Y-randomization

The statistical significance of the model is analyzed by a randomization plot (Figure 4.13). The randomization plot authenticates that the model is not the result of any chance correlation (Rücker et al., 2007). The randomization process provides a number of models by shuffling different combinations of X or Y variables (here Y-variable only) based on the fit of the reordered model. Here we have used 100 permutations for random model generation, though the number of permutations can be changed. To avoid chance correlation, the basic statistics of the randomization models (Q^2 and R^2) should be poor and not within the range of those for acceptable regression models (R_Y^2 intercept should not exceed 0.3 and Q_Y^2 intercept should not exceed 0.05) (Rücker et al., 2007). The randomization plots given in Fig. 4.13 show that the developed models are non-random and robust, and are suitable for prediction of the binding affinity of the imaging agents within the AD of the model.

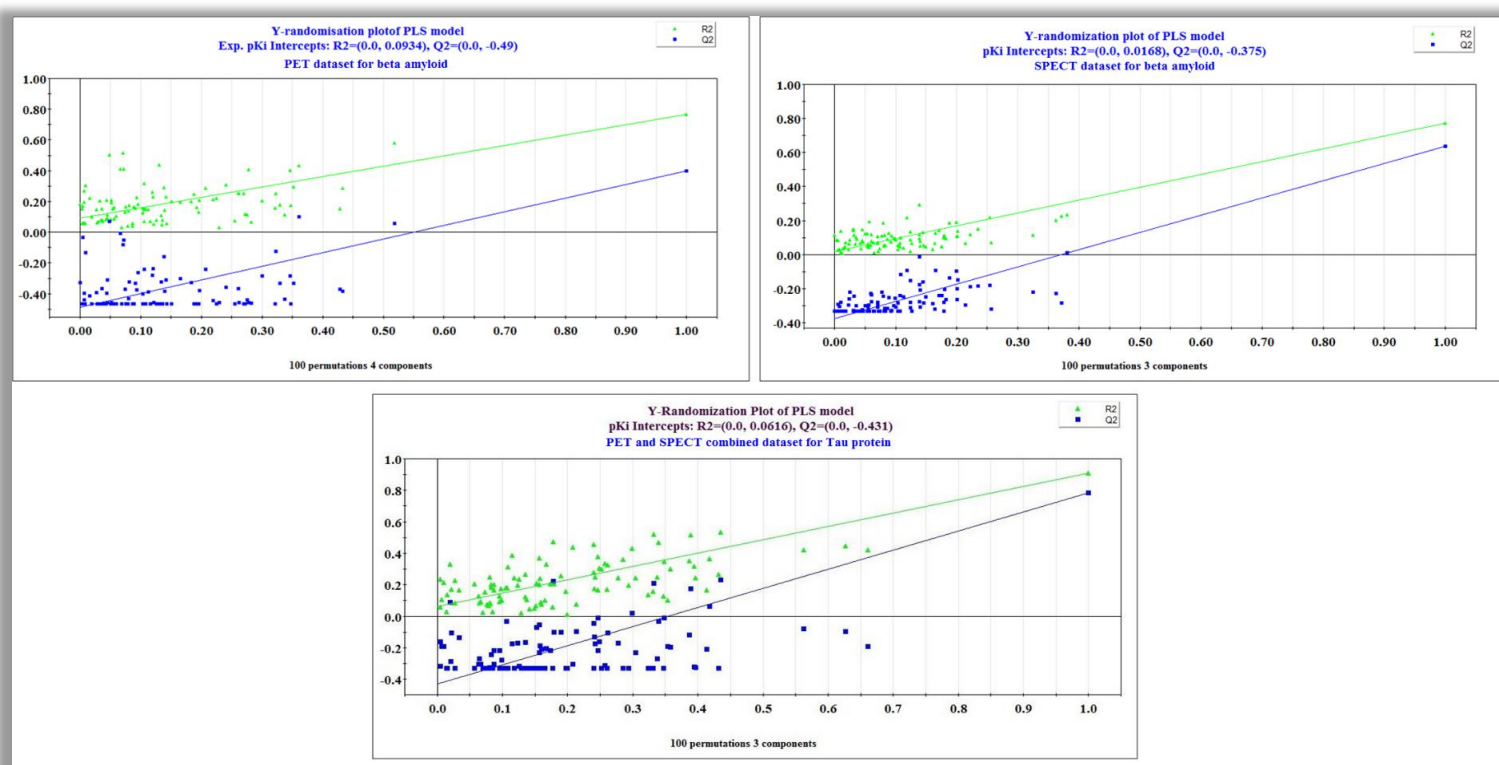


Figure 4.13. Y-randomization plots of 3 models

4.1.3. Molecular docking

Molecular docking studies yield critical information related to the orientation of the imaging agents at the binding zone of the enzyme and the information about the intermolecular interaction between protein and ligands at molecular level. The aim in the present study was to understand the interactions occurring between the two proteins and different PET and SPECT imaging agents and correlate the observations found with the QSAR results. It was found that hydrogen bonding and π bonding interactions were predominant. The ligand-receptor interaction analysis suggests that the imaging agents interact with both polar and non-polar amino acids.

4.1.3.1. Molecular docking for selected PET imaging agents against A β plaques

In cases of compounds **A-P-2** and **A-P-56** which have higher binding affinity (pKi = 5.15 and 4.77 respectively), the interaction forces include hydrogen bonds (carbon-hydrogen bonds (Alkorta et al., 1998), conventional hydrogen bonds and π -donor hydrogen bonds), π interactions (π -sulfur bond, π - π T shaped bond and π -alkyl bonds) and unfavorable acceptor-acceptor bond. The number of interacting residues is higher in case of these compounds thus supporting their high values of binding affinity. The amino acid residues interacting with compound **A-P-2** are **Val D:39**, **His B:13** and **Val D:36**. **Fig. 4.14** shows the interactions obtained for the most stable pose where it is found that Val D:39 and His B:13 show π -alkyl (Echeverría, 2017; Ribas et al., 2002) and π - π T shaped (Martinez & Iverson, 2012) interactions respectively with the ligand due to the presence of unsaturation in the aromatic nuclei. Also sulphur in the thiazole nucleus interacts with the aromatic nucleus (thiazole moiety) of histidine making π -sulphur interaction. On the other hand, Val D:36 makes carbon-hydrogen bond with the ligand. In compound **A-P-56**, the interacting amino acids include **Val A:12**, **His B:13**, **Val D:36** and **Val D:39**. In **Fig. 4.14**, the different interactions are shown. Hydrogen bonds like carbon-hydrogen bonds and π -donor hydrogen bonds are found with Val D:36 and His B:13 respectively. The alkyl part of Val A:12 interacts with Bromine and other π bonds are formed with Val D:39 and His B:13 residues.

PET compounds like **A-P-52** and **A-P-50** having low binding affinity (pKi= 3.19 and 3.34) show similar kind of interactions (hydrogen and π bonds) as in case of higher affinity compounds, but the number of interacting amino acid residues are much less as shown in **Fig. 4.14**. Val D:39 was found to interact with both the ligands forming π -alkyl interactions. In compound **A-P-52**, the nitrogen of cyano group forms hydrogen bond with Val E:36 whereas in compound **A-P-50**, His B:13 shows π interactions (π -donor Hydrogen and π - π interactions) with the ligand. The docking sites for both high and low binding affinity PET imaging agents targeted against A β are given in **Table 4.2**.

Relation with QSAR models

In the docking study, it is observed that formation of hydrogen bonds between the ligands and receptor plays a vital role in binding. This observation supports the occurrence of **TPSA(Tot)** (total polar surface area using N, S, O and P polar contributions) and **nHDon** (number of donor atoms) descriptors in the QSAR models. Furthermore, formation of π -sulfur bonds can be correlated with the **T(O..S)** descriptor, where the presence of sulfur atoms in the molecules is essential. Val A12 residue forms hydrophobic interaction with compound **A-P-56** due to the presence of bromine (halogen) which corroborates with the **nArX** descriptor proving that the presence of halogen groups is beneficial for binding.

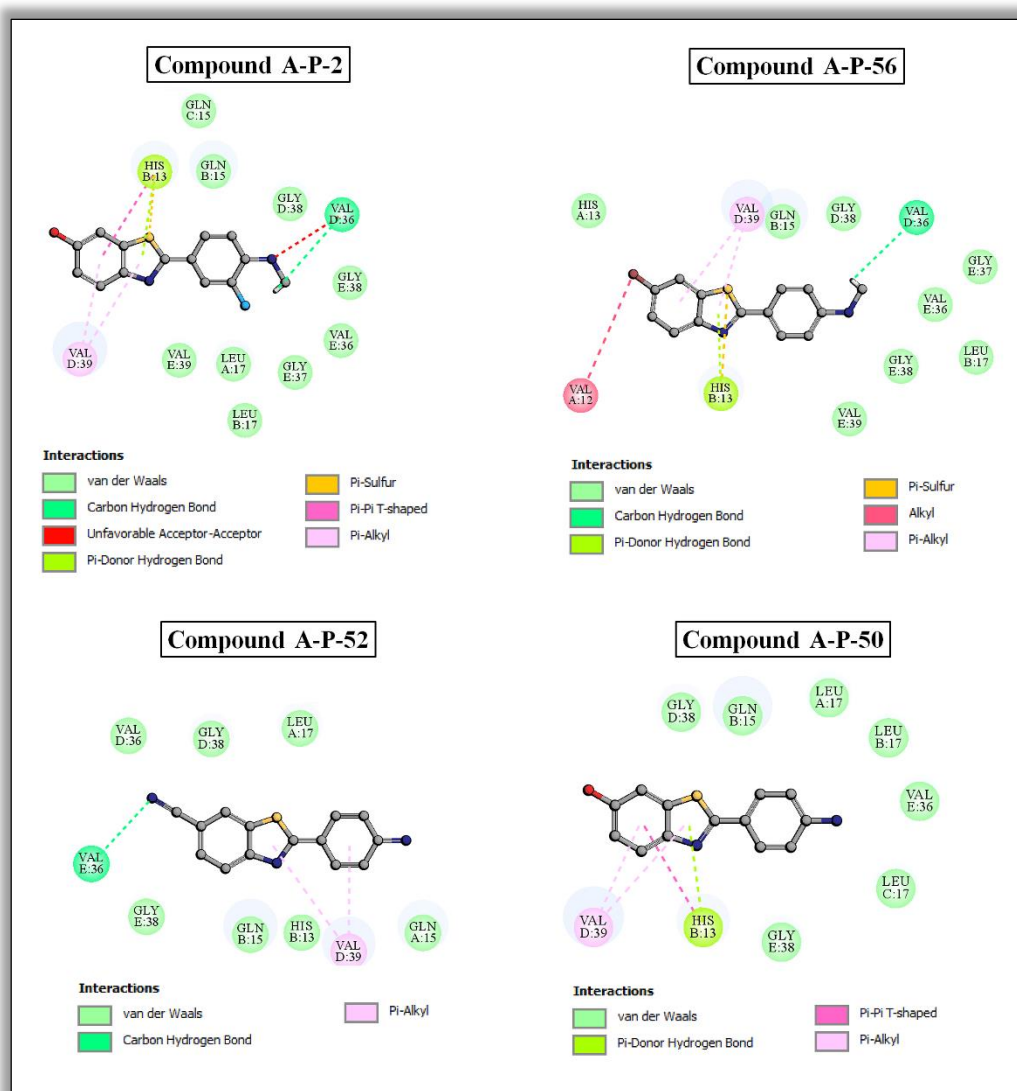


Figure 4.14. Molecular interactions between high and low active PET imaging agents with A β protein

4.1.3.2. Molecular docking for selected SPECT imaging agents against A β plaques

In compounds like **A-S-53** and **A-S-52** having higher binding affinity ($pK_i = 6.0$ and 5.77 respectively), interaction forces include hydrogen bonding (carbon-hydrogen bonding, conventional hydrogen bonding and π -donor hydrogen bonds), π interactions (like π -alkyl, π -sigma interactions, π -lone pair interactions and amide- π interactions) and alkyl interactions. The amino acid residues interacting with compound **A-S-53** are **Gly J:29**, **His B:13**, **Gly D:38**, **Leu A:17**, **Gly D:37**, **Gly C:37**, **Val D:39**, **Val E:39**, **Ile K:31** and **Val D:40**. In **Fig.4.15**, we can see the interactions for the most stable pose, where Val D:40, Ile K:31, Val D:39 and Leu A:17 makes π -alkyl (Echeverría, 2017; Ribas et al., 2002) interactions with the ligand due to the presence of unsaturation in the ligand moiety. Gly D:38 makes π -sigma interaction with the ligand. Hydrogen bond interactions are observed with Gly J:29, Val D:39, Gly D:37 and Gly C:37. In compound **A-S-52**, hydrogen bond interaction such as carbon hydrogen bonds is observed with Gly K:29, Gly J:29 and Ala L:30 whereas Val D:39 makes π -donor hydrogen bond interaction. π -alkyl interaction is observed with His B:13,

Val D:40 and Ile K:31. The interacting amino acid residues are **Ile K:31, Val D:39, Val D:40, Val E:39, His B:13, Ala L:30, Gly J:29** and **Gly K:29**.

In compounds like **A-S-55** and **A-S-20** having low binding affinity ($pK_i = 2.18$ and 2.66 respectively), similar kinds of interactions are observed like hydrogen bond and π interactions but the number of interacting residues are much less (**Fig. 4.15**). The docking sites for both high and low binding affinity SPECT imaging agents targeted against $A\beta$ are given in **Table 4.2**.

Relation with QSAR models

From the docking study it is observed that hydrogen bonding formation between the protein receptor and ligand molecule plays an important role in binding affinity of the later. This observation corroborates with the **SAacc** (denotes the surface area of acceptor atoms) descriptor occurred in the QSAR model.

4.1.3.3. Molecular docking for selected PET and SPECT imaging agents against tau protein

The tau protein (PDB ID:6FAU) was docked with higher and lower active imaging agents in order to study their binding pattern and the molecular interactions occurring between them. In compounds like **T-P-2** and **T-S-29** with high binding affinities ($pK_i = 4.319$ and 3.959 respectively), higher number of hydrogen bonding interactions and π -interactions have been observed. Compound **T-P-2** makes interaction with **Trp A:230, Asn A:226, Leu A:174, Val A:178** and **Leu A:229** amino acid residues (**Fig. 4.16**). The stable pose makes π -alkyl interaction with Val A:178 and Leu A:174. The fluorine atom makes alkyl interaction with Leu A:229 and Val A:178 and halogen interaction with Asn A:226. The amino acid residues interacting with compound **T-S-29** are **Leu A:229, Val A:178, Leu A:174** and **Leu A:222**. From **Fig. 4.16**, it is seen that π -interaction is the predominant binding mode with the protein (as observed with Val A:178, Leu A:174 and Leu A:222). Other interactions noticed are alkyl interaction and various hydrogen bonding interactions. Low affinity compounds include **T-P-7** and **T-P-10** ($pK_i = 1.311$ and 1.957) which showed less number of interactions in comparison to higher affinity compounds (in **Fig.10**). Two π -alkyl interactions is observed for both the compounds, with Leu A:174 and Val A:178 in both the cases. The docking sites for both high and low binding affinity PET and SPECT imaging agents targeted against tau protein are given in **Table 4.2**.

Relation with QSAR models

In the docking study it is observed that π - interactions play a vital role in ligand-receptor binding. This observation supports the occurrence of **D/Dtr09** (distance/detour ring of order 9) descriptor which is a ring descriptor. Increased number of aromatic nuclei will increase the value of this descriptor thereby increasing the binding affinity; also paving way for more π -interactions. Also, π -interactions corroborate with **SaaCH** descriptor, where aaCH represents $-CH$ groups in benzene nucleus. From the observations it is concluded that aromaticity is a major feature regulating the binding affinity of PET and SPECT imaging agents.

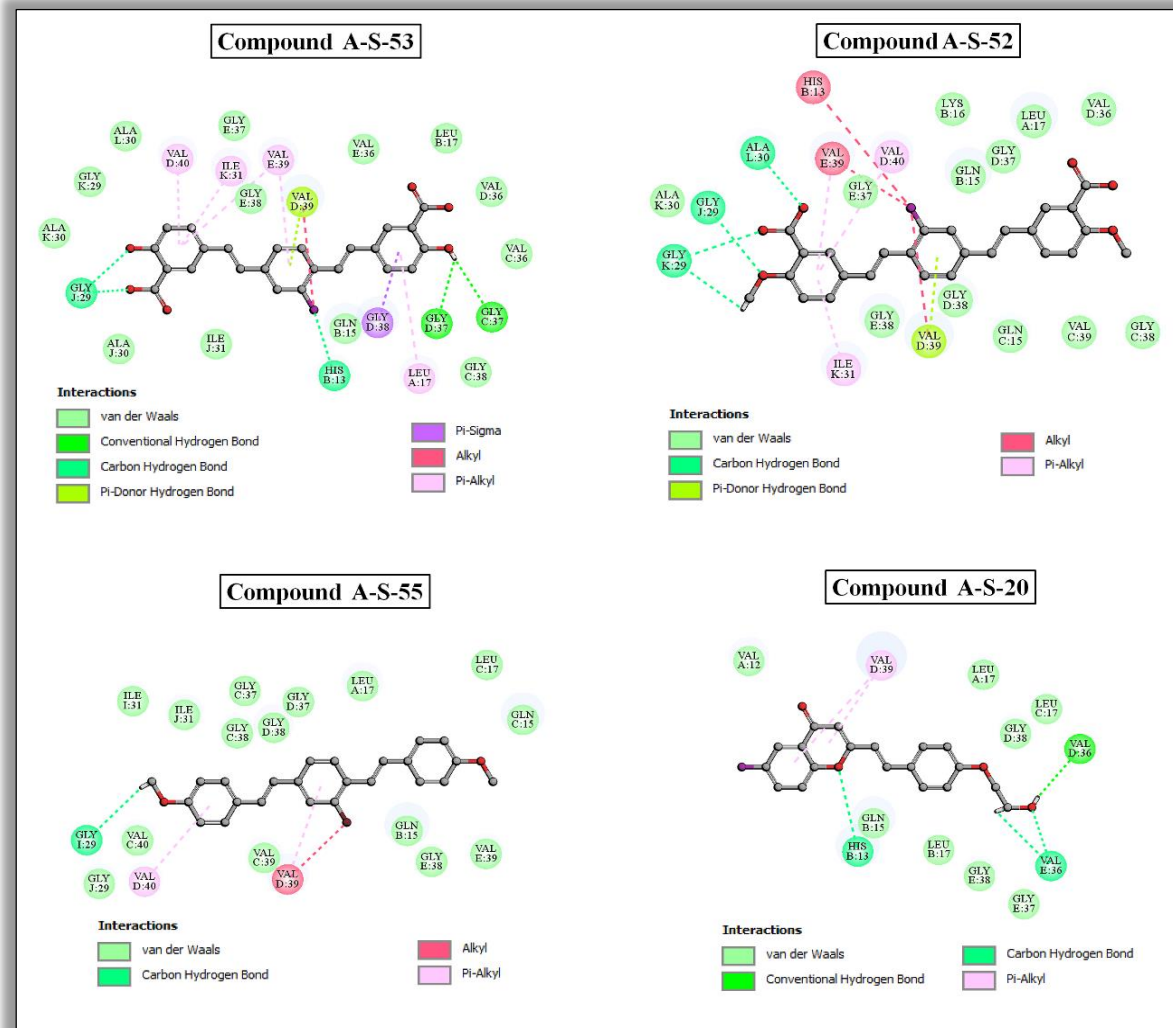


Figure 4.15. Molecular interactions between high and low active SPECT imaging agents with A β protein

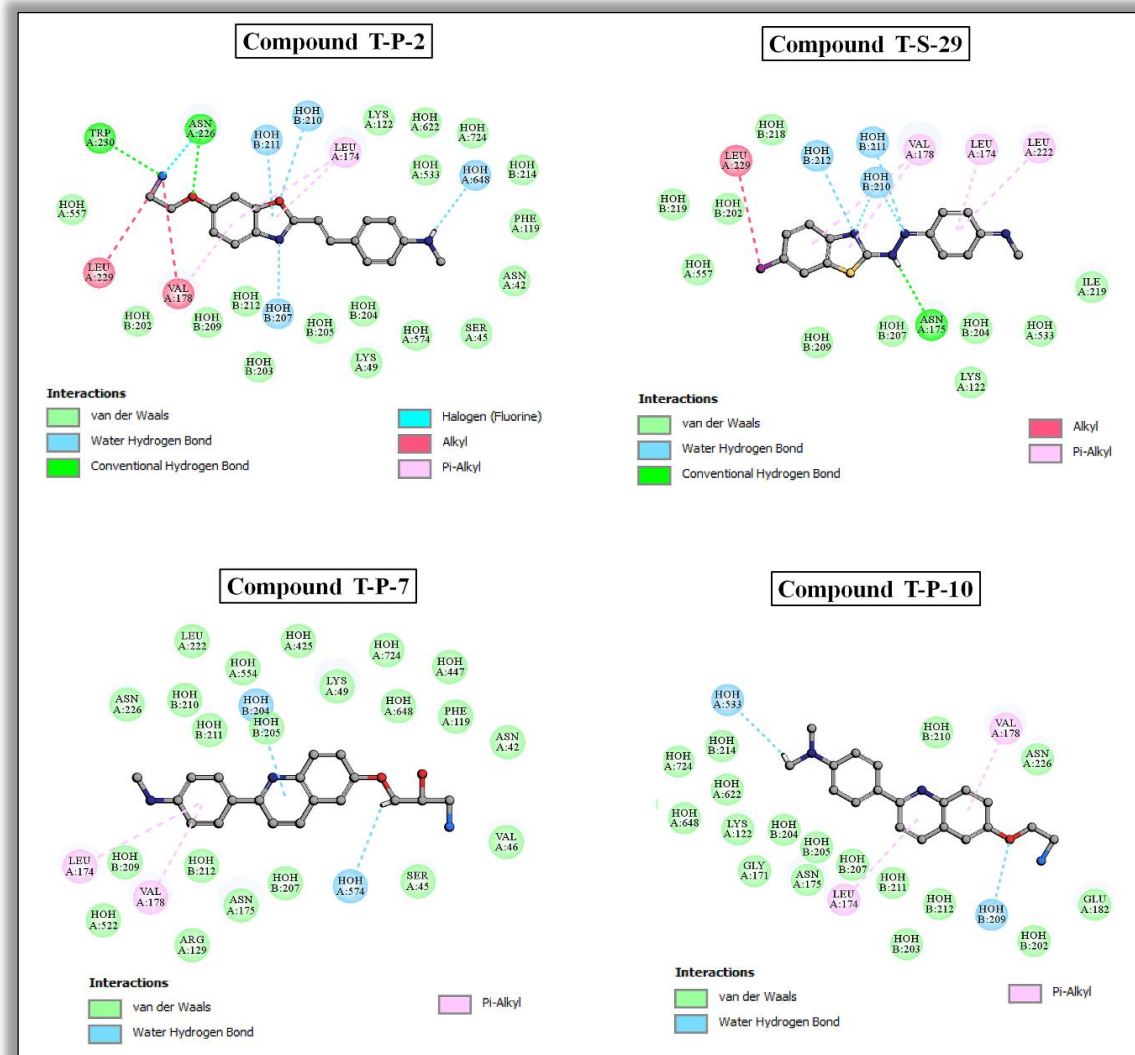


Figure 4.16. Molecular interactions between high and low active PET or SPECT imaging agents with tau protein

Table 4.2: The docking site, interacting residues and different types of binding interaction occurring between the imaging agents and target protein (A β or tau)

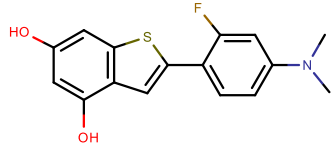
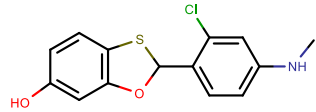
| Dataset | Imaging agents | Compound ID | pKi | (-)Docking interaction energy (kcal/mol) | Docking site | Interacting amino acids | Binding interactions |
|--------------|----------------|-------------|------|--|--|--|--|
| Beta amyloid | PET | A-P-2 | 5.15 | 23.53 | Val D:39, Gln B:13, Leu A:17, His B:13, Leu B:17, Val E:36, Leu C:17, Gly E:38, Val E:39, Val A:12, His A:13. | Val D:39, His B:13, Val D:36 | Carbon hydrogen bond, unfavorable acceptor-acceptor, π -donor hydrogen, π -sulphur, π - π T-shaped, π -alkyl and van der Waals |
| | | A-P-56 | 4.77 | 25.71 | Val A:12, His B:13, Val E:39, Gly E:38, Leu B:17, Val E:36, Gly E:37, Val D:36, Gly D:38, Gln B:15, Val D:39, His A:13. | Val A:12, His B:13, Val D:36, Val D:39 | Carbon hydrogen bond, π -donor hydrogen, π -sulphur, alkyl, π -alkyl and van der Waals |
| | | A-P-52 | 3.19 | 17.95 | Val E:36, Gly E:38, Gln B:15, His B:13, Val D:39, Gln A:15, Leu A:17, Gly D:38, Val D:36 | Val E:36, Val D:39 | Carbon hydrogen bond, π -alkyl and van der Waals |
| | | A-P-50 | 3.34 | 20.68 | Val D:39, His B:13, Gly E:38, Leu C:17, Val E:36, Leu B:17, Leu A:17, Gln B:15, Gly D:38 | Val D:39, His B:13 | Carbon hydrogen bond, π - π T-shaped, π -alkyl and van der Waals |
| | SPECT | A-S-53 | 6.0 | -7.469 | Gly J29, Ala J30, Ile J31, His B13, Gln B15, Gly D38, Leu A17, Gly D37, Gly C38, Gly C37, Val C36, Val D36, Leu B17, Val E36, Val D39, Val E39, Gly E38, Ile K31, Gly E37, Val D40, Ala L30, Gly K29 Ala K30 | Gly J:29, His B:13, Gly D:38, Leu A:17, Gly D:37, Gly C:37, Val D:39, Val E:39, Ile K:31, Val D:40 | Conventional hydrogen bond, carbon hydrogen bond, π -donor hydrogen bond, π -sigma, π -alkyl and van der Waals |
| | | A-S-52 | 5.77 | -23.89 | Ile K31, Gly E38, Val D39, Gly D38, Gln C15, Val C39, Gly C38, Val | Ile K:31, Val D:39, Val D:40, Val E:39, His | Carbon hydrogen bond, π -donor hydrogen bond, alkyl, π -alkyl and |

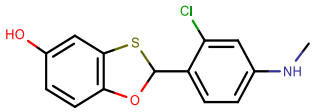
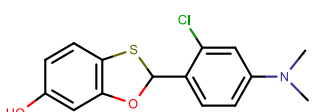
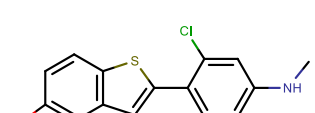
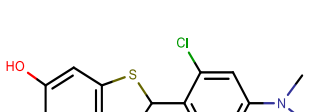
| | | | | | | | |
|-------------|--------------------------|--------|-------|--|---|--|---|
| | | | | D36, Leu A17, Gly D37, Gln B15, Lys B16, Val D40, Gly E37, Val E39, His B13, Ala L30, Gly J29, Gly K29, Ala K30 | B:13, Ala L:30, Gly J:29, Gly K:29 | van der Waals | |
| | A-S-55 | 2.18 | 7.20 | Ile I:31, Ile J:31, Gly C:38, Gly C:37, Gly D:38, Gly D:37, Leu A:17, Leu C:17, Gln C:16, Val E:39, Gly E:38, Gln B:15, Val D:39, Val C:39, Val D:40, Val C:40, Gly J:29, Gly I:29 | Val D:39, Val D:40, Gly I:29 | Carbon hydrogen bond, alkyl, π -alkyl and van der Waals | |
| | A-S-20 | 2.66 | 31.87 | Val A:12, Val D:39, Leu A:17, Gly D:38, Leu C:17, Val D:36, Val E:36, Gly E:37, Gly E:38, Leu B:17, Gln B:15, His B:13 | Val D:39, Val D:36, Val E:36, His B:13 | Conventional hydrogen bond, carbon hydrogen bond, π -alkyl and van der Waals | |
| | T-P-2 | 4.319 | 33.49 | Trp A:230, Asn A:226, Leu A:174, Lys A:122, Phe A:119, Asn A:42, Ser A:45, Lys A:49, Val A:178, Leu A:229 | Trp A:230, Asn A:226, Leu A:174, Val A:178, Leu A:229 | Conventional hydrogen bond, halogen (Fluorine), alkyl, π -alkyl, water hydrogen bond and van der Waals | |
| | T-S-29 | 3.959 | 19.82 | Leu A:229, Val A:178, Leu A:174, Leu A:222, Ile A:219, Lys A:122, Asn A:175 | Leu A:229, Val A:178, Leu A:174, Leu A:222, Asn A:175 | Conventional hydrogen bonds, alkyl, π -alkyl, water hydrogen bond and van der Waals | |
| Tau protein | PET and SPECT (combined) | T-P-10 | 1.311 | 34.12 | Lys A:122, Gly A:171, Asn A:175, Leu A:174, Glu A:182, Asn A:226, Val A:178 | Leu A:174, Val A:178 | π -alkyl, water hydrogen bond and van der Waals |
| | T-P-7 | 1.957 | 29.80 | Leu A:174, Val A:178, Arg A:129, Asn A:175, Ser A:45, Val A:46, Asn A:42, Phe A:119, Lys A:49, Leu A:222, Asn A:226 | Leu A:174, Val A:178 | π -alkyl, water hydrogen bond and van der Waals | |

4.1.4. QSAR modeling and molecular docking studies for newly designed PET and SPECT imaging agents

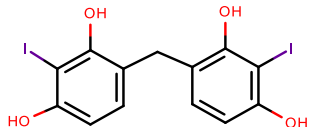
A set of 12 imaging agents (6 each for both PET and SPECT) were designed for and their QSAR prediction and docking studies were performed to understand the binding properties towards A β plaques. Also, another 6 imaging agents (PET and SPECT combined) targeting tau protein were designed for QSAR model prediction and molecular binding. From the QSAR analysis, it was found that all the compounds designed for both A β and tau protein gave good predicted binding affinity (**Table 4.3**) and also falls under the model applicability domain as calculated by DModX method. The docking interactions as given in the **Figures 4.17, 4.18** and **4.19** also support the observations found for the actual dataset compounds. Similar interactions are observed in case of the newly designed compounds, thus ensuring the validity of the new design.

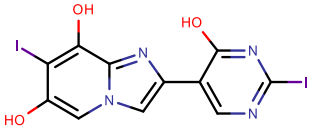
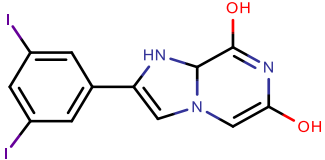
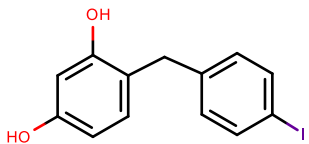
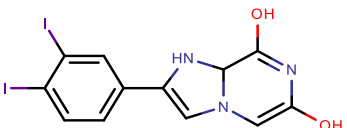
Table 4.3. The predicted binding affinity values and different types of molecular interactions occurring in case of the newly designed PET and SPECT imaging agents

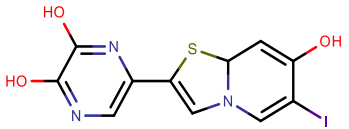
| NEWLY DESIGNED PET IMAGING AGENTS FOR BETA AMYLOID | | | | | | |
|--|---|-------------------------------|--|--|---|---|
| SL NO. | Structure | Predicted pKi from QSAR model | (-)Docking interaction energy (kcal/mol) | Docking Site | Interacting residues | Type of Interactions Present |
| N-1 |  | 5.57 | 34.64 | Gln C15, Gln B15, Val E39, Gly E37, Gly D37, Gly D38, Ile K31, Val D40, Ile J31, Gly J29, Ala J30, Val D39, Gly E38, His B13 | Val E39, Gly E37, Gly D37, Gly D38, Ile K31, Val D40, Val D39, Gly E38, His B13 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, halogen (Fluorine) interaction, pi-donor hydrogen interaction, pi-sigma interaction, amide-pi stacked interaction, pi-alkyl interaction and van der Waals interaction |
| N-2 |  | 5.52 | 32.03 | Gly D37, Gly C37, Val C36, Val D36, Leu A17, Gln B15, Val D39, His B13, Val E39, Gln C15, Gly E38, Gly D38, Val C39, Gly C38 | Gly D37, Gly C37, Val C36, Gln B15, Val D39, His B13, Val E39 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |

| | | | | | | |
|-----|---|------|-------|---|---|---|
| N-3 |  | 5.41 | 31.86 | Val E39, Gly E38, Gln C15, Gln B15, Gly D38, Leu A17, Val D36, Gly C37, Gly D37, Val C36, Gly C38, Val D39 | Val E39, Gly E38, Gln B15, Leu A17, Gly C37, Gly D37, Val C36, Val D39 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |
| N-4 |  | 5.40 | 20.37 | Gly E38, Gly D38, Gly E37, Val D36, Val E36, Lys B16, Leu A17, Leu B17, Gln A15, Val D39, Gln B15, Gln C15 | Gly E38, Gly D38, Val E36, Leu A17, Leu B17, Gln A15, Gln B15, Gln C15 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, pi-lone pair interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |
| N-5 |  | 5.37 | 30.34 | Val E39, Gly E38, Gln C15, Gly D38, Val D36, Gly D37, Gly C37, Val C36, Gly C38, Leu A17, Gln A15, Gln B15, Val D39, His B13 | Val E39, Gln C15, Gly C37, Val C36, Leu A17, Val D39, His B13 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |
| N-6 |  | 5.28 | 29.03 | Gln B15, His B13, Gly E38, Val E39, Ala L30, Gly K29, Ala K30, Gly J29, Ile K31, Val D40, Gly E37, Gly D38, Val E36, Val D39, Gln C15 | Gln B15, Gly E38, Val E39, Gly J29, Ile K31, Val D40, Gly E37, Val D39, Gln C15 | Carbon-hydrogen interaction, halogen (Cl, Br, I) interaction, Sulfur-X interaction, pi-donor hydrogen interaction, pi-alkyl interaction |

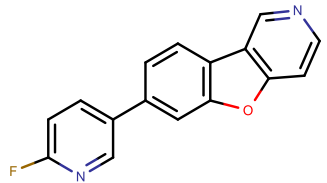
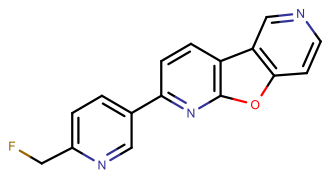
NEWLY DESIGNED SPECT IMAGING AGENTS FOR BETA AMYLOID

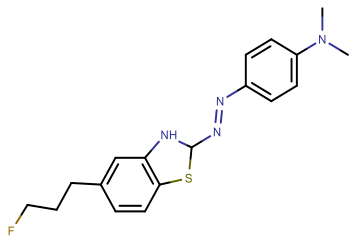
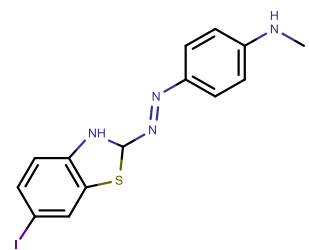
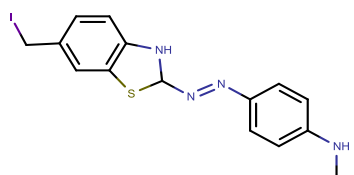
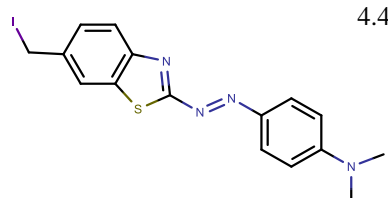
| SL NO. | Structure | Predicted pKi from QSAR model | (-)Docking interaction energy (kcal/mol) | Docking Site | Interacting residues | Type of Interactions Present |
|--------|---|-------------------------------|--|---|---|---|
| N-7 |  | 6.91 | 38.21 | Gly E38, Val E36, Gly E37, Gly D38, Leu B17, Val D39, Leu A17, Gly C38, Gly D37, Gln B15, Val D36, Gln C15, Val E39 | Gly E38, Val E36, Gly E37, Gly D38, Leu A17, Val D36, Gln C15 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, amide-pi stacked interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |

| | | | | | | |
|------|--|------|-------|---|------------------------------------|--|
| N-8 |  | 6.66 | 33.94 | Gly E38, Val E39, Val A12, His A13, Gln A15, Gln B15, His B13, His A14, Val D39, Val E36, Gly D38 | Gln B15, His B13, Val D39, Val E36 | Conventional hydrogen bond interaction, carbon-hydrogen bond interaction, pi-donor hydrogen interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |
| N-9 |  | 6.66 | 27.41 | Val E39, Gly E38, Gly D38, Leu A17, Val C36, Gly C38, Gln A15, Val C39, Gln B15, Val D39, His B13 | Gly E38, Val C36, Val C39, Gln B15 | Carbon-hydrogen bond interaction, halogen (Cl, Br, I) interaction, alkyl interaction and van der Waals interaction |
| N-10 |  | 6.52 | 28.30 | Val E39, Gly E38, Leu C17, Val E36, Leu B17, Leu A17, Gly c37, Gly C38, Gly D37, Gly D38, Val D36, Gln B15 | Val E39, Leu A17, Gly C37 | Conventional hydrogen bond interaction, halogen (Cl, Br, I) interaction, pi-alkyl interaction and van der Waals interaction |
| N-11 |  | 6.28 | 29.66 | Leu A17, Gln A15, Val D36, Gly D38, Gly E38, leu, C17, Gln C15, Gln D15, Val E40, Val E39, Val D39, Gln B15 | Leu A17, Val D36, Val E39 | Unfavorable acceptor-acceptor interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |

| | | | | | | |
|------|---|------|-------|--|------------------|---|
| N-12 |  | 6.17 | 22.66 | His B13, Val E39, Val D39, Gly E38, Gly D38, Val E36, Val D36, Leu B17, Leu A17, Gln B15 | His B13, Val D39 | Conventional hydrogen bond interaction, alkyl interaction, pi-alkyl interaction and van der Waals interaction |
|------|---|------|-------|--|------------------|---|

NEWLY DESIGNED PET AND SPECT IMAGING AGENTS FOR TAU PROTEIN

| SL NO. | Structure | Predicted pKi from QSAR model | (-)Docking interaction energy (kcal/mol) | Docking Site | Interacting residues | Type of Interactions Present |
|--------|--|-------------------------------|--|--|--|---|
| N-13 |  | 3.69 | 23.35 | Ile A:219, Leu A:222, Asn A:175, Arg A:129, Val A:178, Leu A:174, Lys A:49 | Ile A:219, Arg A:129, Val A:178, Leu A:174 | π -cation, π -alkyl, water hydrogen and van der Waals interaction |
| N-14 |  | 3.57 | 24.35 | Arg A:56, Arg A:129, Leu A:222, Ile A:219, Leu A:218, Leu A:174, Asn A:175, Val A:178, Tyr A:130 | Arg A:56, Arg A:129, Leu A:222, Ile A:219, Tyr A:130 | Conventional hydrogen bond, halogen (fluorine), π -alkyl, water hydrogen bond and van der Waals interaction |

| | | | | | | |
|------|---|------|-------|---|---|--|
| N-15 |  | 3.63 | 33.43 | Leu A:218, Ile A:219, Gly A:171, Leu A:222, Asn A:175, Asn A:226, Leu A:229, Val A:178, Leu A:174 | Leu A:218, Ile A:219, Leu A:222, Asn A:175, Asn A:226, Val A:178 | Carbon hydrogen bond, Sulfur – X, alkyl, π -alkyl, water hydrogen bond and van der Waals interaction |
| N-16 |  | 3.87 | 33.32 | Asn A:226, Val A:178, Leu A:222, Gly A:171, Ser A:45, Lys A:49, Lys A:122, Arg A:129, Leu A:174, | Val A:178, Lys A:122 Leu A:174 | π -cation, π -alkyl, water hydrogen and van der Waals interaction |
| N-17 |  | 4.51 | 35.87 | Leu A:229, Tro A:230, Asn A:226, Leu A:174, Gly A:171, Lys A:122, Asn A:175, Val A:178, Glu A:182 | Leu A:229, Tro A:230, Asn A:226, Leu A:174, Gly A:171, Lys A:122, Asn A:175, Val A:178, Glu A:182 | Carbon hydrogen bond, π -donor hydrogen, alkyl, π -alkyl, water hydrogen and van der Waals interaction |
| N-18 |  | 4.44 | 36.92 | Trp A:230, Leu A:229, Val A:178, Gly A:171, Asn A:175, Leu A:174 | Trp A:230, Leu A:229, Val A:178 | Alkyl, π -alkyl, water hydrogen and van der Waals interaction |

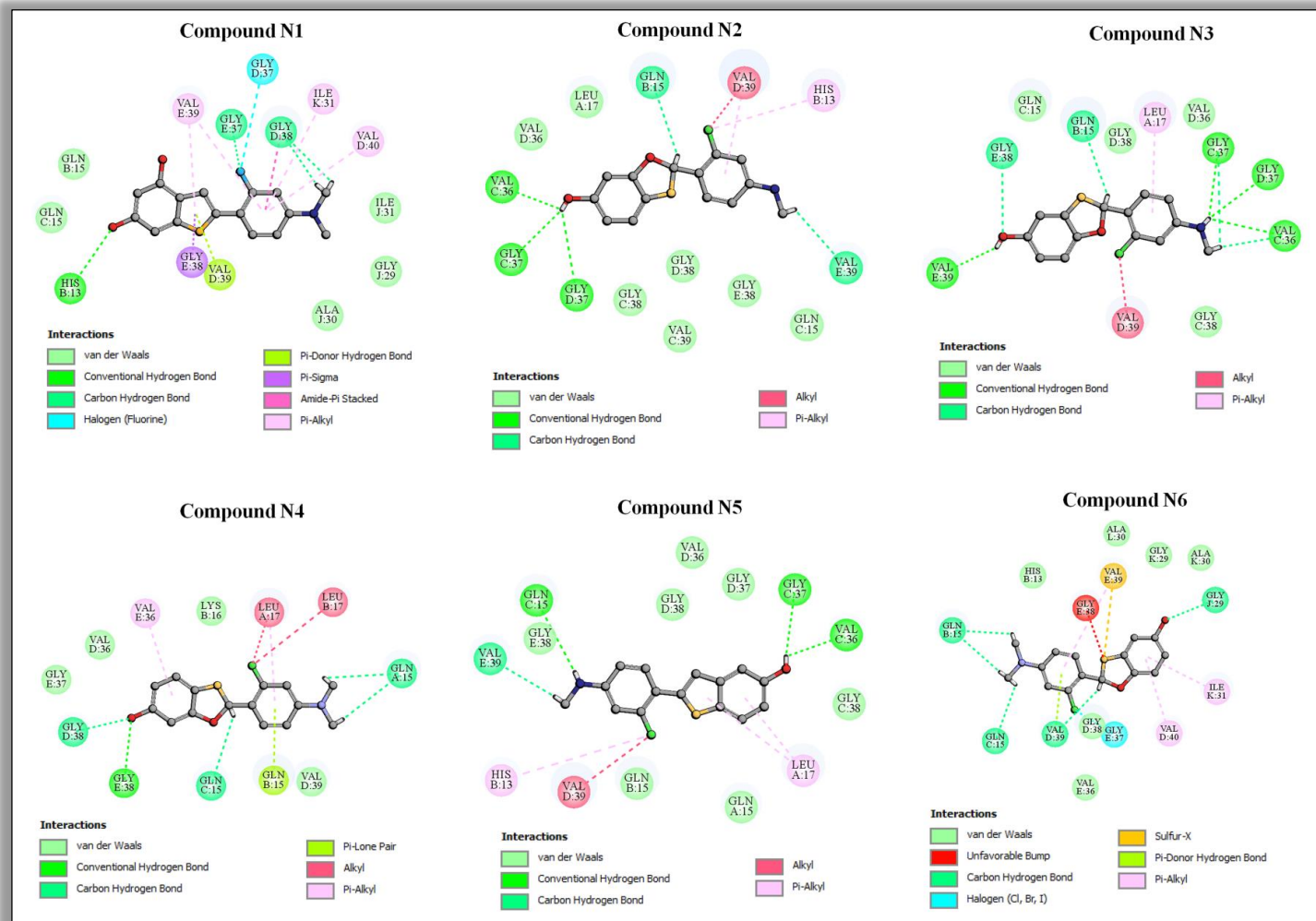


Figure 4.17. Molecular docking results of newly designed PET imaging agents targeted against A β plaques showing the intermolecular interactions

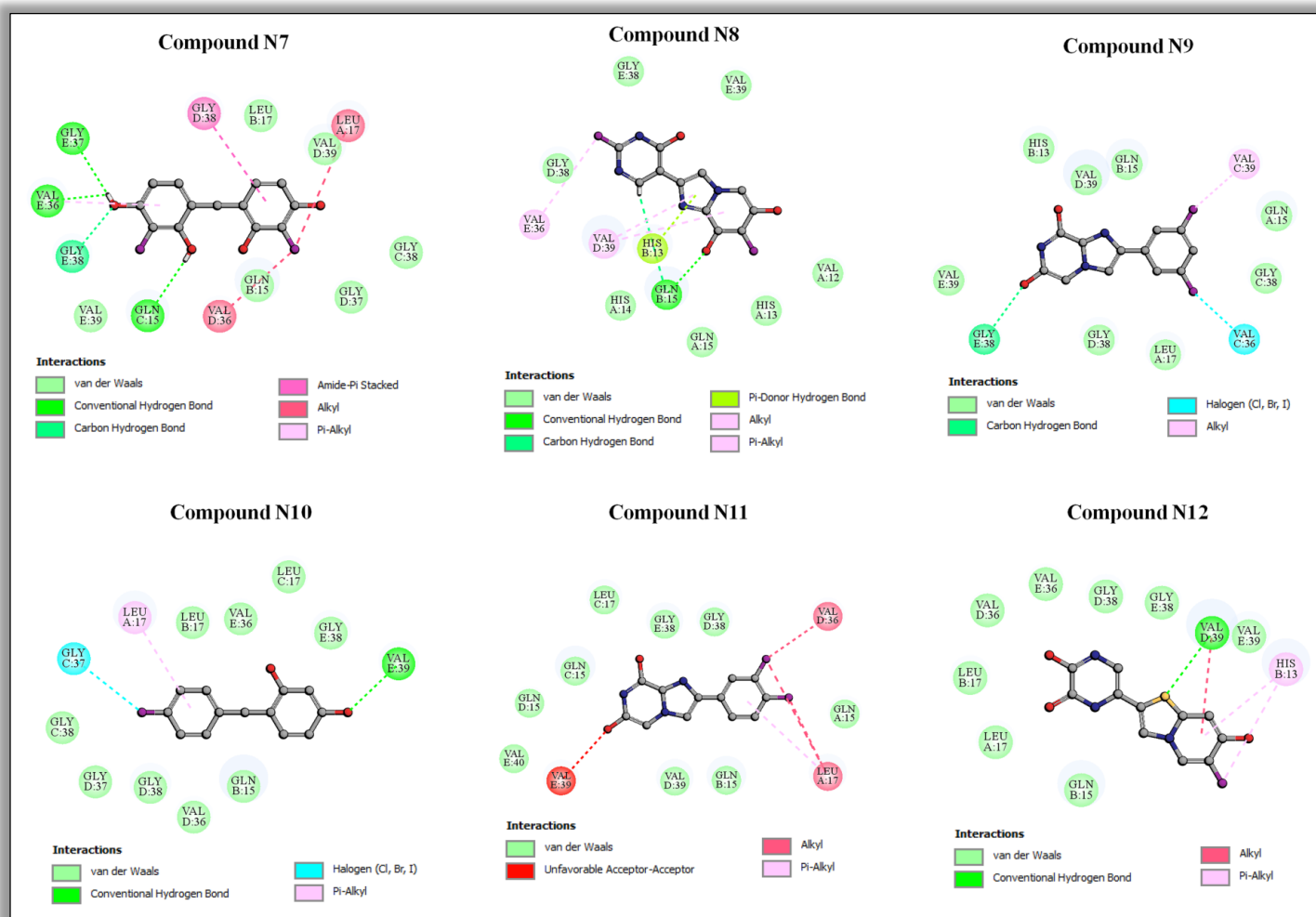


Figure 4.18. Molecular docking results of newly designed SPECT imaging agents targeted against A β plaques showing the intermolecular interactions

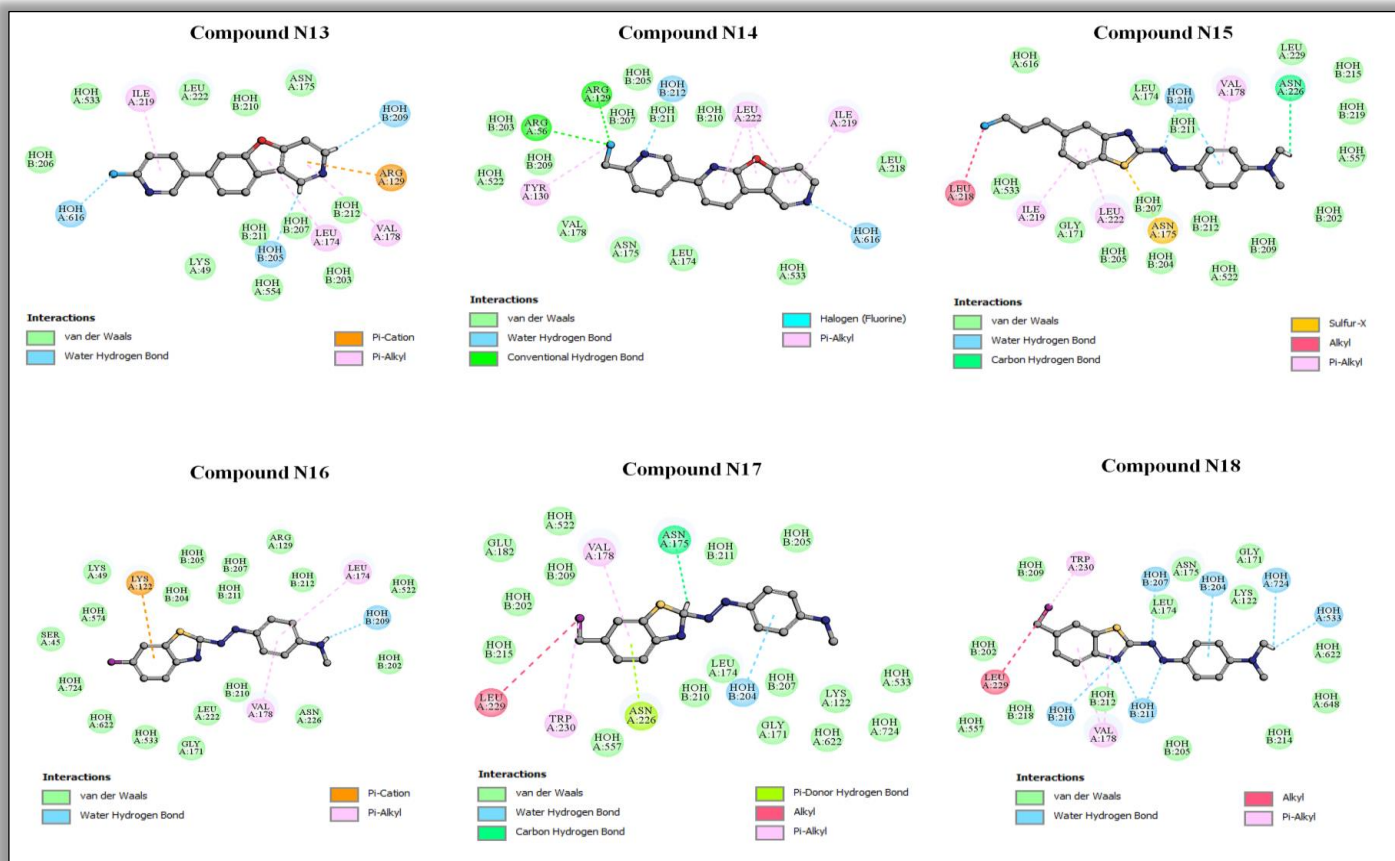


Figure 4.19. Molecular docking results of newly designed PET (first 3) and SPECT (last 3) imaging agents targeted against tau protein showing the intermolecular interactions

4.2. Study 2: Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: A QSAR approach

Based on the binding affinity and selectivity endpoints of 35 xanthine PET tracer antagonists of adenosine A_{2A} receptor, we have developed one model for the binding affinity ($Q^2=0.85$, $R^2=0.90$, $Q^2_{F1}=0.80$) and 4 models ($Q^2=0.80-0.87$, $R^2=0.87-0.91$, $Q^2_{F1}=0.84-0.85$) for selectivity. All the models were externally and internally validated which showed model robustness and good predictivity in terms of the statistical results. We have also checked the r_m^2 parameters for both internal sets ($\overline{r_{m(loo)}^2}$, $\Delta r_{m(loo)}^2$) and external sets ($\overline{r_{m(test)}^2}$ and $\Delta r_{m(test)}^2$), and the statistical results were above the critical point justifying the reliability of the models. To improve the quality of the external prediction for selectivity, we also performed “*Intelligent Consensus Prediction*” of the multiple MLR models using the ICP tool (Kunal Roy et al., 2018), and found that the consensus predictions were better than the individual MLR model derived predictions. The winner model was consensus model 0 (CM0).

4.2.1. Modeling binding affinity of PET tracers towards Adenosine (A_{2A}) receptor

The model for binding affinity consists of five descriptors: C-025, F09 [N-O], nBnz, NRS, and nCIR which significantly influence the binding of the antagonists to the adenosine (A_{2A}) receptor. The 5 descriptor MLR model (Equation 4.1) developed using Genetic Function Algorithm (GFA) could predict 85.0% variance of the training set and 80.0% of the test set. The values of all descriptors appearing in the model for training and test set compounds are given in **Table 4.4**. The observed versus predicted scattered plot is given **Figure 4.20**.

$$pKi(A_{2A}R) = -0.849(\pm 0.2167) - 0.36271(\pm 0.06190) \times C - 025 + 0.17693(\pm 0.05895) \times F09[N - O] - 0.52109(\pm 0.07616) \times NRS + 0.81699(\pm 0.09908) \times nBnz + 0.3024(\pm 0.03363) \times nCIR$$

$$\begin{aligned} n_{training} &= 25, R^2 = 0.901, R_{adj}^2 = 0.875, Q^2 = 0.850, S = 0.170027, F = 34.62, PRESS \\ &= 0.833306, \overline{r_{m(loo)}^2} = 0.790, \Delta r_{m(loo)}^2 = 0.072, MAE \text{ based criteria} \\ &= \text{Moderate} \end{aligned}$$

$$n_{test} = 10, Q_{F1}^2 = 0.80, Q_{F2}^2 = 0.681, \overline{r_{m(test)}^2} = 0.54, \Delta r_{m(test)}^2 = 0.23, MAE \text{ based criteria} = \text{Good}$$

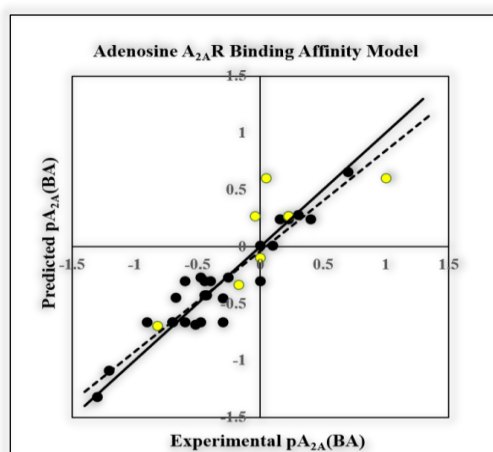


Figure 4.20. Observed vs predicted A_{2A}R binding affinity scatter plot.

Table 4.4. Descriptor values appearing in the model for training and test set compounds including the predicted pKi (A_{2A}R) values.

| No. | nCIR | NRS | nBnz | C-025 | F09[N-O] | pKi(A _{2A} R) | Pred_pKi(A _{2A} R) |
|-----|------|-----|------|-------|----------|------------------------|-----------------------------|
| 1 | 7 | 5 | 2 | 2 | 0 | -0.4472 | -0.4120 |
| 2 | 7 | 5 | 2 | 2 | 0 | -0.4314 | -0.4120 |
| 3* | 7 | 5 | 2 | 1 | 0 | 0.0000 | -0.0493 |
| 4 | 8 | 4 | 1 | 0 | 2 | 0.6990 | 0.6737 |
| 5* | 4 | 2 | 0 | 0 | 0 | -0.8195 | -0.6741 |
| 6 | 4 | 2 | 0 | 0 | 0 | -0.5185 | -0.6741 |
| 7 | 5 | 3 | 1 | 1 | 0 | -0.6721 | -0.4346 |
| 8* | 9 | 4 | 1 | 0 | 0 | 1.0000 | 0.6241 |
| 9* | 9 | 4 | 1 | 0 | 0 | 0.0458 | 0.6241 |
| 10 | 9 | 4 | 1 | 1 | 0 | 0.1549 | 0.2614 |
| 11* | 6 | 4 | 1 | 0 | 0 | -0.6021 | -0.2888 |
| 12 | 6 | 4 | 1 | 1 | 0 | -0.6990 | -0.6515 |
| 13 | 6 | 4 | 1 | 1 | 0 | -0.4771 | -0.6515 |
| 14 | 5 | 3 | 1 | 2 | 2 | -0.3010 | -0.4435 |
| 15 | 6 | 4 | 1 | 1 | 0 | -0.6990 | -0.6515 |
| 16 | 6 | 4 | 1 | 1 | 0 | -0.9031 | -0.6515 |
| 17 | 6 | 4 | 1 | 1 | 0 | -0.3010 | -0.6515 |
| 18 | 6 | 4 | 1 | 1 | 0 | -0.6021 | -0.6515 |
| 19 | 5 | 3 | 1 | 1 | 1 | -0.4771 | -0.2577 |
| 20 | 3 | 3 | 2 | 4 | 0 | -1.3010 | -1.3123 |
| 21* | 9 | 4 | 1 | 1 | 0 | 0.2218 | 0.2614 |
| 22* | 9 | 4 | 1 | 1 | 0 | 0.2218 | 0.2614 |
| 23 | 9 | 4 | 1 | 1 | 0 | 0.2218 | 0.2614 |
| 24* | 9 | 4 | 1 | 1 | 0 | -0.0414 | 0.2614 |
| 25 | 9 | 4 | 1 | 1 | 0 | 0.3979 | 0.2614 |
| 26* | 9 | 4 | 1 | 1 | 0 | 0.2218 | 0.2614 |
| 27 | 5 | 3 | 1 | 1 | 1 | -0.2553 | -0.2577 |
| 28 | 6 | 4 | 1 | 0 | 0 | -0.4472 | -0.2888 |
| 29* | 6 | 4 | 1 | 0 | 0 | -0.1761 | -0.2888 |
| 30 | 6 | 4 | 1 | 0 | 0 | 0.0000 | -0.2888 |

| | | | | | | | |
|----|---|---|---|---|---|---------|---------|
| 31 | 6 | 4 | 1 | 0 | 0 | -0.3979 | -0.2888 |
| 32 | 5 | 3 | 3 | 4 | 1 | 0.3010 | 0.2921 |
| 33 | 5 | 3 | 2 | 2 | 0 | 0.0969 | 0.0216 |
| 34 | 5 | 3 | 2 | 2 | 0 | 0.0000 | 0.0216 |
| 35 | 5 | 3 | 0 | 1 | 1 | -1.2041 | -1.0767 |

‘*’-Compounds in test set

Table 4.5: Definition and contribution of all the descriptors obtained from the MLR models (models developed by using binding affinity)

| Sl. no. | Name of descriptors | Descriptor Type | Contribution | Discussion | Probable mechanism of binding |
|---------|---------------------|-----------------------------------|--------------|--|---|
| 1 | C-025 | Atom centered fragment descriptor | -ve | C-025 can be depicted as R--CR--R , where ‘ R ’ can be any group linked to carbon and ‘--’ is any aromatic bond. It is the number of fragments in which a C (sp ²) aromatic atom is bound to three carbon atoms, two of them by an “aromatic bond” and the third by a simple single bond | Flexibility which helps in accommodating the antagonist well in the receptor pocket |
| 2 | nBnz | Ring descriptor | +ve | Indicates number of benzene-like rings | π - π stacking interaction |
| 3 | F09 [N-O] | 2D atom pair descriptor | +ve | Frequency of N-O fragment at the topological distance 9 | Hydrogen bonding |
| 4 | NRS | Ring descriptor | -ve | A ring descriptor indicates number of ring systems within a molecule | - |
| 5 | nCIR | Ring descriptor | +ve | Number of circuits, i.e., larger loops around two or more rings in a molecule | Hydrophobic interaction/ π - π stacking interaction |

4.2.1.1. Essential features required for binding and receptor interaction

The descriptors obtained in the QSAR model gives an insight regarding the mechanism of interaction occurring during binding of the xanthine PET tracer antagonists to adenosine A_{2A} receptor. Unsaturation and aromaticity play a dominating role in regulating the receptor binding affinity which is evident from the occurrence of descriptors such as **C-025**, **nBnz**, **NRS** and **nCIR**. Descriptors like nBnz and nCIR has positive influences on the adenosine A_{2A} receptor binding (**Figure 4.21**). But on the other hand, descriptors like C-025 and NRS has negative effects on the binding affinity of the PET tracers (**Figure 4.22**). The occurrence of these similar types of descriptors with opposite influence is contradictory and leads to a conclusion that aromaticity provided by benzene nucleus (as seen in

compounds like **A-32** and **A-23**) is more important for binding. On the other hand, the presence of heterocyclic aromatic ring and fused ring system decreases the overall binding affinity of the radiotracer molecule (found in compounds **A-1**, **A-2** and **A-20**).

The 2D atom pair descriptor **F09 [N-O]** gives information about the electronegativity of the compounds and the positive coefficient of the descriptor suggests that higher occurrence of nitrogen and oxygen at topological distance 9 would enhance binding affinity of the compounds as seen in compounds **A-4** and **A-32**. It is found that the presence of electronegative atoms in the compounds or chemical structures can influence the binding to the receptor through hydrogen bonding (Kunal Roy et al., 2018).

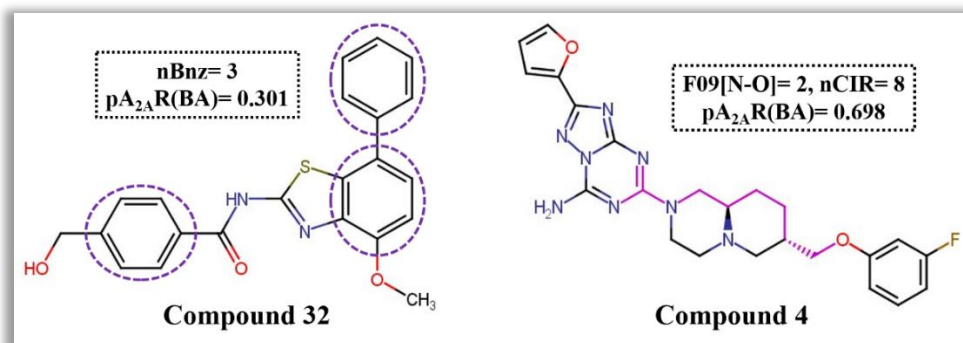


Figure 4.21: Features increasing the binding affinity (pK_i) value.

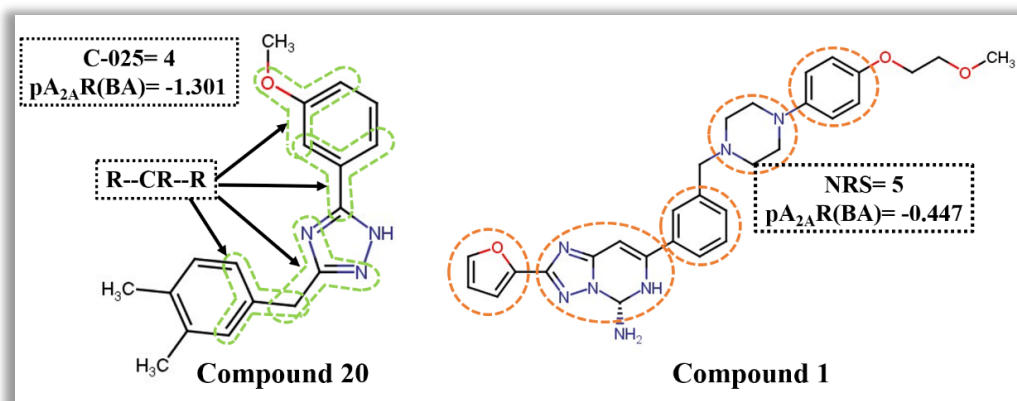


Figure 4.22: Features decreasing the binding affinity (pK_i) value.

4.2.1.2. Molecular docking

Molecular docking helped in understanding the optimized conformation of the complex between the imaging agent and A_{2A} receptor and gave evidences related to the orientation of the imaging agents at the binding zone of the receptor. The major goal was to understand the molecular interactions taking place during radiotracer binding and correlate these findings with QSAR analysis. The docking analysis showed the predominance of different types of π bonding interactions and hydrogen bonding interactions. In higher active compounds like **A-4**, **A-8** and **A-25** ($pA_{2A}R(BA)=0.699$, 1.000 and 0.398 respectively), the interaction forces include mainly hydrogen bonding interactions (conventional hydrogen bond and carbon-hydrogen bond interaction), π interactions (π -cation, π -donor hydrogen, π - π stacked, π - π T-shaped and π -alkyl) (**Figure 4.23**). Other interactions include halogen and alkyl interaction in compound **A-4** and salt bridge formation in compound **A-8**. Higher number of interacting

residues supports the fact that these compounds have higher binding affinity. Compounds having binding affinity in the medium range like compound number **A-14** and **A-27** ($pA_{2A}R(BA) = -0.301$ and -0.255 respectively) makes less number of interactions with the adenosine receptor but the type of interactions remains similar, i.e., π interactions and hydrogen bonding interactions. The lowest active compounds like compound number **A-20** and **A-35** ($pA_{2A}R(BA) = -1.301$ and -1.204 respectively) show the least number of interactions (**Figure 4.24**). All the details of binding including interacting residues and type of binding interactions are given in **Table 2**.

Table 4.6: Details of interacting residues and different types of binding interaction occurring between the PET imaging agents and the target protein (adenosine A_{2A} receptor).

| Compound No. | Activity | Binding affinity [$pA_{2A}R(BA)$] | (-)Docking interaction energy (kcal/mol) | Interacting Residues | Binding Interactions |
|--------------|----------|-------------------------------------|--|---|---|
| A-4 | High | 0.699 | 57.07 | Ala A:88, Val A:186, Leu A:85, Asn A:181, His A:250, Asn A:253, Phe A:168, Ser A:67, Met A:270, Leu A:267, Ile A:274, Ala A:63, Ile A:66, Leu A:249, Met A:177, Trp A:246 | Conventional hydrogen bond, carbon hydrogen bond, halogen (Fluorine), π -cation, π -donor hydrogen bond, π - π stacked, π - π T-shaped, alkyl, π -alkyl |
| A-8 | High | 1.000 | 64.46 | Met A:270, Asn A:253, Leu A:249, Phe A:168, Ala A:81, Ile A:66, Glu A:169 | Conventional hydrogen bond, carbon hydrogen bond, π - π stacked, π -alkyl, salt bridge |
| A-25 | High | 0.398 | 64.43 | Leu A:267, Tyr A:271, Ile A:274, Asn A:181, Gln A:89, Leu A:85, Leu A:249, Val A:84, Ser A:67, Glu A:169 | Conventional hydrogen bond, carbon hydrogen bond, π -sigma, π - π T-shaped, π -alkyl |
| A-14 | Medium | -0.301 | 40.01 | Val A:84, Leu A:249, Met A:270, Ile A:274, Ile A:66, Tyr A:271, Phe A:168 | π -sulfur, π - π T-shaped, π - π stacked, amide- π stacked, π -alkyl |
| A-27 | Medium | -0.255 | 43.06 | Asn A:253, Ser A:67, Ile A:274, Leu A:167, Glu A:169, Ala A:63, Ile A:66, Leu A:249, Val A:84 | Conventional hydrogen bond, carbon hydrogen bond, π -anion, π -alkyl |
| A-20 | Low | -1.301 | 37.90 | Val A:84, Leu A:249, Leu A:267, Tyr A:271, Ser A:67, Ile A:274, Asn A:253 | Conventional hydrogen bond, π - π T-shaped, π -sigma, π -alkyl, alkyl |
| A-35 | Low | -1.204 | 34.29 | Val A:84, Ala A:277, Leu A:249, Ile A:274, Met A:270, Glu A:169 | π -alkyl, alkyl, π -anion |

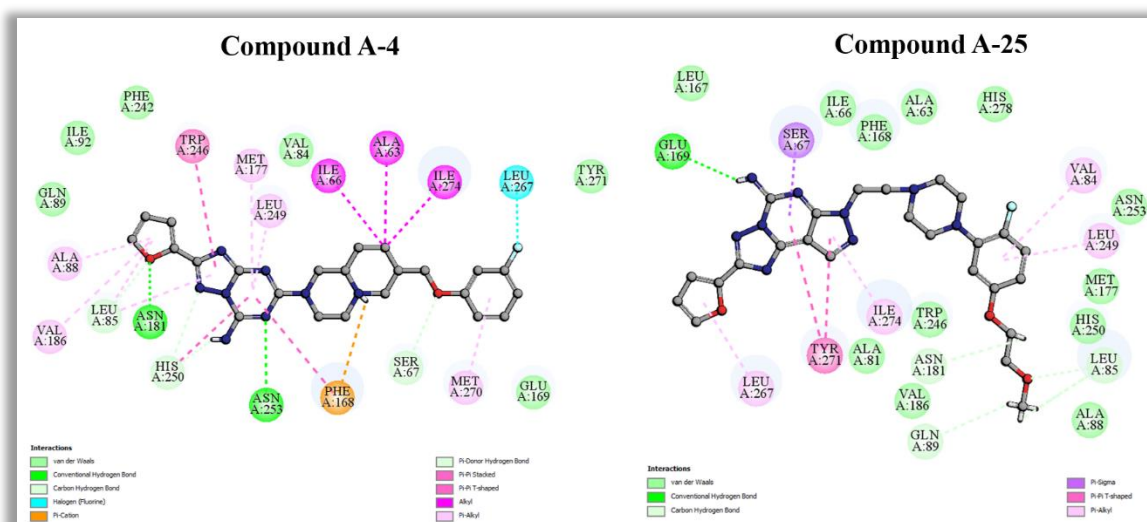


Figure 4.23: Docking interactions for compounds having higher binding affinity (pKi)

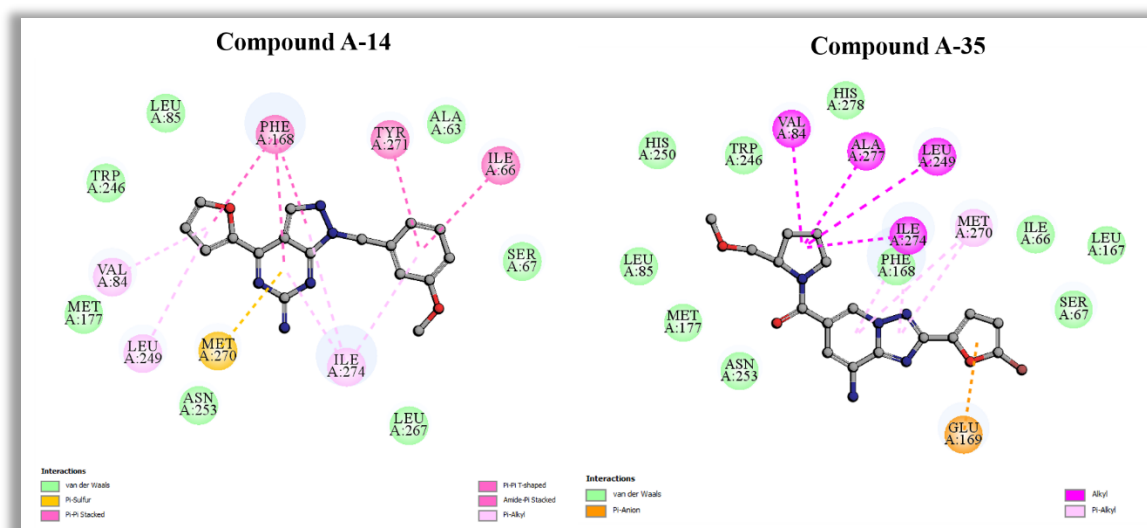


Figure 4.24: Docking interactions for compounds having medium (A-14) and low (A-35) binding affinity (pKi)

4.2.1.3. Relationship with QSAR models

The docking study shows different types of π interactions occurring between the PET radiotracer molecules and adenosine A_{2A} receptor. This observation supports the occurrence of **nBnz** and **nCIR** descriptors obtained in the QSAR models. The presence of aromatic rings like benzene can enhance binding with the receptor through aromatic π - π stacking interaction with the phenyl/imidazole residue of the receptor (Jaakola et al., 2008). The interaction of these antagonists through π - π stacking interaction eventually blocks the receptor in the indirect pathway thus blocking the activity of GABA-mediated influence in the globus pallidus pars externa (GPe). This helps the PD patients to gain the motor function again by regaining the balance between direct and indirect pathway. Nitrogen and oxygen are capable of hydrogen bond formation and various types of hydrogen bonding as observed in both higher active and lower active compounds, and this can be also correlated to **F09[N-O]** descriptor which gives an idea about the electronegativity of the molecule.

4.2.2. Modeling selectivity of PET tracers towards Adenosine (A_{2A}) receptor

In the current work, we have developed four MLR models to understand the selectivity of the PET tracer molecules towards adenosine A_{2A} receptor. A single QSAR model may not be efficient enough for the prediction of activity since the property of molecules cannot be understood by a limited number of features. The use of multiple models for prediction using consensus approach helps in reducing model uncertainty by enhancing the prediction quality of external set and also to reducing the prediction errors (Roy et al., 2016). The four MLR models are given below:

Model 1

$$\log A_{2A} R(Sel) = 0.5875(\pm 0.4130) + 0.4643(\pm 0.1574) C - 027 - 0.8679(\pm 0.1797) C - 040 + 0.7245(\pm 0.1006) F09[N - O] + 0.8382(\pm 0.01749) ETA_Beta_s$$

$$n_{training} = 21, R^2 = 0.915, R_{adj}^2 = 0.893, Q^2 = 0.867, S = 0.234982, F = 42.88,$$

$$PRESS = 1.37546, \overline{r_{m(LOO)}^2} = 0.81227, \Delta r_{m(LOO)}^2 = 0.07373, MAE \text{ based criteria} = \text{Moderate}$$

$$n_{test} = 10, Q_{F1}^2 = 0.84, Q_{F2}^2 = 0.81, \overline{r_{m(test)}^2} = 0.7682, \Delta r_{m(test)}^2 = 0.11949, MAE \text{ based criteria} = \text{Good}$$

Model 2

$$\log A_{2A} R(Sel) = 0.36359(\pm 0.43605) - 0.76227(\pm 0.18863) C - 040 - 0.05224(\pm 0.02421) T(F..Cl) + 0.71046(\pm 0.11057) F09[N - O] + 0.09777(\pm 0.01808) ETA_Beta_s$$

$$n_{training} = 21, R^2 = 0.90, R_{adj}^2 = 0.87, Q^2 = 0.82, S = 0.274853, F = 35.21,$$

$$PRESS = 1.05627, \overline{r_{m(LOO)}^2} = 0.7526, \Delta r_{m(LOO)}^2 = 0.05874, MAE \text{ based criteria} = \text{Moderate}$$

$$n_{test} = 10, Q_{F1}^2 = 0.84, Q_{F2}^2 = 0.82, \overline{r_{m(test)}^2} = 0.7737, \Delta r_{m(test)}^2 = 0.04197, MAE \text{ based criteria} = \text{Good}$$

Model 3

$$\log A_{2A} R(Sel) = 0.9642(\pm 0.4535) + 0.31245(\pm 0.08846) nCIC + 0.4848(\pm 0.1856) C - 027 - 0.9394(\pm 0.2114) C - 040 + 0.6662(\pm 0.1184) F09[N - O]$$

$$n_{training} = 21, R^2 = 0.883, R_{adj}^2 = 0.854, S = 0.274853, F = 30.27,$$

$$PRESS = 1.72765, Q^2 = 0.833, \overline{r_{m(LOO)}^2} = 0.76, \Delta r_{m(LOO)}^2 = 0.12, MAE \text{ based criteria} = \text{Moderate},$$

$$n_{test} = 10, Q_{F1}^2 = 0.84, Q_{F2}^2 = 0.82, \overline{r_{m(test)}^2} = 0.77, \Delta r_{m(test)}^2 = 0.13, MAE \text{ based criteria} = \text{Good}$$

Model 4

$$\log A_{2A} R(Sel) = 1.3245(\pm 0.2988) - 0.6702(\pm 0.2119) C - 040 + 0.10445(\pm 0.04427) SssN + 0.05519(\pm 0.01932) F07[C - C] + 0.5954(\pm 0.1263) F09[N - O]$$

$$n_{training} = 21, R^2 = 0.872, R_{adj}^2 = 0.84, S = 0.287861, F = 27.24,$$

$$PRESS = 2.09555, Q^2 = 0.827, \overline{r_{m(LOO)}^2} = 0.717, \Delta r_{m(LOO)}^2 = 0.131, MAE \text{ based criteria} = \text{Moderate},$$

$$n_{test} = 10, Q_{F1}^2 = 0.85, Q_{F2}^2 = 0.83, \overline{r_{m(test)}^2} = 0.78, \Delta r_{m(test)}^2 = 0.07, MAE \text{ based criteria} = \text{Good}$$

The significant descriptors obtained from the four MLR models (M1-M4) contributing to A_{2A} receptor selectivity are C-040, C-027, F09 [N-O], ETA_Beta_s, nCIC, T (F..Cl), SsssN and F07[C-C]. All the descriptors positively contribute to the A_{2A} receptor selectivity, except C-040, as identified from the regression coefficients of the descriptors and summarized in Table 4.7. We have also checked the applicability domain of the developed MLR models. The models showed good predictive ability as per the statistical results. The details of the descriptors, their contribution and frequency of appearance in all the four models are explained elaborately in Table 4.7. The experimental versus predicted A_{2A} R selectivity scatter plot is given in Figure 4.25.

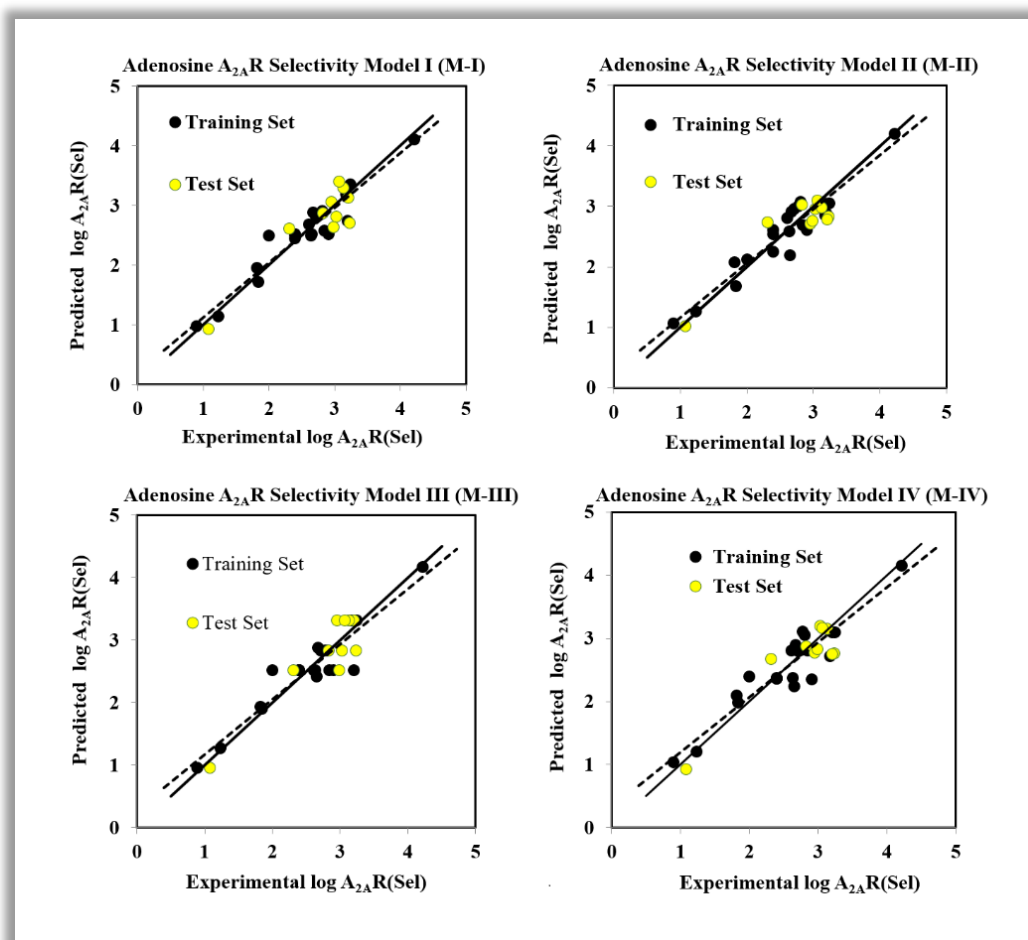


Figure 4.25. Observed vs predicted A_{2A} R selectivity plot for all four MLR models.

Table 4.7: Definition, frequency and contribution of all the descriptors obtained from the MLR models.

| Sl. No. | Name of descriptors | Type of descriptor | Contribution | Discussion | Frequency of descriptors |
|---------|---------------------|---------------------------|--------------|---|--------------------------|
| 1 | C-027 | Atom centred fragment | +ve | Counts for certain structural fragment (R--CH--X) in the antagonist, where 'R' can be any group linked to carbon and '--' is any aromatic bond. X can be any electronegative atom (O, N, S, P, Se, halogens) | 3 |
| 2 | ETA_Beta_s | ETA indices | +ve | Sum of all sigma bond contributions considering non-hydrogen vertices divided by 2. The descriptor deals with the presence of dissimilar heteroatoms. | 1 |
| 3 | F09 [N-O] | 2D atom pairs | +ve | Frequency of the N-O fragment at the topological distance 9 | 4 |
| 4 | SsssN | Atom-type E-state indices | +ve | E-state of ssssN which encodes the intrinsic electronic state of the nitrogen atom as perturbed by the electronic influence of other molecules with the context of topological character within the molecule. SsssN is the atom-type E-state of all tertiary nitrogen in molecules. | 1 |
| 5 | nCIC | Ring descriptors | +ve | Number of rings (cyclomatic number) present in the antagonist | 2 |
| 6 | C-040 | Atom centred fragment | -ve | Represented as R-C(=X)-X / R-C#X / X=C=X fragments where number of carbon atoms are attached to heteroatoms by single/double or triple bonds | 4 |
| 7 | F07[C-C] | 2D atom pairs | +ve | Frequency of C - C at topological distance 7 | 1 |

4.3.3.1. Mechanistic interpretation

All the descriptors obtained in the four models and their frequency gives an idea about their importance in modeling the selectivity of the PET tracers towards adenosine A_{2A} receptor. The descriptors like C-027, F09[N-O], SsssN, T(F..Cl) and ETA_Beta_s appearing in the models give information about the electronic feature of the compounds and are essential when selectivity of receptor is considered (**Figure 4.26**). Electronegativity is a chemical property that describes the tendency of an atom to draw electron towards itself. If a compound contains higher number of electronegative atoms in its structure, then the selectivity of the A_{2A} receptor for that compound also increases.

The presence of atom-centred fragments like **C-027** (R--CH--X) in compounds like **A-23** and **A-25** increases the antagonist selectivity of the PET compounds. Since 'X' represents any electronegative atom like O, N, S, P, Se, halogens, thus the presence of heteroatoms increases the selectivity of the compounds towards A_{2A} receptor. The descriptor **F09[N-O]** explains the frequency of presence of nitrogen and oxygen at topological distance 9 and its positive regression coefficient indicates its influential activity on the antagonistic behavior of the imaging agents (as seen in compounds **A-4** and **A-27**). Another similar kind of descriptor is **T (F..Cl)**, explaining the information about sum of topological distances between F and Cl atoms in the chemical structure. These descriptors too give information about the electronegative atoms, i.e. nitrogen and oxygen in **F09[N-O]** and fluorine and chlorine in **T(F..Cl)**. **ETA_Beta_s** ($\Sigma\beta_s$) is an extended topochemical atom (ETA) descriptor, which can be represented as sum of β_s values of all non-hydrogen vertices by 2. The term ' β_s ' can be denoted as

$$\Sigma\beta_s = \Sigma x\sigma$$

Here, x represents contribution of sigma bonds and σ signifies parameters related to sigma bonds. During the computation of β values, the sigma bond value for two similar types of electronegative atoms should be considered as 0.5 and dissimilar electronegative atoms should be considered as 0.75. This suggests that compounds bearing dissimilar heteroatoms will have greater selectivity to A_{2A} receptor as seen in compound **A-25**, **A-23** and **A-4**. Sigma bonds connected with different heteroatoms will have higher descriptor values indicating that the presence of dissimilar heteroatoms is more favorable for selectivity than similar heteroatoms. E-state descriptor **SsssN** (>N—) encodes the intrinsic electronic state of the nitrogen atom as perturbed by the electronic influence of other molecules with the context of topological character within the molecule. The electronegative contribution of nitrogen is well depicted in this descriptor, and the positive regression coefficient shows that an increase in the number of tertiary nitrogen benefits in receptor selectivity as seen in compounds **A-30** and **A-4**.

Other descriptors which significantly contribute to A_{2A} receptor selectivity are **nCIC**, **F07[C-C]**, **C-040**. These descriptors give information about the number of rings present, type of bonds and size of the antagonists showing selectivity towards the receptor. The number of rings (cyclomatic number) in the structure is indicated by **nCIC** descriptor. The positive regression coefficient of the descriptor suggests that presence of high number of rings increases the selectivity towards the A_{2A} receptor as observed in compounds **A-25** and **A-4**. **F07[C-C]**, a 2D atom pair stands for frequency of C – C fragment at the topological distance 7. It provides information about the size (chain length) of the molecule. This means that with an increase in the number of this fragment, i.e., carbon chain, the selectivity towards the A_{2A} receptor increases (as in compounds **A-4** and **A-25**). The atom-centered fragment descriptor, **C-040** (**Table 4.7**) gives information about the number of carbon atoms that are

attached to heteroatoms by single/double or triple bonds in the straight chain length. The negative regression coefficient suggests that an increase in the number of such fragments decreases the selectivity of the compound towards the A_{2A} receptor as seen in compounds A-6, A-7 and A-35. As this fragment suggests high number of double and triple bonds attached with the carbon, it can be concluded that unsaturation in the straight chain of the antagonists is unfavorable for the receptor selectivity.

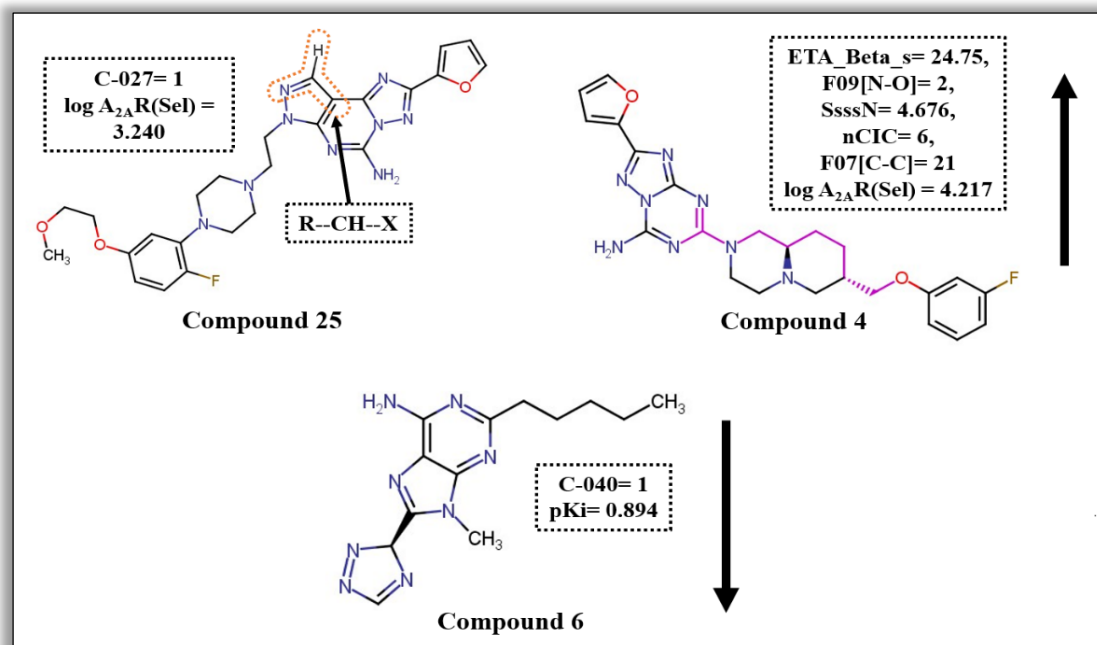


Figure 4.26: Features affecting the adenosine A_{2A} selectivity.

4.2.3. Intelligent Consensus Predictions

For further refinement of the predictions obtained from the individual models we have applied intelligent consensus modeling methods. Consensus modelling helps in enhancing the prediction performance of the models and also reduces the test set errors. It was observed that consensus prediction of the test set compounds (**Table 4.8**) are better in terms of both MAE based criteria and predicted R² parameter. Four different consensus approaches were used employing “Intelligent Consensus Prediction” tool (Kunal Roy et al., 2018): CM0 (simple average of predictions), CM1 (average of predictions from the 'qualified' individual models), CM2 (weighted average predictions (WAPs) from 'qualified' individual models) and CM3 (best selection of predictions (compound-wise) from 'qualified' individual models). From the four consensus model obtained, CM0 was found to be the best.

Table 4.8: Detailed summary of the QSAR models and consensus models obtained for selectivity PET tracer compounds for adenosine A_{2A} selectivity.

| Dataset | Type of model | Training set statistics | | | | | Test set statistics | | | | | | | |
|---------|---------------------------|-------------------------|---------------------------------------|-----------|---------------------------|-----------------------|---|-------------------------------|-------------|----------------------------|------------------------|-------------|-------------|-------------|
| | | Model R ² | Model Q ² _(LOO) | MAE_train | $\overline{r_{m(LOO)}^2}$ | $\Delta r_{m(LOO)}^2$ | R ² _{pred} or Q ² F ₁ | Q ² F ₂ | CCC | $\overline{r_{m(test)}^2}$ | $\Delta r_{m(test)}^2$ | MAE (95%) | MAE | |
| | Individual Models (N1-N5) | IM1 | 0.92 | 0.87 | Good | 0.81 | 0.07 | 0.84 | 0.81 | - | 0.77 | 0.12 | 0.18 | Good |
| | | IM2 | 0.90 | 0.82 | Moderate | 0.75 | 0.06 | 0.84 | 0.82 | - | 0.77 | 0.04 | 0.24 | Good |
| | | IM3 | 0.88 | 0.83 | Moderate | 0.76 | 0.12 | 0.84 | 0.82 | - | 0.77 | 0.13 | 0.18 | Good |
| | | IM4 | 0.87 | 0.83 | Moderate | 0.72 | 0.13 | 0.85 | 0.83 | - | 0.78 | 0.07 | 0.22 | Good |
| | Consensus Models | CM0 | - | - | - | - | - | 0.88 | 0.86 | 0.93 | 0.82 | 0.10 | 0.16 | Good |
| | | CM1 | - | - | - | - | - | 0.86 | 0.84 | 0.92 | 0.82 | 0.11 | 0.18 | Good |
| | | CM2 | - | - | - | - | - | 0.85 | 0.83 | 0.92 | 0.82 | 0.11 | 0.18 | Good |
| | | CM3 | - | - | - | - | - | 0.83 | 0.81 | 0.90 | 0.80 | 0.10 | 0.19 | Good |

4.2.4. Applicability Domain (AD)

Applicability domain (AD) is an important tool for reliable application of QSAR models. It can be considered as a “theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable”. We have checked the AD of all the models using standardization approach to check whether any molecule in the test set lies outside the AD of a model. From, the domain of applicability analysis it was found that there were no test set compounds outside the AD and no compound in the training set came as outlier (see **Table 4.9**).

Table 4.9. Predicted A_{2A}R selectivity of all the four selectivity models along with the AD information.

| <i>Training Set</i> | | | | | | | | | |
|---------------------|-----------------------------------|---|---------|---|---------|---|---------|---|---------|
| Compd. No | Obs A _{2A} R selectivity | Selectivity Model 1 Pred A _{2A} R selectivity | AD Info | Selectivity Model 2 Pred A _{2A} R selectivity | AD Info | Selectivity Model 3 Pred A _{2A} R selectivity | AD Info | Selectivity Model 4 Pred A _{2A} R selectivity | AD Info |
| 1 | 2.779 | 2.851 | In | 3.003 | In | 2.837 | In | 3.108 | In |
| 2 | 2.808 | 2.914 | In | 3.077 | In | 2.837 | In | 3.059 | In |
| 4 | 4.217 | 4.111 | In | 4.204 | In | 4.170 | In | 4.163 | In |
| 6 | 0.894 | 0.977 | In | 1.068 | In | 0.960 | In | 1.041 | In |
| 7 | 1.231 | 1.145 | In | 1.263 | In | 1.273 | In | 1.206 | In |
| 10 | 2.715 | 2.809 | In | 2.955 | In | 2.837 | In | 2.800 | In |
| 12 | 2.000 | 2.494 | In | 2.118 | In | 2.525 | In | 2.399 | In |
| 13 | 2.636 | 2.494 | In | 2.588 | In | 2.525 | In | 2.382 | In |
| 15 | 2.398 | 2.452 | In | 2.539 | In | 2.525 | In | 2.374 | In |
| 16 | 2.398 | 2.515 | In | 2.612 | In | 2.525 | In | 2.362 | In |
| 17 | 2.903 | 2.515 | In | 2.612 | In | 2.525 | In | 2.355 | In |
| 18 | 2.398 | 2.515 | In | 2.247 | In | 2.525 | In | 2.368 | In |
| 19 | 2.653 | 2.522 | In | 2.194 | In | 2.424 | In | 2.246 | In |
| 20 | 1.839 | 1.719 | In | 1.683 | In | 1.900 | In | 1.987 | In |
| 23 | 3.176 | 3.189 | In | 2.857 | In | 3.322 | In | 2.726 | In |
| 25 | 3.240 | 3.357 | In | 3.052 | In | 3.322 | In | 3.107 | In |
| 27 | 2.675 | 2.884 | In | 2.907 | In | 2.879 | In | 2.906 | In |
| 28 | 2.607 | 2.683 | In | 2.808 | In | 2.525 | In | 2.816 | In |
| 30 | 3.199 | 2.746 | In | 2.881 | In | 2.525 | In | 3.104 | In |
| 31 | 2.841 | 2.578 | In | 2.686 | In | 2.525 | In | 2.815 | In |

| 35 | 1.818 | 1.953 | In | 2.072 | In | 1.939 | In | 2.104 | In |
|-----------------|----------------------------------|--|---------|--|---------|--|---------|--|---------|
| <i>Test Set</i> | | | | | | | | | |
| Compd. No | Obs A ₂ R selectivity | Selectivity Model 1 Pred A ₂ R selectivity | AD Info | Selectivity Model 2 Pred A ₂ R selectivity | AD Info | Selectivity Model 3 Pred A ₂ R selectivity | AD Info | Selectivity Model 4 Pred A ₂ R selectivity | AD Info |
| 3 | 3.0249 | 2.809 | In | 2.955 | In | 2.837 | In | 3.206 | In |
| 5 | 1.0763 | 0.935 | In | 1.019 | In | 0.960 | In | 0.930 | In |
| 8 | 3.2292 | 2.704 | In | 2.832 | In | 2.837 | In | 2.771 | In |
| 9 | 2.8254 | 2.872 | In | 3.028 | In | 2.837 | In | 2.881 | In |
| 11 | 2.3118 | 2.620 | In | 2.735 | In | 2.525 | In | 2.682 | In |
| 21 | 2.9513 | 3.063 | In | 2.710 | In | 3.322 | In | 2.782 | In |
| 22 | 3.2041 | 3.126 | In | 2.783 | In | 3.322 | In | 2.758 | In |
| 24 | 3.1271 | 3.294 | In | 2.979 | In | 3.322 | In | 3.154 | In |
| 26 | 3.0637 | 3.399 | In | 3.101 | In | 3.322 | In | 3.164 | In |
| 29 | 2.9845 | 2.641 | In | 2.759 | In | 2.525 | In | 2.834 | In |

4.2.5. Comparison with a previously published model

A direct comparison between the current and a previously published model (Roy et al., 2018) is infeasible due to the differences in the composition of training and test sets. However, the current model can be considered more advantageous since it has been developed using simple and easily interpretable two-dimensional descriptors which does not require any conformational analysis or energy minimization before their calculation.

4.3. Study 3: QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting Dopamine receptor

4.3.1. Modeling binding affinity of PET tracers towards dopamine (D2) receptor

The final PLS model of 3 latent variables (LVs) consisted of five descriptors that explains the binding properties of the PET radioligands towards dopamine receptor. The final model is given below:

$$pKi = 4.512 - 0.184 \times SaaCH - 1.554 \times B08[C - S] + 0.060 \times SsF - 2.350 \times B10[N - F] + 1.425 \times B10[C - O]$$

$$n_{training} = 27, R^2 = 0.731, R_{adj}^2 = 0.696, Q^2 = 0.623, \overline{r_{m(LOO)}^2} = 0.507, \Delta r_{m(LOO)}^2 = 0.159, MAE(train) = 0.528, SD(train) = 0.550, PRESS = 15.392$$

$$n_{test} = 7, Q_{F1}^2 = 0.687, Q_{F2}^2 = 0.664, \overline{r_{m(test)}^2} = 0.742, \Delta r_{m(test)}^2 = 0.116, MAE(test) = 0.505, SD(test) = 0.280, CCC(Test) = 0.812$$

4.3.2. Mechanistic interpretation

The variable importance plot (VIP) (**Figure 4.27**) gives an idea about the influence of the individual descriptors on the model and thereby on the binding affinity (Akarachantachote et al., 2014). The order of importance of the descriptors was found as follows: SaaCH, B10[N-F], B10[C-O], SsF and B08[C-S]. The VIP gives an understanding that descriptors SaaCH and B10[N-F] are highly influential due to their VIP scores being more than 1. The regression coefficient plot (not shown) provides a basic understanding about the contribution of the individual descriptor on the model (Wold et al., 2001). It is seen that the descriptors SaaCH, B08[C-S] and B10[N-F] negatively contribute to the model while the descriptors SsF and B10[C-O] positively contribute to the model. The details of the descriptors and their contributions are given in **Table 4.10** and also explained below in detail. The observed vs predicted scatter plot is shown in **Figure 4.28**.

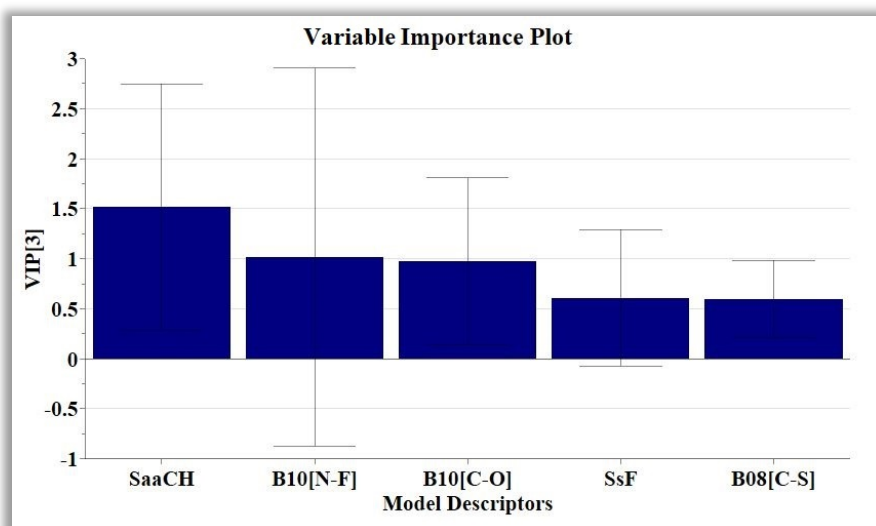


Figure 4.27. Variable importance plot of the PLS model

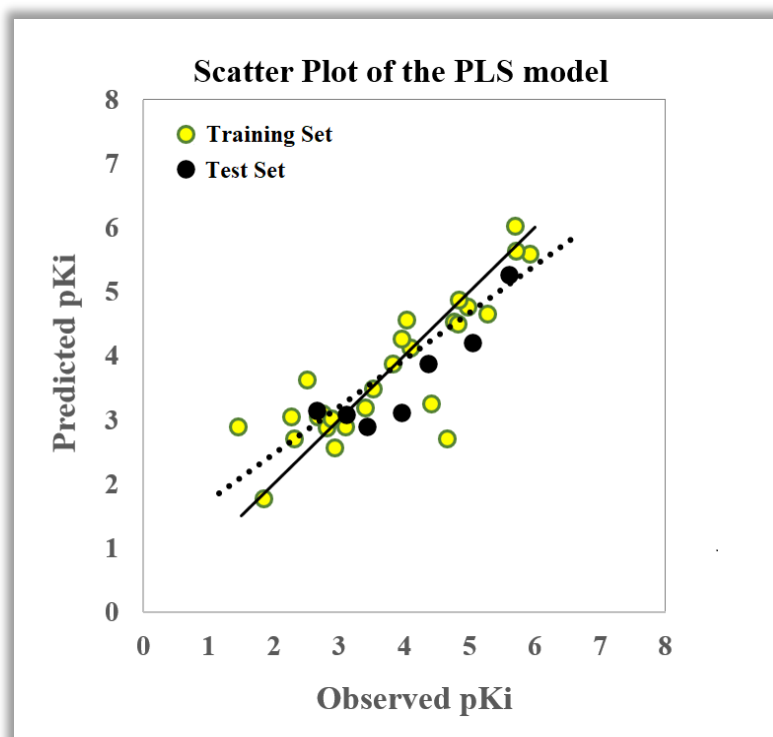


Figure 4.28. Observed vs predicted pKi plot.

Table 4.10. Descriptor meaning and their contribution

| Serial No | Descriptor | Descriptor type | Contribution | Discussion |
|-----------|------------|-------------------|--------------|---|
| 1 | SaaCH | Atom type E state | -ve | Sum of the atom-type E-state values for aromatic -CH group. |
| 2 | B08[C-S] | 2D atom pairs | -ve | Presence or absence of carbon and sulphur at the topological distance 8 |
| 3 | SsF | Atom type E state | +ve | Sum of the atom type E-state values for -F fragments. |
| 4 | B10[N-F] | 2D atom pairs | -ve | Presence or absence of nitrogen and fluorine at the topological distance 10 |
| 5 | B10[C-O] | 2D atom pairs | +ve | Presence or absence of carbon and oxygen at the topological distance 10 |

The E-state indices descriptor SaaCH gives idea on the sum of the atom-type E-state values for aromatic -CH groups. From the regression coefficient of the descriptor, it can be inferred that aromaticity hinders the binding of the PET compounds to the D2 receptor as in compounds **8** (SaaCH = 18.392) (**Figure 4.29**), **10** (SaaCH = 16.63) and **11** (SaaCH= 14.214). These compounds are aromatic and have high SaaCH values, and they have lower binding affinity values (pKi= 2.931, 1.460 and 1.839). Further, in compounds like **29** and **32**, aromaticity is less as compared to the previously mentioned compounds, thus having lower values for the descriptor (SaaCH= 3.583 and 1.640 respectively). These compounds have better binding affinity (compound **29** (pKi= 5.700) and compound **32** (pKi= 5.721)) towards dopamine receptor.

The next important descriptor is B10[N-F] (2D atom pair type) and the negative contribution implies that the presence of nitrogen and fluorine at the topological distance 10 will hinder the binding affinity seen in compounds **11** (B10[N-F]= 1; pKi= 1.838) (**Figure 4.29**) and **33** (B10[N-F]= 1; pKi= 2.886). Further, the absence of this fragment will increase the binding affinity observed in compounds **29** (B10[N-F] = 0; pKi= 5.700) and **32** (B10[N-F] = 0; pKi= 5.721). The effect of the electronegativity of fluorine atom on nitrogen is a determining factor for the good binding which is latter explained while studying the descriptor SsF. The closeness between nitrogen and fluorine atom explains how the binding will occur.

B10[C-O] is another 2D atom pair descriptor representing the presence or absence of C-O fragment at the topological distance 10. The descriptor positively affects the binding affinity of the PET tracers towards dopamine receptor as seen in compounds **18** (B10[C-O] = 1; pKi= 5.921) (**Figure 4.29**), **29** (B10[C-O] = 1; pKi= 5.721) and **32** (B10[C-O] = 1; pKi= 5.700). The presence of this kind of fragment affects the electronegativity of the compounds essential for binding. The absence of this fragment on the other hand decreases the dopamine binding affinity observed in compounds like **1** (pKi= 2.321) and **5** (pKi= 2.262).

The E-state values for the descriptor SsF depends on the number of fluorine atoms present in a PET tracer molecule. From the regression coefficient, it can be understood that with increasing fluorine atoms the binding affinity also increases as observed in **18** (SsF= 14.107; pKi= 5.921), **32** (SsF= 12.490; pKi= 5.698) (**Figure 4.29**) and **31**(SsF= 13.108; pKi= 4.833). The electronegative fluorine atom is presumed to decrease electron charge density on nitrogen groups. This reduces nitrogen basicity and its prospect to get protonated at physiological pH which a basic requirement for good binding to dopamine receptors [34].

The least important descriptor is B08[C-S], which is also a 2D atom pair descriptor and gives an idea of the presence or absence of C-S fragment at a topological distance 8. The negative contribution suggests that the presence of this fragment will result in a decreased binding affinity towards the dopamine receptor which is observed in compounds **21** (pKi= 2.807) and **20** (pKi= 3.107) (**Figure 4.29**). Alternatively, compounds like **18** (pKi= 5.921), **29** (pKi= 5.721) and **32** (pKi= 5.698) have no such fragment thus having higher binding affinity.

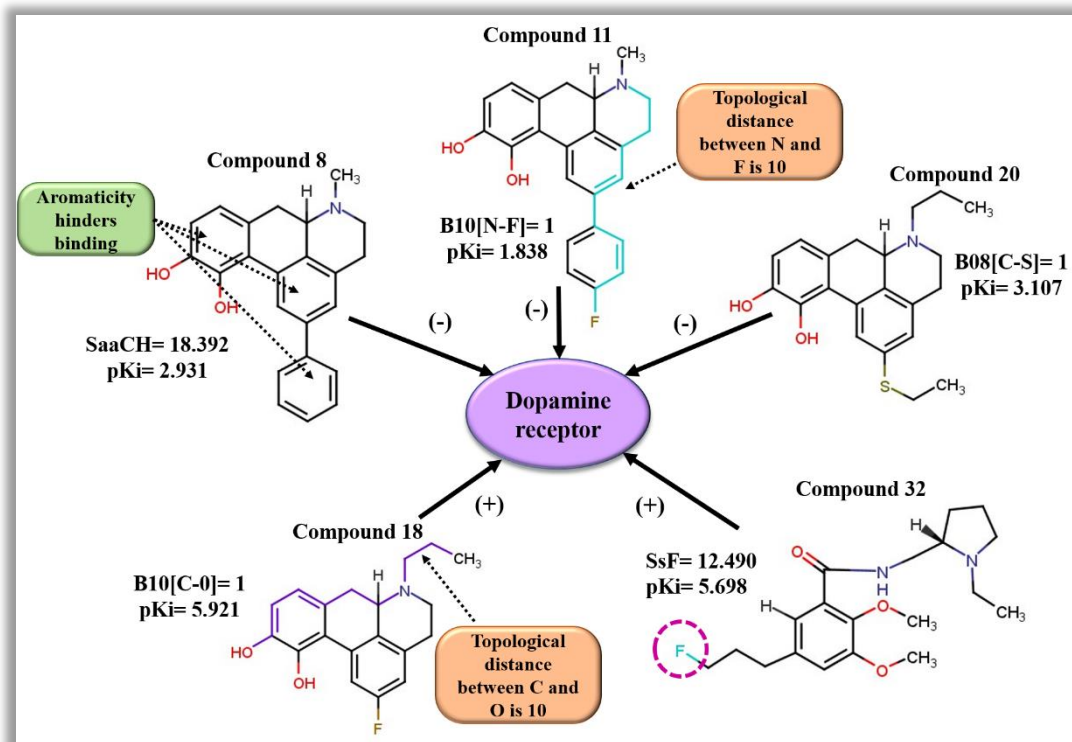


Figure 4.29. Descriptors appearing in the PLS model and their contribution

From the descriptors and their contributions, we can draw an inference that the oxygen for B10[C-O] and fluorine for SsF impart an electronegative character to the PET ligands which plays an essential role for the good dopamine (D2) binding.

4.3.3. Plot Interpretation

4.3.3.1. Loading plot- This plot gives a relationship between the X- variables (i.e., the descriptors) and Y-variable (i.e., response) (De et al., 2018a). In **Figure 4.30**, five X-variables and one Y variable are shown. Generally, the plot is developed with the first and second components. A loading plot provides an insight about how much a variable contributes to a model and which variable provides the maximum footprint. For interpretation, the distance from the origin is taken under consideration. Descriptors which are similar in nature and providing similar contribution are correlated and grouped together. Descriptors which are situated far away from the plot origin are supposed to have greater impact on the Y-response. From the loading plot it is seen that descriptors SaaCH and B10[N-F] are far away from the plot origin supporting their higher influence also explained by the VIP. The positive or negative algebraic symbol is also taken under consideration in a PLS plot. Features explained by descriptors SsF and B10[C-O] are beneficial for binding because of their closeness to pKi in the plot. On the other hand, SaaCH, B10[N-F] and B08[C-S] are present in the negative side of the plot origin and are detrimental for good binding.

4.3.3.2. Score Plot- **Figure 4.31** shows the distribution of the compounds in the latent variable space as defined by the scores. We have plotted the scores of the first two components t1 and t2. The applicability domain of the model is designated by the ellipse, as defined by Hotelling's t2. Hotelling's t2 defines multivariate generalization of Student's t-test. The method offers a check for compounds adhering to multivariate normality (Jackson, 2005). Compounds which are situated near each other in

the plot have similar properties, whereas compounds which are far from each other have dissimilar properties with respect to their binding affinity towards dopamine receptor. As an example, we can take compounds 14, 15, 16 and 17 are clubbed together as a group on the plot space and can be considered to be with similar properties. On the other hand, compounds 18 and 12 are completely located on the opposite side of the origin and far from each other and they represent heterogeneity in their properties. Since there are no compounds out of the ellipse, we can conclude that there are no outliers according to this method.

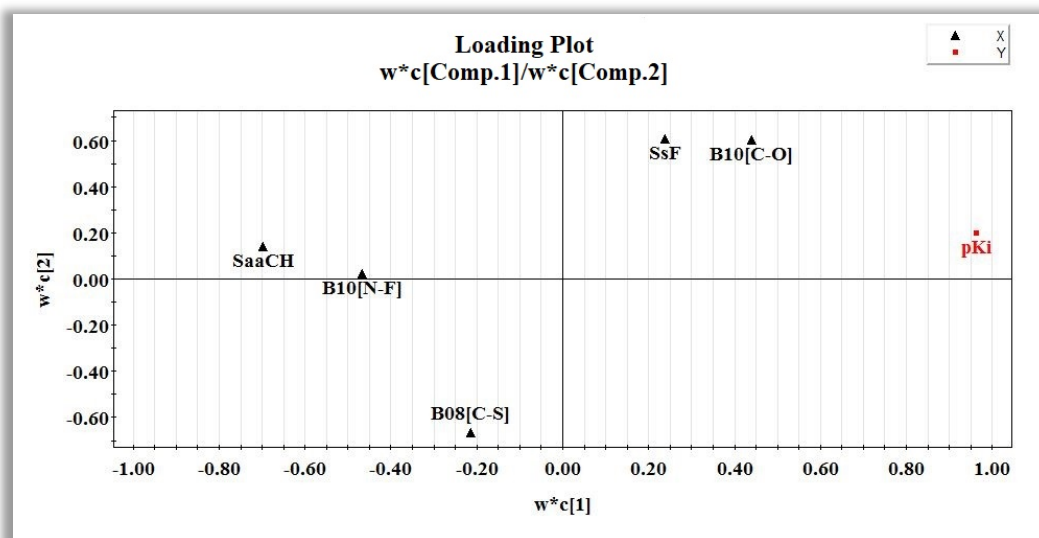


Figure 4.30. Loading plot of the PLS model

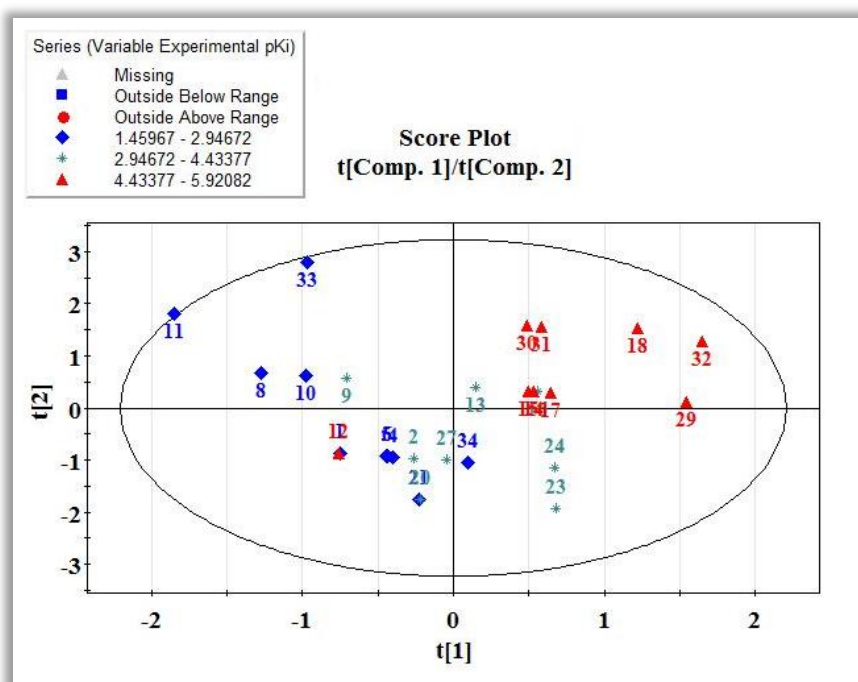


Figure 4.31. Score plot of the PLS model

4.3.3.3. Y-Randomization Plot- Model randomization gives a notion about the model significance and ensures that the model is not an outcome of a chance correlation (Topliss & Edwards, 1979). A randomized model is generated by the development of multiple models by shuffling or reordering different combinations of X or Y variables (here Y variable only) and based on the fit of the reordered model. In the present study we have used 100 permutations which can be changed according to the choice of the user. A randomized model should have very poor statistics. The R^2 and Q^2 values for the random models (Y-axis) are plotted against correlation coefficient between the original Y values and the permuted Y values (X axis); the R^2_y intercept should not exceed 0.3 and the Q^2_y intercept should not exceed 0.05. **Figure 4.32** shows the correlation between original Y-vector and permuted Y-vector versus cumulative R^2_y , cumulative Q^2_y plot where R^2_y intercept = 0.09 and Q^2_y intercept = -0.393 proving the model is robust and non-random.

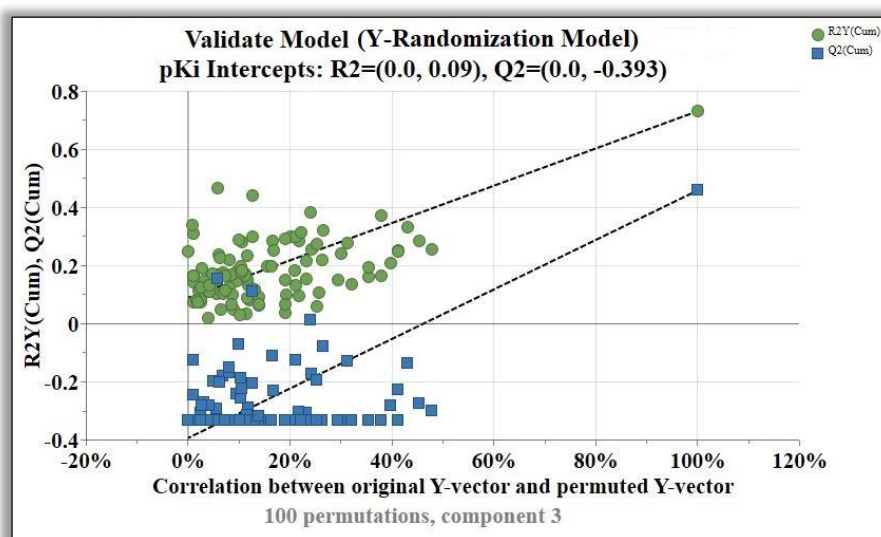


Figure 4.32. Y-Randomization plot of the PLS model

4.3.3.4. Applicability Domain (AD)- The prediction reliability of a particular model is dependent on its applicability domain (AD) assessment. Applicability domain (AD) “represents a chemical space from which a model is derived and where a prediction is considered to be reliable” (Gadaleta et al., 2016). The AD evaluation was done using the DModX (distance to model) in the X-space using SIMCA 16.0.2 software available at <https://landing.umetrics.com/downloads-simca>. The AD plots are given in **Figures 4.33** and **4.34** and for training and test sets respectively, and it is found that there are no outliers in case of training set, and none of the compounds are outside AD in case of the test set at 99% confidence level (D-crit= 0.009999, M-Dcrit[3]= 3.213).

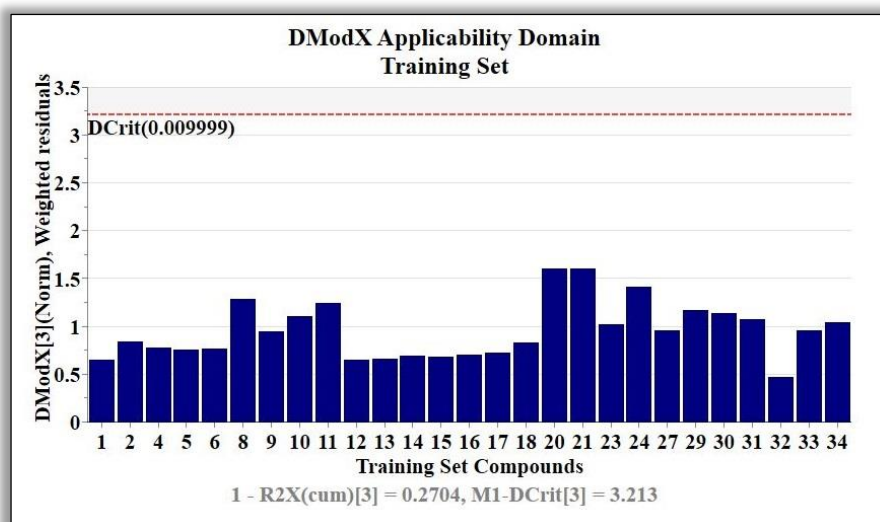


Figure 4.33. DModX Applicability Domain of the training set.

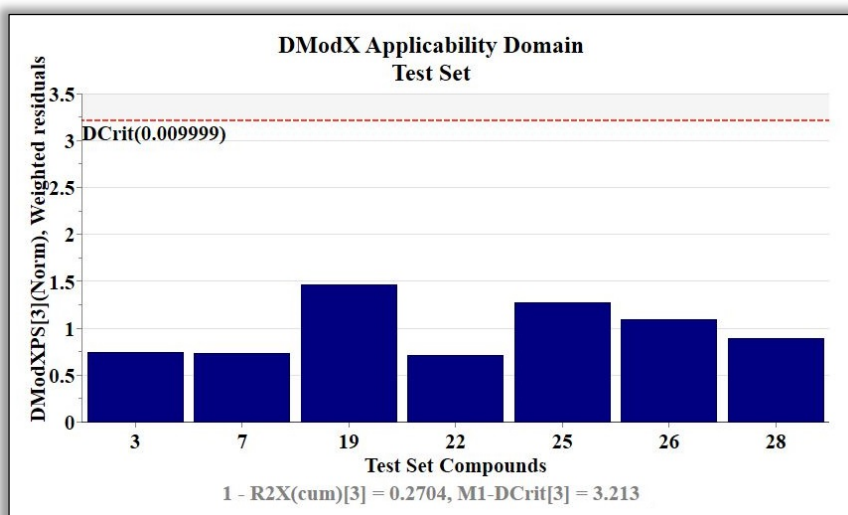


Figure 4.34. DModX Applicability Domain of the test set.

4.4. Study 4: Computational modeling of PET imaging agents against vesicular acetylcholine transporter (VACHT) protein binding affinity: Application of 2D-QSAR modeling and molecular docking techniques

The present work demonstrates the contribution of different structural attributes of PET imaging agents required for binding to and quantifying the presence of vesicular acetylcholine transporter. The main work is focused on the development of a simple 2D-QSAR model to obtain the major structural features responsible for binding. These features were further validated using structural similarity-based read across study as well as molecular docking techniques.

4.4.1. QSAR modeling of binding affinity of PET imaging agents towards VACHT

The dataset procured for this study consisted of 19 compounds. A three-descriptor partial least squares (PLS) model with two latent variables (LVs) was developed which could explain 71.77% of the variance. The leave-one-out cross-validated determination coefficient (i.e., $Q_{LOO}^2 = 0.523$) is above the critical threshold value fulfilling the statistical reliability of the model. The observed versus predicted pKi scatter plot is shown in **Figure 4.35**.

$$pKi = 2.018 - 0.831 \times B06[N - O] + 0.757 \times F08[C - N] - 0.812 \times F09[N - F]$$

$$N = 19, R^2 = 0.718, Q_{(LOO)}^2 = 0.523, \overline{r_{m(LOO)}^2} = 0.439, \Delta r_{m(LOO)}^2 = 0.027, MAE = 0.335, SD = 0.273$$

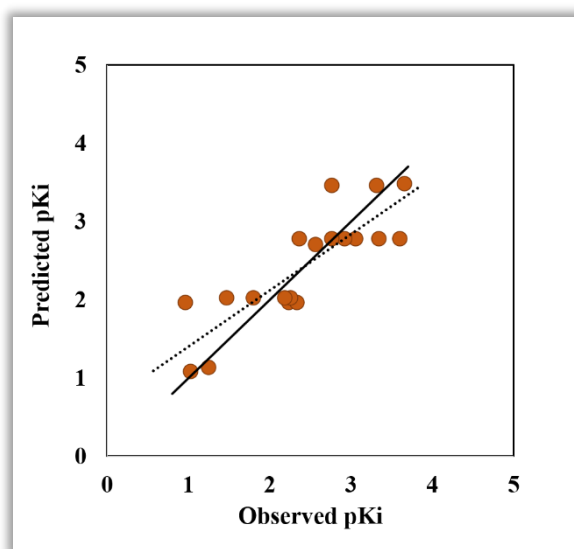


Figure 4.35: Observed versus predicted scatter plot of the PLS model

The descriptors appearing in the final PLS model are all 2D atom pair descriptors suggesting the importance of the presence of a particular atom pair in the PET tracer molecule. The variable importance plot (Akarachantachote et al., 2014) given in Figure 2 shows the significance level of each descriptor toward VACHT binding affinity. Descriptor **F09[N-F]** was the most significant descriptor with VIP Score > 1 (VIP = 1.289) followed by **F08[C-N]** (VIP = 1.043) and **B06[N-O]** (VIP = 0.502). **F09[N-F]** which contributes negatively to the binding affinity, is the frequency of the N-F fragment at a topological distance 9. Compounds like **10** and **11** (**Figure 4.36**) have nitrogen and fluorine at a topological distance 9, thereby decreasing the binding affinity towards VACHT, whereas in compounds like **21** and **23**, the N-F fragment at 9 distance is absent and the pKi values are high.

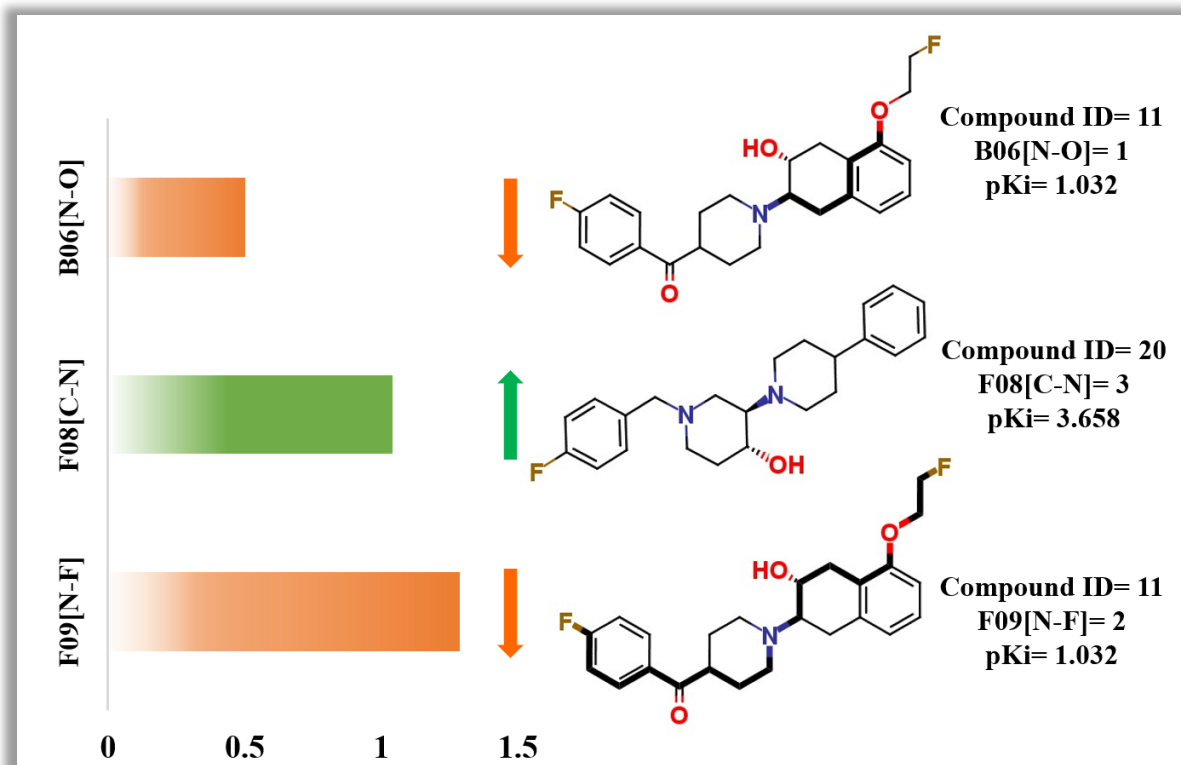


Figure 4.36: Variable importance plot and significance of the descriptors appearing in the PLS model.

The next important 2D atom pair descriptor is F08[C-N] which denotes the frequency of C-N fragment at a topological distance 8. The positive regression coefficient indicates that with an increase in the frequency of C-N at 8 distances, the binding affinity will increase as observed in compounds like **20** (Figure 4.36), **23**, and **25**. These compounds have three such fragments and have high pKi values of 3.658, 3.319, and 2.700 respectively.

The least important among all the descriptors is B06[N-O] which implies the presence or absence of an N-O fragment at a topological distance 6. The negative contribution indicates that the presence of such a fragment will decrease the VAcHT binding of the PET imaging agents as seen in compounds like **10** and **11** (Figure 4.36). These compounds have a very low binding affinity towards (1.251 and 1.032 respectively) VAcHT receptor.

The significance and validity of the developed model were further analyzed using some important PLS plots, namely, the loading plot, randomization plot, and applicability domain which are described below.

A **loading plot** (Figure 4.37) explains the relationship between the independent variables or descriptors (X-variables) with the dependent variable or pKi values (Y-variable). The influence of the descriptors on the developed model can be recognized from the loading plot. Descriptors that are far from the plot origin (like F08[C-N] and F09[N-F]) contribute significantly more toward the binding affinity. Descriptors with different meanings appear distantly from each other in the loading plot.

Model randomization confirms that the model is not the outcome of any chance correlation (Topliss & Edwards, 1979). The **randomization plot** determines the statistical significance of the model. Multiple models are generated during a randomization plot development by shuffling different

combinations of either X-variables (X-randomization) or Y-response (Y-randomization). Y-randomization was performed in the present study with 100 permutations for each model for random model generation. For a non-random model R_y^2 intercept should not exceed 0.3 and Q_y^2 intercept should not exceed 0.05. The randomization plot given in **Figure 4.38** shows that the developed model is non-random and robust and is suitable for prediction.

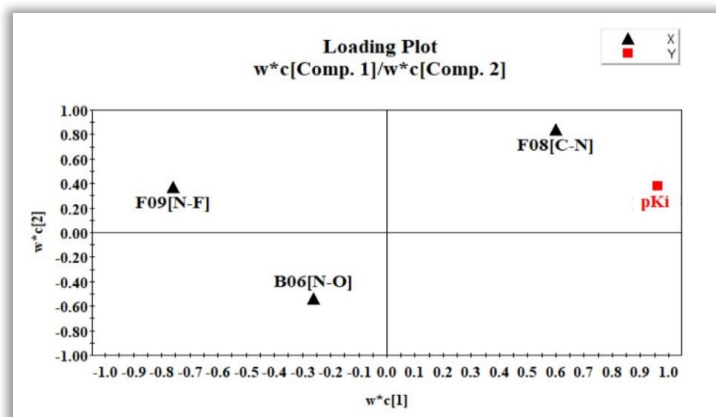


Figure 4.37: Loading plot of the PLS model

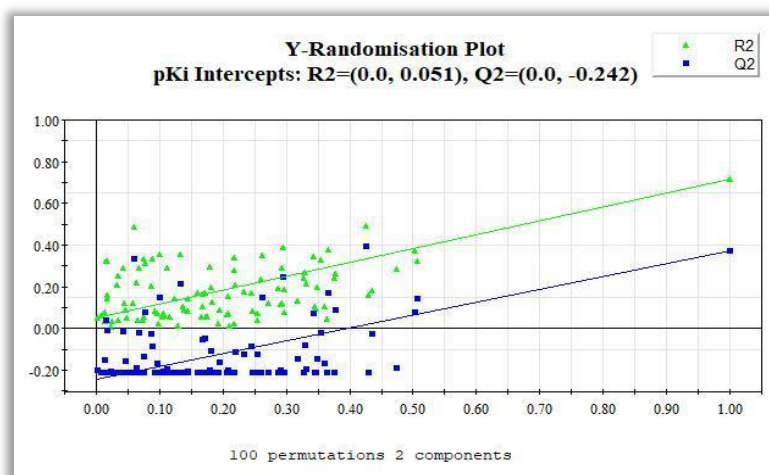


Figure 4.38: Y-randomization of developed PLS model

According to OECD guideline 3, a developed QSAR model should possess a defined chemical domain of applicability. AD can be interpreted as a chemical space defined by the structural information or molecular properties of the chemicals used in the model development (Gadaleta et al., 2016). Compounds present within this chemical space can only be properly predicted. In this study, the DModX (distance to model in X-space) method of AD determination at a 99% confidence interval (D-crit= 0.009999) was applied using SIMCA 16.0.2 software (<https://landing.umetrics.com/downloads-simca>). AD plot (**Figure 4.39**) shows none of the compounds was an outlier.

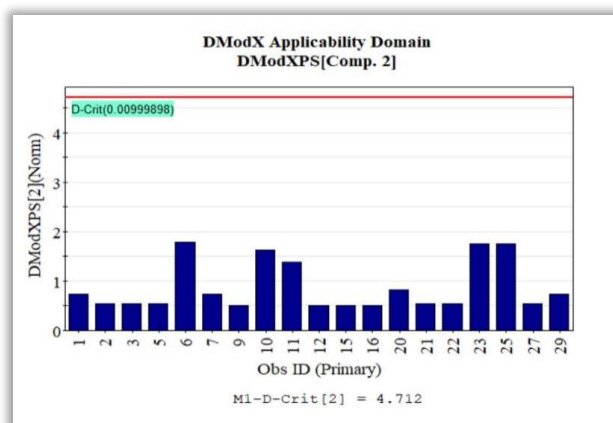


Figure 4.39. DModX AD plot of the PLS model

4.4.2. Read-Across based prediction

To explore the predictivity of the selected features used for QSAR modeling, a similarity-based read-across prediction was performed by using a group of five compounds (compound ID: 3, 11, 12, 21, and 27) as the test set (Chatterjee et al., 2022a). Three types of similarity were measured: the Euclidean Distance-based, the Gaussian Kernel Similarity-based, and the Laplacean Kernel Similarity based predictions using Read-Across-v4.1 (<https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>) tool and after hyperparameter optimization using Auto_RA_Optimizer-v1.0 tool (<https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>) it was found that the external validation results obtained from quantitative Read-Across algorithm using Gaussian Kernel Similarity-based functions ($Q_{F1}^2 = 0.763$, $Q_{F2}^2 = 0.763$, $RMSE = 0.414$, $MAE = 0.331$) was better compared to the results obtained with the other two read-across approaches (Table 4.11).

Table 4.11. Comparison between three types of read-across predictions

| Method | N_{train} | R^2 | $Q_{(LOO)}^2$ | MAE | N_{test} | Q_{F1}^2 | Q_{F2}^2 | MAE |
|--------------------------------|-------------|-------|---------------|-------|------------|--------------|--------------|--------------|
| QSAR | 19 | 0.718 | 0.523 | 0.335 | - | - | - | - |
| Euclidean distance | - | - | - | - | - | 0.189 | 0.189 | 0.596 |
| Read-Across Gaussian Kernel | 14 | - | - | - | 5 | 0.763 | 0.763 | 0.331 |
| Laplacian Kernel | - | - | - | - | - | 0.719 | 0.719 | 0.380 |

Note: **Bold** values indicate best prediction

4.4.3. Molecular Docking

Molecular docking must include a reasonably accurate model of energy and should be able to deal with the combinatorial complexity experienced by the molecular flexibility of the docking partners. In the present research, molecular docking studies were performed to understand the individual molecular interactions and orientation of the imaging agents occurring at the binding zone of the VAcHT receptor (Figure 4.40). Initially, the standard compound, i.e., vesamicol was docked at the

binding site to understand its nature of interactions. Further, both high and low-active compounds were also used for the docking study. In the case of vesamicol (compound 9), which has a moderate binding affinity (pKi= 2.261), the interaction forces include hydrogen bond interactions (both conventional and carbon-hydrogen bond interactions) and π -anion interactions. The amino acid residues engaged in vesamicol binding are Asp A:202, Asp A:483, and Ser A:480. Comparing vesamicol-VACHT binding interactions with highly active compounds like compound ID **20** (pKi= 3.658), **21** (pKi= 3.602), and **22** (pKi= 3.347), it was observed that similar interactions were also involved in their binding. However, it was found that these highly active compounds were docked with more interactions at their binding site with far better binding (**Table 4.12**). For compound **20**, halogen (fluorine) interactions, attractive charge, π -cation, and π -alkyl interactions were formed along with hydrogen bond interactions. In the case of compound 21, additional interactions include attractive charge, π -anion, and π -cation interactions. Similarly, in the case of compound **22**, attractive charges, alkyl, and π -alkyl interactions were formed along with conventional hydrogen bond and carbon-hydrogen bond interactions. The formation of attractive charge interaction of **Asp A:483** amino acid with the nitrogen of piperidine moiety of all three high active compounds was a noteworthy finding inferring the importance of the fragment in VACHT binding.

In the case of lower active compounds like compound **10** (pKi= 1.032) and compound **29** (pKi= 0.967), the number of molecular interactions was much less than the higher active ones (**Table 4.12**). Conventional hydrogen bond and carbon-hydrogen bond interactions were prevalent, with additional halogen and π -alkyl in the case of compound 29.

Table 4.12: The interacting residues and different types of binding interaction occurring between the PET imaging agents and VACHT.

| Compound | Category | pKi | (-)Docking interaction energy (kcal/mol) | Binding amino acids | Types of interactions |
|---------------|---------------|-------|--|---|---|
| 9 (Vesamicol) | Standard | 2.261 | 27.57 | Asp A:202, Asp A:483, Ser A:480 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, and π -anion interaction |
| 20 | Highly active | 3.658 | 37.00 | Ser A:415, Asp A:410, Tyr A:417, Arg A:477, Arg A:479, Asp A:483, Arg A:482 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, attractive charge, halogen (fluorine) interaction, π -cation, and π -alkyl interaction |
| 21 | | 3.602 | 38.27 | Arg A:477, Asp A:483, Ser A:480, Arg A:482, Asp A:202 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, attractive charge, π -cation, and π -anion interactions |
| 22 | | 3.347 | 43.95 | Pro A:205, Tyr A:494, Arg A:482, Asp A:483, Ser A:480, Pro A:490 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, attractive charge, alkyl, and π -alkyl interactions |
| 10 | Least active | 1.032 | 33.87 | Asp A:202, Arg A:482, Pro A:490 | Conventional hydrogen bond interactions and carbon-hydrogen interactions |
| 29 | | 0.967 | 31.41 | Arg A:479, Arg A:477, Tyr A:417 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, halogen (fluorine), and π -alkyl interactions |

Relationship with QSAR features

From QSAR modeling it was found that F08[C-N] is the only positively correlated descriptor. Therefore, the presence of nitrogen in the PET imaging agent is very essential for good VACHT binding. In the case of highly active compounds (compounds **20**, **21**, and **22**) used for molecular docking, it was found that attractive charge interaction was prevalent in all three compounds which occurred between **Asp A:483** amino acid with the nitrogen of piperidine moiety of the PET tracer. These two observations correlate with each other and thus can be inferred nitrogen (as piperidine moiety) is essential for good VACHT binding.

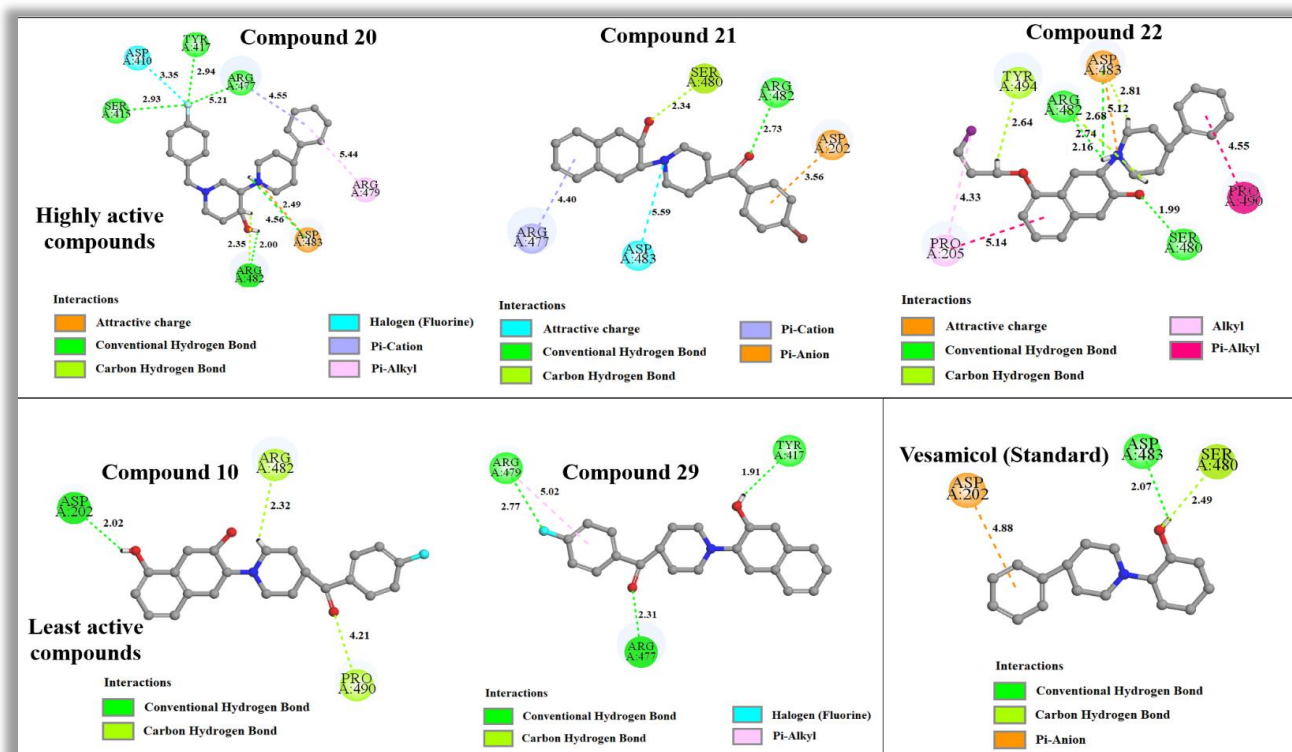


Figure 4.40. Molecular docking interactions of highly active, least active and standard compounds against VACHT binding.

4.5. Study 5: Exploration of nitroimidazoles as radiosensitizers: Application of multi-layered feature selection approach in QSAR modeling

Statistically significant 2D-QSAR models using Dragon and simplex (SiRMS) descriptors explaining the chemical features required for good radiosensitization are presented in the following section. The observed versus predicted $pC_{1.6}$ values are plotted for both the models is shown in **Figure 4.41**.

4.5.1. 2D-QSAR model using Dragon descriptors

$$pC_{1.6} = 3.612 + 0.613C - 035 - 0.285nCp - 1.129C - 043 + 0.068H - 052 - 1.630C - 042 + 0.295nRNHR$$

$$N_{train} = 63, R^2 = 0.773, R_{adj}^2 = 0.757, Q_{(LOO)}^2 = 0.746, \overline{r_m^2(Train)} = 0.647, \Delta r_m^2(Train) = 0.173, MAE(Train) = 0.246, SD(Train) = 0.195, RMSEC = 0.30, Quality = Good$$

$$N_{test} = 21, Q_{F1}^2 = 0.752, Q_{F2}^2 = 0.724, \overline{r_m^2(Test)} = 0.608, \Delta r_m^2(Test) = 0.216, CCC (Test): 0.831, MAE(Test) = 0.240, SD(Test) = 0.204, RMSEP = 0.31, Quality = Moderate$$

Model 1

The PLS model with 4 latent variables (LVs) could predict 74.6% variance of the training set and 75.2% of the test set. Important internal and external metrics used to determine the quality of a QSAR model are listed in equation 1. Mechanistic interpretation of the six descriptors obtained in the model would give us an insight about the structural features of the nitroimidazoles which are likely to influence their radiosensitization effectiveness. The obtained descriptors are C-035, nCp, C-043, H-052, C-042 and nRNHR. The model contains four atom centred fragments **C-035** (R--CX..X; positive contribution), **C-043** (X--CR..X, negative contribution), **H-052** (hydrogen (H^e) attached to sp³ carbon (C⁰) with one X attached to next carbon, 'e' represents the formal oxidation number; positive contribution) and **C-042** (X--CH..X; negative contribution). These descriptors are further explained with molecular structures from the dataset in **Figure 4.42**. The other two descriptor belonging to functional group counts are **nCp** (number of terminal primary C(sp³); negative contribution) and **nRNHR** (number of secondary amines (aliphatic); positive contribution). The descriptors obtained in the model gives us an idea regarding the vital features essential for better radiosensitization which includes the position of nitro group in the imidazole moiety. Atom centred fragment-based descriptors like C-042 and C-043 could explain that presence of nitro group at position 4 and position 5 would decrease the $pC_{1.6}$.

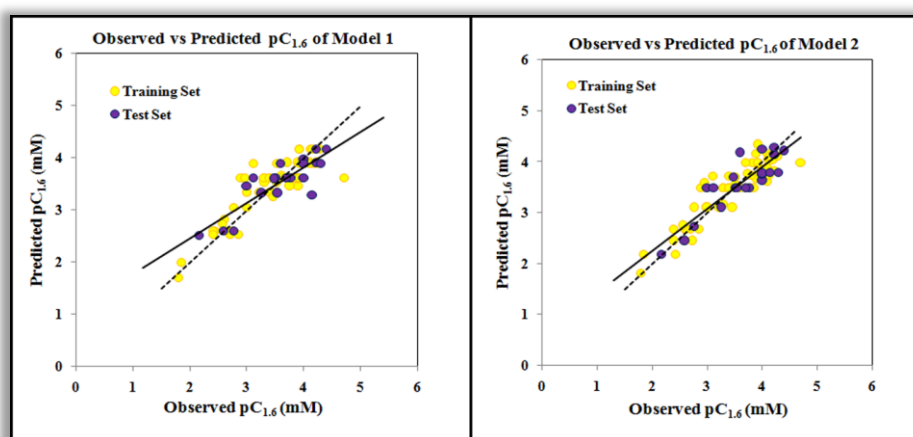


Figure 4.41. Scatter plots for observed vs predicted $pC_{1.6}$ values for Model 1 and Model 2.

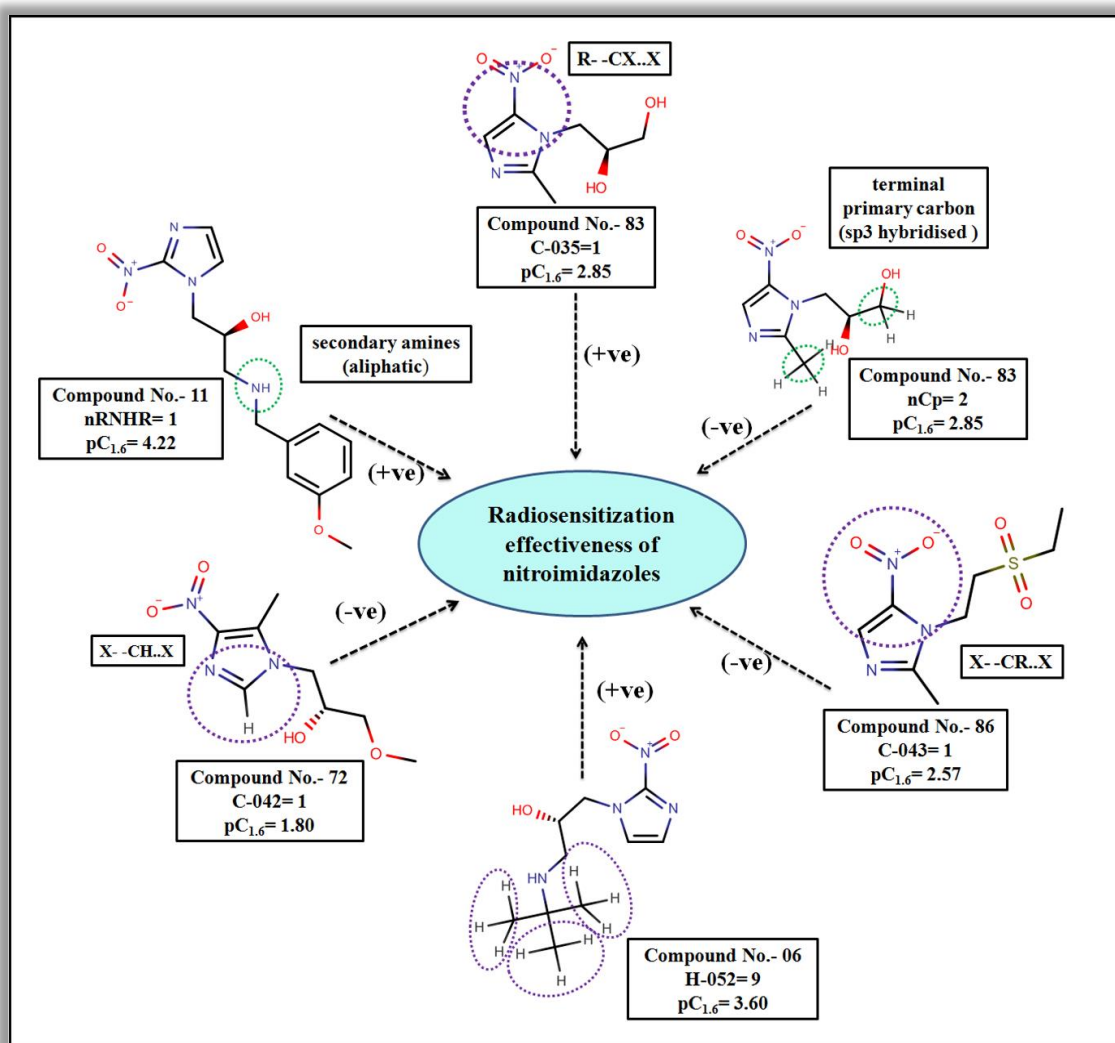


Figure 4.42. Descriptor features obtained from Dragon controlling the radiosensitization effectiveness of nitroimidazoles.

The variable importance plot (VIP) analysis gives us a premonition that C-042 and C-035 are the most important descriptors ($VIP > 1$) and contributing mostly towards the radiation enhancement of the compounds. The loading plot gives the relationship between the Y-variable ($pC_{1,6}$) and the X-variables (descriptors). For interpretation of the loading, the distance from the plot origin is considered, where similar types of descriptors with similar properties are located together. The variables which are far away from the plot origin are considered to have stronger impact on the model. This statement is verified by descriptors C-042 and C-035 which are proved to have higher impact from the VIP values also. The closeness of any descriptor to the Y-variable signifies its higher influence on the response. The VIP and loading plot are shown in **Figure 4.43**.

The 2D-QSAR model with Dragon descriptors gives an insight about the importance of the position of nitro group in the nitroimidazole compounds. Also, it is found that the presence of secondary aliphatic amine has significant importance on radiosensitization.

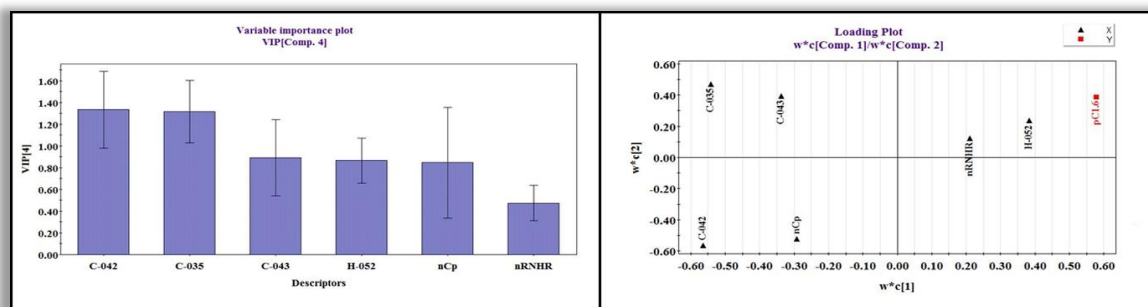


Figure 4.43. VIP and loading plot of Model 1.

4.5.2. 2D-QSAR model using SiRMS descriptors

We have further tried to improve the quality of the model by the use of SiRMS descriptors. The obtained 2D-QSAR model using SiRMS descriptors for radiosensitization effectiveness of nitroimidazoles was highly robust in terms of the statistical parameters as the values of quality metrics were above the recommended threshold as currently practiced (Ojha et al., 2011).

$$pC_{1.6} = 1.381 + 0.802Fr3(elm)/C_N_N/1_2s,1_3a/ + 0.494S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6 + 0.004S_A(chg)/B_C_C_C/1_4s,3_4s/4 - 0.377Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/ + 0.269Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/$$

$$N_{train} = 63, R^2 = 0.82, R_{adj}^2 = 0.81, Q_{(LOO)}^2 = 0.79, \overline{r_{m(loo)}^2} = 0.70, \Delta r_{m(loo)}^2 = 0.14, MAE_{train} = 0.22, SD_{train} = 0.18, RMSEC = 0.26, Quality_{(Train)} = Moderate$$

$$N_{test} = 21, Q_{F1}^2 (or R_{pred}^2) = 0.80, Q_{F2}^2 = 0.77, \overline{r_{m(Test)}^2} = 0.70, \Delta r_{m(Test)}^2 = 0.05, CCC(Test) = 0.88, MAE_{test} = 0.23, SD_{test} = 0.16, RMSEP = 0.28, Quality_{(Test)} = Moderate$$

Model 2

The PLS equation with 3 LVs is able to predict 79% variance of the training set (Q^2) and 80% of the test set (R_{pred}^2). The various internal and external metric values obtained are given in equation 2. The observed and predicted radiosensitization effectiveness values of the nitroimidazoles are listed in Table S1 in the Supplementary Section.

From VIP (Figure 4.44) the descriptors from highest to lowest order of significance are as follows: Fr3(elm)/C_N_N/1_2s,1_3a/, S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, S_A(chg)/B_C_C_C/1_4s,3_4s/4, Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/ and Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/. The loading plot developed using first two components describe the relationship between the X-variables and Y-variable is shown in Figure 4.45.

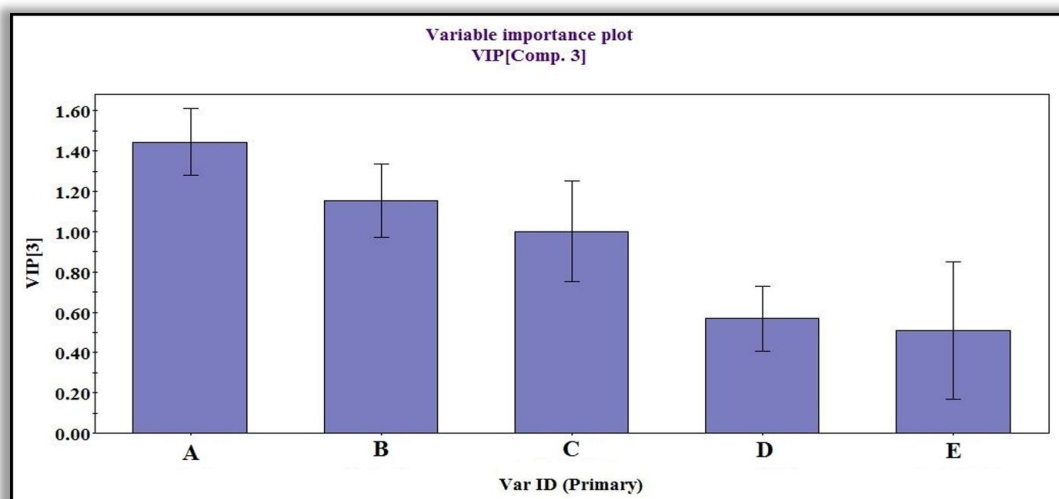


Figure 4.44: Variable importance plot of SiRMS model. (A- Fr3(elm)/C_N_N/1_2s,1_3a/, B- S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, C- S_A(chg)/B_C_C_C/1_4s,3_4s/4, D- Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/, E- Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/)

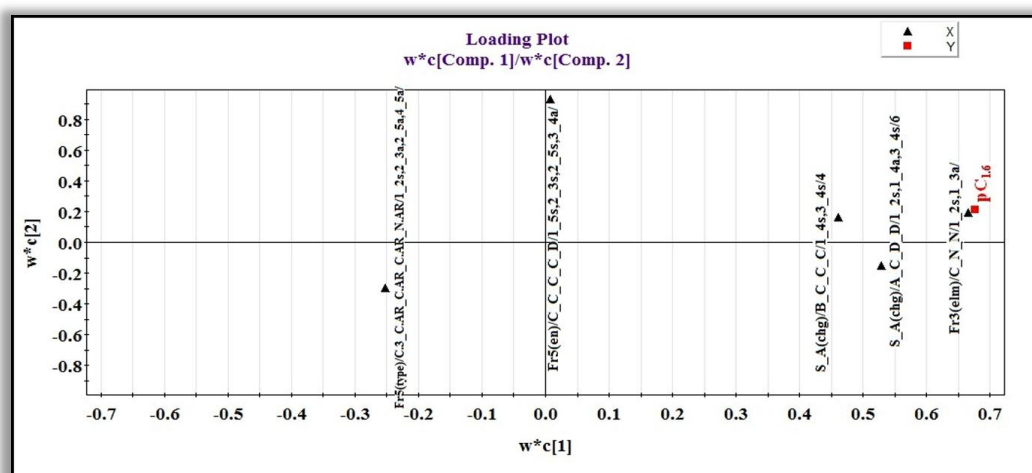


Figure 4.45: Loading plot of SiRMS model.

The highest contributing descriptor is **Fr3(elm)/C_N_N/1_2s,1_3a/** which is a three atomic fragment depicted by N-C=N (**Figure 4.46: Box 1**). Here, the unsaturation between carbon and nitrogen takes place within the imidazole moiety and the other nitrogen is from the nitro group. This descriptor has a positive impact on the radiosensitization of the nitroimidazoles thus with higher number of such fragments increases the pC_{1.6} value. All the compounds in the dataset have this particular group once or twice. Compounds with two fragments of this kind has higher pC_{1.6} values as prominently seen in compounds like **63, 47, 11, 53, 46, 51, 43, 45, 10, 22, 54**, etc. Compounds with only one fragment have considerably lower pC_{1.6} values as observed in **72, 71, 82, 78, 75, 86, 80, 81, 85, 84**, etc. Thus, the importance of this fragment leads us to a conclusion that the presence of nitro groups in nitroimidazole should be between N1 and N3 position of imidazole moiety so as to show better radiosensitization property.

The second important descriptor is $S_A(\text{chg})/A_C_D_D/1_2s,1_4a,3_4s/6$ that represents the partial charge of any of the four atom fragment as given in **Figure 4.46: Box 2**. The fragment here has two possibilities, one with single nitrogen present within the imidazole moiety and another with two nitrogens (one from the imidazole moiety and another from the nitro group) (given in Box 2). Most of the compounds having this fragment have a nitro group attached at position 2 of the imidazole ring. Thus, the position of nitro group plays a vital role in controlling the $pC_{1.6}$ value. This fragment has a positive influence on the radiosensitization effectiveness observed in compounds like **63, 66, 65, 68, 47, 11** and **53**. Compounds which are devoid of these kinds of fragments have considerably low $pC_{1.6}$ value (such as in **74, 77, 80, 75, 78, 71** and **72**).

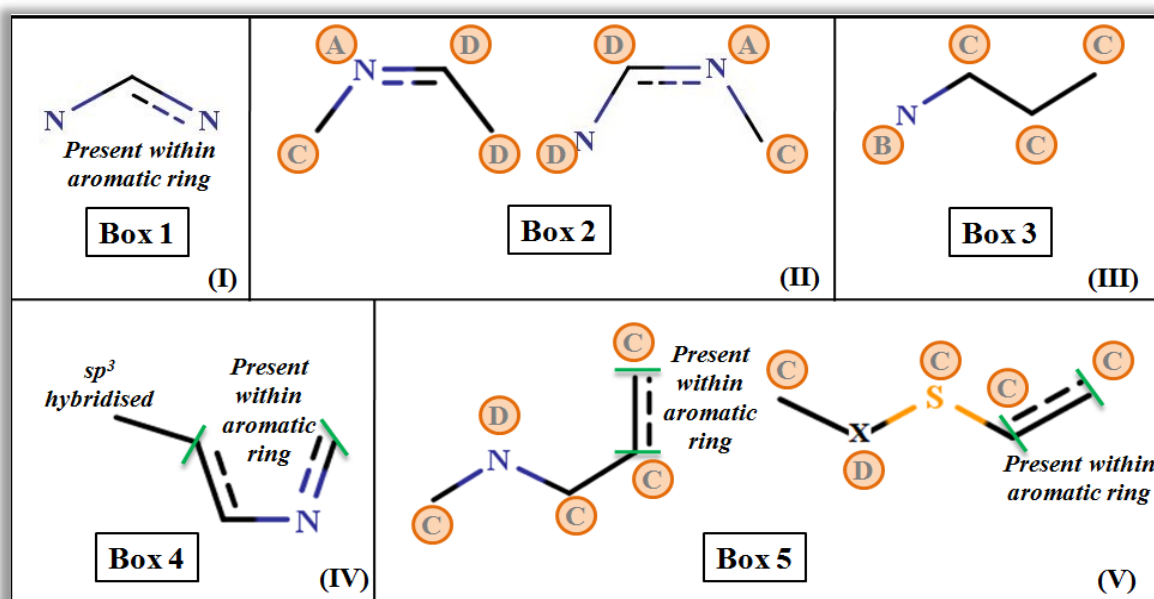


Figure 4.46. Simplex representation of molecular structures (SiRMS) fragments appearing in the nitroimidazole dataset. (I- $Fr3(\text{elm})/C_N_N/1_2s,1_3a/$, II- $S_A(\text{chg})/A_C_D_D/1_2s,1_4a,3_4s/6$, III- $S_A(\text{chg})/B_C_C_C/1_4s,3_4s/4$, IV- $Fr5(\text{type})/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/$, V- $Fr5(\text{en})/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/$).

The next important descriptor is $S_A(\text{chg})/B_C_C_C/1_4s,3_4s/4$ which represents the partial charge of a four atom fragments as given in **Figure 4.46: Box 3**. The presence of the mentioned fragment (i.e., three carbon chain attached to nitrogen from a cyclic nucleus) would increase the radiosensitization effectiveness due to the positive influence of the descriptor. Compounds like **47, 51, 43, 46, 55, 49, 54** and **53** have higher partial charges due to the presence of the mentioned fragments thereby increasing the radiosensitization effectiveness whereas in compounds with no such fragments (like in **71, 72, 82, 78, 75, 80** and **81**) the effect of such charges is not observed thereby the $pC_{1.6}$ value is less.

The next important descriptor $Fr5(\text{type})/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/$ is a five atomic fragment signifying the following formula: $C(sp^3)-C(\text{aromatic})-C(\text{aromatic})-C(\text{aromatic})-N(\text{aromatic})$. The structure of the possible fragment is given in **Figure 4.46: Box 4**. The presence of

this type of fragment reduces the radiosensitization effectiveness as indicated by the negative influence of the descriptor on $pC_{1.6}$ value. This is well observed in compounds like **72**, **59**, **57**, **61**, **69**, **62**, **41** and **70**. On the other hand, absence of this fragment increases the radiosensitization property as seen in compounds such as **43**, **45**, **51**, **46**, **11**, **53**, **47** and **63**.

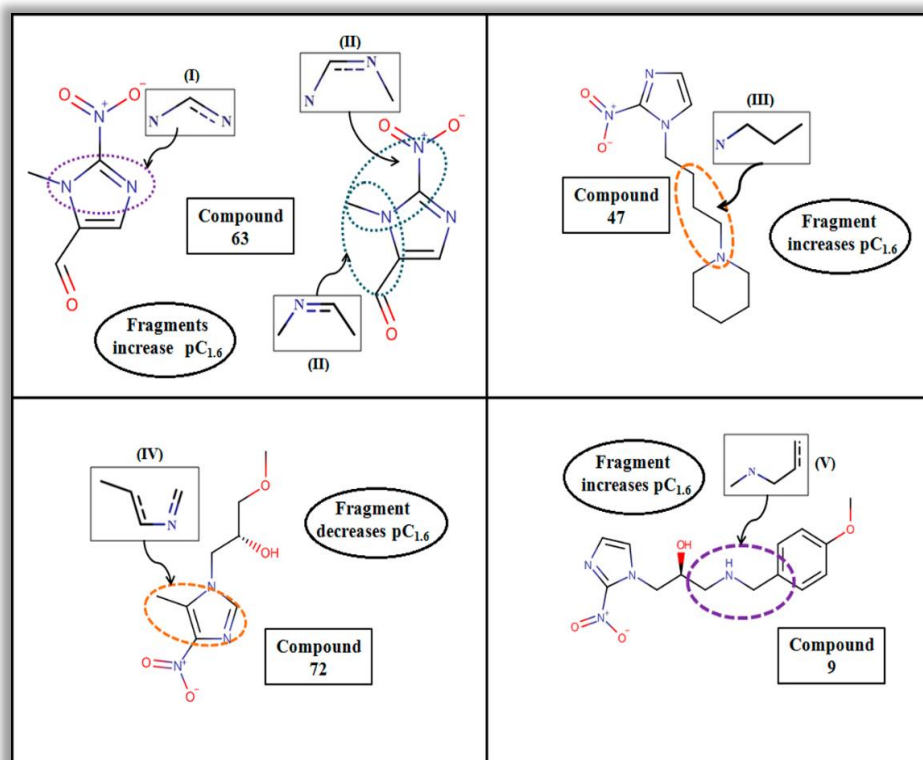


Figure 4.47: Features controlling the increase or decrease in $pC_{1.6}$.

The descriptor with the least significance is **Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/** which denotes the electronegativity of the compound due to the presence of a four atomic fragment given in **Figure 4.46: Box 5**. The positive contribution suggested that the presence of any of the given fragments will influence the electronegativity of the compound thereby increasing the $pC_{1.6}$ value. Compounds **9**, **10** and **11** have been reported to have two such fragments and thereby increase the radiosensitization effectiveness.

4.5.3. Applicability Domain Assessment

The prediction reliability of both the 2D-QSAR models is determined by the applicability domain (AD) assessment. AD gives a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable (Gadaleta et al., 2016). AD assessment for both the models was performed using DModX (distance to model in the X-space) approach at 99% confidence level (**Figure 4.48** and **4.49**). Both the models displayed good coverage of domain of applicability showing maximum number of compounds in the AD (only compound **6** is outside the AD in case of Model 1, i.e., 2D-QSAR model with Dragon descriptors). There was no outlier obtained from the test set for both the models. We have also performed AD assessment at 95% confidence level for both the models and found that in this case three compounds in the test set were outside AD for the model with Dragon descriptors and two compounds in the test set for the model with SiRMS descriptors.

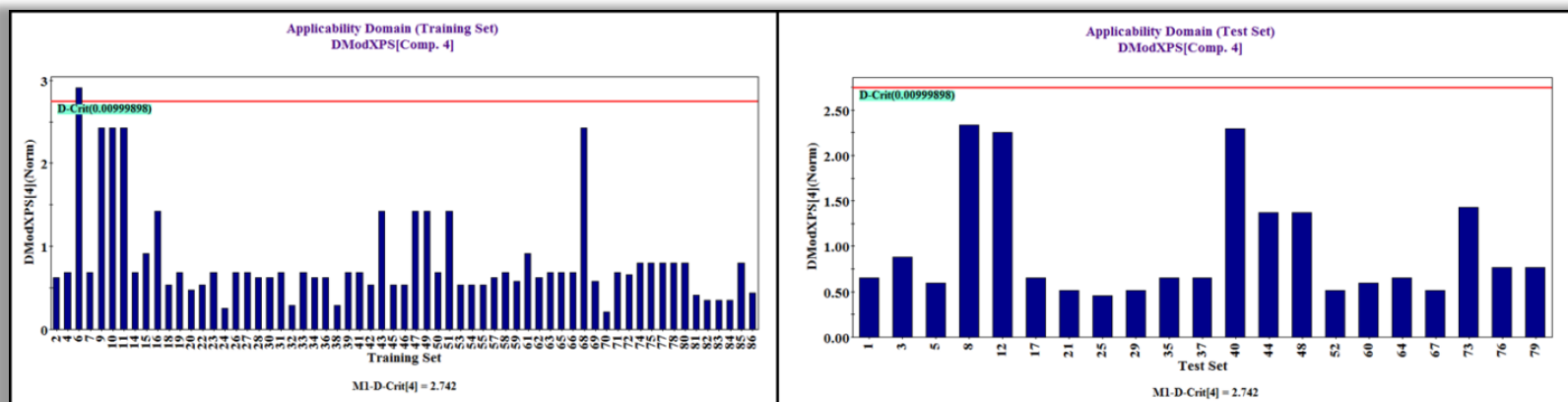


Figure 4.48: Applicability Domain of training and test set of Model 1 (with dragon descriptors) at 99% confidence level.

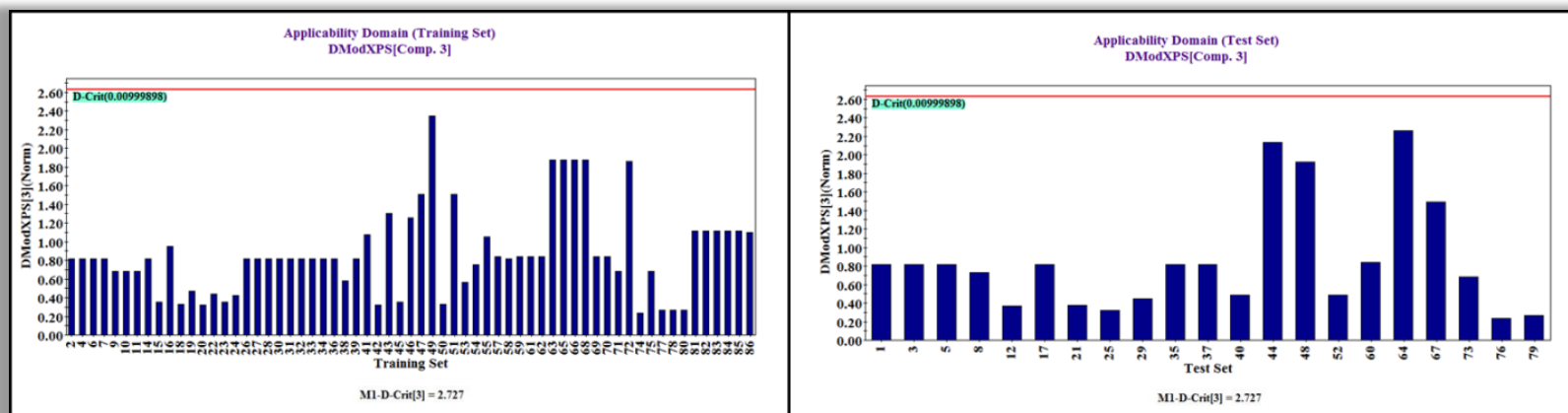


Figure 4.49: Applicability Domain of training and test set of Model 2 (with SiRMS descriptor) at 99% confidence level.

4.5.4. Y-randomization

The Y-randomization plot analysis helps to understand the statistical significance of the model. The randomization plot confirms that the model is not the result of any chance correlation (Rücker et al., 2007). In this process, a number of models are generated by shuffling different combinations of X or Y variables (here Y variable only) based on the fit of the reordered model. In our work, we have used 100 permutations for random model generation. A model with no chance correlation would show very poor statistics for the randomized models, i.e., R_Y^2 intercept should not exceed 0.3 and Q_Y^2 intercept should not exceed 0.05 (Rücker et al., 2007). The randomization plots given in **Figure 4.50** show that the developed models are non-random and robust (as understood from their R_Y^2 and Q_Y^2 values) and are suitable for prediction of the radiosensitization effectiveness within the AD of the model.

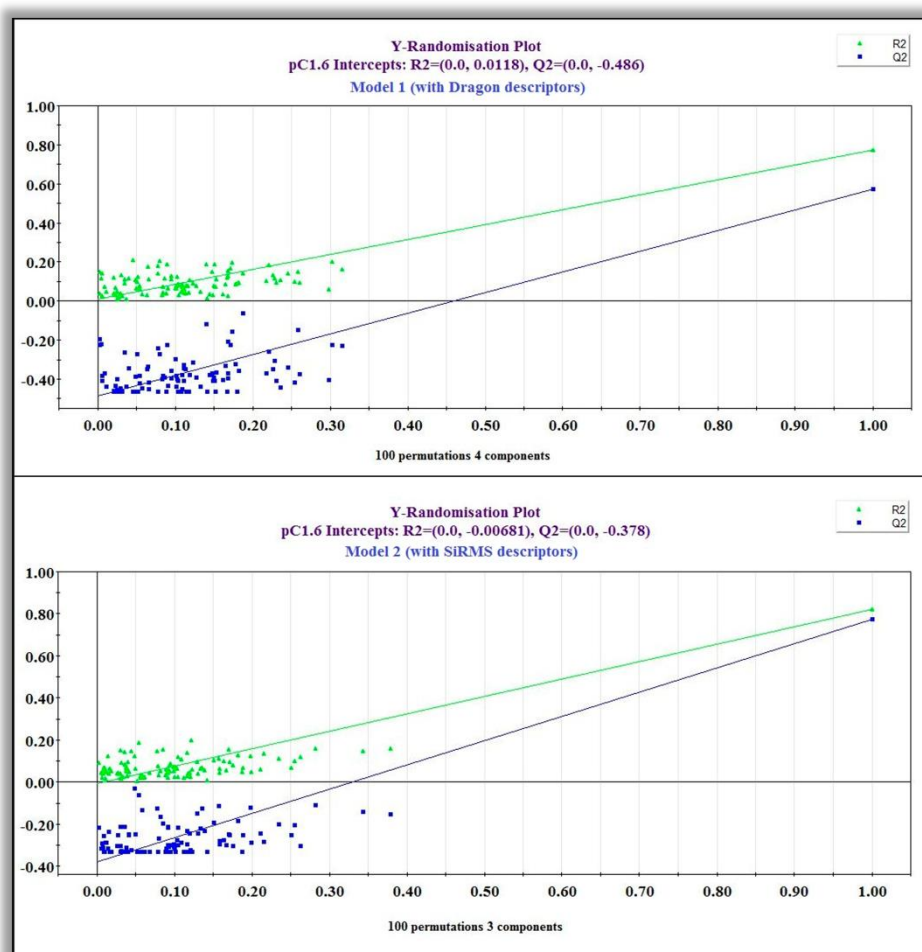


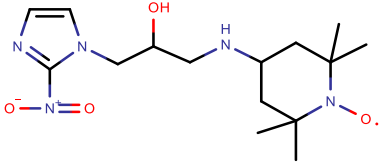
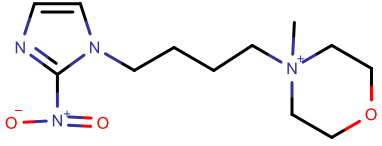
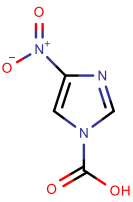
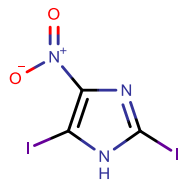
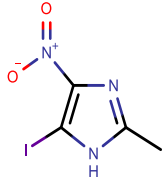
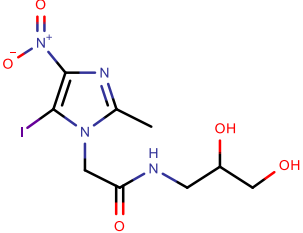
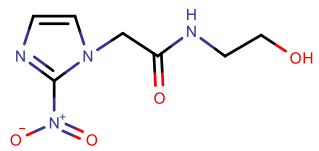
Figure 4.50. Y-randomization plots for Model 1 and Model 2.

4.5.5. True External Predictions

The prediction of responses for external compounds based on their molecular features using chemometric methods can reduce the experiment costs and animal handling. To verify the predictive power of both the models, we have used a set of eight nitroimidazole derivatives (**Table 4.13**) as an external prediction set (Krause et al., 2005; Long & Liu, 2010; Brown et al., 1981). The original dataset in the source literature contains 86 nitroimidazoles but we have removed two of them and used the rest 84 for modeling. These two compounds are now used for prediction purpose. In addition to this, the domain of applicability and their predictive reliability are analyzed using '*prediction*

reliability indicator' tool (Roy et al., 2018). The prediction quality and domain of applicability are given in **Table 4.14**. From the prediction status, it can be inferred that model with fragment-based SiRMS descriptors provides better prediction than model with dragon descriptors.

Table 4.13. External dataset and their predicted $pC_{1.6}$ values

| Compound Number | Structure | Observed $pC_{1.6}$ | Predicted $pC_{1.6}$ using model 1 | Predicted $pC_{1.6}$ using model 2 | Reference |
|-----------------|---|---------------------|------------------------------------|------------------------------------|-----------------------|
| P-1 |  | 4.05 | 3.58 | 3.67 | (Long & Liu, 2010) |
| P-2 |  | 2.89 | 3.88 | 3.82 | (Long & Liu, 2010) |
| P-3 |  | - | 1.98 | 2.18 | (Krause et al., 2005) |
| P-4 |  | - | 4.22 | 2.18 | (Krause et al., 2005) |
| P-5 |  | - | 2.81 | 2.18 | (Krause et al., 2005) |
| P-6 |  | - | 2.53 | 2.18 | (Krause et al., 2005) |
| P-7 |  | - | 3.33 | 3.48 | (Brown et al., 1981) |

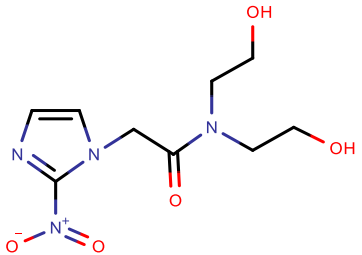
| | | | | | |
|-----|---|---|------|------|----------------------|
| P-8 |  | - | 3.04 | 3.48 | (Brown et al., 1981) |
|-----|---|---|------|------|----------------------|

Table 4.14. Prediction quality (Roy et al., 2018) for the true external dataset.

| Compound Number | Prediction Status of model with Dragon descriptors | | | Prediction Status of model with SiRMS descriptors | | |
|-----------------|--|--------------------|--|---|--------------------|--|
| | Composite Score | Prediction Quality | AD status (using standardization approach) | Composite Score | Prediction Quality | AD status (using standardization approach) |
| P-1 | 3 | Good | Outside AD | 3 | Good | In |
| P-2 | 3 | Good | In | 3 | Good | In |
| P-3 | 2 | Moderate | In | 3 | Good | In |
| P-4 | 3 | Good | In | 3 | Good | In |
| P-5 | 3 | Good | In | 3 | Good | In |
| P-6 | 3 | Good | Outside AD | 3 | Good | In |
| P-7 | 3 | Good | In | 3 | Good | In |
| P-8 | 3 | Good | In | 3 | Good | In |

4.5.6. Comparison with the previously published research

In the previously published research by Long and Liu (Long & Liu, 2010), the authors developed MLR and projection pursuit regression (PPR) (Du et al., 2002; Friedman & Stuetzle, 1981; Liu et al., 2007) models using complex descriptors such as geometrical, electrostatic and quantum chemical descriptors. The models developed by us cannot be critically compared to the previously published since the calibration and validation set compositions are different. However, it can be found that our MLR model developed using SiRMS descriptor is better in terms of both training and test set validation metrics if we consider their MLR model (**Table 4.15**). Also, the current model comes with an added advantage of presence of lower number of simple descriptors and non-requirement of conformation analysis or energy minimization prior to their calculation. Furthermore, the PPR based model reported in the previous study is derived from a more complicated process which uses projection-based approach to convert high dimensional data to lower dimension. Moreover, 3D descriptors were used in the previous work. MLR or PLS models are more straight-forward and reproducible as used in the current work. In addition, 2D descriptors used in the present work are easy to compute and do not need any conformation analysis or energy minimization process.

Table 4.15. Comparison of the current SiRMS model with previously developed MLR model.

| Model | Total no. of compounds used | No. of compounds in the training set | No. of compounds in the test set | Descriptor type | No. of descriptors in final model | Training Set | | | Test Set | |
|---------------------------|-----------------------------|--------------------------------------|----------------------------------|------------------------------|-----------------------------------|--------------|-------|-------|------------|-------|
| | | | | | | R^2 | Q^2 | RMSEC | Q_{F1}^2 | RMSEP |
| Current study | 84 | 63 | 21 | 2D (fragment based SiRMS) | 5 (3 LVs) | 0.82 | 0.79 | 0.26 | 0.80 | 0.28 |
| Long and Liu, 2010 | 86 | 68 | 18 | 3D | 6 | 0.80 | 0.76 | 0.28 | 0.76 | 0.28 |

4.6. Study 6: QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: Application of Small Dataset Modelling

2D-QSAR models using Dragon and SiRMS descriptors explaining chemical features required for good drug radiosensitization (both SER and logSR) are shown in the following section. There are 4 models developed of which two are QSAR models and the rest two are QSAAR models. All the models are three-descriptor PLS models with 2 latent variables (LVs) showing acceptable values for all validation metrics as shown in **Table 4.16**. The validation metrics included R^2 , Q^2 , $Q_{LMO}^2(20\%)$, $\overline{r_{m(LOO)}^2}$, $\Delta r_{m(LOO)}^2$, SD (95% data; Training), MAE (95% data; Training) and RMSE. Furthermore, we have calculated the Q_{F1}^2 metric for the validation set in each iteration cycle for each model during the calculation of $Q_{LMO}^2(20\%)$ (Table 4.17). The experimental and predicted values for all the models are given in **Table 4.18** and the observed versus predicted plots for all the developed QSAR and QSAAR models are shown in **Figure 4.51**. The different PLS plots including variable importance plot, loading plot, regression coefficient plot and randomisation plot are discussed later in Section 4.6.4.

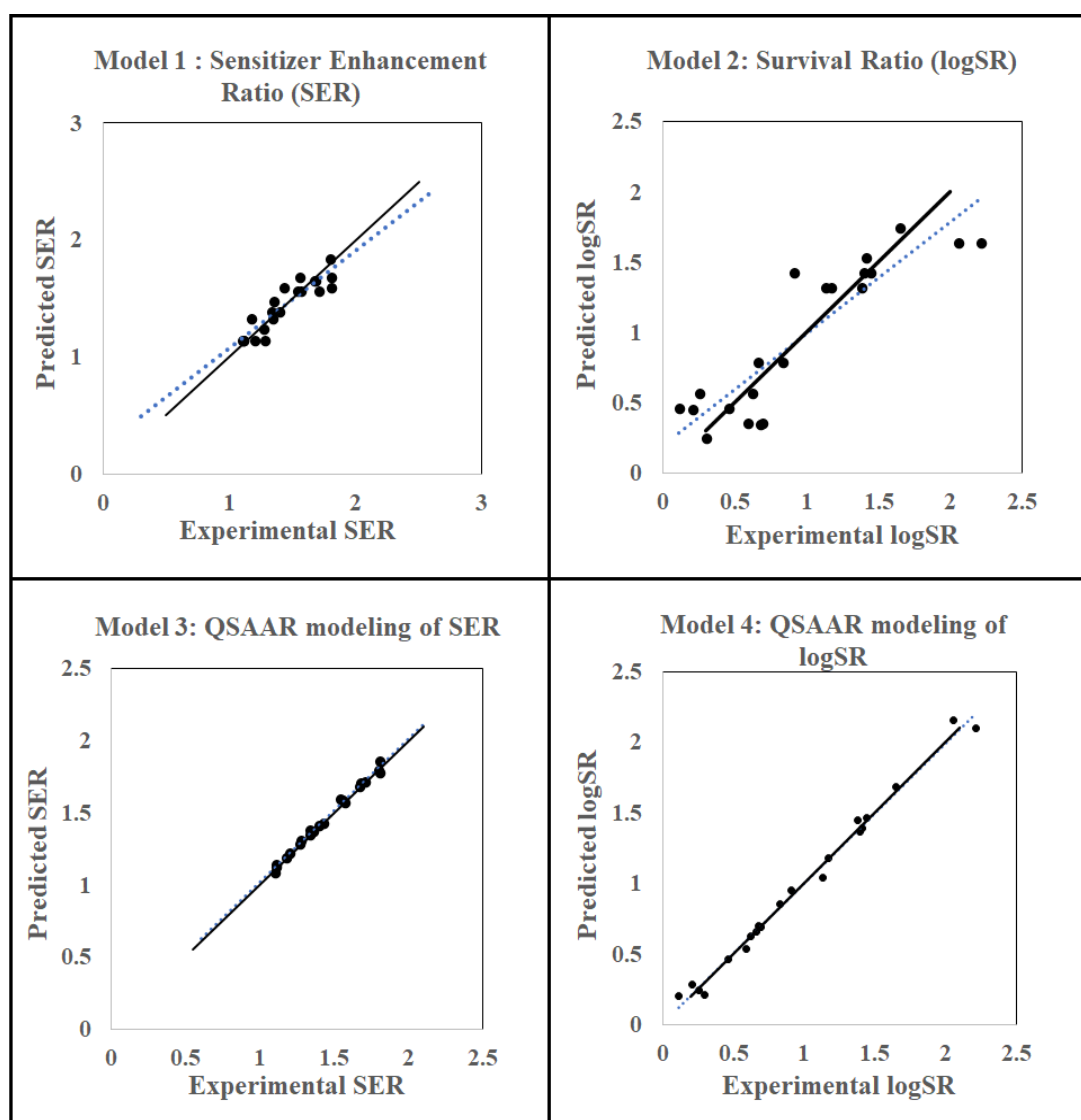


Figure 4.51: Scatter plots for QSAR and QSAAR models.

4.6.1. Model 1: Modeling Drug Sensitizer Enhancement Ratio (SER)

$$SER = 0.931 + 0.452 \times H - 049 - 0.238 \times B05[O - S] + 0.09 \times F05[C - S]$$

The first descriptor **H-049** belongs to atom-centred fragment type, which indicates H atom attached to C^3 (sp^3)/ C^2 (sp^2)/ C^3 (sp). The descriptor symbolizes the hydrogen of a CH group with the carbon bonded to varying numbers of heteroatoms in a variety of hybridizations. The descriptor has a positive contribution towards the response (**Figure 4.52**) which is well understood from certain higher active compounds in the dataset like compounds **19** (SER=1.81) and **24** (SER=1.81), each of which has two H-049 fragments. On the other hand, compounds like **12** (SER=1.11) and **16** (SER=1.105) having only one such fragments have low SER values.

The next descriptor is **B05[O-S]**, which is a 2D atom pair descriptor demonstrating the presence or absence of oxygen and sulphur atoms at the topological distance 5. The negative contribution explains that presence of oxygen and sulphur atoms at the topological distance 5 will lower the SER values (**Figure 4.52**) as observed in compounds **7** (SER=1.11) and **16** (SER=1.105). On the other hand, in compounds like **4** (SER=1.835) and **30** (SER=1.687), the absence of such fragment does not lower the SER value.

The descriptor **F05[C-S]**, another 2D atom pair descriptor, denotes the frequency of C - S at the topological distance 5. The positive contribution of the descriptor indicates that higher frequency of the C-S fragment at the topological distance 5 will increase the SER value (**Figure 4.52**) as seen in compounds **30** (F05[C-S]= 3, SER=1.68) and **38** (F05[C-S]=3, SER=1.67).

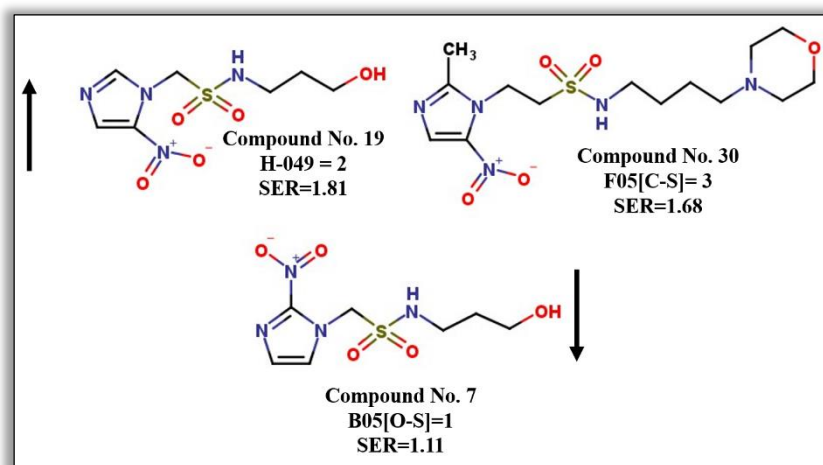


Figure 4.52. Features increasing or decreasing SER values as explained in Model 1.

4.6.2. Model 2: Modeling Drug Survival Ratio (logSR)

$$\log SR = 1.965 - 1.08 \times S_A(chg)/A_B_B_D/1_4s, 3_4s/4 - 1.073 \times C - 033 - 0.108 \times F07[C - C]$$

S_A(chg)/A_B_B_D/1_4s,3_4s/4 represents a four atomic fragment labeled by partial charges, and its negative regression coefficient indicates that it reduces the radiosensitization property with the presence of such fragment (shown in **Figure 4.53**). In compounds like **26** and **28**, presence of such fragment reduces the radiosensitization (logSR= 0.681 and 0.208).

C-033 is an atom-centred fragment descriptor represented by R--CH..X fragment. 'R' denotes any group linked through carbon, '-' represents an aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group, '..' represents aromatic single bonds as the C-N bond in pyrrole and 'X' is any electronegative atom (O, N, S, P, Se, halogens) (R Todeschini & Consonni, 2009). The negative coefficient indicates that presence of this type of fragment lowers logSR (**Figure 4.53**) values as observed in compounds **6** (C-033= 1, logSR= 0.462) and **7** (C-033= 1, logSR= 0.255).

F07[C-C] is a 2D atom pair descriptor, which signifies the frequency of the C-C fragment at the topological distance 7. The negative coefficient indicates that a higher value of the descriptor may decrease the radiosensitization (logSR value) (**Figure 4.53**). This is observed in compounds like **12** and **8** where F07[C-C] are high (6 and 5 respectively) and their logSR values are low (0.301 and 0.591 respectively).

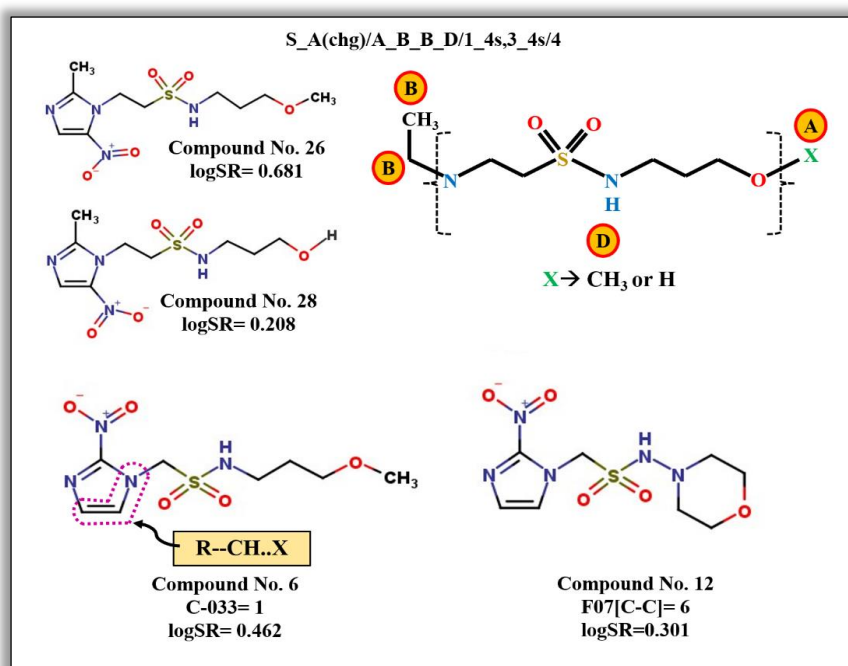


Figure 4.53. Factors decreasing logSR values as explained in Model 2

4.6.3. Quantitative Structure Activity-Activity Relationship (QSAAR) models

QSAAR models are mathematical expressions correlating two biological end points, here SER and logSR, with the aim to extrapolate any one explicit activity endpoint when the experimental data is not available. This advanced technique can overcome the additional cost of manifold experimental procedures. In the present study, we have developed two QSAAR models, one taking SER as the endpoint and logSR as an independent variable and another taking logSR as the endpoint and SER as an independent variable. It was found that these two endpoints had positive correlation between themselves explaining that increase in experimental value of any of the endpoints would increase the other endpoint and vice versa.

4.6.1.1. Model 3: QSAAR modeling of SER

$$SER = 1.084 + 0.018 \times F03[C - C] + 0.363 \times \log SR - 0.001 \times T(N..O)$$

Model 3 is a PLS model with 2 latent variables and shows acceptable values of the validation metrics. Here, logSR has been used as an independent variable to produce a QSAAR model for drug SER.

Thus, for any compound, if survival ratio (SR) value is known, the SER value can be extrapolated using model 3. This reduces time and experimental expenses. In the model, logSR shows a positive regression coefficient; hence, a higher value of logSR will increase SER values as observed in compounds like **19** (logSR= 2.212, SER= 1.81) and **24** (logSR= 2.057, SER= 1.81).

The descriptor **F03[C-C]** is a 2D atom pair descriptor signifying the frequency of C-C fragments at the topological distance 3. This makes a positive contribution to the endpoint thus indicating that with an increase in the F03[C-C] descriptor value, SER value will also increase as seen in compounds **30** (F03[C-C]=14, SER=1.68) and **35** (F03[C-C]=13, SER=1.71). Another 2D atom pair descriptor **T(N..O)** appears in the model signifying the sum of topological distances between N..O. This descriptor has a negative influence on the SER values indicating that the total distance between nitrogen and oxygen should be low for higher SER values as in compound **4** (T(N..O)=51, SER=1.8). Compounds with higher T(N..O) values will have lower SER values as observed in compounds **8** (T(N..O)=130, SER=1.28) and **12** (T(N..O)=106, SER=1.11). Features increasing and decreasing SER values are shown in **Figure 4.54**.

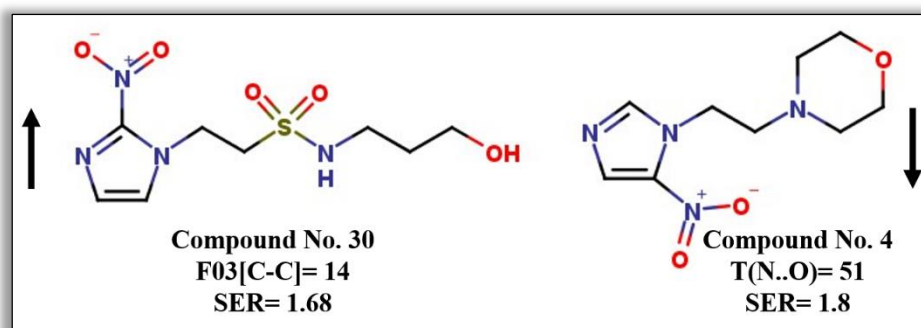


Figure 4.54. Features increasing or decreasing SER value as explained in Model 3

4.6.1.2. Model 4: QSAAR modeling of logSR

$$\log SR = -3.364 + 2.735 \times SER - 0.028 \times F03[C - C] + 0.125 \times nO$$

In Model 4, SER has been used as an independent variable for modeling logSR. SER makes a positive contribution to logSR, proving the authenticity of the previously developed model 3 and this can be explained by the same compounds **19** and **24**.

F03[C-C] is a 2D atom pair descriptor symbolizing the frequency of the C-C fragment at the topological distance 3. The descriptor shows a negative regression coefficient, thus signifying that with an increase in F03[C-C] values, logSR value will decrease and vice versa. It is observed that in compounds **15** and **34**, the F03[C-C] values are high (10 and 11 respectively) and their logSR values are low (log SR= 0.699 and 1.134 respectively). The opposite is observed in compounds **19** (F03[C-C]= 2, logSR= 2.057) and **24** (F03[C-C]= 4, logSR= 2.212) having lower values for F03[C-C]. Descriptor **nO** is a constitutional descriptor meaning the number of oxygen atoms present in a molecule. The positive regression coefficient indicates that presence of oxygen atoms is beneficial for the in vitro radiosensitization (logSR). In compounds like **19** (logSR= 2.057) and **24** (logSR= 2.212), higher number of oxygen (nO=5) contributes to a higher value of logSR. Features increasing and decreasing logSR value are shown in **Figure 4.55**.

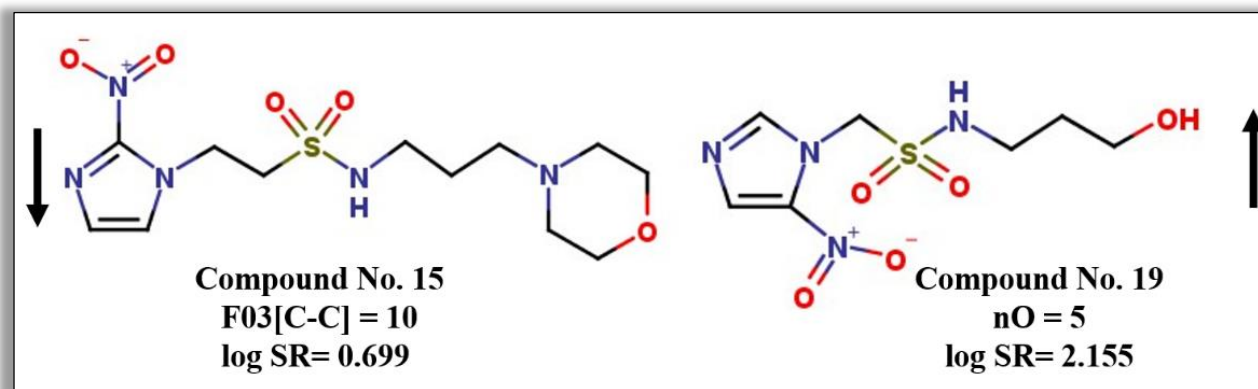


Figure 4.55. Features increasing or decreasing logSR value as explained in Model 5.

Table 4.16: Validation metrics of the four models developed using the Small Dataset Modeler.

| Model Number | Endpoint | Number of descriptors | LV | R ² | Q ² | Q _{LMO} ² (20%) | $\overline{r_m^2(L00)}$ | $\Delta r_m^2(L00)$ | SD (95% data; TRAIN) | MAE (95% data; TRAIN) | RMSE |
|--------------|-------------|-----------------------|----|----------------|----------------|-------------------------------------|-------------------------|---------------------|----------------------|-----------------------|-------|
| 1 | SER | 3 | 2 | 0.834 | 0.746 | 0.712 | 0.660 | 0.134 | 0.066 | 0.073 | 0.096 |
| 2 | logSR | 3 | 2 | 0.798 | 0.660 | 0.665 | 0.563 | 0.109 | 0.189 | 0.216 | 0.261 |
| 3 | QSAAR_SER | 3 | 2 | 0.993 | 0.985 | 0.982 | 0.972 | 0.012 | 0.013 | 0.016 | 0.027 |
| 4 | QSAAR_logSR | 3 | 2 | 0.991 | 0.983 | 0.983 | 0.968 | 0.014 | 0.037 | 0.046 | 0.055 |

Table 4.17: Q_{F1}^2 metric for the validation set for each model in each iteration during the calculation of Q_{LMO}^2 (20%).

| Model No. | Iteration | Compound numbers belonging to validation set | Q_{F1}^2 |
|-----------|-----------|--|------------|
| M1 | Step 1 | 1, 8, 19, 28, 38 | 0.522 |
| | Step 2 | 2, 12, 21, 30 | 0.669 |
| | Step 3 | 4, 14, 22, 31 | 0.800 |
| | Step 4 | 6, 15, 24, 34 | 0.720 |
| | Step 5 | 7, 16, 26, 35 | 0.896 |
| M2 | Step 1 | 1, 8, 19, 28, 38 | 0.720 |
| | Step 2 | 2, 12, 21, 30 | 0.559 |
| | Step 3 | 4, 14, 22, 31 | 0.901 |
| | Step 4 | 6, 15, 24, 34 | 0.709 |
| | Step 5 | 7, 16, 26, 35 | 0.561 |
| M3 | Step 1 | 1, 8, 19, 28, 38 | 0.966 |
| | Step 2 | 2, 12, 21, 30 | 0.993 |
| | Step 3 | 4, 14, 22, 31 | 0.983 |
| | Step 4 | 6, 15, 24, 34 | 0.980 |
| | Step 5 | 7, 16, 26, 35 | 0.995 |
| M4 | Step 1 | 1, 8, 19, 28, 38 | 0.987 |
| | Step 2 | 2, 12, 21, 30 | 0.965 |
| | Step 3 | 4, 14, 22, 31 | 0.996 |
| | Step 4 | 6, 15, 24, 34 | 0.980 |
| | Step 5 | 7, 16, 26, 35 | 0.991 |

Table 4.18: Experimental SER and logSR values and Predicted SER and logSR values for all four models

| Compound ID | Exp SER | Exp logSR | M1 | M2 | M3 | M4 |
|-------------|---------|-----------|----------|------------|------------------------|---------------------------|
| | | | Pred SER | Pred logSR | Pred SER (QSAAR model) | Pred log SR (QSAAR model) |
| 1 | 1.400 | 0.833 | 1.383 | 0.833 | 1.417 | 0.833 |
| 2 | 1.339 | 0.663 | 1.383 | 0.663 | 1.351 | 0.663 |
| 4 | 1.800 | 1.652 | 1.835 | 1.652 | 1.795 | 1.652 |
| 6 | 1.200 | 0.462 | 1.145 | 0.462 | 1.223 | 0.462 |
| 7 | 1.110 | 0.255 | 1.145 | 0.255 | 1.130 | 0.255 |
| 8 | 1.280 | 0.591 | 1.145 | 0.591 | 1.313 | 0.591 |
| 12 | 1.110 | 0.301 | 1.145 | 0.301 | 1.141 | 0.301 |
| 14 | 1.270 | 0.623 | 1.235 | 0.623 | 1.288 | 0.623 |
| 15 | 1.357 | 0.699 | 1.473 | 0.699 | 1.375 | 0.699 |
| 16 | 1.105 | 0.114 | 1.145 | 0.114 | 1.089 | 0.114 |

| | | | | | | |
|----|-------|-------|-------|-------|-------|-------|
| 19 | 1.810 | 2.057 | 1.597 | 2.057 | 1.782 | 2.057 |
| 21 | 1.430 | 0.914 | 1.597 | 0.914 | 1.428 | 0.914 |
| 22 | 1.560 | 1.415 | 1.687 | 1.415 | 1.592 | 1.415 |
| 24 | 1.810 | 2.212 | 1.687 | 2.212 | 1.863 | 2.212 |
| 26 | 1.340 | 0.681 | 1.325 | 0.681 | 1.379 | 0.681 |
| 28 | 1.176 | 0.208 | 1.325 | 0.208 | 1.190 | 0.208 |
| 30 | 1.680 | 1.447 | 1.653 | 1.447 | 1.708 | 1.447 |
| 31 | 1.570 | 1.173 | 1.563 | 1.173 | 1.574 | 1.173 |
| 34 | 1.540 | 1.134 | 1.563 | 1.134 | 1.595 | 1.134 |
| 35 | 1.710 | 1.380 | 1.563 | 1.380 | 1.721 | 1.380 |
| 38 | 1.670 | 1.398 | 1.653 | 1.398 | 1.683 | 1.398 |

4.6.4. Plot interpretation

4.6.4.1. Variable importance plot (VIP)

A VIP can provide with a better knowledge about the descriptors and their contribution in controlling the radiosensitization properties of nitroimidazole sulfonamides. The plot signifies the order of contribution of each descriptor appearing in the model. The most and least important descriptors can be identified using this plot. A variable with VIP score > 1 indicates the descriptor has higher statistical significance as compared to the one with a lower VIP value (Akarachantachote et al., 2014). The VIP plot showing the descriptors from higher to lower significance is given in the **Figure 4.56** and **4.57**.

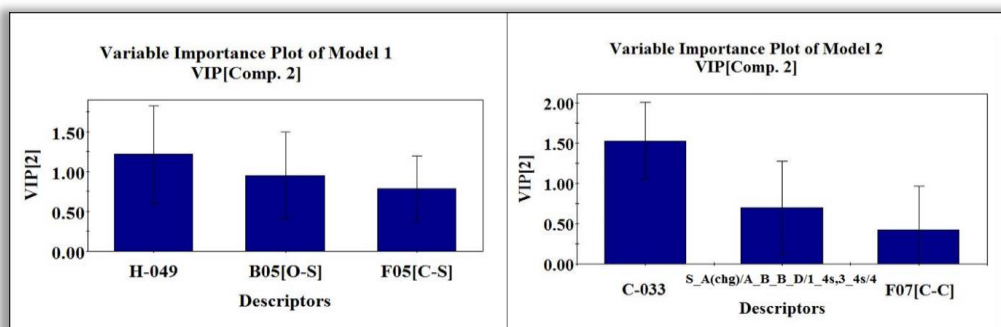


Figure 4.56. VIP of Model 1 and Model 2 (QSAR models)

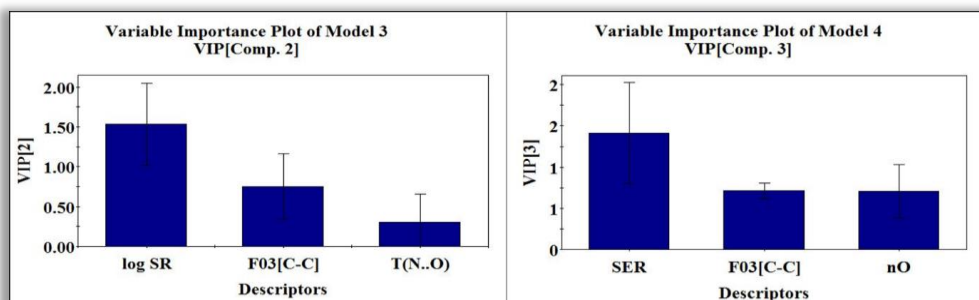


Figure 4.57: VIP of Model 3 and Model 4 (QSAAR models)

4.6.4.2. Loading Plot

The loading plot defines the relationship between X variables and Y variables (Wold et al., 2001). The plot was developed using the two latent variables for all the four models. The plot describes the impact of the different variables. Descriptors that are grouped together have similar meanings and similar effects on the response whereas descriptors with different meanings are situated at a considerable distance from each other. Descriptors which are situated far from the plot origin have greater impact on the response. The loading plots of the four models are given in the **Figures 4.58** and **4.59**.

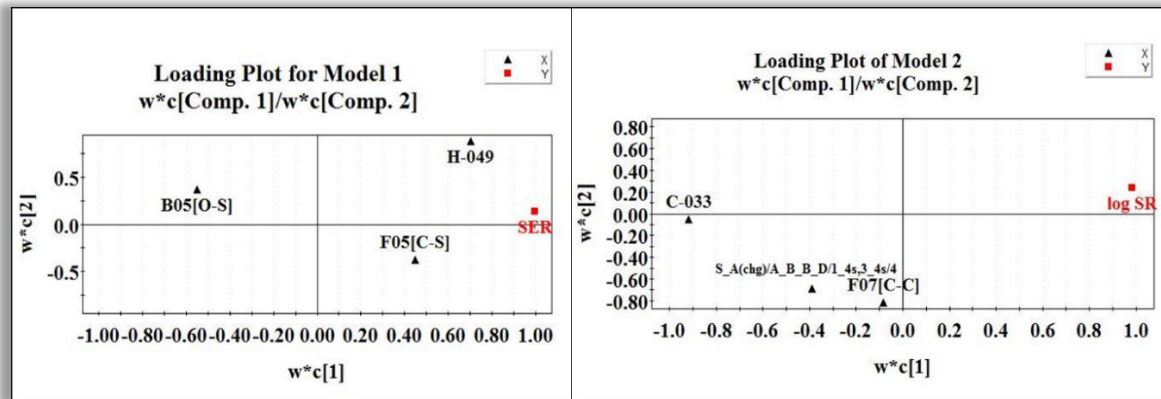


Figure 4.58. Loading Plot of Model 1 and Model 2

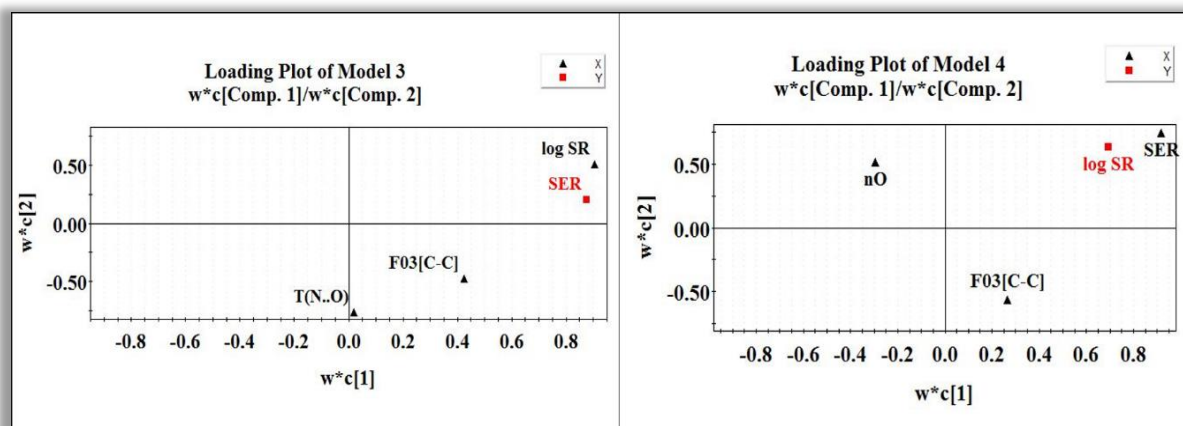


Figure 4.59. Loading Plot of Model 3 and Model 4

4.6.4.3. Randomization Plot

Model randomization is done to ensure that the model is not the result of any chance correlation (Rücker et al., 2007). The statistical significance of the model is determined by a randomisation model. During the model randomisation, multiple models are generated by shuffling different combinations of X or Y variables (here Y variable) based on the fit of the reordered model. Here, we have used 100 permutations for each model for random model generation. A model not generated out of chance correlation should have poor statistics (R_y^2 intercept should not exceed 0.3 and Q_y^2 intercept should not exceed 0.05). The randomization plots given in **Figures 4.60** and **4.61** show that the developed models are non-random and robust and are suitable for prediction.

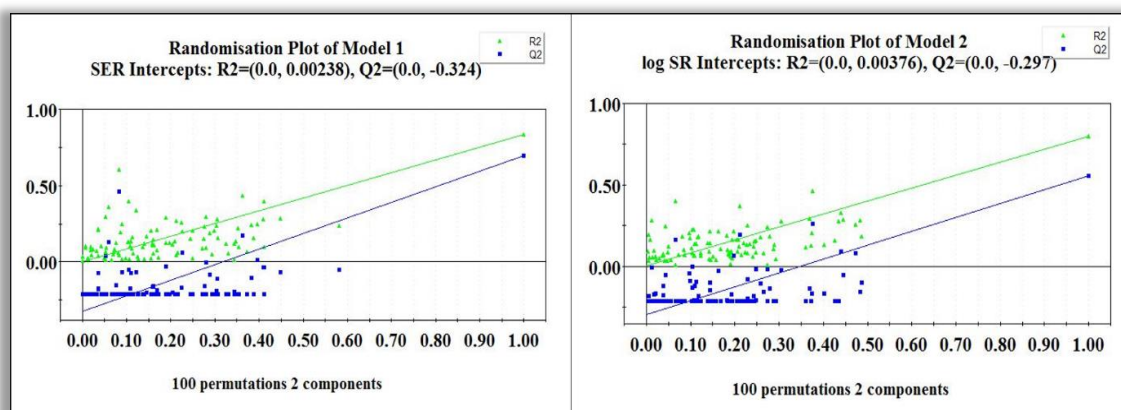


Figure 4.60. Y-Randomisation Plot of Model 1 and Model 2

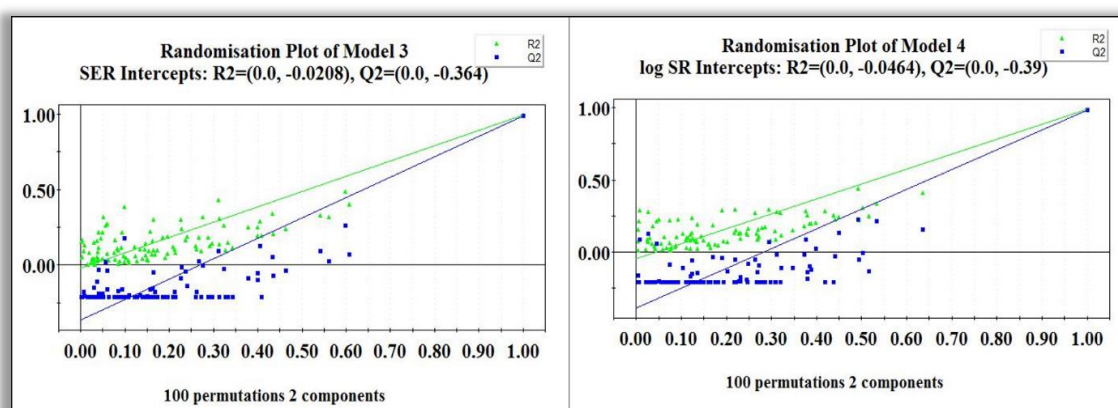


Figure 4.61. Y-Randomisation Plot of Model 3 and Model 4

4.6.5. Applicability Domain assessment

Applicability Domain (AD) explains the prediction reliability of a particular model. It is the “chemical space from which a model is derived and where a prediction is considered to be reliable” (Gadaleta et al., 2016). AD evaluation was done using DModX (distance to model) in the X-space using SIMCA 16.0.2 software (<https://landing.umetrics.com/downloads-simca>). The AD plots are given in **Figures 4.62** and **4.63**. It is found that there is no outlier in any of the four models developed at 95% confidence level ($D\text{-crit}= 0.009999$).

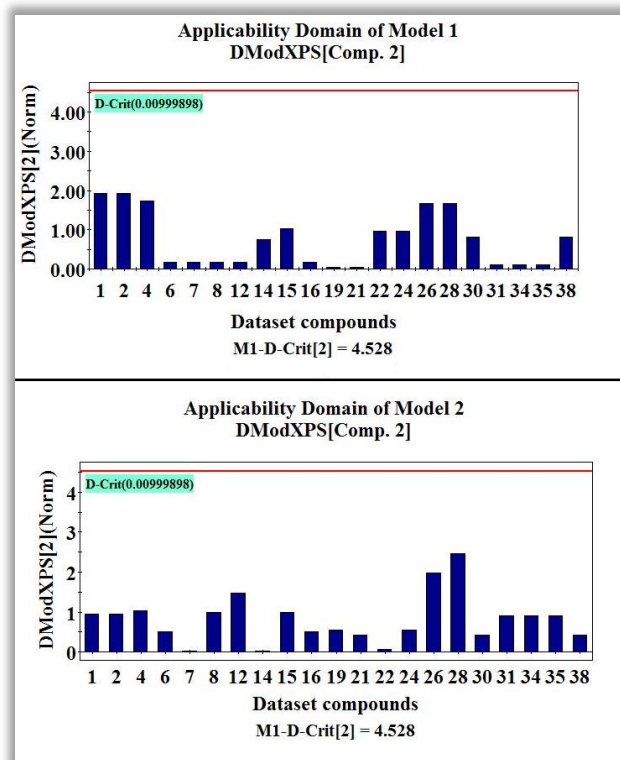


Figure 4.62. DModX Applicability Domain plot of Model 1 and Model 2

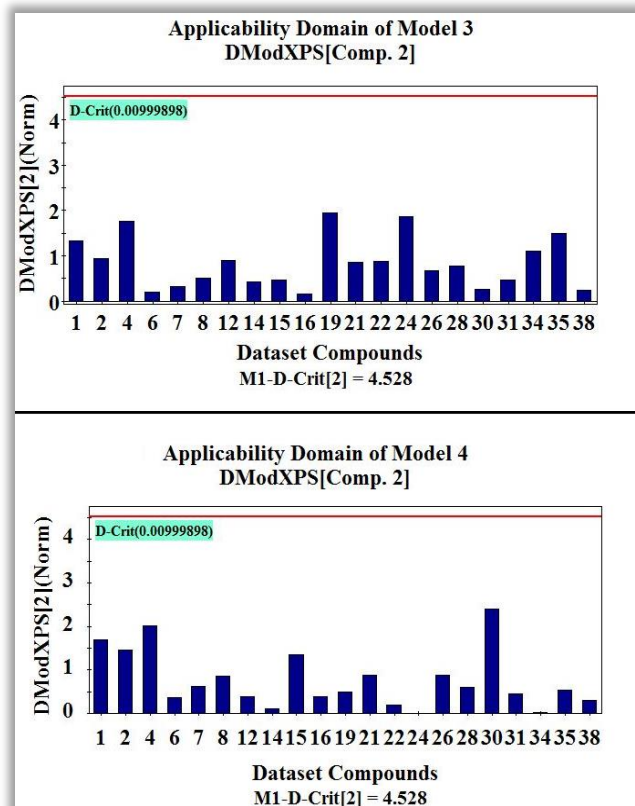


Figure 4.63. DModX Applicability Domain of Model 3 and Model 4.

4.6.6. Prediction dataset

A QSAR model helps in the prediction of external datasets based on their molecular features thereby reducing the experiment costs and animal handling. To study the predictive power of the developed models, we have used 14 compounds whose SER and logSR values have been predicted. These 14 compounds were selected from Table 1 of the source article (Bonnet et al., 2018). This table contained about 36 nitroimidazole sulphonamides out of which 21 compounds were used for QSAR and QSAAR modelling and rest 14 compounds were used as an external set for prediction. Further, we have analysed the prediction quality and domain of applicability using *Prediction Reliability Indicator* tool (Roy et al., 2018). The prediction status and domain of applicability are given in **Table 4.19**. Prediction was possible for model 1 (M1), model 2 (M2) and model 3 (M3). In M1 and M2, the predicted SER and predicted logSR values were calculated for 14 compounds. In case of M3 (QSAAR-SER), SR₁₅ values were obtained from source article (Bonnet et al., 2018) and the values were converted to logarithmic form and used as an independent variable for the calculation of predicted SER values. Prediction for model M4 was not possible since experimental SER values for the prediction compounds are not available. During prediction with model M1, three compounds had bad/unreliable predictions. This is due to the difference between the mean of the training set response and predicted value of the query compound being considerably higher. However, these compounds fall inside the AD of the model. In case of M2, one compound (compound no. **25**) is outside AD, however it shows moderate prediction quality. During prediction with model M3, all the compounds are found to have ‘moderate’ prediction quality and are inside the model AD.

Table 4.19. Prediction dataset and their predicted SER and logSR values along with prediction quality and AD status obtained from ‘Prediction Reliability Indicator’ tool.

| Serial No. | Compound No. | Structure (SMILES) | M1 (SER) | | | M2 (logSR) | | | M3 (QSAAR-SER) | | |
|------------|--------------|--|-----------|--------------------|-----------|-------------|--------------------|------------|----------------|--------------------|-----------|
| | | | Pred_SE R | Prediction Quality | AD status | Pred_logS R | Prediction Quality | AD status | Pred_SE R | Prediction Quality | AD status |
| 1 | 9 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCN1CCCC1)[N+](=O)[O-]</chem> | 0.694 | Bad/Unreliable | In | 0.355 | Moderate | In | 1.174 | Moderate | In |
| 2 | 10 | <chem>c1(n(ccn1)CS(=O)(=O)NCCC(=O)O)[N+](=O)[O-]</chem> | 0.694 | Bad/Unreliable | In | 0.570 | Moderate | In | 1.207 | Moderate | In |
| 3 | 11 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCC(=O)O)[N+](=O)[O-]</chem> | 0.784 | Moderate | In | 0.570 | Moderate | In | 1.106 | Moderate | In |
| 4 | 13 | <chem>c1(n(ccn1)CCS(=O)(=O)NCCCOC)[N+](=O)[O-]</chem> | 0.784 | Moderate | In | 0.462 | Moderate | In | 1.356 | Moderate | In |
| 5 | 18 | <chem>c1n(cc(n1)[N+](=O)[O-])CS(=O)(=O)NCCCO</chem> | 0.694 | Bad/Unreliable | In | 0.570 | Moderate | In | - | - | - |
| 6 | 20 | <chem>c1n(cc(n1)[N+](=O)[O-])CS(=O)(=O)NCCCN1CCOCC1</chem> | 0.931 | Moderate | In | 0.355 | Moderate | In | 1.371 | Moderate | In |
| 7 | 23 | <chem>c1n(cc(n1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC</chem> | 0.784 | Moderate | In | 0.462 | Moderate | In | 1.401 | Moderate | In |
| 8 | 25 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC)C</chem> | 0.874 | Moderate | In | 0.456 | Moderate | Outside AD | 1.387 | Moderate | In |
| 9 | 27 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC)C</chem> | 1.201 | Moderate | In | 1.428 | Moderate | In | 1.382 | Moderate | In |
| 10 | 29 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN1CCOCC1)C</chem> | 1.111 | Moderate | In | 1.320 | Moderate | In | 1.304 | Moderate | In |
| 11 | 32 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCCN(C)C)C</chem> | 1.201 | Moderate | In | 1.428 | Moderate | In | 1.471 | Moderate | In |
| 12 | 33 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN(CC)CC)C</chem> | 1.111 | Moderate | In | 1.320 | Moderate | In | 1.548 | Moderate | In |
| 13 | 36 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NN1CCOCC1)C</chem> | 0.874 | Moderate | In | 1.535 | Moderate | In | 1.130 | Moderate | In |
| 14 | 37 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NN1CCN(CC1)C)C</chem> | 1.111 | Moderate | In | 1.428 | Moderate | In | 1.215 | Moderate | In |

4.7. Study 7: Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness

4.7.1. Modeling local nitro datasets

2D-QSAR models for explaining radiosensitization effectiveness ($pC_{1.6}$) are discussed in this section. The QSAR models from individual class of nitro compounds were found to have good and acceptable values for all validation metrics. The validation metrics included R^2 , Q^2 , Q_{LMO}^2 , $\overline{\Delta r_{m(LOO)}^2}$, $\Delta r_{m(LOO)}^2$, MAE and RMSE. The current work proposes statistically robust and acceptable local models employing simple 2D descriptors. The observed versus predicted $pC_{1.6}$ plots for the local models are given in **Figure 4.64**

4.7.1.1. QSAR model studying radiosensitization effectiveness of Nitrofurans

$$pC_{1.6} = -0.617(\pm 0.249) - 0.361(\pm 0.039)nTA + 0.127(\pm 0.028)nCrS + 1.050(\pm 0.110)DBI$$

$$N = 18, R^2 = 0.911, R_{adj}^2 = 0.892, Q_{LOO}^2 = 0.842, Q_{LMO(20\%)}^2 = 0.780, \overline{r_{m(LOO)}^2} = 0.786, \Delta r_{m(LOO)}^2 = 0.037, MAE(95\%) = 0.078, RMSE = 0.090, Prediction\ Quality = Moderate$$

The number of data points in case of nitrofurans was very less and not suitable for data set division into training and test sets. Thus, small dataset modeling was used for robust model development where data set division is not worthy. The MLR model developed showed good determination coefficient (R^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) for internal validation. The leave-many-out predicted variance (Q_{LMO}^2) was also calculated. The descriptors appearing in the model are: **nTA** (number of terminal atoms), **nCrS** (number of ring secondary C(sp³)) and **DBI** (Dragon branching index).

The descriptor **nTA** belonging to the constitutional type has a negative contribution towards radiosensitization effectiveness; thus, compounds having higher number of terminal atoms will have lower $pC_{1.6}$ value and vice versa. This can be explained with compounds **NF-10** and **NF-12**. In compound **NF-10**, which has a lower value for $pC_{1.6}$ ($pC_{1.6} = 1.523$), the number of terminal atoms is 7 (in higher side) (**Figure 4.65-a**). Again, compound **NF-12** which has a lower number of terminal atoms ($nTA = 3$) shows higher $pC_{1.6}$ value ($pC_{1.6} = 2.09691$).

nCrS represents the number of sp³ hybridised secondary carbon present in a ring system. The positive contribution implicates that with an increase in $nCrS$ values, radiosensitization effectiveness will increase. This has been observed in compounds like **NF-18** ($nCrS = 3$) (**Figure 4.65-a**) and **NF-12** ($nCrS = 1$) where presence of such secondary carbon has increased the $pC_{1.6}$ value ($pC_{1.6} = 2$ and 2.097 respectively).

Another positively correlated descriptor, **DBI**, represents the branching nature of the compound. With an increase in the branching index, the radiosensitization will increase as observed in compound **NF-13** ($DBI = 4.301$, $pC_{1.6} = 2.097$) (**Figure 4.65-a**).

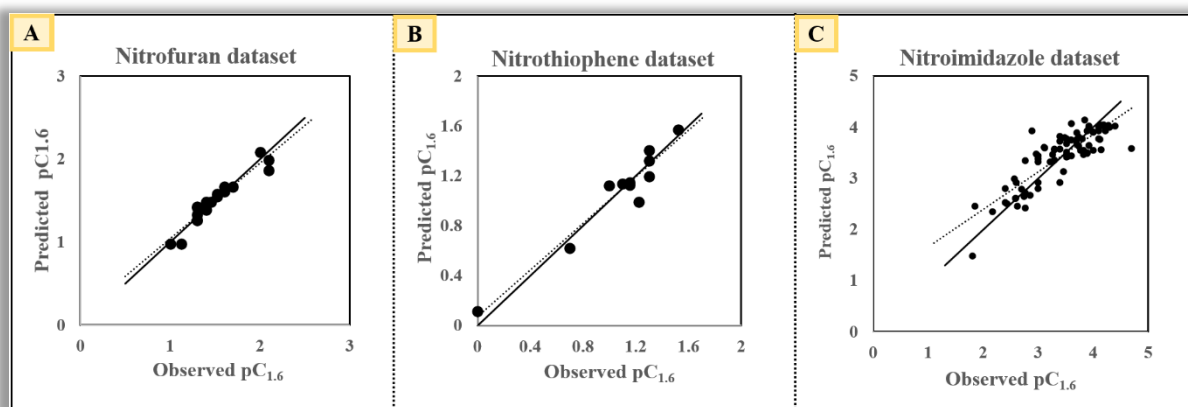


Figure 4.64. Observed vs predicted $pC_{1.6}$ scatter plot for the local nitro datasets.

4.7.1.2. QSAR model studying radiosensitization effectiveness of Nitrothiophenes

$$pC_{1.6} = 0.816 + 0.191nCs + 4.555MATS4v$$

$$N = 11, R^2 = 0.933, Q_{LOO}^2 = 0.807, Q_{LMO(20\%)}^2 = 0.896, \overline{\Delta r_{m(LOO)}^2} = 0.660, \Delta r_{m(LOO)}^2 = 0.178, MAE_{Fitted} = 0.081, MAE_{LOO} = 0.124, RMSE = 0.101, Prediction\ Quality = Moderate$$

Small dataset modeling was utilised again owing to the limited number of compounds in the dataset. The developed PLS model has two descriptors and one latent variable: **nCs** (total number of secondary carbon (sp^3)) and **MATS4v** (Moran autocorrelation of lag 4 weighted by van der Waals volume). From the VIP plot (**Figure 4.66-a**), it was found that MATS4v has higher VIP score than nCs denoting that MATS4v is of higher significance than nCs. The predicted variance explained by specific features for each latent variable is given in the Supplementary section.

MATS4v is a 2D autocorrelation descriptor, which represents the distribution mode of the atomic van der Waals volumes along the topological structure of nitrothiophenes. Here, the path connecting a pair of atoms has length 4 and applies the atomic van der Waals volumes as weighting scheme. The positive regression coefficient advocates that a higher positive value of the descriptor enhances the radiosensitivity as observed in compound **NT-9** ($MATS4v = 0.068914$, $pC_{1.6} = 1.30103$) (**Figure 4.65-b**).

nCs is a functional group count descriptor and has a positive correlation with radiosensitization effectiveness. A secondary carbon is one which is bound by two other carbon atoms. Increase in the number of such fragments in nitrothiophenes will increase their radiosensitivity. This is observed in compounds like **NT-5** ($nCs = 3$, $pC_{1.6} = 1.522879$) (**Figure 4.65-b**) and **NT-6** ($nCs = 4$, $pC_{1.6} = 1.30103$).

4.7.1.3. QSAR model studying radiosensitization effectiveness of Nitroimidazoles

$$pC_{1.6} = 0.873 - 1.267C - 042 - 0.227H - 051 + 0.287B09[C - C] + 3.115PDI$$

$$N = 84, R^2 = 0.733, R_{adj}^2 = 0.723, Q_{LOO}^2 = 0.701, Q_{LMO(20\%)}^2 = 0.696, \overline{r_{m(LOO)}^2} = 0.588, \Delta r_{m(LOO)}^2 = 0.193, MAE = 0.251, RMSE = 0.321, Prediction\ Quality = Moderate$$

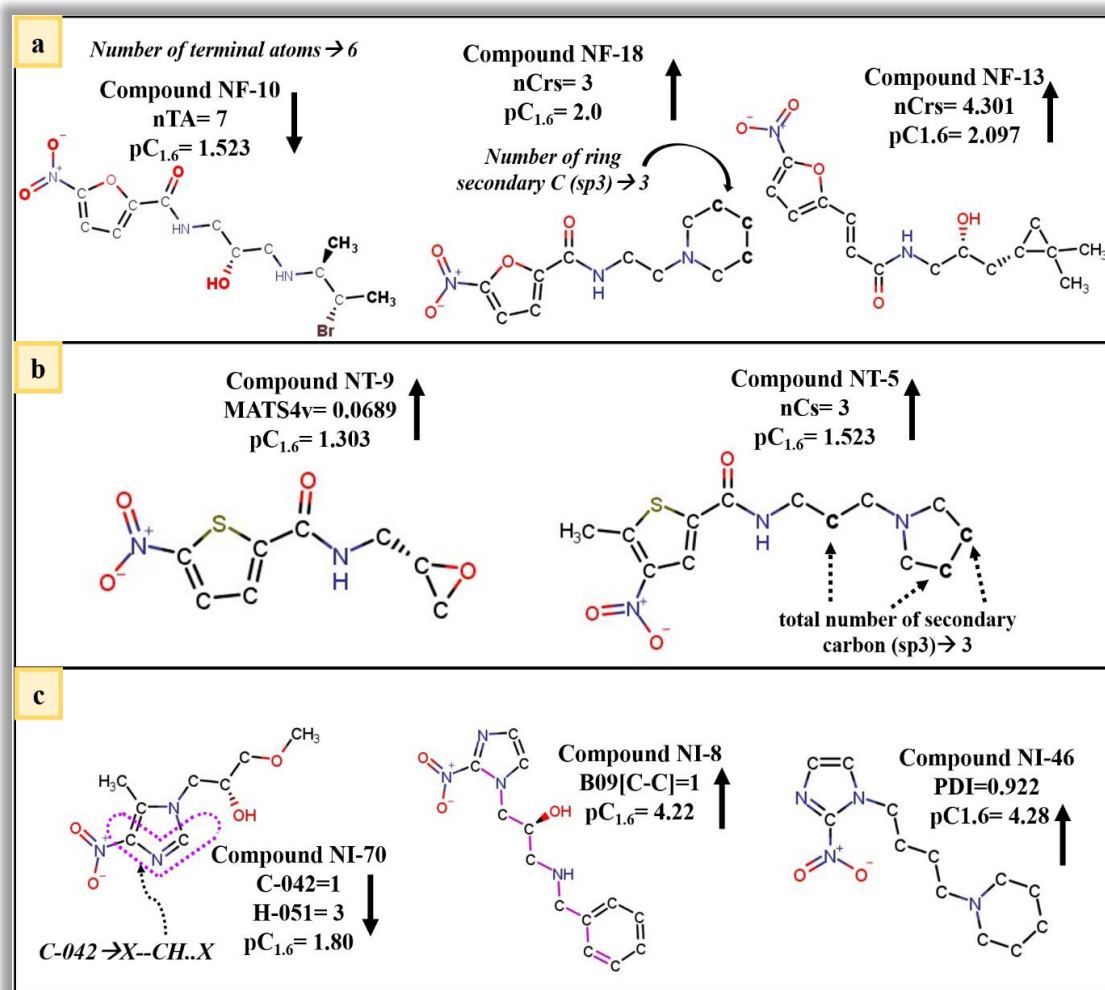


Figure 4.65. Contribution of the descriptors obtained in local nitro dataset modeling towards radiosensitization effectiveness ($pC_{1.6}$ values).

A four descriptor PLS model with 3 latent variables (LVs) was developed for the nitroimidazole dataset. Here, the number of compounds in the dataset was relatively higher and could be divided into training and test sets for model development. However, this dataset was earlier used by our group for model development in a previously published literature [37] where division of this dataset provided acceptable results. Here, we have tried modeling for the whole dataset using GA-MLR method of variable selection followed by BSS method of model development. The descriptors selected in the best MLR model was further subjected to PLS regression with 3 LVs which showed good determination coefficient (R^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) as given in model 3. From the VIP plot (Figure 4.66-b), the significance of the descriptors are as follows: C-042, B09[C-C], H-051 and PDI.

The descriptor **C-042** is an atom-centred fragment descriptor representing the fragment X--CH..X (Figure), where X is any electronegative atom (O, N, S, P, Se, halogens); '--' is an aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group; and '..' is an aromatic single bond as the C-N bond in pyrrole. The negative regression coefficient implicates that increase in the number of such type of fragments in the nitroimidazole analogues will hinder its radiosensitivity. This has been observed in compounds like **NI-69** (C-042=1, $pC_{1.6}= 1.85$) and **NI-70** (C-042=1, $pC_{1.6}= 1.80$) (Figure 4.65-c) where the presence of C-042 fragment caused a lowering in $pC_{1.6}$ values.

The next descriptor is **B09[C-C]**, a 2D atom pair descriptor denoting the presence or absence of C-C fragment at the topological distance 9. The positive coefficient of this descriptor implies that the value of B09[C-C] is directly proportional to the radiosensitization effectiveness, which is established from the presence of such fragments in most of the active compounds (e.g., compounds **NI-8** and **NI-11**) (Figure 4.65-c).

Another atom-centred fragment descriptor, **H-051** corresponds to H attached to alpha carbon (where alpha carbon is any carbon attached through a single bond with -C=X, -C#X, -C-X). This descriptor also contributes negatively towards the radiosensitive effectiveness; thus, with an increase in the descriptor value, $pC_{1.6}$ value will decrease. This can be explained with compound number **NI-70** where there are three such H-051(H-051= 3) fragments and $pC_{1.6}$ is low ($pC_{1.6}= 1.80$).

The last descriptor for this model is **PDI** or packing density index is a molecular property descriptor. PDI is described as the ratio between the McGowan volume and the total surface area (Pirovano et al., 2015). The descriptor has a positive correlation with $pC_{1.6}$ thereby implicating an enhancing effect on radiosensitivity. This is observed in compounds **NI-46** (PDI=0.922; $pC_{1.6}= 4.28$) and **NI-44** (PDI=0.927; $pC_{1.6}= 4.12$) (Figure 4.65-c).

The loading plot of the two local PLS models are given in the Figure 4.67.

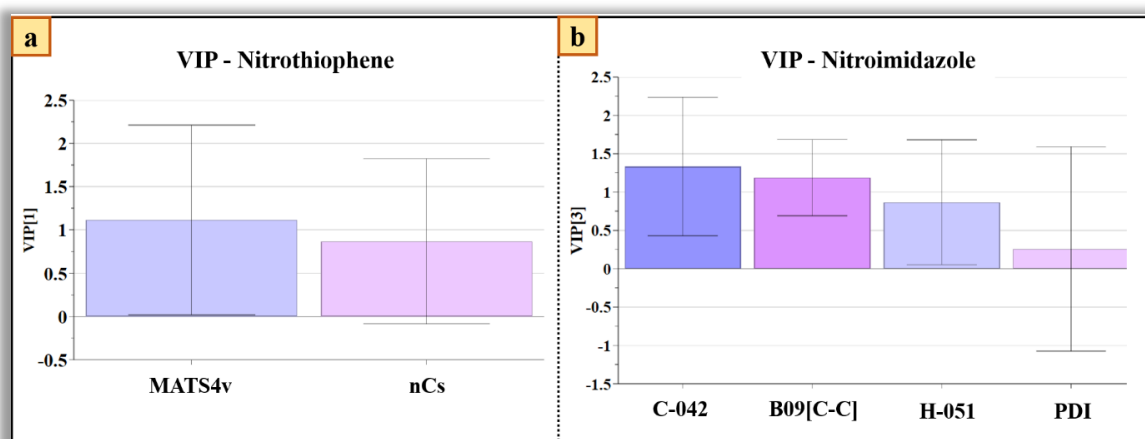


Figure 4.66. Variable importance plot of local nitrothiophene and nitroimidazole datasets.

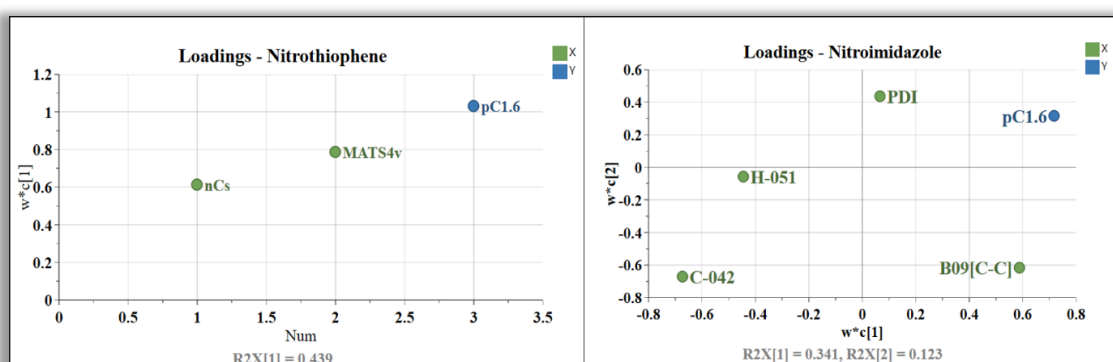


Figure 4.67. Loading plot of local nitrothiophene and nitroimidazole datasets

4.7.2. Modeling the global nitroaromatics dataset

The global dataset, i.e., the dataset containing all the compounds from the individual local datasets was subjected to modeling using Dragon descriptors. The dataset was divided into training and test sets by the Kennard-Stone method of data division, and then the DCV-GA method was utilised for feature selection. The final model was developed using the Best Subset Selection (BSS) method followed by PLS regression. The PLS model with 3 LVs derived exhibited 88.1% variance for the training set (86.5% in terms of leave one out variance) and 92.5% for the test set variance (in terms of Q_{F1}^2 or R_{pred}^2). The observed versus predicted $pC_{1.6}$ plot for the global model is given in **Figure 4.68**. The residuals of the observed and predicted $pC_{1.6}$ values for some compounds were on the higher side as evident from the scatter plot. However, it was found that all the training set and test set compounds were inside the domain of applicability which will be discussed in later section.

$$pC_{1.6} = 1.256 + 1.318nImidazole + 0.951C - 044 + 0.354B09[C - C] - 0.459B03[O - S]$$

$$N_{train} = 79, R^2 = 0.881, R_{adj}^2 = 0.876, Q_{LOO}^2 = 0.865, Q_{LMO(20\%)}^2 = 0.866, \\ \overline{r_{m(LOO)}^2} = 0.806, \Delta r_{m(LOO)}^2 = 0.101, MAE(train) = 0.262, SD(train) \\ = 0.256, RMSEC = 0.344, Quality_{Train} = Good$$

$$N_{test} = 34, Q_{F1}^2 = 0.925, Q_{F2}^2 = 0.899, \overline{r_{m(LOO)}^2} = 0.863, \Delta r_{m(LOO)}^2 = 0.006, CCC \\ = 0.948, MAE(Test) = 0.301, SD(Test) = 0.211, RMSEP = 0.366, Quality_{Test} \\ = Good$$

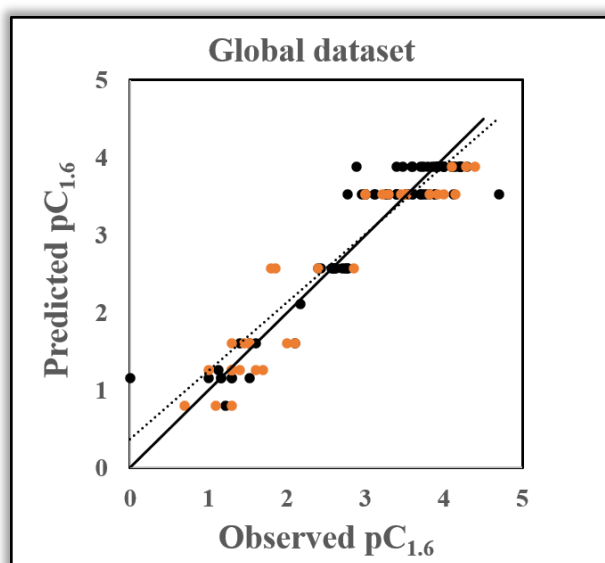


Figure 4.68: Scatter plot of the global model

The model is constituted of four descriptors, viz., nImidazole, C-044, B09[C-C] and B03[O-S]. From the VIP plot (**Figure 4.69-a**), the descriptors are in the following order of significance: C-044, nImidazole, B03[O-S] and B09[C-C].

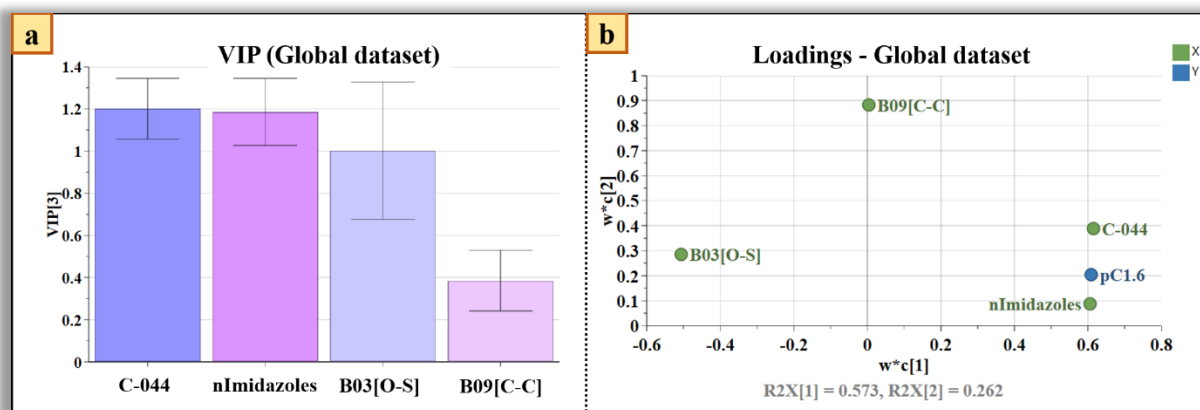


Figure 4.69. Variable importance plot and loading plot of the global model

The first descriptor is **C-044**, which is an atom centre fragment descriptor and represented as X--CX..X, where X is any electronegative atom (O, N, S, P, Se, halogens); '--' is an aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group; and '..' is an aromatic single bond as the C-N bond in pyrrole. Compounds showing positive values for the C-044 descriptor were found to have a specific fragment in their structure, i.e., O=NC-N. Here the O=N fragment represents the delocalized bonds in the nitro group and C-N is an aromatic single bond in pyrrole giving an idea of the 2-nitroimidazole fragment (**Figure 4.70**). Hence, the descriptor C-044 provides a knowledge that nitroimidazoles are better radiosensitizers having higher radiosensitization effectiveness. Next, **nImidazole** is a functional group descriptor indicating the number of imidazole present in the compound. The positive correlation gives an idea that imidazole group will increase the compounds' radiosensitivity, leading to a conclusion that nitroimidazoles are better radiosensitizers than nitrofurans or nitrothiophenes (**Figure 4.70**).

Another 2D atom pair descriptor **B03[O-S]** describes the presence or absence of O-S fragment at a topological distance 3. It has a negative correlation with radiosensitization effectiveness denoting that with the presence of such fragment pC_{1.6} value decreases as in compounds **NS-1** (pC_{1.6} = 1.0) (**Figure 4.70**) and **NS-2** (pC_{1.6} = 0.0).

The next descriptor is **B09[C-C]**, a 2D atom pair descriptor, which denotes the presence or absence of C-C fragment at the topological distance 9. The positive coefficient indicates that presence of C-C fragment at distance 9 will enhance pC_{1.6} values as seen in compounds like **NI-51** (pC_{1.6} = 4.3) (**Figure 4.70**) and **NI-8** (pC_{1.6} = 4.22).

From the descriptors obtained in the global dataset, it can be inferred that nitroimidazoles are better radiosensitizers than nitrofurans or nitrothiophene analogues. Although the global model gives any idea regarding the superiority of nitroimidazoles giving better radiosensitization, the model actually takes into account a diverse group of chemicals. Also, division of dataset gives us more reliance regarding the predictivity of the model.

The loading plot explains the relationship between the descriptors (or the X-variables) with the response (or the Y-variable). The first two latent variables were utilised for the development of the plot. Through a loading plot, the impact of the descriptors on the response can be understood. Descriptors having similar meaning are grouped together close to one another. This can be explained by **Figure 4.69-b** where descriptors C-044 and nImidazole are grouped together and they almost

impart the same meaning (contribution of imidazole group). Compounds with high impact on the model are situated far from the plot origin (e.g., C-044 and nImidazole).

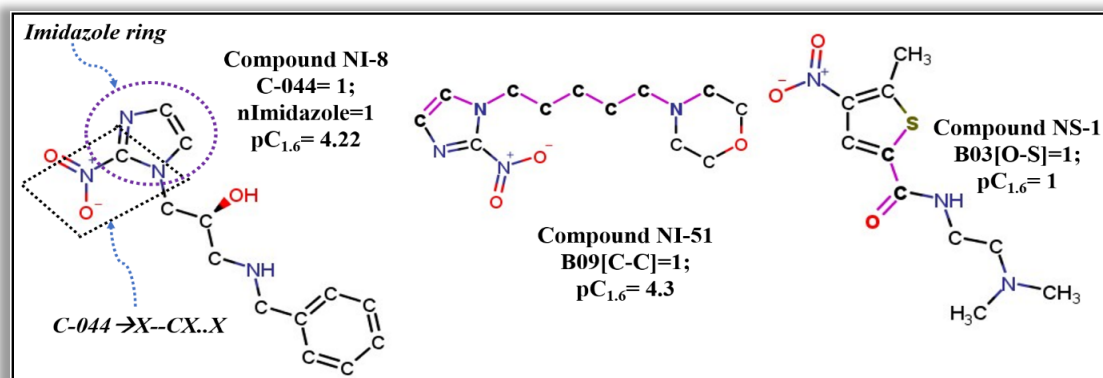


Figure 4.70. Contribution of the descriptors appearing in the global model

Golbraikh and Tropsha's Criteria

We have calculated the Golbraikh-Tropsha's criteria (Golbraikh & Tropsha, 2002) for all the local models as well as for the global model and reported in **Table 4.20**. All the models developed in the present study passed the criteria.

Table 4.20. Golbraikh and Tropsha's criteria for all local and global models.

| Metrics | Acceptable range | Local Nitrofurans | Local Nitrothiophene | Local Nitroimidazole | Global dataset |
|------------------------|------------------------------|-------------------|----------------------|----------------------|----------------|
| r^2 | >0.6 | 0.911 | 0.933 | 0.733 | 0.881 |
| Q^2 | >0.5 | 0.842 | 0.896 | 0.701 | 0.865 |
| $ r_0 - r_0'^2 $ | <0.3 | 0.008 | 0.067 | 0.094 | 0.016 |
| k | 0.85 < k < 1.15 | 1 | 1 | 1 | 1 |
| $[(r^2 - r_0^2)/r^2]$ | $[(r^2 - r_0^2)/r^2] < 0.1$ | 0 | 0 | 0 | 0 |
| k' | 0.85 < k' < 1.15 | 0.997 | 0.992 | 0.992 | 1 |
| $[(r^2 - r_0'^2)/r^2]$ | $[(r^2 - r_0'^2)/r^2] < 0.1$ | 0.009 | 0.0045 | 0.129 | 0.018 |

4.7.3. Applicability Domain (AD) assessment

In accordance with OECD guideline 3, any QSAR model should hold a defined domain of applicability. AD can be interpreted as a chemical space defined by the structural information or molecular properties of the chemicals used in the model development (Gadaleta et al., 2016). Any compound which is present within the chemical space can only be properly predicted. In the present study, for the nitrofurans data set, we have used the standardization approach (Kunal Roy et al., 2015). There was no outlier found for the nitrofurans dataset. In case of local nitrothiophenes, local nitroimidazoles and global datasets, we have applied the DModX (distance to model in X-space) method of AD determination at 99% confidence interval ($D\text{-crit} = 0.009999$) using SIMCA 16.0.2 software (<https://landing.umetrics.com/downloads-simca>). The AD plots for the two local datasets given in **Figure 4.71** and **4.72** show that there was no outlier. In case of the global dataset as shown in **Figure 4.73**, it was observed that there was neither any outlier in the training set nor any compound was outside the AD in the test set.

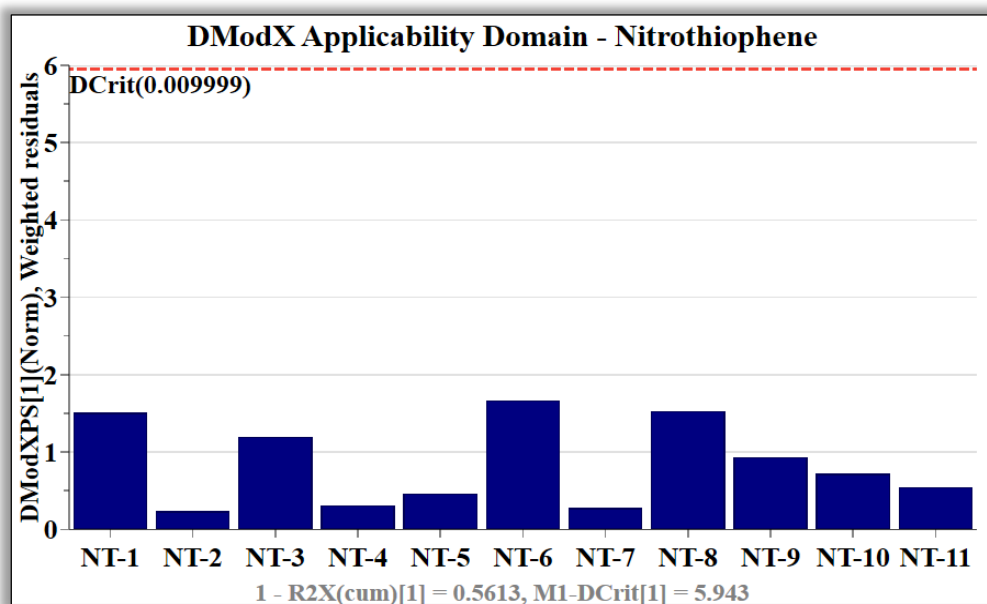


Figure 4.71. DModX Applicability Domain plot of local nitrothiophene dataset

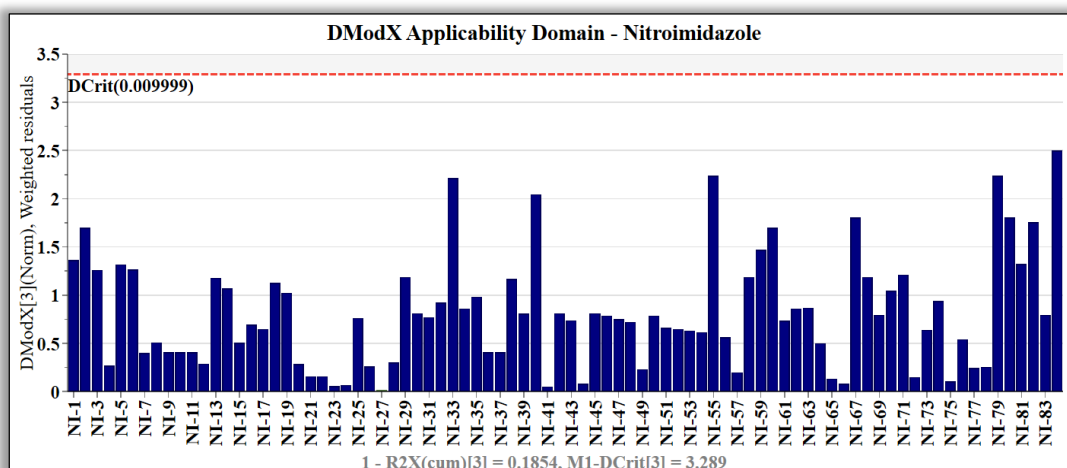


Figure 4.72. DModX Applicability Domain plot of local nitroimidazole dataset

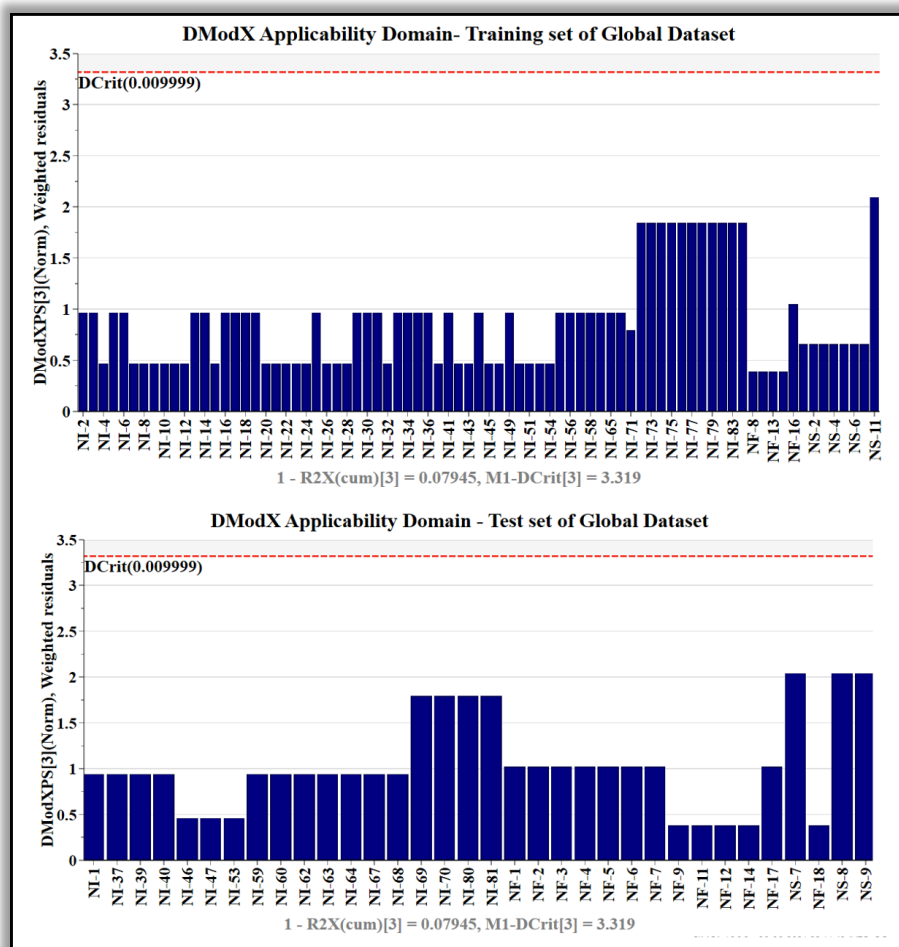


Figure 4.73. Applicability domain plot for the global dataset.

4.7.4. Y-randomization test

The significance of a developed QSAR model is understood by a model randomisation test, and it ensures that the model is not an outcome of a chance correlation (Topliss & Edwards, 1979). During the development of a randomized model, many models are generated by reordering or shuffling different combination of X- or Y-variables (Y-variable here) and accordingly are called X-randomization or Y-randomisation. In the present work, we have used 100 permutations for all the developed models; however, this can be changed according to the choice of the user. Models which are randomly developed with y-variable shuffling should have very poor statistics. The R_y^2 intercept should not exceed 0.3 and the Q_y^2 intercept should not exceed 0.05. The metrics for the randomised models given in **Table 4.21.** and Supplementary **Figures 4.74, 4.75** and **4.76** indicate that the local and global models developed are not out of chance correlation and are robust for suitable predictions.

Table 4.21. Y-Randomization model metrics for the developed local and global models.

| Models | | R_y^2 | $Q_{(LOO)y}^2$ |
|--------|----------------|---------|----------------|
| Local | Nitrofuran | 0.1727 | -0.3979 |
| | Nitrothiophene | -0.0311 | -0.262 |
| | Nitroimidazole | -0.0132 | -0.248 |
| Global | | -0.0305 | -0.246 |

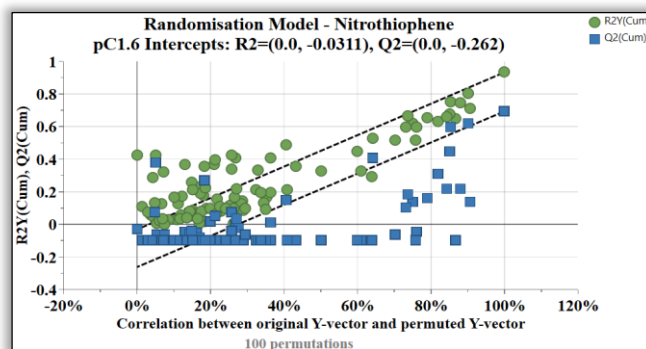


Figure 4.74. Y-randomization plot of local nitrothiophene dataset

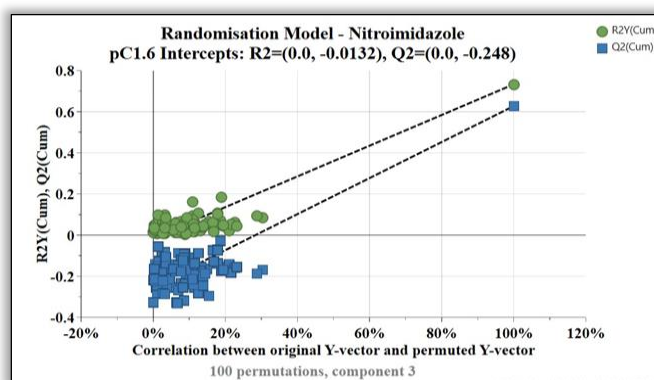


Figure 4.75. Y-randomization plot of local nitroimidazole dataset

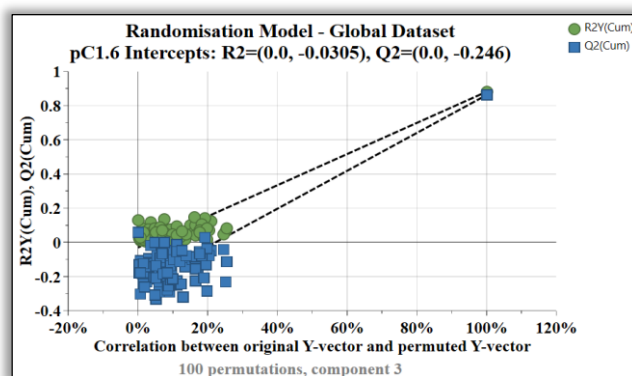
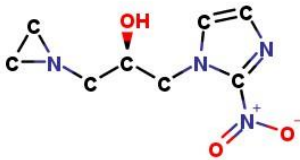
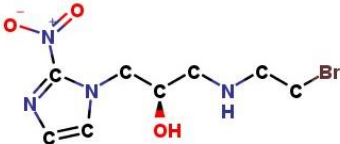


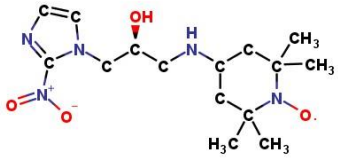
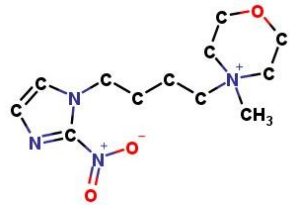
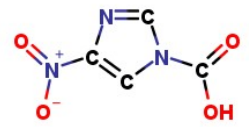
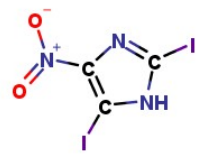
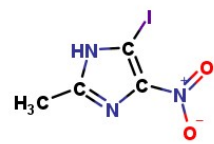
Figure 4.76. Y-randomization plot of global dataset

4.7.5. True External Prediction using the global model

The global model can be considered the best model here, owing to the diversity of the nitro compounds used for modeling. Further, to analyse the predictivity of the developed global model, we have considered a set of external compounds for prediction (**Table 4.22**). Predictions for these compounds were further verified by the application of “*Prediction Reliability Indicator*” tool (Roy et al., 2018) available from <https://dtclab.webs.com/software-tools>. The PRI results showed that predictions for all the 10 compounds were ‘Good’ (with Composite Score 3) and all the compounds were inside the AD of the model (**Table 4.22**). Based on the insights obtained, it can be inferred that developed global model can be used for the prediction of radiosensitization effectiveness in nitro compounds, especially for nitroimidazole derivatives. We have further computed predictions using the global model for another external dataset retrieved from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>) and checked the quality of predictions using the PRI tool. We have reported the results in the Supplementary Section (**Table 4.23**). Of note, the prediction quality was found to be good for all the external compounds. It will be interesting to verify the predictions experimentally in the future.

Table 4.22. Predicting $pC_{1.6}$ values of a true external dataset using the global model.

| Compound ID | Structure | Observed $pC_{1.6}$ | Predicted $pC_{1.6}$ (Global model) | Composite Score | Prediction Quality | AD status |
|-------------|---|---------------------|-------------------------------------|-----------------|--------------------|-----------|
| 1 |  | - | 3.879 | 3 | Good | In |
| 2 |  | - | 3.879 | 3 | Good | In |

| | | | | | | |
|---|---|------|-------|---|------|----|
| 3 |  | 4.05 | 3.879 | 3 | Good | In |
| 4 |  | 2.89 | 3.879 | 3 | Good | In |
| 5 |  | - | 2.574 | 3 | Good | In |
| 6 |  | - | 3.525 | 3 | Good | In |
| 7 |  | - | 2.574 | 3 | Good | In |

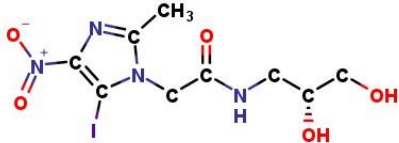
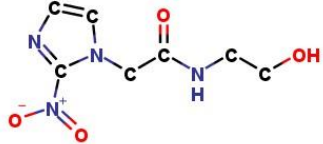
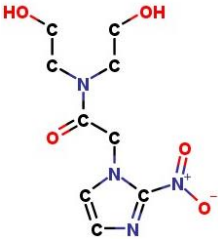
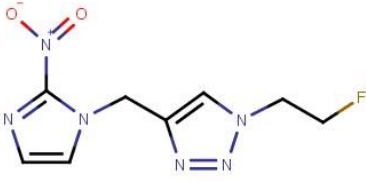
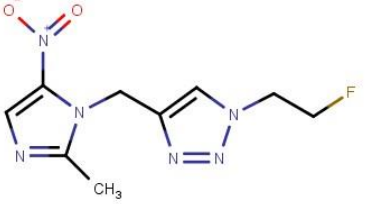
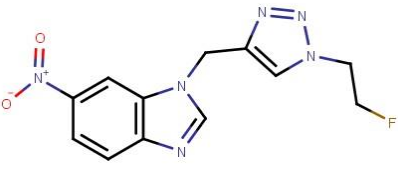
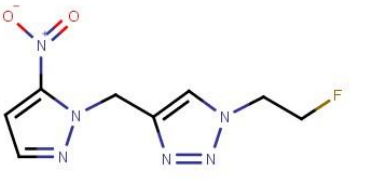
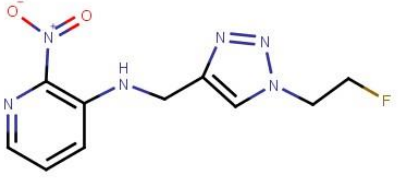
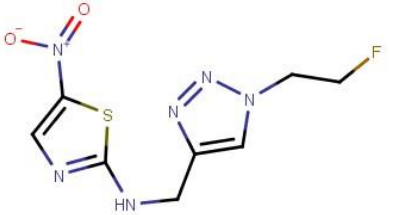
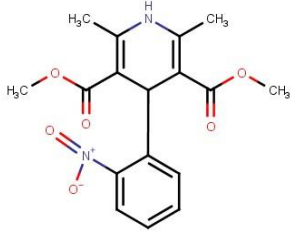
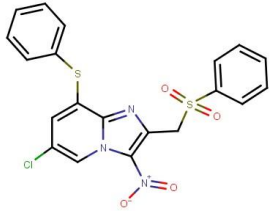
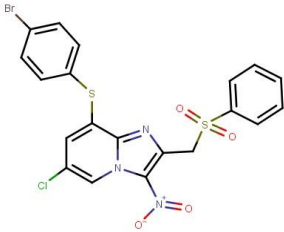
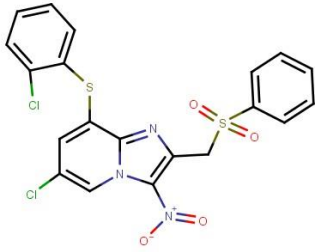
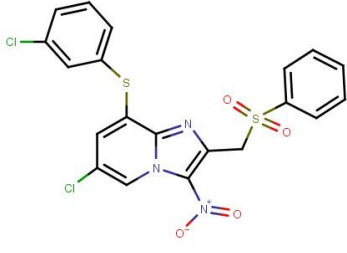
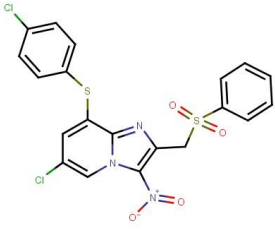
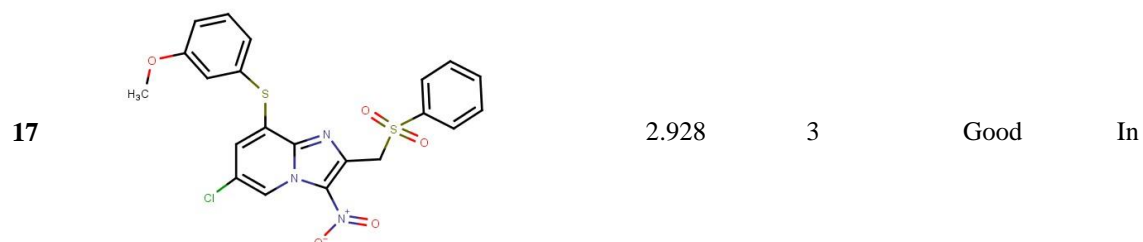
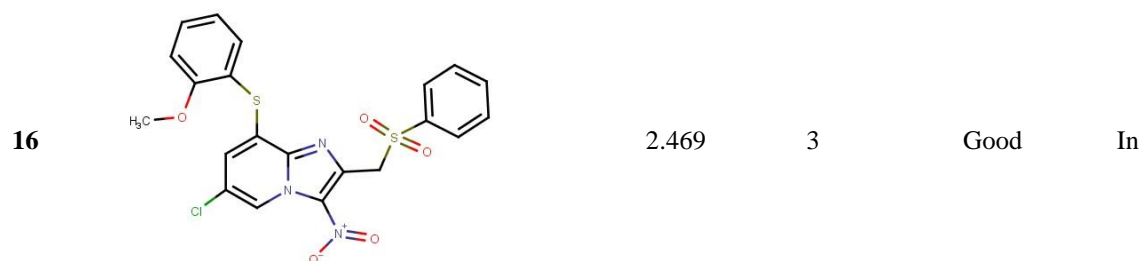
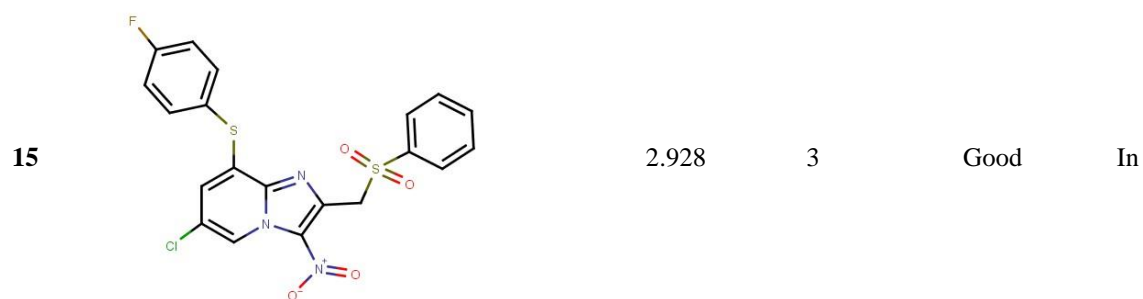
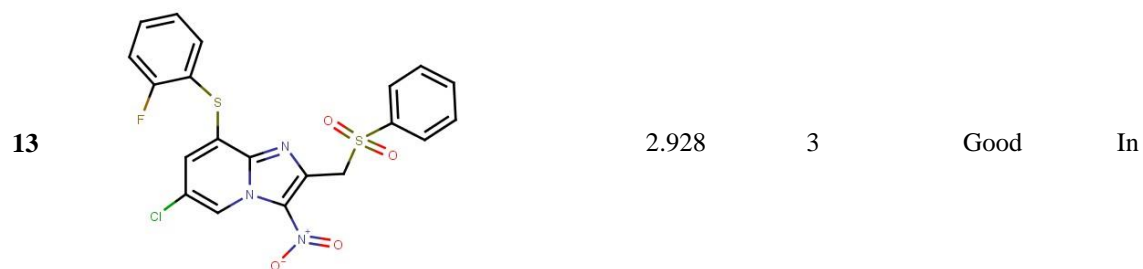
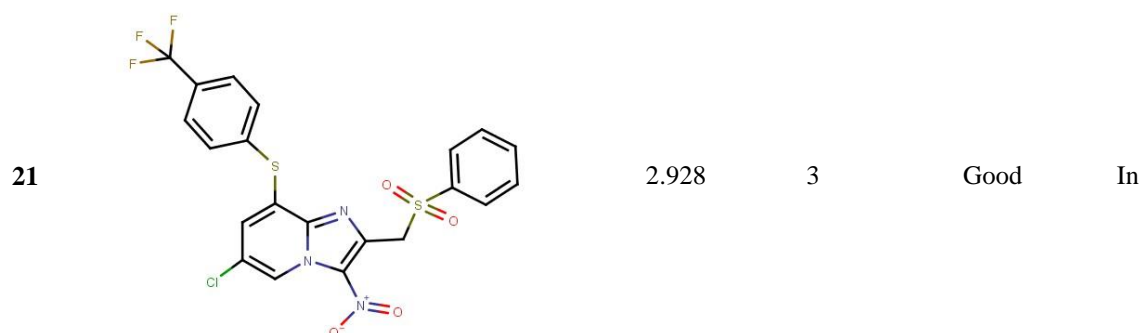
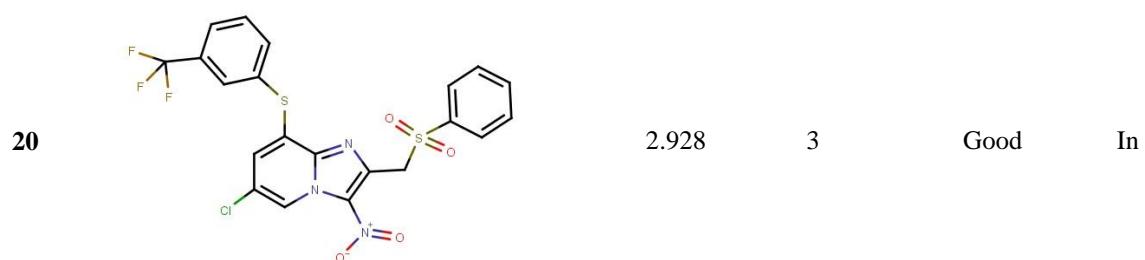
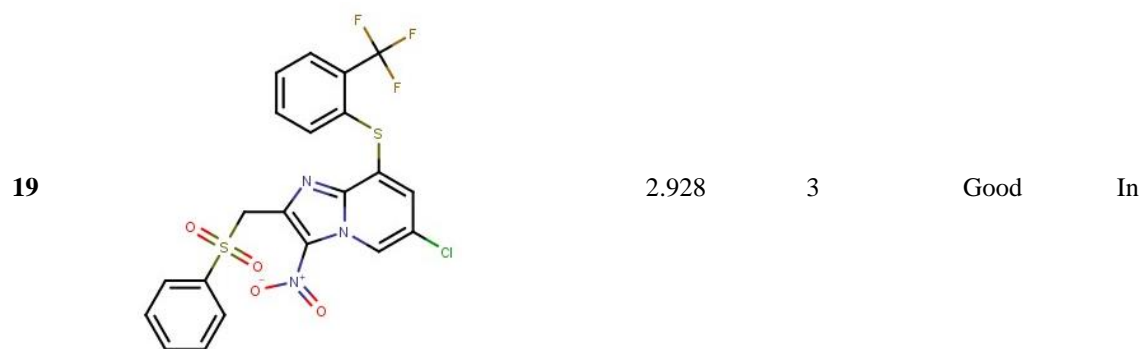
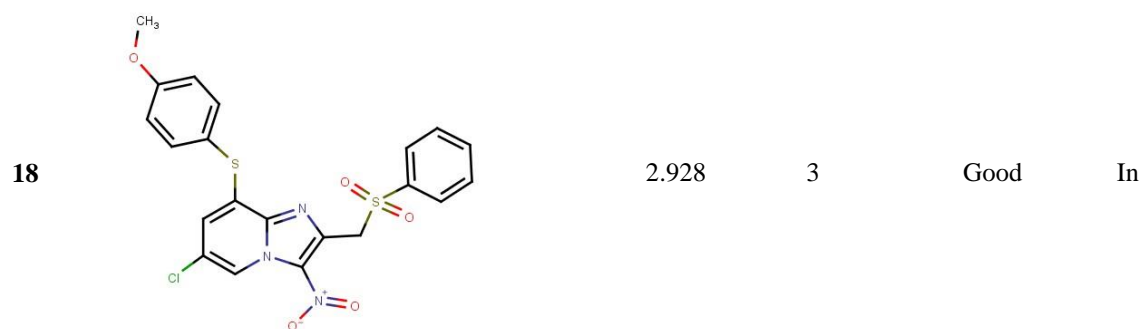
| | | | | | | |
|----|---|---|-------|---|------|----|
| 8 |  | - | 2.928 | 3 | Good | In |
| 9 |  | - | 3.879 | 3 | Good | In |
| 10 |  | - | 3.879 | 3 | Good | In |

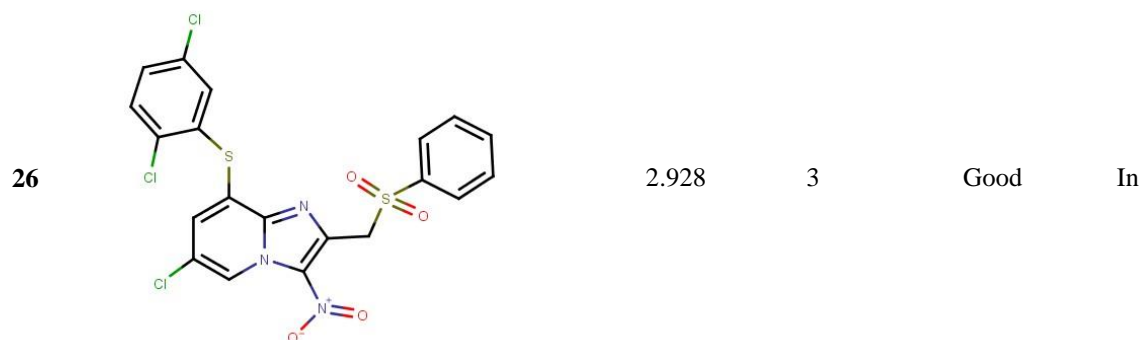
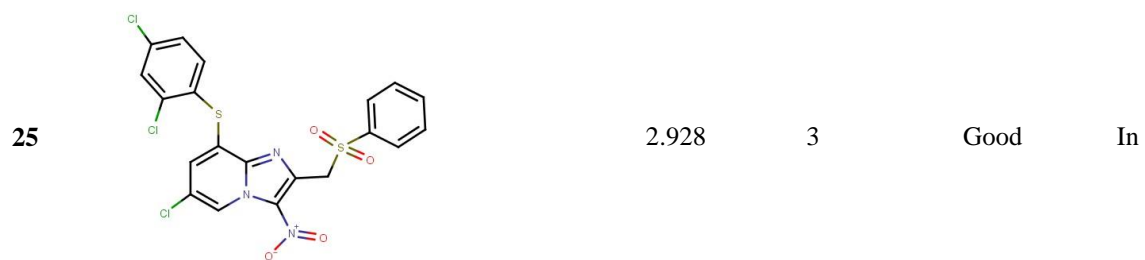
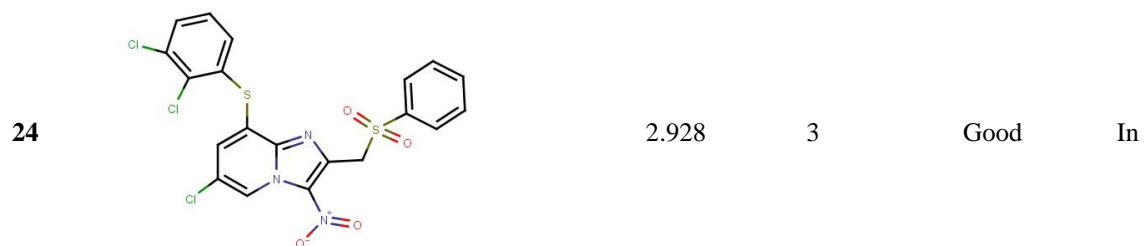
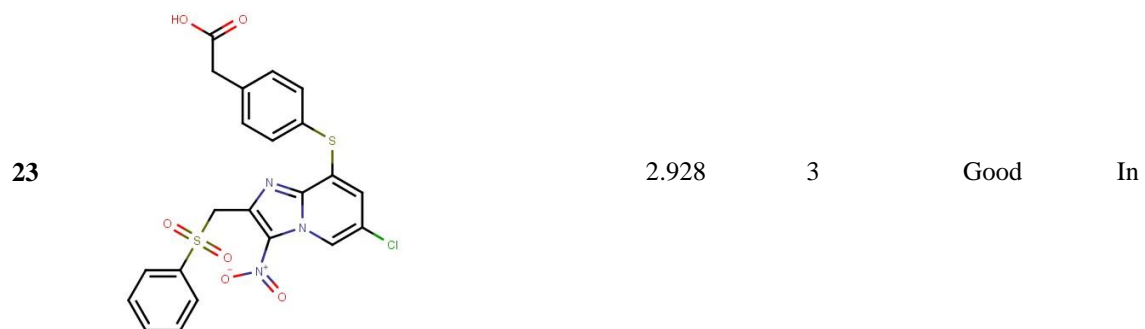
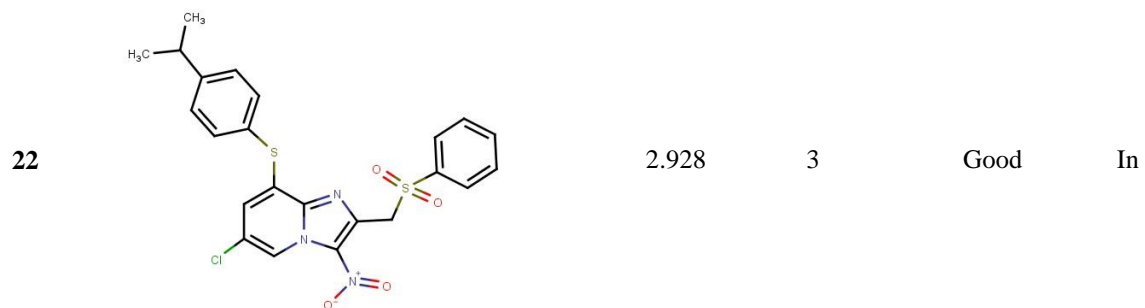
Table 4.23. Prediction quality of external nitroaromatic dataset retrieved from ChEMBL database.

| Serial No | Structure | Predicted pC1.6 | Composite Score | Prediction Quality | AD status |
|-----------|---|-----------------|-----------------|--------------------|-----------|
| 1 |  | 3.525 | 3 | Good | In |
| 2 |  | 2.574 | 3 | Good | In |
| 3 |  | 2.928 | 3 | Good | In |
| 4 |  | 1.256 | 3 | Good | In |
| 5 |  | 1.610 | 3 | Good | In |
| 6 |  | 2.103 | 3 | Good | In |

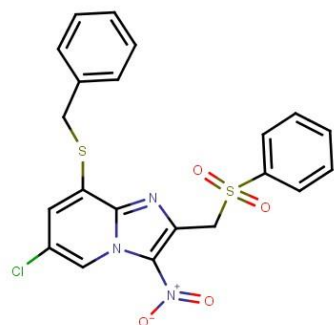
| | | | | | |
|----|---|-------|---|------|----|
| 7 |  | 1.256 | 3 | Good | In |
| 8 |  | 2.928 | 3 | Good | In |
| 9 |  | 2.928 | 3 | Good | In |
| 10 |  | 2.928 | 3 | Good | In |
| 11 |  | 2.928 | 3 | Good | In |
| 12 |  | 2.928 | 3 | Good | In |







27



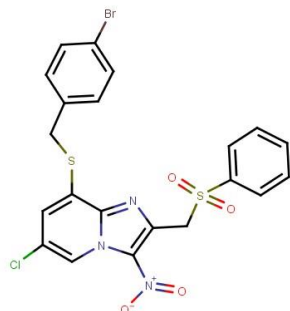
2.928

3

Good

In

28



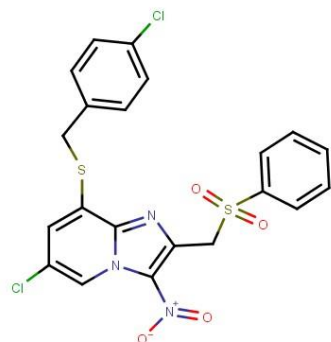
2.928

3

Good

In

29



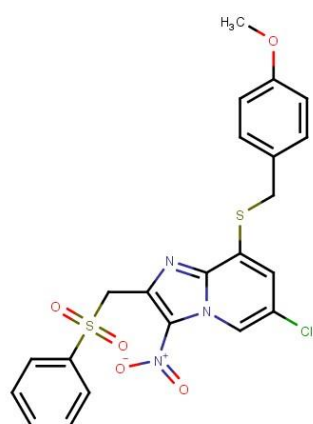
2.928

3

Good

In

30

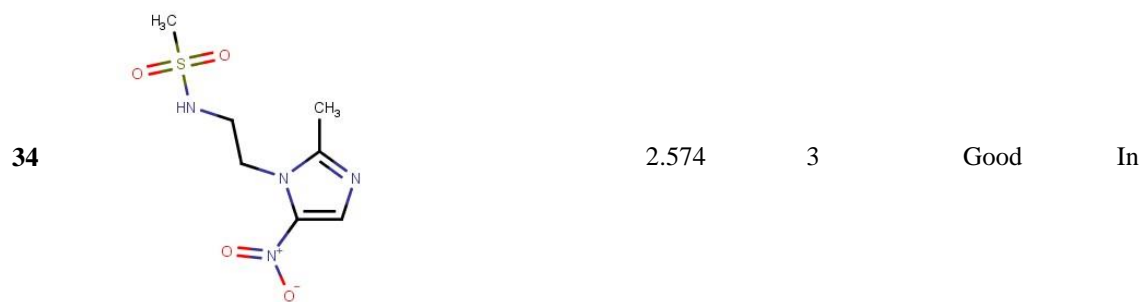
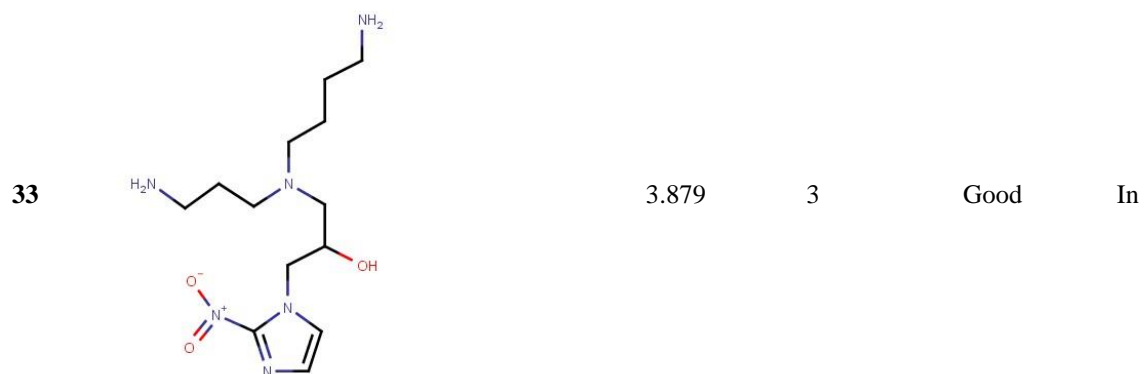
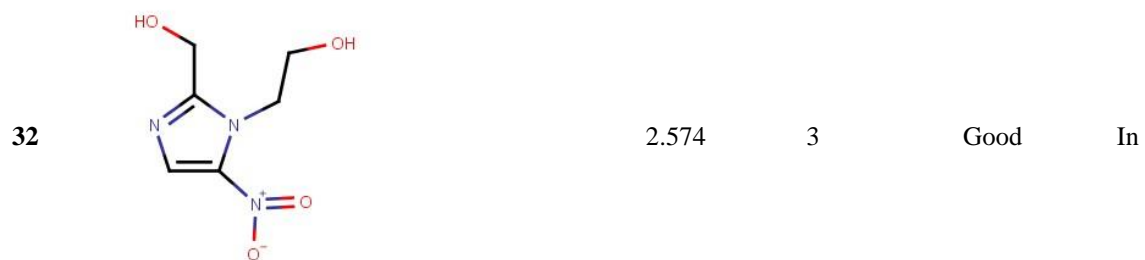
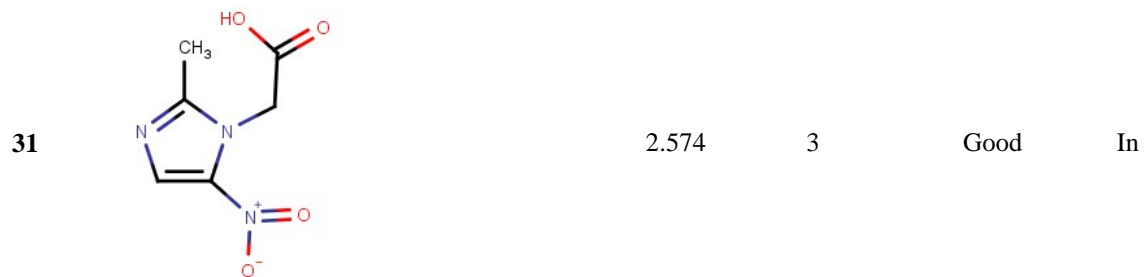


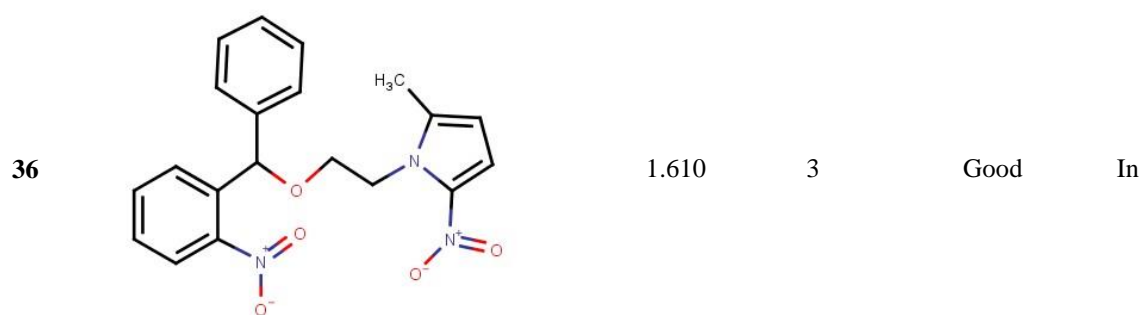
2.928

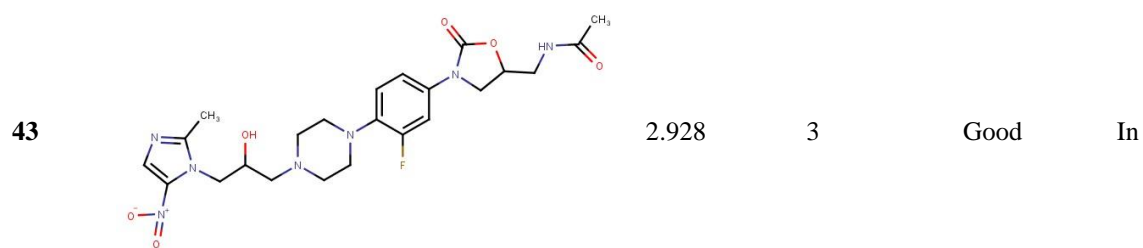
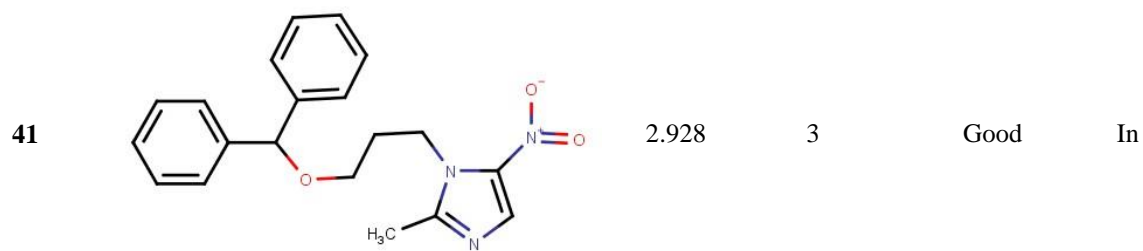
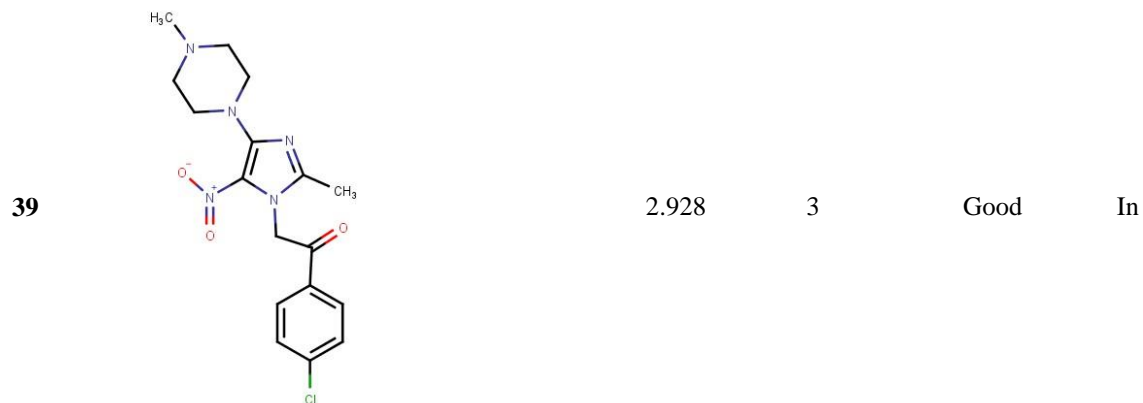
3

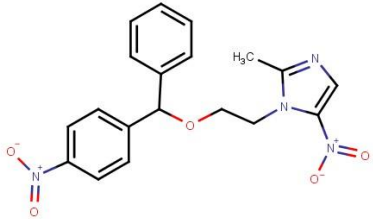
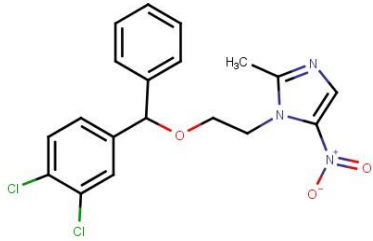

Good

In







| | | | | | |
|----|---|-------|---|------|----|
| 44 |  | 2.928 | 3 | Good | In |
| 45 |  | 2.928 | 3 | Good | In |
| 46 |  | 2.928 | 3 | Good | In |
| 47 |  | 2.928 | 3 | Good | In |

CHAPTER 5

CONCLUSIONS

Chapter 5: Conclusions

Innovative scientific solutions provide the central strength for meeting the needs of novel applications. It is admissibly true that the success of any research is the portrayal of the conclusion drawn from the results obtained of the analysis addressing the given problem which can explore the revealed and/or unrevealed scientific explanation. This can guide in developing a better understanding of the problem analysed and has the potential to contribute ideas of new avenues associated with the areas of interest. In the present research several *in silico* techniques were employed to study the nature and chemistry of PET and SPECT imaging. The application of QSARs in the design of these imaging agents led us to understand the structural features essential in imaging agent binding essential for the diagnosis of various neurodegenerative diseases and cancer. In this research, exploration of PET and SPECT imaging of various neurodegenerative diseases like Alzheimer's disease and Parkinson's disease was enacted, targeting their concerned receptors like amyloid beta, tau protein, adenosine ($A_{2A}R$) receptor, dopamine (D2) and vesicular acetylcholine transporter (VACHT). **Studies 1-4** concern study of the binding affinity of PET or SPECT imaging agents against different aforementioned receptors associated with neurodegenerative diseases. Further, the effect of nitroaromatic compounds as potential radiosensitizer molecules in the treatment of hypoxia, a common pathophysiology of cancer, was also studied. **Studies 5-7** explore different structural features of nitroaromatics like nitroimidazoles, nitrofurans, nitrothiophenes to study various radiosensitization parameters like radiosensitization effectiveness ($pC_{1.6}$), sensitivity enhancement ratio (SER) and survival ratio (SR).

Further, all the computational models (QSAR models) developed in this work were appropriately validated through stringent internal and external validation techniques. These models were also subjected to randomization tests to avoid chance correlation. Thus, these models can be efficiently used in screening query molecules or databases, predicting the biological activity of newly designed PET or SPECT compounds, and last but not the least, for designing new analogues with improved activity. Finally, the precise information revealed or knowledge gained from all of the studies that were performed in this thesis work is described individually as follows:

5.1. Study 1: Application of multi-layered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease

The present research used chemometric tools for investigating the binding affinity of PET and SPECT against $A\beta$ plaques and tau protein. The three QSAR models developed through DCV method in this study give knowledge about the essential structural requirements necessary for improved binding affinity against $A\beta$ plaques and tau fibril. Many of the imaging agents used for modeling inhibits plaque formation, in addition to just binding to β -amyloid. Thus, these compounds can also be considered as multifunctional imaging agents (useful for both binding and inhibition) (Darras & Pang, 2017). Double Cross Validation proved its efficacy in modeling large dataset compounds previously (De & Roy, 2018; Khan et al., 2019). In the present study we have utilized small size datasets (<50 compounds in two cases) where DCV has proved its competence in searching for optimum combination of descriptors for generating models with good predictive ability. Thus, it can be concluded that DCV can not only be applied in modeling of large datasets but it is also suitable for modeling smaller dataset compounds. Furthermore, new sets of designed PET and SPECT imaging agents with better predicted binding properties are reported in the current report. Further experiments might be conducted in future on these potential compounds.

5.2. Study 2: Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: A QSAR approach

Parkinson's disease is a neurodegenerative disease affecting the elderly person around the world. An important target for its treatment is blocking adenosine A_{2A} receptor which is co-located with the D_2 receptor and is pharmacologically opposite in motor function. Many studies hint that blocking A_{2A} receptor would be a beneficial strategy in the treatment of PD. Thus, this work endeavours exploring QSAR analysis to correlate the chemical structures with their biological activity with the aim to filter the essential chemical features of an antagonist for selectivity and binding affinity to A_{2A} receptor. The computational approach used in this work consists firstly the calculation of the molecular descriptors, and secondly, correlating these descriptors with the binding affinity and selectivity using different chemometric tools such as Genetic Function Algorithm (GFA), Best Subset Selection (BSS) method and Intelligent consensus predictor (ICP) tools. The statistical quality of the models was checked using traditional metrics both internally and externally. We have also discussed about the contributions of the descriptors in the light of known binding mechanisms such as π - π stacking interaction, hydrophobic interaction and hydrogen bonding with the different protein residues present in the receptor binding sites. From the insights obtained from such mechanism, we found that electronegative atoms and presence of aromatic ring like benzene are favorable for enhancing the binding affinity to the A_{2A} receptor. Further the docking studies supported with the conclusions found in the QSAR studies. In conclusion, the study highlights the pharmacophoric features mainly responsible for antagonizing adenosine receptors that can be further modified for better binding and selectivity to A_{2A} receptor. In case of selectivity also, electronegativity and aromaticity of the compounds play essential and influential roles. The simple two-dimensional (2D) descriptors appeared in all the models are easier to compute requiring no conformation analysis or energy minimization process. Thus, this information would help in the future development and synthesis of newer PET tracer targeted towards adenosine receptor.

5.3. Study 3: QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting Dopamine receptor

In vivo imaging targeting dopamine receptor is a subject of extensive studies nowadays. Dopamine plays a vital role controlling the pathophysiology of Parkinson's disease. Hence, it can be treated as a suitable target in controlling the disease. The present study aims in the development of a 2D QSAR model of a group of 34 PET imaging agents having affinity towards dopamine D_2 receptor. The 2D QSAR model developed is simple and interpretable and provides knowledge about the basic structural features required for good dopamine binding. The use of simple two-dimensional descriptors reduces the need of time-consuming computational approaches of conformational analysis or energy minimization; thus, the developed model may be suitable for the quick screening purposes.

5.4. Study 4: Computational modeling of PET imaging agents against vesicular acetylcholine transporter (VACHT) protein binding affinity: Application of 2D-QSAR modeling and molecular docking techniques

The neurotransmitter acetylcholine (ACh) plays a ubiquitous role in cognitive functions including learning and memory with widespread innervation in the cortex, subcortical structures and the cerebellum. Cholinergic receptors, transporters, or enzymes associated with many neurodegenerative diseases, including Alzheimer's disease (AD) and Parkinson's disease (PD), are potential imaging targets. In the present study, we have developed 2D quantitative structure-activity relationship (2D-

QSAR) models for 19 positron emission tomography (PET) imaging agents targeted against presynaptic vesicular acetylcholine transporter (VACHT). In our work, we aimed to understand the important structural features of the PET imaging agents required for their binding with VACHT. This was done by feature selection using Genetic Algorithm followed by Best Subset Selection method and developing a Partial Least Squares- based 2D QSAR model using the best feature combination. The developed QSAR model showed significant statistical performance and reliability. Using the features selected in the 2D-QSAR analysis, we have also performed similarity-based chemical read-across predictions and obtained encouraging external validation statistics. From the developed QSAR model, it was found that the presence of nitrogen in the PET tracer molecule potentiates the binding affinity towards VACHT receptor. This was further confirmed by molecular docking studies where nitrogen in piperidine moiety produced attractive charge interaction with **Asp A:483** amino acid of VACHT. In future this study will help in the prediction of newly developed compounds targeted towards VACHT.

5.5. Study 5: Exploration of nitroimidazoles as radiosensitizers: Application of multi-layered feature selection approach in QSAR modeling

This study targets for the development of fragment based 2D-QSAR models for predicting radiosensitization of nitroimidazole derivatives. The simplex descriptors give an insight about the fragments and their proper position in the nitroimidazole ring that enhance or decline the radiosensitization effectiveness. Also reduction in the large data pool by using multi-layered variable selection is shown for better handling of large pool of descriptors and removing chances of intercorrelation among them. Further, the newly developed models were used for prediction of eight external compounds and their prediction reliability was checked.

5.6. Study 6: QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: Application of Small Dataset Modelling

This study aims at developing 2D-QSAR models with the notion to investigate the essential features in nitroimidazole sulphonamide analogues to show radiosensitization properties with respect to sensitizer enhancement ratio and survival ratio endpoints. The different descriptors obtained give an idea about the position of the features and type of chemical groups required to enhance or hinder these properties. Moreover, QSAAR modelling helps in correlating two endpoints (SER and logSR) and suggests how to extrapolate an endpoint if the experimental information is unavailable. The current study emphasizes on the application of ‘Small Dataset Modeller’ software when the dataset is small and splitting of dataset is not worthy. Further, the newly developed models were used for prediction of 14 compounds and their prediction reliability was checked. These developed QSAR and QSAAR models are able to predict newly developed nitroimidazole sulphonamide derivatives with known structural features. The complete overview of the work is explained in **Figure 5.1**.

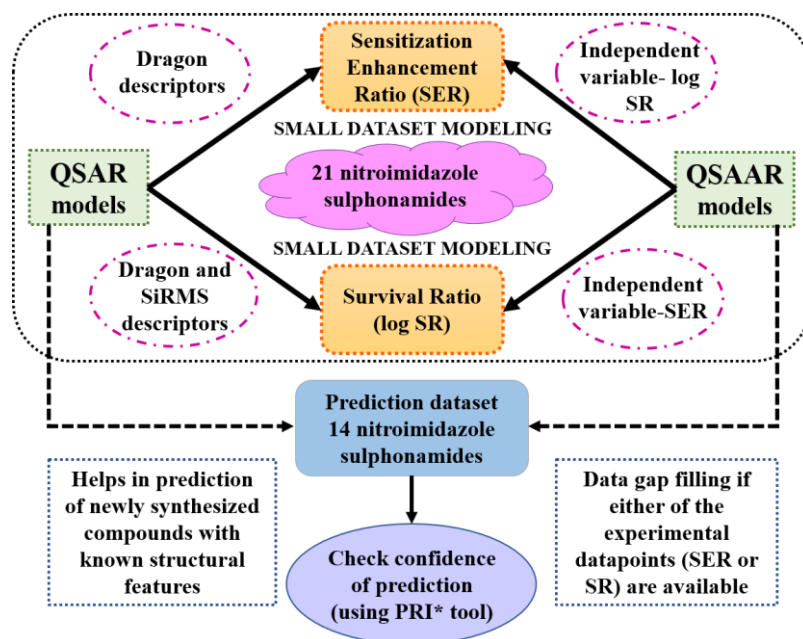


Figure 5.1. Overview of the present work involving the development of QSAR and QSAAR model using Small Dataset Modeller.

5.7. Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness

The present study targets for the development of 2D-QSAR models for three nitroaromatics datasets both locally and globally predicting radiosensitization effectiveness. The local models gave us an idea about the structural features required for effective radiosensitization within their own group while the global model imparted an insight regarding which type of nitroaromatic compounds are more efficient to produce better radiosensitization. The descriptors obtained in the global model clearly implicated that nitroimidazoles are better radiosensitizers as compared with nitrofuran or nitrothiazole derivatives. Moreover, all the developed local and global models were statistically sound and well validated. The global model was further used for the prediction of ten true external set compounds, and their prediction reliability was analysed using the PRI tool.

REFERENCES

References

- Adak, S., Bhalla, R., Vijaya Raj, K. K., Mandal, S., Pickett, R., & Luthra, S. K. (2012). Radiotracers for SPECT imaging: Current scenario and future prospects. *Radiochimica Acta*, *100*(2), 95–107.
- Agarwal, S., & Mehrotra, R. (2016). An overview of Molecular Docking. *JSM Chem*, *4*(2), 1024.
- Aiken, L. S., West, S. G., & Pitts, S. C. (2003). Multiple Linear Regression. *Handbook of Psychology*, 481–507.
- Akarachantachote, N., Saithanu, K., Chadcham, S., Akarachantachote, N., Chadcham, S., & Saithanu, K. (2014). Cutoff threshold of variable importance in projection for variable selection. *International Journal of Pure and Applied Mathematics*, *94*(3), 307–322.
- Alagille, D., Dacosta, H., Baldwin, R. M., & Tamagnan, G. D. (2011). 2-Arylimidazo[2,1-b]benzothiazoles: A new family of amyloid binding agents with potential for PET and SPECT imaging of Alzheimer's brain. *Bioorganic & Medicinal Chemistry Letters*, *21*(10), 2966–2968.
- Alkorta, I., Rozas, I., & Elguero, J. (1998). Non-conventional hydrogen bonds. *Chemical Society Reviews*, *27*(2), 163–170.
- Ambure, P., Gajewicz-Skretna, A., Cordeiro, M. N. D. S., & Roy, K. (2019). New workflow for QSAR model development from small data sets: Small dataset curator and small dataset modeler. Integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. *Journal of Chemical Information and Modeling*, *59*(10), 4070–4076.
- Ambure, P., & Roy, K. (2015). Exploring structural requirements of imaging agents against A β plaques in Alzheimer's disease: a QSAR approach. *Combinatorial Chemistry & High Throughput Screening*, *18*(4), 411–419.
- Ametamey, S. M., Honer, M., & Schubiger, P. A. (2008). Molecular imaging with PET. *Chemical Reviews*, *108*(5), 1501–1516.
- Anderson, C. J., & Ferdani, R. (2009). Copper-64 Radiopharmaceuticals for PET Imaging of Cancer: Advances in Preclinical and Clinical Research. <https://Home.Liebertpub.Com/Cbr>, *24*(4), 379–393.
- Andrei, S. A., Meijer, F. A., Neves, J. F., Brunsveld, L., Landrieu, I., Ottmann, C., & Milroy, L. G. (2018). Inhibition of 14-3-3/Tau by Hybrid Small-Molecule Peptides Operating via Two Different Binding Modes. *ACS Chemical Neuroscience*, *9*(11), 2639–2654.
- Ankrah, A. O., Span, L. F. R., Klein, H. C., de Jong, P. A., Dierckx, R. A. J. O., Kwee, T. C., Sathekge, M. M., & Glaudemans, A. W. J. M. (2019). Role of FDG PET/CT in monitoring treatment response in patients with invasive fungal infections. *European Journal of Nuclear Medicine and Molecular Imaging*, *46*(1), 174–183.
- Auletta, S., Varani, M., Horvat, R., Galli, F., Signore, A., & Hess, S. (2019). PET Radiopharmaceuticals for Specific Bacteria Imaging: A Systematic Review. *Journal of Clinical Medicine* 2019, Vol. 8, Page 197, *8*(2), 197.
- Baldessarini, R. J., Kula, N. S., Gao, Y., Campbell, A., & Neumeyer, J. L. (1991). R(-)-2-fluoro-n-n-propyl-norapomorphine: A very potent and D2-selective dopamine agonist. *Neuropharmacology*, *30*(1), 97–99.
- Baumann, D., & Baumann, K. (2014). Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *Journal of Cheminformatics*, *6*(1), 1–19.
- Benfenati, E. (2011). *Quantitative structure-activity relationships (QSAR) for pesticide regulatory*

- purposes. Elsevier.
- Blazhenets, G., Frings, L., Ma, Y., Sörensen, A., Eidelberg, D., Wiltfang, J., & Meyer, P. T. (2021). Validation of the Alzheimer Disease Dementia Conversion-Related Pattern as an ATN Biomarker of Neurodegeneration. *Neurology*, *96*(9), e1358–e1368.
- Bonnet, M., Hong, C. R., Wong, W. W., Liew, L. P., Shome, A., Wang, J., Gu, Y., Stevenson, R. J., Qi, W., Anderson, R. F., Pruijn, F. B., Wilson, W. R., Jamieson, S. M. F., Hicks, K. O., & Hay, M. P. (2018). Next-Generation Hypoxic Cell Radiosensitizers: Nitroimidazole Alkylsulfonamides. *Journal of Medicinal Chemistry*, *61*(3), 1241–1254.
- Brown, A. C., & Fraser, T. R. (1868). On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *Journal of Anatomy and Physiology*, *2*(2), 224.
- Cascini, G. L., Niccoli Asabella, A., Notaristefano, A., Restuccia, A., Ferrari, C., Rubini, D., Altini, C., & Rubini, G. (2014). ¹²⁴Iodine: A longer-life positron emitter isotope - New opportunities in molecular imaging. *BioMed Research International*, 2014.
- Chartrand, G., Johns, G. L., & Tian, S. (1993). Detour Distance in Graphs. *Annals of Discrete Mathematics*, *55*(C), 127–136.
- Chatterjee, M., Banerjee, A., De, P., Gajewicz-Skretna, A., & Roy, K. (2022a). A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environmental Science: Nano*, *9*(1), 189–203.
- Chatterjee, M., Banerjee, A., De, P., Gajewicz-Skretna, A., & Roy, K. (2022b). A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environmental Science: Nano*, *9*(1), 189–203.
- Chen, C. J., Bando, K., Ashino, H., Taguchi, K., Shiraishi, H., Shima, K., Fujimoto, O., Kitamura, C., Matsushima, S., Uchida, K., Nakahara, Y., Kasahara, H., Minamizawa, T., Jiang, C., Zhang, M. R., Ono, M., Tokunaga, M., Suhara, T., Higuchi, M., ... Ji, B. (2015). In Vivo SPECT Imaging of Amyloid- β Deposition with Radioiodinated Imidazo[1,2-a]Pyridine Derivative DRM106 in a Mouse Model of Alzheimer's Disease. *Journal of Nuclear Medicine*, *56*(1), 120–126.
- Chen, Z. Y., Wang, Y. X., Lin, Y., Zhang, J. S., Yang, F., Zhou, Q. L., & Liao, Y. Y. (2014). Advance of molecular imaging technology and targeted imaging agent in imaging and therapy. *BioMed Research International*, 2014.
- Chételat, G., Arbizu, J., Barthel, H., Garibotto, V., Law, I., Morbelli, S., van de Giessen, E., Agosta, F., Barkhof, F., Brooks, D. J., Carrillo, M. C., Dubois, B., Fjell, A. M., Frisoni, G. B., Hansson, O., Herholz, K., Hutton, B. F., Jack, C. R., Lammertsma, A. A., ... Drzezga, A. (2020). Amyloid-PET and ¹⁸F-FDG-PET in the diagnostic investigation of Alzheimer's disease and other dementias. *The Lancet Neurology*, *19*(11), 951–962.
- Chin Chung, M., Longhin Bosquesi, P., & Leandro dos Santos, J. (2011). A Prodrug Approach to Improve the Physico-Chemical Properties and Decrease the Genotoxicity of Nitro Compounds. *Current Pharmaceutical Design*, *17*(32), 3515–3526.
- Chirico, N., & Gramatica, P. (2012). Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *Journal of Chemical Information and Modeling*, *52*(8), 2044–2058.
- Chumpradit, S., Kung, H. F., & Kung, M. P. (1993). Fluorinated and iodinated dopamine agents (FIDA): D2 imaging agents for PET and SPECT. *Journal of Labelled Compounds and Radiopharmaceuticals*, *32*(1–12), 223–224.

- Clark, D. E. (2008). What has virtual screening ever done for drug discovery?
- Cohen, A. D., Rabinovici, G. D., Mathis, C. A., Jagust, W. J., Klunk, W. E., & Ikonovic, M. D. (2012). Using Pittsburgh Compound B for In Vivo PET Imaging of Fibrillar Amyloid-Beta. *Advances in Pharmacology*, *64*, 27–81.
- Congreve, M., Andrews, S. P., Doré, A. S., Hollenstein, K., Hurrell, E., Langmead, C. J., Mason, J. S., Ng, I. W., Tehan, B., Zhukov, A., Weir, M., & Marshall, F. H. (2012). Discovery of 1,2,4-triazine derivatives as adenosine A_{2A} antagonists using structure based drug design. *Journal of Medicinal Chemistry*, *55*(5), 1898–1903.
- Consonni, V., & Todeschini, R. (2010). Molecular descriptors. In *Challenges and Advances in Computational Chemistry and Physics* (Vol. 8, pp. 29–102). Springer.
- Cronin, M. T. D., Walker, J. D., Jaworska, J. S., Comber, M. H. I., Watts, C. D., & Worth, A. P. (2003). Use of QSAR in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environmental Health Perspectives*, *111*(10), 1376–1390.
- Dar, A. M., & Mir, S. (2017). Molecular docking: approaches, types, applications and basic challenges. *Journal of Analytical & Bioanalytical Techniques*, *8*(2), 1–3.
- Darras, F. H., & Pang, Y. P. (2017). On the use of the experimentally determined enzyme inhibition constant as a measure of absolute binding affinity. *Biochemical and Biophysical Research Communications*, *489*(4), 451–454.
- De, P., Kar, S., Roy, K., & Leszczynski, J. (2018). Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environmental Science: Nano*, *5*(11), 2742–2760.
- De, P., & Roy, K. (2018a). Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. *SAR and QSAR in Environmental Research*, *29*(4), 319–337.
- De, P., Aher, R. B., & Roy, K. (2018). Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector *Aedes aegypti* : application of ETA indices. *RSC Advances*, *8*(9), 4662–4670.
- De, P., Bhattacharyya, D., & Roy, K. (2020). Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling. *Structural Chemistry*, *31*(3), 1043–1055.
- De, P., Kar, S., Ambure, P., & Roy, K. (2022). Prediction reliability of QSAR models: an overview of various validation tools. *Archives of Toxicology*, *96*(5), 1279–1295.
- De, P., & Roy, K. (2021). QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: application of small dataset modeling. *Structural Chemistry*, *32*(2), 631–642.
- Dearden, J. C., Cronin, M. T. D., & Kaiser, K. L. E. (2010). How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR).
- Declercq, L., Celen, S., Lecina, J., Ahamed, M., Tousseyn, T., Moechars, D., Alcazar, J., Ariza, M., Fierens, K., Bottelbergs, A., Mariën, J., Vandenberghe, R., Andres, I. J., Van Laere, K., Verbruggen, A., & Bormans, G. (2016). Comparison of new tau PET-tracer candidates with [18F]T808 and [18F]T807. *Molecular Imaging*, *15*, 1–15.
- Devillers, J. (1996). *Genetic algorithms in molecular modeling*. Academic press.
- Donnelly, D. J., Preshlock, S., Kaur, T., Tran, T., Wilson, T. C., Mhanna, K., Henderson, B. D., Batalla, D., Scott, P. J. H., & Shao, X. (2022). Synthesis of Radiopharmaceuticals via “In-Loop” ¹¹C-Carbonylation as Exemplified by the Radiolabeling of Inhibitors of Bruton’s Tyrosine

- Kinase. *Frontiers in Nuclear Medicine*, 1.
- Douglas, A. P., Thursky, K. A., Worth, L. J., Drummond, E., Hogg, A., Hicks, R. J., & Slavin, M. A. (2019). FDG PET/CT imaging in detecting and guiding management of invasive fungal infections: a retrospective comparison to conventional CT imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(1), 166–173.
- Du, Y., Liang, Y., & Yun, D. (2002). Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *Journal of Chemical Information and Computer Sciences*, 42(6), 1283–1292.
- Echeverría, J. (2017). Alkyl groups as electron density donors in π -hole bonding. *CrystEngComm*, 19(42), 6289–6296.
- Free, S. ., & Wilson, J. W. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, 7(4), 395–399.
- Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376), 817–823.
- Fuchigami, T., Yamashita, Y., Kawasaki, M., Ogawa, A., Haratake, M., Atarashi, R., Sano, K., Nakagaki, T., Ubagai, K., Ono, M., Yoshida, S., Nishida, N., & Nakayama, M. (2015). Characterisation of radioiodinated flavonoid derivatives for SPECT imaging of cerebral prion deposits. *Scientific Reports 2015 5:1*, 5(1), 1–11.
- Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A., & Nicolotti, O. (2016). Applicability Domain for QSAR Models: Where Theory Meets Reality. *IJQSPR*, 1(1), 45–63.
- Gao, Y., Baldessarini, R., Kula, N., & Neumeyer, J. (1990). Synthesis and dopamine receptor affinities of enantiomers of 2-substituted apomorphines and their N-n-propyl analogs. *Journal of Medicinal Chemistry*, 33(6), 1800–1805.
- Gao, Y., Ram, V., Campbell, A., Kula, N., Baldessarini, R., & Neumeyer, J. (1990). Synthesis and structural requirements of N-substituted norapomorphines for affinity and activity at dopamine D-1, D-2, and agonist receptor sites in rat brain. *Journal of Medicinal Chemistry*, 33(1), 39–44.
- Gajewicz-Skretna, A., Kar, S., Piotrowska, M., & Leszczynski, J. (2021). The kernel-weighted local polynomial regression (KwLPR) approach: an efficient, novel tool for development of QSAR/QSAAR toxicity extrapolation models. *Journal of cheminformatics*, 13, 1-20.
- Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. De, Lee, K. H., & Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design 2003 17:2*, 17(2), 241–253.
- Golbraikh, A., & Tropsha, A. (2000). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Molecular Diversity 2000 5:4*, 5(4), 231–243.
- Golbraikh, A., & Tropsha, A. (2002). Beware of q²! *Journal of Molecular Graphics and Modelling*, 20(4), 269–276.
- Golmohammadi, H., Dashtbozorgi, Z., & Acree, W. E. (2012). Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *European Journal of Pharmaceutical Sciences*, 47(2), 421–429.
- Gong, L., Zhang, Y., Liu, C., Zhang, M., & Han, S. (2021). Application of Radiosensitizers in Cancer Radiotherapy. *International Journal of Nanomedicine*, 16, 1083. <https://doi.org/10.2147/IJN.S290438>
- Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR &*

- Combinatorial Science*, 26(5), 694–701.
- Grassi, I., Nanni, C., Allegri, V., Morigi, J. J., Montini, G. C., Castellucci, P., & Fanti, S. (2012). The clinical use of PET with ¹¹C-acetate. *American Journal of Nuclear Medicine and Molecular Imaging*, 2(1), 33.
- Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* 1962 194:4824, 194(4824), 178–180.
- Harada, R., Okamura, N., Furumoto, S., & Yanai, K. (2018). Imaging protein misfolding in the brain using β -sheet ligands. *Frontiers in Neuroscience*, 12(AUG), 585.
- Hashimoto, H., Kawamura, K., Takei, M., Igarashi, N., Fujishiro, T., Shiomi, S., Watanabe, R., Muto, M., Furutsuka, K., Ito, T., Yamasaki, T., Yui, J., Nemoto, K., Kimura, Y., Higuchi, M., & Zhang, M. R. (2015). Identification of a major radiometabolite of [¹¹C]PBB3. *Nuclear Medicine and Biology*, 42(12), 905–910.
- Hatori, A., Yui, J., Yamasaki, T., Xie, L., Kumata, K., Fujinaga, M., Yoshida, Y., Ogawa, M., Nengaki, N., Kawamura, K., Fukumura, T., & Zhang, M. R. (2012). PET Imaging of Lung Inflammation with [¹⁸F]FEDAC, a Radioligand for Translocator Protein (18 kDa). *PLOS ONE*, 7(9), e45065.
- Herholz, K., & Ebmeier, K. (2011). Clinical amyloid imaging in Alzheimer's disease. *The Lancet Neurology*, 10(7), 667–670.
- Hocke, C., Prante, O., Salama, I., Hübner, H., Löber, S., Kuwert, T., & Gmeiner, P. (2008). ¹⁸F-Labeled FAUC 346 and BP 897 Derivatives as Subtype-Selective Potential PET Radioligands for the Dopamine D3 Receptor. *ChemMedChem*, 3(5), 788–793.
- Jaakola, V. P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y. T., Lane, J. R., Ijzerman, A. P., & Stevens, R. C. (2008). The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*, 322(5905), 1211–1217.
- Jackson, J. E. (2005). *A user's guide to principal components*. John Wiley & Sons.
- Kar, S., Gajewicz, A., Roy, K., Leszczynski, J., & Puzyn, T. (2016). Extrapolating between toxicity endpoints of metal oxide nanoparticles: Predicting toxicity to Escherichia coli and human keratinocyte cell line (HaCaT) with Nano-QTTR. *Ecotoxicology and environmental safety*, 126, 238-244.
- Khan, K., Benfenati, E., & Roy, K. (2019). Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the DrugBank database compounds. *Ecotoxicology and Environmental Safety*, 168, 287–297.
- Khan, P. M., & Roy, K. (2018). Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR).
- Körner, W. (1874). Studi sulla Isomeria delle Così Dette Sostanze Aromatiche a Sei Atomi di Carbonio. *Gazz Chim It*, 4, 242.
- Kovac, M., Mavel, S., Deuther-Conrad, W., Méheux, N., Glöckner, J., Wenzel, B., Anderluh, M., Brust, P., Guilloteau, D., & Emond, P. (2010). 3D QSAR study, synthesis, and in vitro evaluation of (+)-5-FBVM as potential PET radioligand for the vesicular acetylcholine transporter (VACHT). *Bioorganic & Medicinal Chemistry*, 18(21), 7659–7667.
- Krause, W., Jordan, A., Scholz, R., & Jimenez, J. L. M. (2005). Iodinated Nitroimidazoles as Radiosensitizers. *Anticancer Research*, 25(3B), 2145–2151.
- Kubinyi, H. (1976). Quantitative structure-activity relationships. IV. Non-linear dependence of

- biological activity on hydrophobic character: a new model. *Arzneimittel-Forschung*, 26(11), 1991–1997. <https://europepmc.org/article/med/1037231>
- Kung, H. F., Choi, S. R., Qu, W., Zhang, W., & Skovronsky, D. (2010). 18F stilbenes and styrylpyridines for PET imaging of A β plaques in Alzheimer's disease: A miniperspective. *Journal of Medicinal Chemistry*, 53(3), 933–941.
- Kung, M. P., Hou, C., Zhuang, Z. P., Skovronsky, D. M., Zhang, B., Gur, T. L., Trojanowski, J. Q., Lee, V. M. Y., & Kung, H. F. (2002). Radioiodinated styrylbenzene derivatives as potential SPECT imaging agents for amyloid plaque detection in Alzheimer's disease. *Journal of Molecular Neuroscience* 2002 19:1, 19(1), 7–10.
- Kuz'min, V. E., Artemenko, A. G., Polischuk, P. G., Muratov, E. N., Hromov, A. I., Liahovskiy, A. V., Andronati, S. A., & Makan, S. Y. (2005). Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. *Journal of Molecular Modeling*, 11(6), 457–467.
- Lau, J., Rousseau, E., Kwon, D., Lin, K. S., Bénard, F., & Chen, X. (2020). Insight into the Development of PET Radiopharmaceuticals for Oncology. *Cancers* 2020, Vol. 12, Page 1312, 12(5), 1312.
- Lee, H. W., Hong, S. B., & Tae, W. S. (2000). Opposite ictal perfusion patterns of subtracted SPECT Hyperperfusion and hypoperfusion. *Brain*, 123(10), 2150–2159.
- Lessigiarska, I., Worth, A. P., Netzeva, T. I., Dearden, J. C., & Cronin, M. T. D. (2006). Quantitative structure–activity–activity and quantitative structure–activity investigations of human and rodent toxicity. *Chemosphere*, 65(10), 1878–1887.
- Li, Y., Zhang, W., Wu, H., & Liu, G. (2014). Advanced tracers in PET imaging of cardiovascular disease. *BioMed Research International*, 2014.
- Liu, H., Yao, X., Liu, M., Hu, Z., & Fan, B. (2007). Prediction of gas-phase reduced ion mobility constants (K₀) based on the multiple linear regression and projection pursuit regression. *Talanta*, 71(1), 258–263.
- Long, W., & Liu, P. (2010). Quantitative Structure Activity Relationship Modeling for Predicting Radiosensitization Effectiveness of Nitroimidazole Compounds. *Journal of Radiation Research*, 51(5), 563–572.
- Luo, S., Zhang, E., Su, Y., Cheng, T., & Shi, C. (2011). A review of NIR dyes in cancer targeting and imaging. *Biomaterials*, 32(29), 7127–7138.
- Martin Brown, J., Yu, N. Y., Brown, D. M., & Lee, W. W. (1981). SR-2508: A 2-nitroimidazole amide which should be superior to misonidazole as a radiosensitizer for clinical use. *International Journal of Radiation Oncology Biology Physics*, 7(6), 695–703.
- Martin, Y., & Stouch, T. (2011). In tribute to Corwin Hansch, father of QSAR. *Journal of Computer-Aided Molecular Design* 2011 25:6, 25(6), 491–491.
- Martinez, C. R., & Iverson, B. L. (2012). Rethinking the term “pi-stacking.” *Chemical Science*, 3(7), 2191–2201.
- Mathis, C. A., Wang, Y., Holt, D. P., Huang, G. F., Debnath, M. L., & Klunk, W. E. (2003). Synthesis and evaluation of 11C-labeled 6-substituted 2-arylbenzothiazoles as amyloid imaging agents. *Journal of Medicinal Chemistry*, 46(13), 2740–2754.
- Matsumura, K., Ono, M., Hayashi, S., Kimura, H., Okamoto, Y., Ihara, M., Takahashi, R., Mori, H., & Saji, H. (2011). Phenyl diazenyl benzothiazole derivatives as probes for in vivo imaging of neurofibrillary tangles in Alzheimer's disease brains. *MedChemComm*, 2(7), 596–600.

- Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular descriptors. In *Handbook of Computational Chemistry* (pp. 2065–2093). Springer International Publishing.
- Maya, Y., Okumura, Y., Kobayashi, R., Onishi, T., Shoyama, Y., Barret, O., Alagille, D., Jennings, D., Marek, K., Seibyl, J., Tamagnan, G., Tanaka, A., & Shirakami, Y. (2016). Preclinical properties and human in vivo assessment of 123 I-ABC577 as a novel SPECT agent for imaging amyloid- β . *Brain*, *139*(1), 193–203.
- Maya, Y., Ono, M., Watanabe, H., Haratake, M., Saji, H., & Nakayama, M. (2009). Novel radioiodinated aurones as probes for SPECT imaging of β -amyloid plaques in the brain. *Bioconjugate Chemistry*, *20*(1), 95–101.
- Meikle, S. R., Kench, P., Kassiou, M., & Banati, R. B. (2005). Small animal SPECT and its place in the matrix of molecular imaging technologies. *Physics in Medicine & Biology*, *50*(22), R45.
- Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design*, *7*(2), 146–157.
- Meyer, H. (1899). Welche Eigenschaft der Anaesthetica bedingt ihre narkotische Wirkung? *Naunyn-Schmiedeberg's Arch Exp Pathol Pharmacol*, *42*, 109–118.
- Mills, B., Awais, R. O., Lockett, J., Turton, D., Williams, P., Perkins, A. C., & Hill, P. J. (2015). [18F]FDG-6-P as a novel in vivo tool for imaging staphylococcal infections. *EJNMMI Research*, *5*(1), 1–11.
- Mills, E. J. (2009). XXIII. On melting-point and boiling-point as related to chemical composition. <https://doi.org/10.1080/14786448408627502> *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *17*(105), 173–187.
- Minoshima, S., Drzezga, A. E., Barthel, H., Bohnen, N., Djekidel, M., Lewis, D. H., Mathis, C. A., McConathy, J., Nordberg, A., Sabri, O., Seibyl, J. P., Stokes, M. K., & Van Laere, K. (2016). SNMMI Procedure Standard/EANM Practice Guideline for Amyloid PET Imaging of the Brain 1.0. *Journal of Nuclear Medicine*, *57*(8), 1316–1322.
- Minoshima, S., Mosci, K., Cross, D., & Thientunyakit, T. (2021). Brain [F-18]FDG PET for Clinical Dementia Workup: Differential Diagnosis of Alzheimer's Disease and Other Types of Dementing Disorders. *Seminars in Nuclear Medicine*, *51*(3), 230–240.
- Murphy, R., Kung, H., Kung, M., & Billings, J. (1990). Synthesis and characterization of iodobenzamide analogs: potential D-2 dopamine receptor imaging agents. *Journal of Medicinal Chemistry*, *33*(1), 171–178.
- Naylor, M. A., Stephens, M. A., Cole, S., Threadgill, M. D., Stratford, I. J., O'Neill, P., Fielden, E. M., & Adams, G. E. (1990). Synthesis and evaluation of novel electrophilic nitrofurans carboxamides and carboxylates as radiosensitizers and bioreductively activated cytotoxins. *Journal of Medicinal Chemistry*, *33*(9), 2508–2513.
- Nesterov, S. V., Deshayes, E., Sciagrà, R., Settimo, L., Declerck, J. M., Pan, X. B., Yoshinaga, K., Katoh, C., Slomka, P. J., Germano, G., Han, C., Aalto, V., Alessio, A. M., Ficaro, E. P., Lee, B. C., Nekolla, S. G., Gwet, K. L., De Kemp, R. A., Klein, R., ... Knuuti, J. M. (2014). Quantification of Myocardial Blood Flow in Absolute Terms Using 82Rb PET Imaging: The RUBY-10 Study. *JACC: Cardiovascular Imaging*, *7*(11), 1119–1127.
- Ojha, P. K., Mitra, I., Das, R. N., & Roy, K. (2011). Further exploring rm2 metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems*, *107*(1), 194–205.
- Okamura, N., Furumoto, S., Harada, R., Tago, T., Yoshikawa, T., Fodero-Tavoletti, M., Mulligan, R. S., Villemagne, V. L., Akatsu, H., Yamamoto, T., Arai, H., Iwata, R., Yanai, K., & Kudo, Y. (2013). Novel 18F-Labeled Arylquinoline Derivatives for Noninvasive Imaging of Tau

- Pathology in Alzheimer Disease. *Journal of Nuclear Medicine*, 54(8), 1420–1427.
- Okamura, N., Suemoto, T., Furumoto, S., Suzuki, M., Shimadzu, H., Akatsu, H., Yamamoto, T., Fujiwara, H., Nemoto, M., Maruyama, M., Arai, H., Yanai, K., Sawada, T., & Kudo, Y. (2005). Quinoline and Benzimidazole Derivatives: Candidate Probes for In Vivo Imaging of Tau Pathology in Alzheimer's Disease. *Journal of Neuroscience*, 25(47), 10857–10862.
- Ollinger, J. M., & Fessler, J. A. (1997). Positron-emission tomography. *IEEE Signal Processing Magazine*, 14(1), 43–55.
- Ono, M., Cheng, Y., Kimura, H., Watanabe, H., Matsumura, K., Yoshimura, M., Iikuni, S., Okamoto, Y., Ihara, M., Takahashi, R., & Saji, H. (2013). Development of Novel ¹²³I-Labeled Pyridyl Benzofuran Derivatives for SPECT Imaging of β -Amyloid Plaques in Alzheimer's Disease. *PLOS ONE*, 8(9), e74104.
- Ono, M., Hayashi, S., Matsumura, K., Kimura, H., Okamoto, Y., Ihara, M., Takahashi, R., Mori, H., & Saji, H. (2011). Rhodanine and thiohydantoin derivatives for detecting tau pathology in Alzheimer's brains. *ACS Chemical Neuroscience*, 2(5), 269–275.
- Ono, M., Kawashima, H., Nonaka, A., Kawai, T., Haratake, M., Mori, H., Kung, M. P., Kung, H. F., Saji, H., & Nakayama, M. (2006). Novel benzofuran derivatives for PET imaging of β -amyloid plaques in Alzheimer's disease brains. *Journal of Medicinal Chemistry*, 49(9), 2725–2730.
- Oomen, A. G., Bleeker, E. A. J., Bos, P. M. J., van Broekhuizen, F., Gottardo, S., Groenewold, M., Hristozov, D., Hund-Rinke, K., Irfan, M. A., Marcomini, A., Peijnenburg, W. J. G. M., Rasmussen, K., Sánchez Jiménez, A., Scott-Fordsmand, J. J., van Tongeren, M., Wiench, K., Wohlleben, W., & Landsiedel, R. (2015). Grouping and Read-Across Approaches for Risk Assessment of Nanomaterials. *International Journal of Environmental Research and Public Health* 2015, Vol. 12, Pages 13415-13434, 12(10), 13415–13434.
- Overton, E. (1899). Ueber die allgemeinen osmotischen Eigenschaften der Zelle, ihre vermutlichen Ursachen u. ihre Bedeutg fd Physiologie. *Fäsi & Beer*.
- Pan, J., Mason, N. S., Debnath, M. L., Mathis, C. A., Klunk, W. E., & Lin, K. S. (2013). Design, synthesis and structure–activity relationship of rhenium 2-arylbenzothiazoles as β -amyloid plaque binding agents. *Bioorganic & Medicinal Chemistry Letters*, 23(6), 1720–1726. h
- Paravastu, A. K., Leapman, R. D., Yau, W. M., & Tycko, R. (2008). Molecular structural basis for polymorphism in Alzheimer's β -amyloid fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 105(47), 18349–18354.
- Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2), 3336–3341.
- Pimlott, S. L., & Sutherland, A. (2010). Molecular tracers for the PET and SPECT imaging of disease. *Chemical Society Reviews*, 40(1), 149–162.
- Pirovano, A., Brandmaier, S., Huijbregts, M. A. J., Ragas, A. M. J., Veltman, K., & Hendriks, A. J. (2015). The utilisation of structural descriptors to predict metabolic constants of xenobiotics in mammals. *Environmental Toxicology and Pharmacology*, 39(1), 247–258.
- Pope, P. T., & Webster, J. T. (2012). The Use of an F-Statistic in Stepwise Regression Procedures.
- Pysz, M. A., Gambhir, S. S., & Willmann, J. K. (2010). Molecular imaging: current status and emerging strategies. *Clinical Radiology*, 65(7), 500–516.
- Qi, Y., Liu, X., Li, J., Yao, H., & Yuan, S. (2017). Fluorine-18 labeled amino acids for tumor PET/CT imaging. *Oncotarget*, 8(36), 60581.
- Qu, W., Kung, M. P., Hou, C., Jin, L. W., & Kung, H. F. (2007). Radioiodinated aza-

- diphenylacetylenes as potential SPECT imaging agents for β -amyloid plaque detection. *Bioorganic & Medicinal Chemistry Letters*, 17(13), 3581–3584.
- Rangger, C., & Haubner, R. (2020). Radiolabelled Peptides for Positron Emission Tomography and Endoradiotherapy in Oncology. *Pharmaceuticals 2020*, Vol. 13, Page 22, 13(2), 22.
- Rezazadeh, F., & Sadeghzadeh, N. (2019). Tumor targeting with ^{99m}Tc radiolabeled peptides: Clinical application and recent development. *Chemical Biology & Drug Design*, 93(3), 205–221.
- Ribas, J., Cubero, E., Luque, F. J., & Orozco, M. (2002). Theoretical study of alkyl- π and aryl- π interactions. Reconciling theory and experiment. *Journal of Organic Chemistry*, 67(20), 7057–7065.
- Richet, C. (1893). On the relationship between the toxicity and the physical properties of substances. *Compt Rendus Seances Soc Biol*, 9, 775–776.
- Ripphausen, P., Nisius, B., Peltason, L., & Bajorath, J. (2010). Quo vadis, virtual screening? A comprehensive survey of prospective applications. *Journal of Medicinal Chemistry*, 53(24), 8461–8467.
- Roy, K., Kar, S., & Das, R. N. (2015a). A primer on QSAR/QSPR modeling: fundamental concepts. In *Springer*. Springer.
- Roy, K., Kar, S., & Das, R. N. (2015b). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press.
- Roy, K. (2015). *Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment*. IGI Global.
- Roy, Kunal. (2007). On some aspects of validation of predictive quantitative structure–activity relationship models.
- Roy, Kunal, & Ambure, P. (2016). The “double cross-validation” software tool for MLR QSAR model development. *Chemometrics and Intelligent Laboratory Systems*, 159, 108–126.
- Roy, Kunal, Ambure, P., & Kar, S. (2018a). How Precise Are Our Quantitative Structure-Activity Relationship Derived Predictions for New Query Chemicals? *ACS Omega*, 3(9), 11392–11406.
- Roy, Kunal, Ambure, P., Kar, S., & Ojha, P. K. (2018). Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *Journal of Chemometrics*, 32(4), e2992.
- Roy, Kunal, Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 152, 18–33.
- Roy, Kunal, Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 145, 22–29.
- Roy, Kunal, & Mitra, I. (2011). On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design. *Combinatorial Chemistry & High Throughput Screening*, 14(6), 450–474.
- Roy, Kunal, Mitra, I., Kar, S., Ojha, P. K., Das, R. N., & Kabir, H. (2012). Comparative studies on some metrics for external validation of QSPR models. *Journal of Chemical Information and Modeling*, 52(2), 396–408.
- Rücker, C., Rücker, G., & Meringer, M. (2007). Y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, 47(6), 2345–2357.

- Sai, K. K. S., Zachar, Z., Bingham, P. M., & Mintz, A. (2017). Metabolic PET Imaging in Oncology.
- Salahinejad, M., & Mirshojaei, S. F. (2016). Quantitative structure–activity relationship analysis to elucidate the clearance mechanisms of Tc-99m labeled quinolone antibiotics. *Journal of Radioanalytical and Nuclear Chemistry*, 307(1), 437–445.
- Salahinejad, Maryam. (2015). Quantitative structure property relationships on formation constants of radiometals for radiopharmaceuticals applications. *Journal of Radioanalytical and Nuclear Chemistry*, 303(1), 671–680.
- Samnick, S., Al-Momani, E., Schmid, J. S., Mottok, A., Buck, A. K., & Lapa, C. (2018). Comparison Initial Clinical to Investigation[124I]Iodine PET/CT of [18F]Tetrafluoroboratefor Imaging Thyroid PET/CTCancer in. *Clinical Nuclear Medicine*, 43(3), 162–167.
- Saporo, A., Tadé, M. O., & Vuthaluru, H. (2012). A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chemical Product and Process Modeling*, 7(1).
- Schilling, L. P., Zimmer, E. R., Shin, M., Leuzy, A., Pascoal, T. A., Benedet, A. L., Borelli, W. V., Palmi, A., Gauthier, S., & Rosa-Neto, P. (2016). Imaging Alzheimer's disease pathophysiology with PET. *Dementia & Neuropsychologia*, 10(2), 79–90.
- Sipos, A., Kiss, B., Schmidt, É., Greiner, I., & Berényi, S. (2008). Synthesis and neuropharmacological evaluation of 2-aryl- and alkylapomorphines. *Bioorganic & Medicinal Chemistry*, 16(7), 3773–3779.
- Søndergaard, K., Kristensen, J. L., Palner, M., Gillings, N., Knudsen, G. M., Roth, B. L., & Begtrup, M. (2005). Synthesis and binding studies of 2-aryl apomorphines. *Organic & Biomolecular Chemistry*, 3(22), 4077–4081.
- Spanoudaki, V. C., & Levin, C. S. (2010). Photo-Detectors for Time of Flight Positron Emission Tomography (ToF-PET). *Sensors 2010, Vol. 10, Pages 10484-10505*, 10(11), 10484–10505.
- Sukumar, N., Prabhu, G., & Saha, P. (2014). Applications of genetic algorithms in QSAR/QSPR modeling. In *Applications of Metaheuristics in Process Engineering* (Vol. 9783319065, pp. 315–324). Springer International Publishing.
- Tago, T., Furumoto, S., Okamura, N., Harada, R., Adachi, H., Ishikawa, Y., Yanai, K., Iwata, R., & Kudo, Y. (2016). Structure–Activity Relationship of 2-Arylquinolines as PET Imaging Tracers for Tau Pathology in Alzheimer Disease. *Journal of Nuclear Medicine*, 57(4), 608–614.
- Tago, T., Furumoto, S., Okamura, N., Harada, R., Ishikawa, Y., Arai, H., Yanai, K., Iwata, R., & Kudo, Y. (2014). Synthesis and preliminary evaluation of 2-arylhydroxyquinoline derivatives for tau imaging. *Journal of Labelled Compounds and Radiopharmaceuticals*, 57(1), 18–24.
- Tamiji, Z., Salahinejad, M., & Niazi, A. (2018). Molecular modeling of potential PET imaging agents for adenosine receptor in Parkinson's disease. *Structural Chemistry*, 29(2), 467–479.
- Threadgill, M. D., Webb, P., O'Neill, P., Naylor, M. A., Stephens, M. A., Stratford, I. J., Cole, S., Adams, G. E., & Fielden, E. M. (1991). Synthesis of a series of nitrothiophenes with basic or electrophilic substituents and evaluation as radiosensitizers and as bioreductively activated cytotoxins. *Journal of Medicinal Chemistry*, 34(7), 2112–2120.
- Todeschini, R., & Consonni, V. (2008). *Handbook of Molecular Descriptors*. Wiley and Sons.
- Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices*. John Wiley & Sons.
- Todeschini, Roberto, Ballabio, D., & Grisoni, F. (2016). Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *Journal of Chemical*

- Information and Modeling*, 56(10), 1905–1913.
- Topliss, J. G., & Edwards, R. P. (1979). Chance Factors in Studies of Quantitative Structure-Activity Relationships. *Journal of Medicinal Chemistry*, 22(10), 1238–1244.
- Tóth, M., Berényi, S., Csutorás, C., Kula, N. S., Zhang, K., Baldessarini, R. J., & Neumeyer, J. L. (2006). Synthesis and dopamine receptor binding of sulfur-containing aporphines. *Bioorganic & Medicinal Chemistry*, 14(6), 1918–1923.
- Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6–7), 476–488.
- Tu, Z., Efange, S. M. N., Xu, J., Li, S., Jones, L. A., Parsons, S. M., & Mach, R. H. (2009). Synthesis and in vitro and in vivo evaluation of ¹⁸F-labeled positron emission tomography (PET) ligands for imaging the vesicular acetylcholine transporter. *Journal of Medicinal Chemistry*, 52(5), 1358–1369.
- Tu, Z., Zhang, X., Jin, H., Yue, X., Padakanti, P. K., Yu, L., Liu, H., Flores, H. P., Kaneshige, K., Parsons, S. M., & Perlmutter, J. S. (2015). Synthesis and biological characterization of a promising F-18 PET tracer for vesicular acetylcholine transporter. *Bioorganic & Medicinal Chemistry*, 23(15), 4699–4709.
- Tute, M. (1990). History and objectives of quantitative drug design. *Comprehensive Medicinal Chemistry*, 4, 1–31.
- Valotassiou, V., Malamitsi, J., Papatriantafyllou, J., Dardiotis, E., Tsougos, I., Psimadas, D., Alexiou, S., Hadjigeorgiou, G., & Georgoulas, P. (2018). SPECT and PET imaging in Alzheimer's disease. *Annals of Nuclear Medicine* 2018 32:9, 32(9), 583–593.
- Van Paesschen, W., Dupont, P., Van Driel, G., Van Billoen, H., & Maes, A. (2003). SPECT perfusion changes during complex partial seizures in patients with hippocampal sclerosis. *Brain*, 126(5), 1103–1111.
- Vasdev, N., Natesan, S., Galineau, L., Garcia, A., Stableford, W. T., McCormick, P., Seeman, P., Houle, S., & Wilson, A. A. (2006). Radiosynthesis, ex vivo and in vivo evaluation of [¹¹C]preclamol as a partial dopamine D2 agonist radioligand for positron emission tomography. *Synapse*, 60(4), 314–318.
- Wardman, P. (2007). Chemical Radiosensitizers for Use in Radiotherapy. *Clinical Oncology*, 19(6), 397–417.
- Watson, D. D., & Glover, D. K. (2010). Overview of Tracer Kinetics and Cellular Mechanism of Uptake - Google Scholar. *Clinical Nuclear Cardiology*, 3–13.
- Weissleder, R., & Pittet, M. J. (2008). Imaging in the era of molecular oncology. *Nature* 2008 452:7187, 452(7187), 580–589.
- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130.
- Wright, B. D., & Lapi, S. E. (2013). Designing the Magic Bullet? The Advancement of Immuno-PET into Clinical Use. *Journal of Nuclear Medicine*, 54(8), 1171–1174.
- Wu, G., Robertson, D. H., Brooks, C. L., & Vieth, M. (2003). Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMM-based MD docking algorithm. *Journal of Computational Chemistry*, 24(13), 1549–1562.
- Yang, Y., Cui, M., Jin, B., Wang, X., Li, Z., Yu, P., Jia, J., Fu, H., Jia, H., & Liu, B. (2013). ^{99m}Tc-labeled dibenzylideneacetone derivatives as potential SPECT probes for in vivo imaging of β -amyloid plaque. *European Journal of Medicinal Chemistry*, 64, 90–98.

- Yang, Y., Zhang, X., Cui, M., Zhang, J., Guo, Z., Li, Y., Zhang, X., Dai, J., & Liu, B. (2015). Preliminary Characterization and In Vivo Studies of Structurally Identical ^{18}F - and ^{125}I -Labeled Benzyloxybenzenes for PET/SPECT Imaging of β -Amyloid Plaques. *Scientific Reports* 2015 5:1, 5(1), 1–11. <https://doi.org/10.1038/srep12084>
- Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474.
- Yap, S. Y., Frias, B., Wren, M. C., Schöll, M., Fox, N. C., Årstad, E., Lashley, T., & Sander, K. (2021). Discriminatory ability of next-generation tau PET tracers for Alzheimer's disease. *Brain*, 144(8), 2284–2290.
- Zhang, X. M., Zhang, H. H., McLeroth, P., Berkowitz, R. D., Mont, M. A., Stabin, M. G., Siegel, B. A., Alavi, A., Barnett, T. M., Gelb, J., Petit, C., Spaltro, J., Cho, S. Y., Pomper, M. G., Conklin, J. J., Bettgowda, C., & Saha, S. (2016). [^{124}I]FIAU: Human dosimetry and infection imaging in patients with suspected prosthetic joint infection. *Nuclear Medicine and Biology*, 43(5), 273–279.
- Zhu, L., Ploessl, K., & Kung, H. F. (2014). PET/SPECT imaging agents for neurodegenerative diseases. *Chemical Society Reviews*, 43(19), 6683–6691.

APPENDIX: REPRINTS



Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease

Priyanka De¹ · Dhananjay Bhattacharyya² · Kunal Roy¹

Received: 20 May 2019 / Accepted: 10 June 2019 / Published online: 19 June 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Non-invasive imaging of amyloid beta ($A\beta$) and tau fibrils in the brain may support an early and precise diagnosis of Alzheimer's disease. Molecular imaging technologies involving radionuclides such as positron emission tomography (PET) and single-photon emission computed tomography (SPECT) against beta amyloid plaques and tau fibrils are among emerging research areas in the field of medicinal chemistry. In the current study, we have developed partial least square (PLS) regression-based two-dimensional quantitative structure-activity relationship (2D-QSAR) models using datasets of 38 PET and 73 SPECT imaging agents targeted against $A\beta$ protein and 31 imaging agents (both PET and SPECT) targeted against tau protein. Following the strict Organization for Economic Co-operation and Development (OECD) guidelines, we have strived to select significant descriptors from the large initial pool of descriptors using multilayered variable selection strategy using the double cross-validation (DCV) method followed by the best subset selection (BSS) method prior to the development of the final PLS models. The developed models showed significant statistical performance and reliability. Molecular docking studies have been performed to understand the molecular interactions between the ligand and receptor, and the results are then correlated with the structural features obtained from the QSAR models. Furthermore, we have also designed some imaging agents based on the information provided by the models developed and some of them are predicted to be similar to or more active than the most active imaging agents present in the original dataset.

Keywords Alzheimer's disease · Imaging agents · Double cross-validation · QSAR · PLS

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder which affects older individuals producing neurobehavioral linked symptoms. It has been conceived as an active pathophysiological process characterized by preclinical, mild

cognitive impairment (MCI), and dementia stages. The neuropathological features of AD constitute the deposition of β -amyloid ($A\beta$) protein in the form of extracellular senile (amyloid) plaques and formation of intracellular neurofibrillary tangles (tau aggregates), brain atrophy, and cell depletion [1–3]. It is estimated that by 2030, the disease will afflict 63 million people and, by 2050, 114 million people worldwide [4]. In the USA, approximately 5.5 million individuals of all ages have Alzheimer's disease. Among these, 5.3 million people are above 65 years age and about 20,000 are younger than 65 [5]. Senile plaques (SPs) and neurofibrillary tangles (NFTs) are important neuropathological hallmarks in Alzheimer's disease which are considered to be specific targets for therapeutic intervention as well as biomarkers for imaging agents acting in vivo [6, 7]. SPs are aggregation of amyloid beta ($A\beta$) protein and peptides of 36–43 amino acids derived from the amyloid precursor protein (APP) through cleavage by beta secretase and gamma secretase [8]. Overproduction of $A\beta$ protein leads to aggregation leading to formation of senile protein, followed by NFT formations.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11224-019-01376-z>) contains supplementary material, which is available to authorized users.

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

- ¹ Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
- ² Computational Science Division, Saha Institute of Nuclear Physics, Kolkata 700064, India

NFTs are accumulation of hyperphosphorylated tau, a neuronal microtubule-associated protein, forming insoluble fibers (also known as paired helical filaments) [9, 10]. Severity of NFT deposition related to neuronal loss leads to severe cognitive impairment [11, 12]. The development of biological markers such as PET and SPECT imaging agents are capable of making detection and quantification of fibrillary A β in vivo [13]. This approach acts as a very useful method in early diagnosis of AD and also to identify preclinical effectiveness of newer drug candidates [13].

Recent drug discovery technologies have focused on developing radiotracer imaging agents for imaging beta amyloid and tau pathology in the human brain [13–20]. Several trials have been made to visualize changes in AD pathologies in living brain involving distribution of intravenously administered radiotracers bound to the SPs and NFTs. A good A β imaging agent should have some basic properties including the following: (i) prominent penetration ability through blood-brain barrier (BBB), (ii) selective binding to A β plaques, and (iii) it should produce prominent and distinctive signals between plaques and non-plaques [21]. For tau binding, NFT-specific imaging probes need to be lipophilic to cross the blood-brain barrier and neuronal membranes, and they should also have a high binding affinity to NFTs with minimal non-specific binding [22, 23]. Positron emission tomography (PET) and single-photon emission computed tomography (SPECT) are non-invasive methodologies which makes use of the dynamic distribution of radiotracer to quantify biological processes.

Quantitative structure-activity relationships (QSAR) have become a recognized tool in the field of molecular modeling. They have found applications in the prediction of biological activity and lately in the prediction of the absorption, distribution, metabolism, excretion, and toxicological (ADMET) properties of organic drug-like compounds [24–26]. The large number of candidate molecules that are considered in the drug discovery pipeline and the high failure rate at the later stages of drug development make the computational approaches inevitable for the early predictions of pharmacokinetic and pharmacodynamics end points, thus facilitating the screening process and reducing the cost and time of high end experiments for unsuccessful compounds [27]. Monte Carlo method is one of the promising methods in QSAR which uses optimal molecular descriptors as a tool to predict different end points [28]. The method has been used widely in predicting and understanding the major biochemical features associated with Alzheimer's disease [29, 30].

Descriptor selection for QSAR model development plays an essential role for unbiased prediction of the response. Feature selection by the utilization of multilayered variable selection strategy has been proven to be an effective method in model development where the pool of descriptors is reduced to a small number which can be statistically handled. Further reduction in the descriptor pool through feature

selection helps to obviate the chances of intercorrelation among the descriptors. Double cross-validation (DCV) [31] is a method that involves two nested loops, constructed from the training set, referred to as the calibration set and the validation set for model building and model selection respectively, while the test set is solely used for model assessment. Thus, it precludes any bias introduced in variable selection in case of usage of a single training set of fixed composition. The present study deals with 2D-QSAR studies in order to determine the chemical features contributing to the binding affinity of the imaging agents against beta amyloid (A β) plaques and tau protein aggregation. Three small datasets (two acting against A β and one against tau) with experimental binding affinity (K_i) as the response variable were used for this computational study. The total A β dataset consists of 111 compounds with 38 PET and 73 SPECT imaging agents modeled individually, and the tau dataset consists of 31 compounds. The data were compiled from various literatures as cited in “Materials and methods.” DCV was utilized in the feature selection while modeling the small datasets starting with a large initial number of descriptors. QSAR model development using the DCV tool available at http://teqip.jdvu.ac.in/QSAR_Tools/ helps in removing any bias in descriptor selection from a fixed composition of a training set and often provides an optimum solution in terms of predictivity. The developed models are intended to provide statistically robust predictions for the binding affinity of the imaging agents.

Materials and methods

The dataset

The experimental binding affinity (K_i) data for 38 PET [1, 16, 32–36] and 73 SPECT imaging agents [15, 35, 37–44] against beta amyloid (A β) plaques and 31 (25 PET compounds and 6 SPECT compounds) imaging agents [6, 13, 44–50] against tau protein were obtained from different literatures. Due to limited number of data available for tau protein, the PET and SPECT data were combined to form a single dataset. In the present study, the binding affinity values for both the PET and SPECT dataset compounds expressed as K_i (nM) were converted to negative logarithm of K_i (pK_i) values. All the structures for both the datasets were drawn in MarvinSketch software (version 14.10.27) [51] with proper aromatization and hydrogen bond addition. The dataset is composed of various classes of heterogeneous molecular structures as given in [Supplementary Materials](#) along with their pK_i values.

Molecular descriptors

The molecular descriptor is the result of a logic and mathematical procedure which transforms chemical information

encoded within a symbolic representation of a molecule into a useful number. QSAR models were developed with a selected class of molecular descriptors (two-dimensional) comprising E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom centered fragments and molecular property descriptors, calculated using Dragon 7 software [52]. Intercorrelated descriptors (intercorrelation values larger than 0.9) were removed from the descriptor pool to reduce the size of the descriptor matrix. Finally, a pool of 335 descriptors was obtained for PET imaging agents and a pool of 529 descriptors was obtained for SPECT imaging agents targeted against A β protein. For the tau dataset, a reduced pool of 263 descriptors from 418 descriptors was employed for model development. A descriptor pool of 633 descriptors was obtained for the A β dataset. In order to reduce the redundant and incompetent data, intercorrelated descriptors (correlation value larger than 0.9) were removed from the descriptor pool, and finally, we took 539 descriptors for modeling. For the tau dataset, a reduced pool of 263 descriptors from 418 descriptors was employed for model development.

Dataset splitting

The main objective in QSAR study is to obtain a well-validated QSAR model which is possible with proper division or splitting of the dataset into training and test set. Ideally, the division must be executed in such a way so that points representing both training and test set are well distributed within the whole descriptor space occupied by the entire dataset. Rational data division helps in providing an unbiased external validation with uniform distribution of compounds into training and test sets [53]. One of the extensively used methods is the Euclidean distance-based division [54], which was used for division of the A β imaging dataset (for both PET and SPECT datasets) into training (~75%) and a test set (~25%). The combined PET and SPECT dataset targeted against the tau protein was divided into a training set (~70%) and a test set (~30%) based on *k*-medoids division method [55]. The *k*-medoids algorithm is a local heuristic method that runs just like *k*-means clustering when updating the medoids. This method tends to select *k* most middle objects as initial medoids. The algorithm involves calculation of the distance matrix once and uses it for finding new medoids at every iterative step.

Model development

A critical evaluation procedure was carried out in order to have the best model with good statistical significance for both internal and external validation metrics. During the development of models for individual subsets, i.e., for PET and SPECT imaging agents targeted against A β , we have used stepwise multiple linear regression (S-MLR) [56, 57] method

implemented in double cross-validation (DCV) tool (version 1.2) [31]. Finally, partial least squares (PLS) regression [56, 58] was used to develop the models. In case of the tau dataset, a descriptor pool 26 descriptors was selected using genetic algorithm (GA) [56] modeling implemented in double cross-validation (DCV) tool (version 1.2). Then, the final model was generated using PLS regression method using descriptors selected from best subset selection (BSS).

In both the cases, the double cross-validation (DCV) method helped in the generation of the most statistically significant and robust models. DCV aids in the generation and selection of models to produce a better predictive model. DCV is a method where the training set compounds are further divided into “*n*” calibration and validation sets and can result in diverse compositions of the modeling set, thus removing any bias in descriptor selection. Additionally, a model with the lowest prediction errors in the validation set is selected, thus providing an optimum solution in terms of predictivity in most cases. The tool comprises two nested cross-validation loops: the internal and external cross-validation loops. In the external loop, the compounds in the dataset are divided into training set compounds and test set compounds. The training set compounds are used in the internal loop for the purpose of model development and model selection, and the test set is used exclusively for checking model predictivity. In the internal loop, the training set is further split into calibration and validation sets repetitively by employing the *k*-fold cross-validation technique (in this study, *k* = 10) [59] and producing *k* iterations to construct calibration and validation sets. At the end, the best models were selected based on various validation metrics.

Statistical validation metrics

In the current study, we have utilized multiple approaches for assessment of model quality for measurement of the fitness, stability, robustness, and predictivity of the developed models. The validation was done using both internal and external validation metrics [60]. The fitting potential of the model is established by the determination coefficient (R^2), whereas internal validation dealing with the predictive ability of the model based on training set compounds is usually established by a cross-validated squared correlation coefficient, Q_{LOO}^2 (leave-one-out or LOO). However, Q^2 is not the ultimate quality measuring metric to determine the performance of the model for a new set of compounds. Thus, for new external compounds (or test compounds), various external validation metrics are used such as Q_{F1}^2 and Q_{F2}^2 [61, 62]. Additionally, r_m^2 metrics [63], root mean square error (RMSE), and mean absolute error (MAE) are also calculated [64].

Molecular docking studies

In the present work, we have implemented molecular docking analysis to understand the intermolecular interactions occurring between the PET and SPECT imaging agents and protein beta amyloid and tau proteins separately. The protein structures in the present case are retrieved from the Protein Data Bank with PDB ID: **2LMN** [65] for A β protein and PDB ID: **6FAU** [66] for tau protein. Docking was performed in CDOCKER module of receptor-ligand interaction implemented in BIOVIA Discovery Studio 2018 [67, 68].

The crystal structure of the beta amyloid protein does not contain any bound ligand; therefore, the active site was defined in the BIOVIA Discovery Studio platform in receptor-ligand interaction section using the option “define site from receptor cavities” before docking.

In case of tau, the X-ray crystal structure of the protein consists of two chains A and C and four bound ligands (two peptide residues, Ace-Arg-Thr-Pro-Sep-Leu-Pro-Gly in chain A, Thr-Pro-Sep-Leu-Pro-Gly in chain C and two instances of D3W ((2~{R})-2-[(~{R})-(2-methoxyphenyl)-phenyl-methyl]pyrrolidine) one in each chain. Due to structural similarity between the chain structures, we have used only one chain (chain A) for our docking purpose. Before docking the target ligands, the protein was prepared by removing the duplicate amino acid conformers, addition of hydrogen, and generation of docking site. The active site was defined in the BIOVIA Discovery Studio platform from the ligand binding domain of the bound peptide residue and D3W by selecting them and generating site “from current selection” program in receptor-ligand interaction section of the software. The bound ligands were then removed for new molecule docking purpose.

The target ligands (imaging agents) were subjected to ligand preparation to obtain a series of ligand conformers in both cases using the small molecules module in Discovery Studio. Each of these conformers was used in the CDOCKER module involving CHARMM interaction energy for molecular docking [68]. The ligand poses were ranked using the CDOCKER interaction energy parameters (kJ/mol), and the top scoring (most negative, thus favorable to binding) poses are kept. The best pose obtained was further analyzed by considering intermolecular polar and non-polar interactions.

Results and discussions

In the present study, PET and SPECT imaging agent datasets for both A β plaques and tau fibrils were modeled for their binding affinity using the PLS regression method. For the

A β dataset, the models for the individual PET and SPECT datasets were developed using PLS regression method after stepwise multiple linear regression (S-MLR) method. In case of the tau dataset, the final descriptors for the PLS model were obtained from best subset selection (BSS) which was carried out on a pool of descriptors obtained from double cross-validation-genetic algorithm (DCV-GA) method of model development. The developed models are statistically robust and predictive to be used for data gap filling as suggested by the obtained values of the different validation metrics as given later.

Descriptor interpretation from QSAR models

Modeling of PET imaging agents against A β plaques

The PLS model 1 having 4 latent variable (LV) shown in Table 1 gives acceptable values of the determination coefficient R^2 (0.766) and cross-validated determination coefficient ($Q_{\text{LOO}}^2=0.600$). The predictivity of the model was analyzed by predictive r^2 (or $r_{\text{pred}}^2 = 0.534$) or Q_{F1}^2 which shows acceptable predictivity for the test set compounds. The experimental and predicted pKi values for model 1 are given in the [Supplementary Materials](#). The scatter plot of observed versus predicted pKi values is given in Fig. 1a.

The descriptor TPSA(Tot) (a molecular property related descriptor) representing the topological polar surface area using N, O, S, and P polar contributions shows a negative correlation to the binding affinity of PET imaging agents. The TPSA descriptor shows the importance of interaction of the O-, N-, S-, and P-centered fragments towards beta amyloid plaques (Fig. 2). For example, compounds like **A-P-30** (TPSA(Tot) = 122.79), **A-P-29** (TPSA(Tot) = 104.33) and **A-P-52** (TPSA(Tot) = 90.94) having more number of O-, N-, S-, and P-centered fragments have low pKi values (3.20, 3.91 and 3.19 respectively). On the other hand, compounds like **A-P-63** (TPSA(Tot) = 36.61), **A-P-31** (TPSA(Tot) = 32.26), and **A-P-21** (TPSA(Tot) = 30.49) having lower number of aforementioned fragments have high pKi values (4.55, 4.62, and 4.54 respectively). From this observation, we can conclude that hydrophobicity enhances the binding of PET imaging agents to amyloid plaques.

The descriptor **T(O..S)**, a 2D atom pair descriptor, denotes the sum of topological distances between oxygen and sulfur. This descriptor has a positive contribution to the binding affinity of the imaging agents, thus with an increase in the total sum of topological distances between oxygen and sulfur atoms, the binding affinity will increase and vice versa (Fig. 2). In compounds like **A-P-1**, **A-P-51**, **A-P-48**, and **A-P-49**, the high values for T(O..S) (T(O..S) = 4) contribute to higher pKi values (5.07, 4.36, 4.31, and 4.72 respectively), whereas in compounds like **A-P-43**, **A-P-30**, and **A-P-52**, the

Table 1 QSAR models for PET and SPECT imaging agents

| Model no. | Target | Dataset | Model | Latent variables (LVs) |
|-----------|---------------------------------|---|-------|------------------------|
| 1 | Amyloid beta (A β) PET | <p> $\text{pKi} = 3.987 - 0.012\text{TPSA}(\text{Tot}) + 0.112\text{T}(\text{O..S}) + 0.685\text{B10}[\text{C-C}] + 0.382\text{nArX} + 0.224\text{hHDon}$ $N_{\text{train}} = 29, R^2 = 0.766, R^2_{\text{adj}} = 0.727, Q^2 = 0.600, \text{MAE}(\text{Train}) = 0.236, \text{RMSEC} = 0.219$ $N_{\text{test}} = 9, Q^2_{\text{FI}} = 0.534, Q^2_{\text{F2}} = 0.534, \text{MAE}(\text{Test}) = 0.296, \text{RMSEP} = 0.393$ </p> <p> $\text{pKi} = 3.536 + 0.015\text{SAacc} - 0.042\text{F05}[\text{C-C}] - 0.284\text{F09}[\text{C-F}] - 0.911\text{nR10} + 0.252\text{F03}[\text{C-I}]$ $N_{\text{train}} = 55, R^2 = 0.771, R^2_{\text{adj}} = 0.758, Q^2 = 0.700, \text{MAE}(\text{Train}) = 0.367, \text{RMSEC} = 0.394,$ $N_{\text{test}} = 18, Q^2_{\text{FI}} = 0.739, Q^2_{\text{F2}} = 0.736, \text{MAE}(\text{Test}) = 0.369, \text{RMSEP} = 0.421$ </p> <p> $\text{pKi} = 0.157 + 0.0185\text{D}/\text{Dir09} + 0.139\text{SaaCH} - 0.176\text{SssCH2} - 0.467\text{B08}[\text{N-F}]$ $N_{\text{train}} = 22, R^2 = 0.910, R^2_{\text{adj}} = 0.889, Q^2 = 0.839, \text{MAE}(\text{Train}) = 0.229, r^2_{\text{m}} = 0.781, \Delta r^2_{\text{m}} = 0.032,$ $\text{RMSEC} = 0.198, \text{MAE}(\text{train}) = 0.229$ $N_{\text{test}} = 9, Q^2_{\text{FI}} = 0.865, Q^2_{\text{F2}} = 0.850, \text{MAE}(\text{Test}) = 0.275, r^2_{\text{m}} = 0.768, \Delta r^2_{\text{m}} = 0.114, \text{RMSEP} = 0.325,$ $\text{MAE}(\text{test}) = 0.275$ </p> | 4 | |
| 2 | Amyloid beta (A β) SPECT | | | 3 |
| 3 | Tau | PET and SPECT combined | | 3 |

descriptor value is low ($\text{T}(\text{O..S}) = 0$ for all), resulting in low pKi values (3.43, 3.20, and 3.19 respectively).

The descriptor **B10[C-C]**, another 2D atom pair descriptor, denotes the presence or absence of C-C at topological distance 10. The positive regression coefficient of this parameter suggested that the presence of such fragment at the topological distance 10 enhances the binding affinity (Fig. 2) as shown in compounds like **A-P-1**, **A-P-51**, **A-P-48**, and **A-P-49**. On the other hand, compounds like **A-P-52**, **A-P-43**, and **A-P-59** show poor binding affinity due to the absence of such fragments. Here, size (the distance between C and C atoms at 10 reflects the size of the molecules) plays an important role for the binding affinity.

The descriptor **nArX** (functional group count descriptor) represents the number of halogen (X) on the aromatic ring contributing positively towards the binding affinity of the PET imaging agents (Fig. 2). In compounds like **A-P-8**, **A-P-56**, and **A-P-57**, the presence of one halogen on the aromatic ring contributes for the high binding affinity ($\text{pKi} = 4.64, 4.77, \text{ and } 4.56$ respectively), whereas in compounds like **A-P-43**, **A-P-52**, and **A-P-59**, the absence of halogen group on the aromatic ring reduces the pKi value (3.43, 3.19, and 3.94 respectively).

The descriptor **nHDon** (functional group count descriptor) denotes the number of donor atoms for H bonds (N and O). The descriptor shows a positive contribution towards binding affinity (pKi) as shown in compounds like **A-P-1** (Fig. 2), **A-P-8**, **A-P-31**, and **A-P-58**, all having two hydrogen bond donor sites and hence having higher pKi values (5.07, 4.64, 4.62, and 4.64 respectively). On the other hand, in compounds like **A-P-30** (**nHDon** = 1) and **A-P-62** (**nHDon** = 0) (Fig. 2), the pKi values are low (3.20 and 3.92 respectively).

Modeling SPECT imaging agents against A β plaques

The PLS model 2 with 3 LVs (in Table 1) could describe 77.1% of the variance (adjusted determination coefficient). The leave one out (LOO) cross-validated determination coefficient ($Q^2 = 0.758$) above the critical value of greater than 0.5 suggests the statistical reliability of the model. The experimental and predicted pKi values for model 2 are given in the [Supplementary Materials](#). The scatter plot of observed versus predicted pKi values are given in Fig. 1b.

The descriptor **SAacc**, a molecular property type descriptor, denotes the surface area of acceptor atoms from P_VSA-like descriptors. It shows a positive contribution to the binding affinity of SPECT imaging agents as shown in Fig. 3. The positive regression coefficient indicates that with an increase in the descriptor value, the binding affinity will increase as seen in compounds like **A-S-54**, **A-S-5**, and **A-S-35** and vice versa as seen in compounds like **A-S-73**, **A-S-76**, and **A-S-85**. Thus, the presence of hydrogen bond donor atoms is beneficial for good binding to beta amyloid plaques.

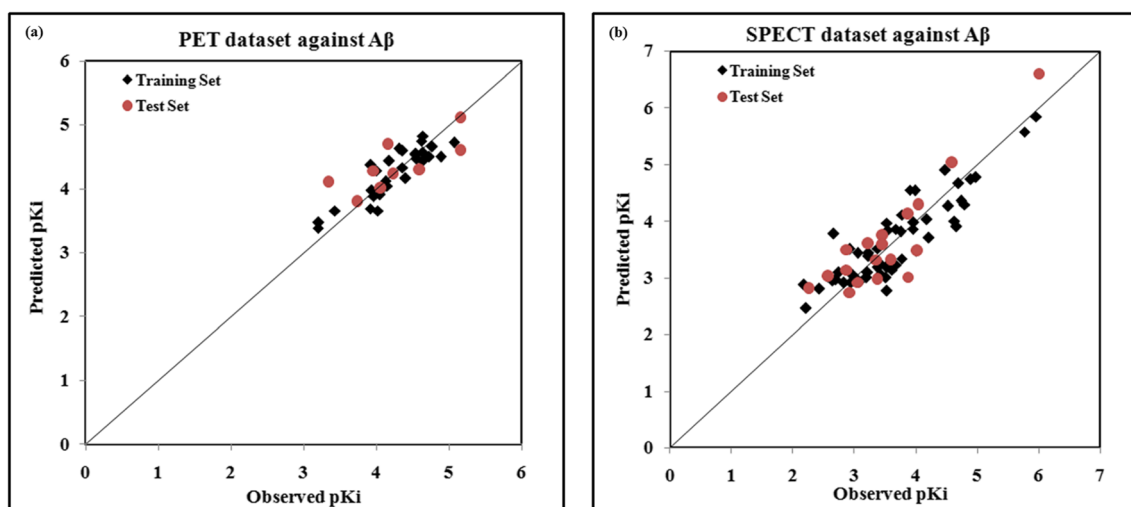


Fig. 1 Scatter plots of observed versus predicted pKi values (Amyloid beta data set)

The descriptor **F05[C-C]** (a 2D atom pair descriptor) depicts the frequency of C-C at the topological distance 5, and it has a negative contribution towards the binding affinity pKi. This indicates that with an increase in the descriptor value (which is an indicator of size and shape), the pKi value will decrease and vice versa as shown in Fig. 3. In compounds like **A-S-5**, **A-S-65**, **A-S-64**, and **A-S-63**, the values for the

descriptors are high, thus making the pKi values low (3.23, 3.52, 3.07, and 2.21 respectively), whereas in compounds like **A-S-6**, **A-S-15**, and **A-S-16** (having low F05[C-C] values), the pKi values are high (4.74, 4.53, and 4.63).

The descriptor **F09[C-F]** (a 2D atom pair descriptor) denotes the frequency of C-F at the topological distance 9 and shows a negative correlation with the binding affinity. This

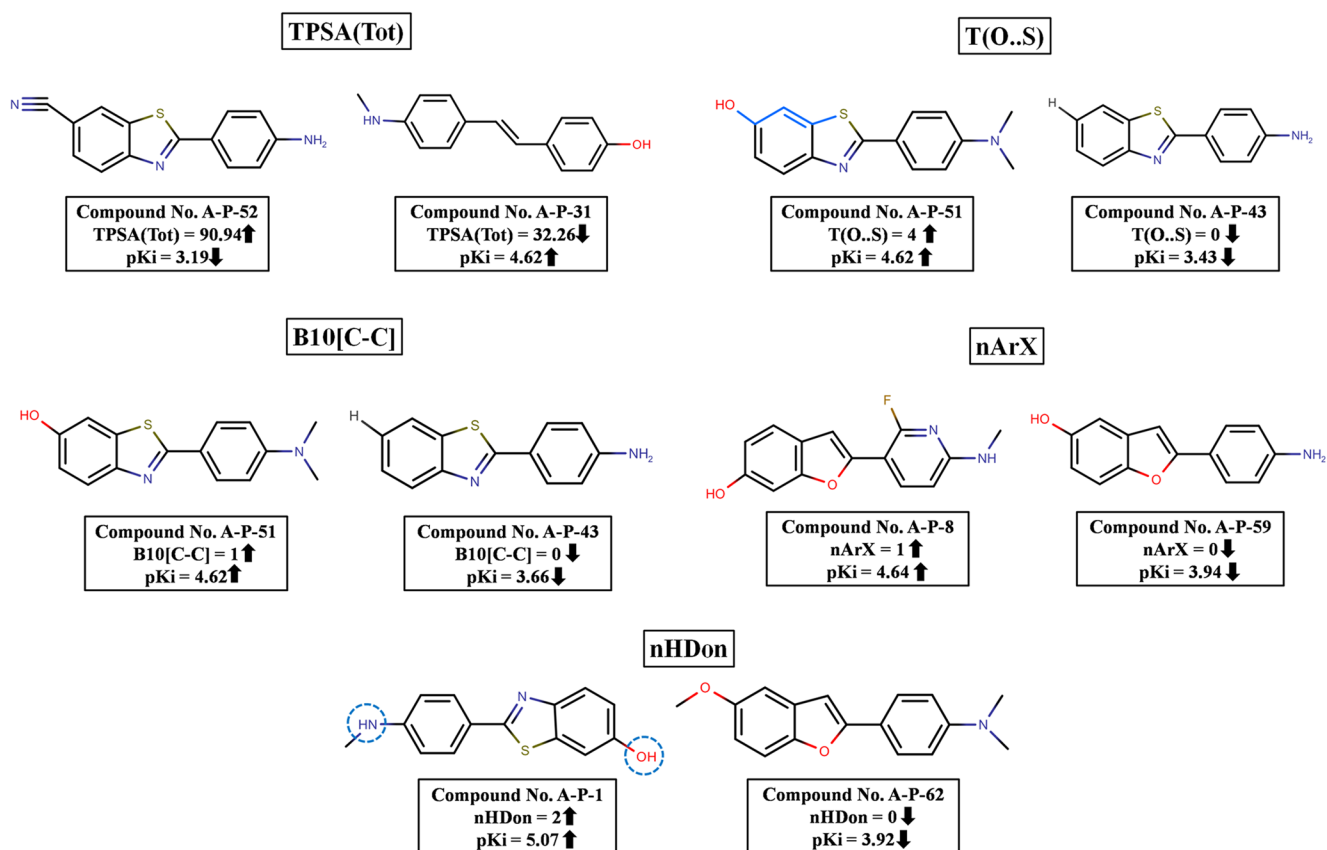


Fig. 2 Descriptor contributions to the binding affinity with respect to model 1 (PET dataset against beta amyloid)

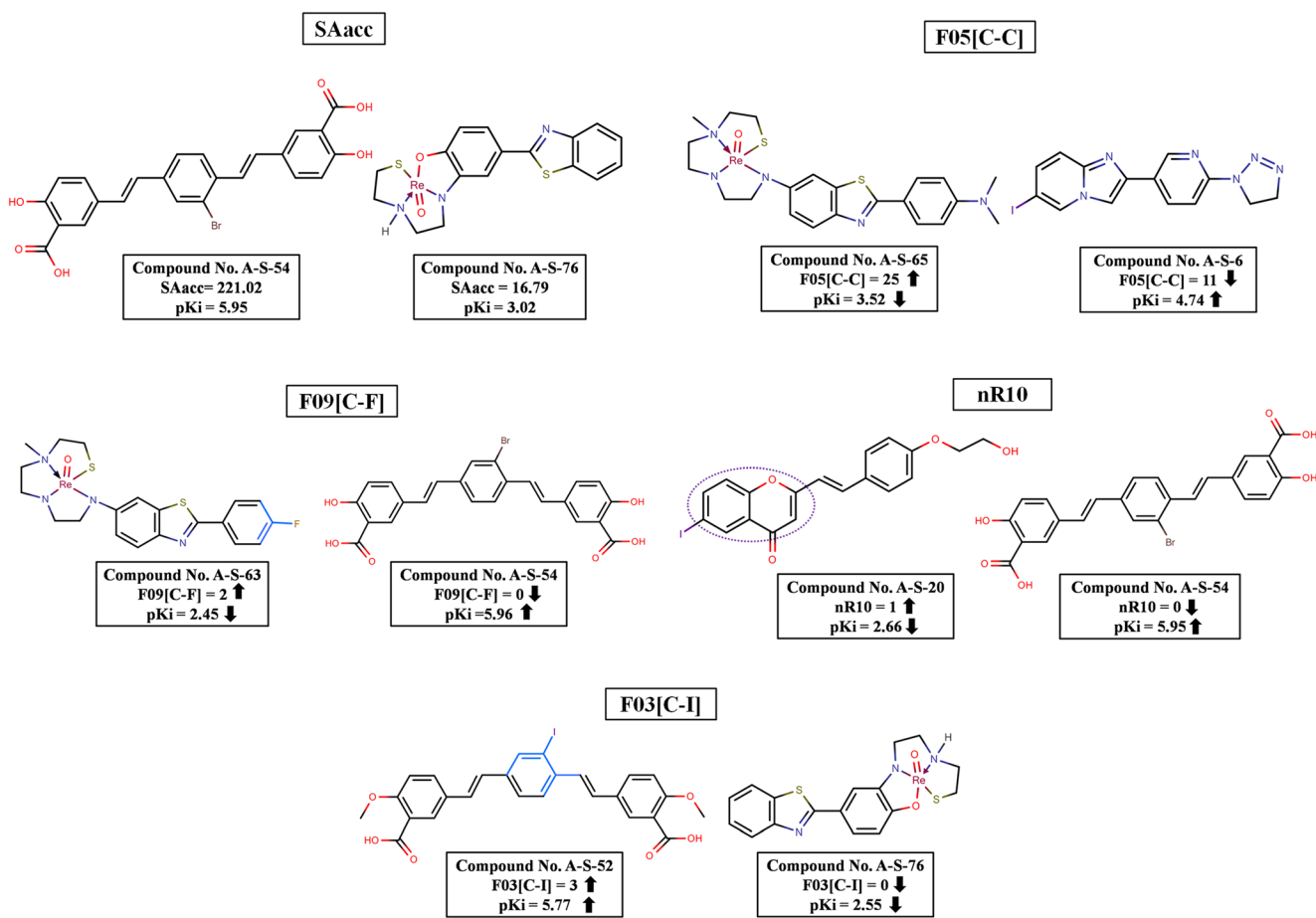


Fig. 3 Descriptor contributions to the binding affinity with respect to model 2 (SPECT dataset against beta amyloid)

descriptor indicates both presence of a fluorine atom and size of the compound. A higher occurrence of C-F at topological distance 9 will decrease the binding affinity as observed in compounds **A-S-41** (pKi = 3.42), **A-S-47** (pKi = 3.38), and **A-S-63** (pKi = 2.21), whereas in compounds like **A-S-52**, **A-S-54**, **A-S-34**, and **A-S-33** with the absence of such groups, the binding affinity is high (5.77, 5.96, 4.98, and 4.89 respectively) (shown in Fig. 3).

The descriptor **nR10**, a ring descriptor, indicates the number of 10-membered rings present in the compounds (here **4H-1-benzopyran** ring), and the descriptor provides a negative contribution to the binding affinity. Compounds like **A-S-17**, **A-S-18**, and **A-S-20** each containing one 10-membered ring have a low binding affinity value (pKi = 3.68, 3.58, and 2.66 respectively), whereas in compounds like **A-S-52**, **A-S-54**, and **A-S-34**, the absence of any 10-membered ring contributes to higher values of the binding affinity (shown in Fig. 3).

The descriptor **F03[C-I]** (a 2D atom pair descriptor) represents the frequency of C-I at the topological distance 3 has a positive contribution towards the binding affinity. Thus, with an increase in the value for this descriptor, the pKi value will increase as seen in compounds **A-S-52**, **A-S-33**, and **A-S-34**

(5.77, 4.89, and 4.98 respectively), whereas with a decrease in the value of **F03[C-I]**, the pKi value will also decrease as seen in **A-S-76**, **A-S-78**, and **A-S-85** (2.55, 2.85, and 2.70 respectively) (Fig. 3).

Modeling PET and SPECT imaging agents against tau protein

The PLS model 3 with 3 latent variables (LVs) evolved as the best model, and it could show good statistical robustness and predictivity. Acceptable values for determination coefficient R^2 (0.910) and cross-validated determination coefficient ($Q_{LOO}^2 = 0.899$) were obtained. The predictivity of the model was analyzed by predictive r^2 (or $r_{pred}^2 = 0.865$) or Q_{F1}^2 which shows good predictivity for the test set compounds. The scatter plot of observed versus predicted pKi values is given in Fig. 4.

The descriptor with the highest contribution, **D/Dtr09** (i.e., distance/detour ring of order 9), is a ring descriptor which is based on operations made on distance or detour matrix D/Δ . The detour matrix is square symmetric matrix that contains the ratios of the lengths of the shortest to the longest path between

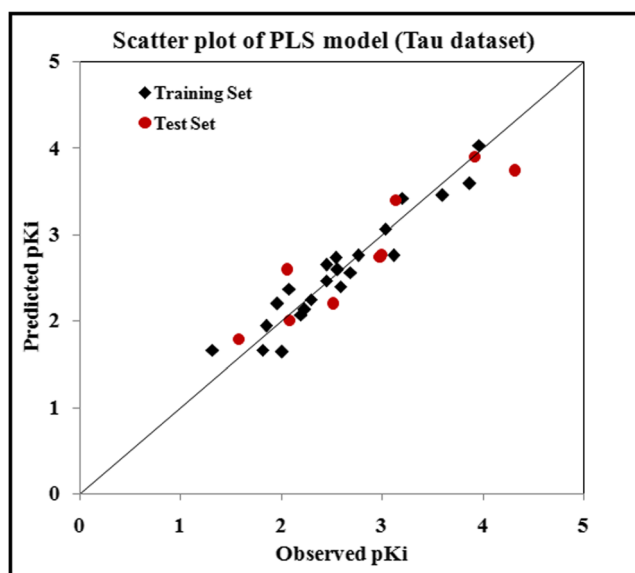


Fig. 4 Scatter plot of observed versus predicted pKi values (Tau data set)

any pair of vertices. The term D/Δ is calculated by the following:

$$D/\Delta = \sum_{i=1}^A \sum_{j=1}^A (D/\Delta)_{ij}$$

Here, Δ is the detour distance [69, 70]. This descriptor shows a positive contribution which indicates its positive influence on the binding affinity of the imaging agents as observed in compounds like **T-P-3** ($D/Dtr09 = 112.603$), **T-P-1** ($D/Dtr09 = 96.788$), and **T-P-30** ($D/Dtr09 = 87.895$) having higher binding affinities (3.96, 3.92, and 4.32 respectively). On the other hand, compounds like **T-S-25** ($pKi = 1.31$), **T-P-9** ($pKi = 1.58$), and **T-S-26** ($pKi = 1.81$) have low pKi values corresponding to low values for the descriptor ($D/Dtr09 = 0$ for all three compounds). Figure 5 shows how the descriptor $D/Dtr09$ contributes towards the binding affinity of the imaging agents.

The next important descriptor **SaaCH**, an E-state descriptor, denotes the sum of E-state of atom type aaCH where aaCH represents $-CH$ groups in benzene nucleus. The descriptor shows a positive contribution to the binding affinity suggesting that the presence of such groups would increase the binding affinity as seen in compounds like **T-P-5** ($SaaCH = 17.67$, $pKi = 3.11$) and **T-P-1** ($SaaCH = 16.71$, $pKi = 3.92$), while in compounds like **T-S-25** ($SaaCH = 11.67$, $pKi = 1.31$) and **T-S-26** ($SaaCH = 11.59$, $pKi = 1.81$), the occurrence of such fragment is low resulting in less binding affinity (Fig. 5).

The third descriptor **SssCH2** is also an E-state descriptor, signifying the sum of E-state of atom type ssCH2 ($-CH_2-$), which has a negative regression coefficient. This indicates that with an increasing descriptor value, the binding affinity

will decrease as seen in compounds **T-P-19** ($SssCH2 = 0.78$, $pKi = 2.01$) and **T-S-25** ($SssCH2 = 0.65$, $pKi = 1.31$) (in Fig. 5). The opposite occurs when the descriptor value is less, i.e., the binding affinity becomes higher as observed in compounds like **T-P-2** ($SssCH2 = -0.48$, $pKi = 3.19$) and **T-P-5** ($SssCH2 = -0.88$, $pKi = 3.11$) (in Fig. 5).

The least important descriptor is **B08[N-F]**, a 2D atom pair descriptor, which denotes the presence or absence of N-F at the topological distance 8. The negative regression coefficient of this parameter suggests that the presence of such fragment at the topological distance 8 is detrimental to the binding affinity as shown in compounds like **T-P-8** ($pKi = 2.23$) and **T-P-18** ($pKi = 1.85$). On the other hand, compounds like **T-S-28**, **T-S-29**, and **T-S-30** show good binding affinity due to the absence of such fragments. Figure 5 shows the contribution of **B08[N-F]** descriptor.

Interpretation of PLS plots

Variable importance plot

The variable importance plot (VIP) [71] signifies the order of contribution of each descriptor. The most and least important descriptors can be identified using this plot. A variable with VIP score > 1 indicates the descriptor's higher statistical significance as compared to the one with a lower VIP value. The descriptors from higher to lower contribution for all the three models are given in Fig. 6.

Regression coefficient plot

The regression coefficient plot [58] (Fig. S1) gives information about the positive or negative contribution of descriptors towards the activity of the compounds. In case of model 1 for the PET dataset against $A\beta$ fibrils, the descriptors like $T(O..S)$, $B10[C-C]$, $nArX$ and $nHDon$ having a positive regression coefficient signify that with an increase in the descriptor value the binding affinity increases, whereas descriptor having negative coefficients like $TPSA(Tot)$ decreases the binding affinity with their increasing numerical values. In case of model 2 for the SPECT dataset against $A\beta$ fibrils, **SAacc** and $F03[C-I]$ descriptors have positive contributions (positive regression coefficients), whereas other three descriptors ($F05[C-C]$, $F09[C-F]$ and $nR10$) have negative regression coefficients. For model 3, i.e., in case of tau protein dataset, it was found that descriptors $D/Dtr09$ and **SaaCH** have a positive regression coefficient and other two descriptors like **SssCH2** and **B08[N-F]** have negative coefficients thereby decreasing pKi values significantly.

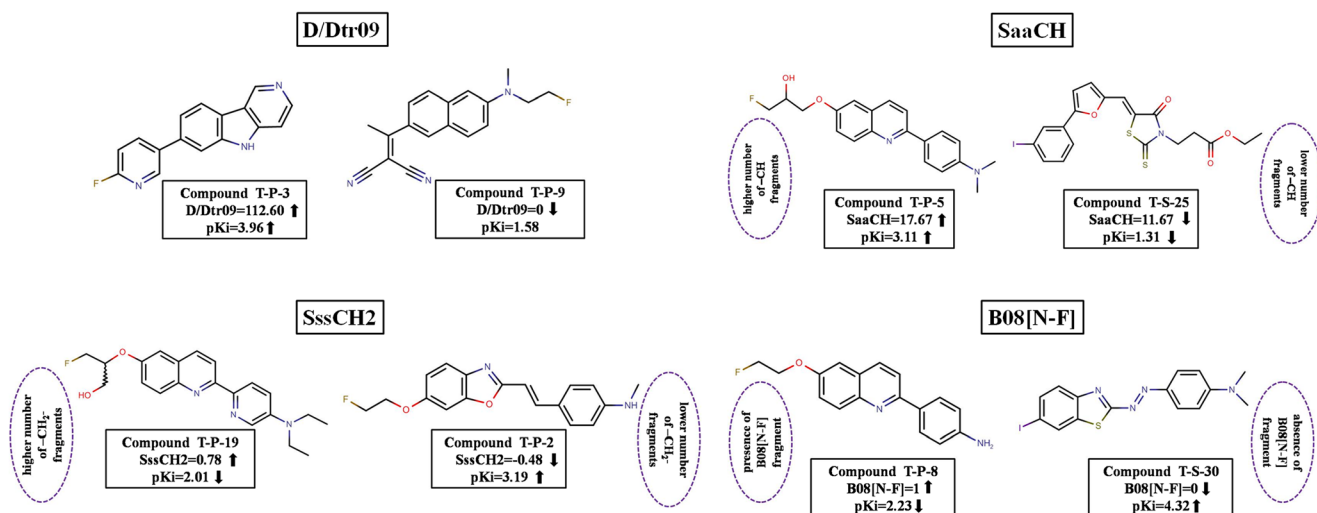


Fig. 5 Descriptor contributions to the binding affinity with respect to model 3 (PET and SPECT dataset against tau protein)

Score plot

The distribution of the compounds in the latent variable space as defined by the scores is expressed in a score plot (Fig. S2) [72]. From the plot, one can conclude that compounds that are situated near each other have similar characteristics or properties, whereas compounds which are far from each other have dissimilar properties with respect to their binding affinity. Compounds which are outside the ellipse in the plot are outliers. The score plots for the derived models are shown in Fig. S5.

Loading plot

The relationship between the X variables and Y variables can be understood by the loading plot (Fig. 7) [58]. The loading

plot was developed using the first two PLS components in all the three cases. The influence of different variables on the model can be understood from the loading plot. Descriptors that are grouped together have similar meanings and similar effects on the response. Descriptors with different meanings are situated at a considerable distance from each other. Any descriptor situated far from the plot origin is considered to have a greater impact on the response.

Applicability domain

The applicability domain (AD) provides a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable [73]. The AD assessment of the proposed model for the imaging agents

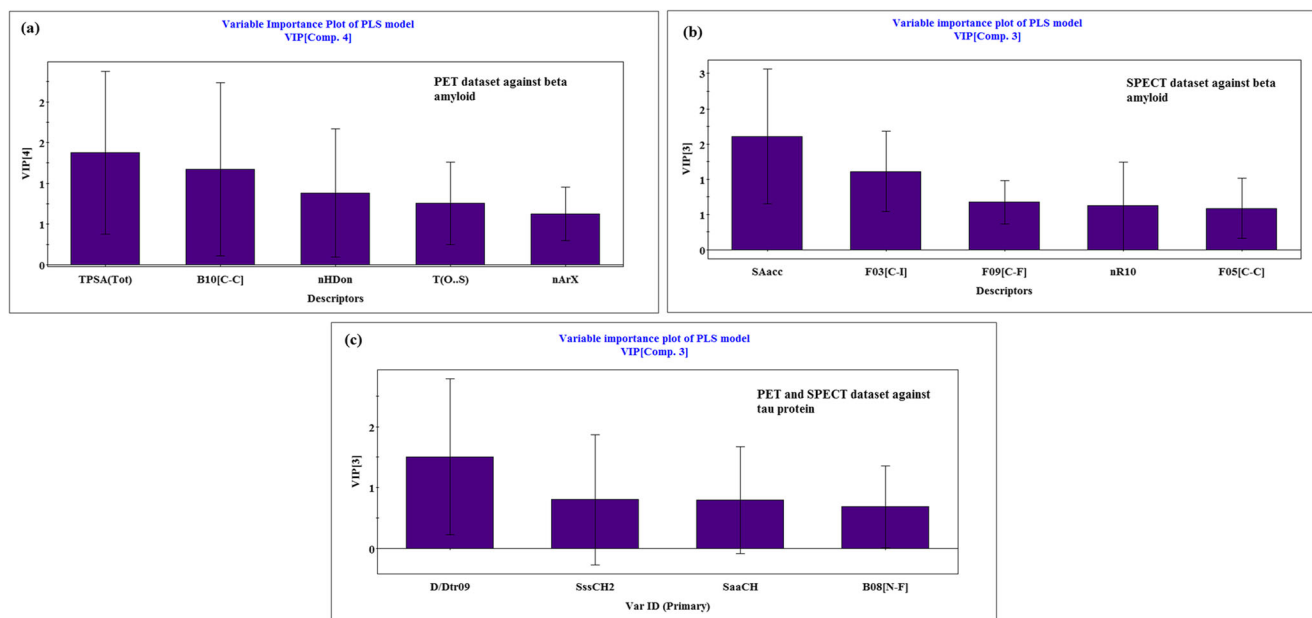


Fig. 6 Variable importance plot (VIP) of the three PLS models

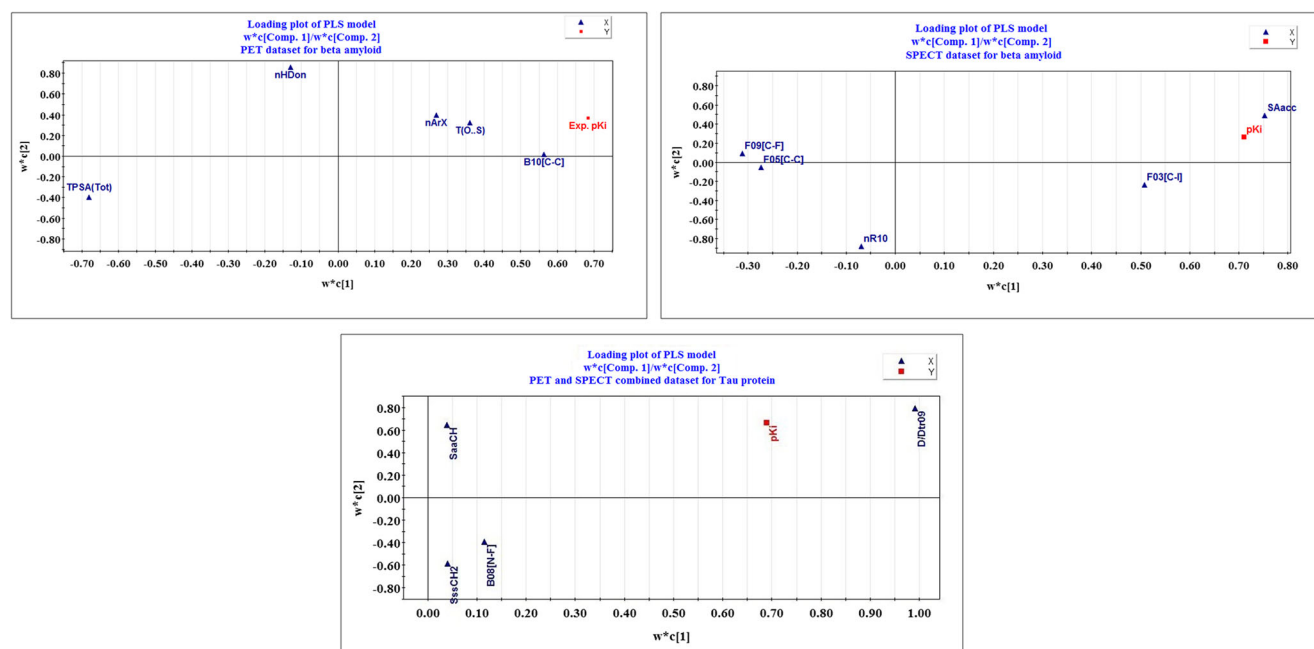


Fig. 7 Loading plots of the relationship between the X variables and Y variables

were performed according to the DModX (distance to model in the X-space) approach using SIMCA-P [74] software. Figures S3, S4, and S5 in the Supplementary Materials show the AD plots of the three models.

Y-randomization

The statistical significance of the model is analyzed by a randomization plot (Fig. S6). The randomization plot authenticates that the model is not the result of any chance correlation [75]. The randomization process provides a number of models by shuffling different combinations of X or Y variables (here Y variable only) based on the fit of the reordered model. Here we have used 100 permutations for random model generation, though the number of permutations can be changed. To avoid chance correlation, the basic statistics of the randomization models (Q^2 and R^2) should be poor and not within the range of those for acceptable regression models (R_Y^2 intercept should not exceed 0.3 and Q_Y^2 intercept should not exceed 0.05) [75]. The randomization plots given in Fig. S8 show that the developed models are non-random and robust and are suitable for prediction of the binding affinity of the imaging agents within the AD of the model.

Molecular docking

Molecular docking studies yield critical information related to the orientation of the imaging agents at the binding zone of the target protein and the information about the intermolecular interaction between protein and ligands at molecular level. The aim in the present study was to understand the

interactions occurring between the two proteins and different PET and SPECT imaging agents and correlate the observations found with the QSAR results. It was found that hydrogen bonding and π bonding interactions were predominant. The ligand-receptor interaction analysis suggests that the imaging agents interact with both polar and non-polar amino acids.

Molecular docking for selected PET imaging agents against A β plaques

In cases of compounds **A-P-2** and **A-P-56** which have higher binding affinity ($pK_i = 5.15$ and 4.77 respectively), the interaction forces include hydrogen bonds (carbon-hydrogen bonds [76], conventional hydrogen bonds and π -donor hydrogen bonds), π interactions (π -sulfur bond, π - π T shaped bond and π -alkyl bonds), and unfavorable acceptor-acceptor bond. The number of interacting residues is higher in case of these compounds thus supporting their high values of binding affinity. The amino acid residues interacting with compound **A-P-2** are Val D:39, His B:13, and Val D:36. Figure 8 shows the interactions obtained for the most stable pose where it is found that Val D:39 and His B:13 show π -alkyl [77, 78] and π - π T [79] shaped interactions respectively with the ligand due to the presence of the aromatic nuclei. Also sulfur in the thiazole nucleus interacts with the aromatic nucleus (thiazole moiety) of histidine making π -sulfur interaction [80]. On the other hand, Val D:36 makes carbon-hydrogen bond with the ligand. In compound **A-P-56**, the interacting amino acids include Val A:12, His B:13, Val D:36, and Val D:39. In Fig. 8, the different interactions are shown. Hydrogen bonds like carbon-hydrogen bonds and π -donor hydrogen bonds are found with

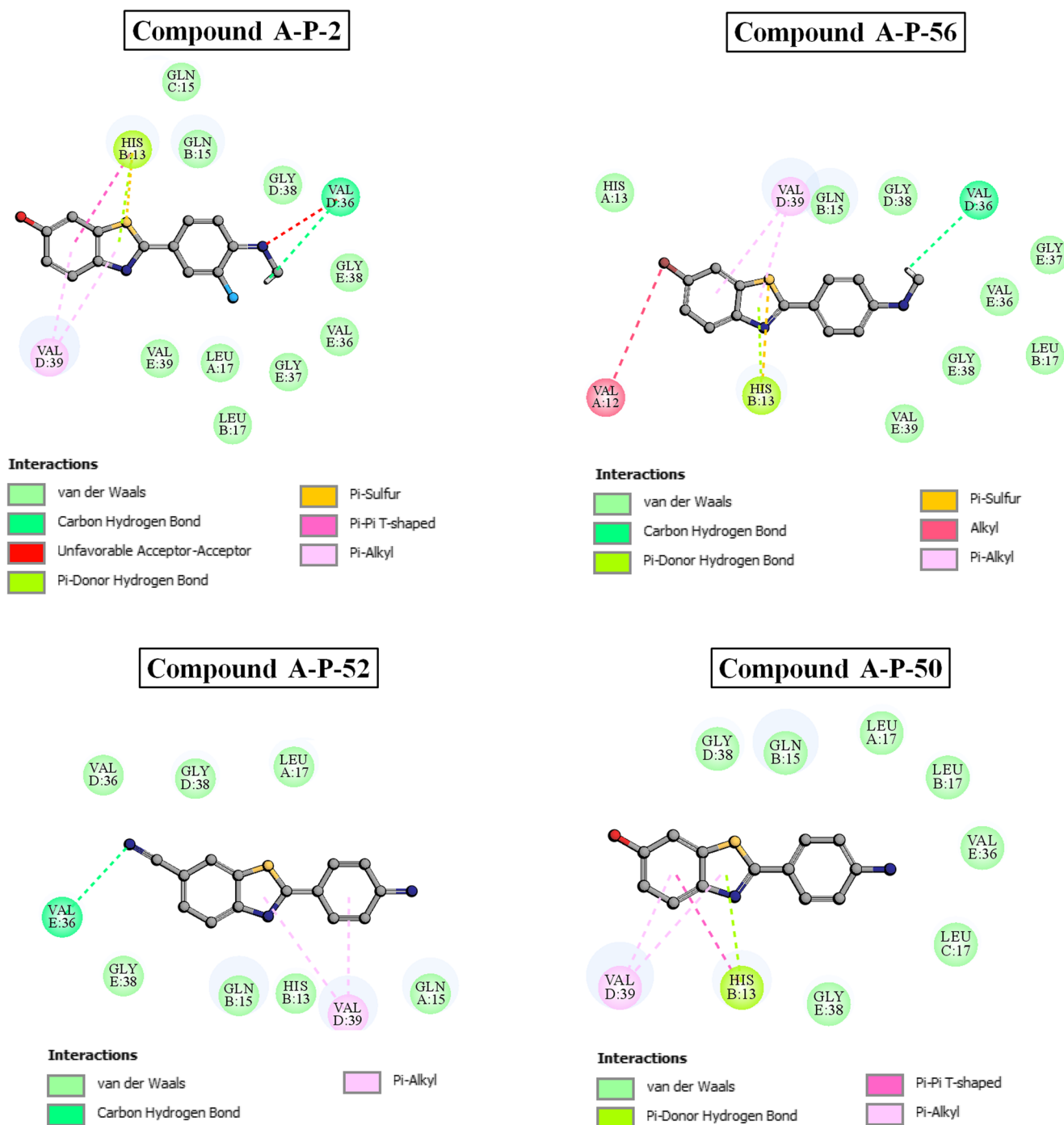


Fig. 8 Molecular interactions between high / low active PET imaging agents and A β protein

Val D:36 and His B:13 respectively. The alkyl part of Val A:12 interacts with Bromine and other π bonds are formed with Val D:39 and His B:13 residues.

PET compounds like **A-P-52** and **A-P-50** having low binding affinity ($pK_i = 3.19$ and 3.34) show similar kind of interactions (hydrogen and π bonds) as in case of higher affinity compounds, but the number of interacting amino acid residues is much less as shown in Fig. 8. Val D:39 was found to interact with both the ligands forming π -alkyl interactions. In

compound **A-P-52**, the nitrogen of cyano group forms hydrogen bond with Val E:36, whereas in compound **A-P-50**, His B:13 shows π interactions (π -donor Hydrogen and π - π interactions) with the ligand. The docking sites for both high and low binding affinity PET imaging agents targeted against A β are given in Table 2.

Relation with QSAR models In the docking study, it is observed that formation of hydrogen bonds between the ligands

Table 2 The docking site, interacting residues, and different types of binding interaction occurring between the imaging agents and the target protein (A β or tau)

| Dataset | Imaging agents | Compound ID | pKi | Docking site | Interacting amino acids | Binding interactions | | |
|--------------|----------------|---------------|--------------------------|--|--|---|---|--|
| Beta amyloid | PET | A-P-2 | 5.15 | Val D:39, Gln B:13, Leu A:17, His B:13, Leu B:17, Val E:36, Leu C:17, Gly E:38, Val E:39, Val A:12, His A:13. | Val D:39, His B:13, Val D:36 | Carbon hydrogen bond, unfavorable acceptor-acceptor, π -donor hydrogen, π -sulfur, π - π T-shaped, π -alkyl and van der Waals | | |
| | | A-P-56 | 4.77 | Val A:12, His B:13, Val E:39, Gly E:38, Leu B:17, Val E:36, Gly E:37, Val D:36, Gly D:38, Gln B:15, Val D:39, His A:13. | Val A:12, His B:13, Val D:36, Val D:39 | Carbon hydrogen bond, π -donor hydrogen, π -sulfur, alkyl, π -alkyl and van der Waals | | |
| | | A-P-52 | 3.19 | Val E:36, Gly E:38, Gln B:15, His B:13, Val D:39, Gln A:15, Leu A:17, Gly D:38, Val D:36 | Val E:36, Val D:39 | Carbon hydrogen bond, π -alkyl and van der Waals | | |
| | | A-P-50 | 3.34 | Val D:39, His B:13, Gly E:38, Leu C:17, Val E:36, Leu B:17, Leu A:17, Gln B:15, Gly D:38 | Val D:39, His B:13 | Carbon hydrogen bond, π - π T-shaped, π -alkyl and van der Waals | | |
| | SPECT | A-S-53 | 6.0 | Gly J29, Ala J30, Ile J31, His B13, Gln B15, Gly D38, Leu A17, Gly D37, Gly C38, Gly C37, Val C36, Val D36, Leu B17, Val E36, Val D39, Val E39, Gly E38, Ile K31, Gly E37, Val D40, Ala L30, Gly K29 Ala K30 | Gly J:29, His B:13, Gly D:38, Leu A:17, Gly D:37, Gly C:37, Val D:39, Val E:39, Ile K:31, Val D:40 | Conventional hydrogen bond, carbon hydrogen bond, π -donor hydrogen bond, π -sigma, π -alkyl and van der Waals | | |
| | | A-S-52 | 5.77 | Ile K31, Gly E38, Val D39, Gly D38, Gln C15, Val C39, Gly C38, Val D36, Leu A17, Gly D37, Gln B15, Lys B16, Val D40, Gly E37, Val E39, His B13, Ala L30, Gly J29, Gly K29, Ala K30 | Ile K:31, Val D:39, Val D:40, Val E:39, His B:13, Ala L:30, Gly J:29, Gly K:29 | Carbon hydrogen bond, π -donor hydrogen bond, alkyl, π -alkyl and van der Waals | | |
| | | A-S-55 | 2.18 | Ile I:31, Ile J:31, Gly C:38, Gly C:37, Gly D:38, Gly D:37, Leu A:17, Leu C:17, Gln C:16, Val E:39, Gly E:38, Gln B:15, Val D:39, Val C:39, Val D:40, Val C:40, Gly J:29, Gly I:29 | Val D:39, Val D:40, Gly I:29 | Carbon hydrogen bond, alkyl, π -alkyl and van der Waals | | |
| | | A-S-20 | 2.66 | Val A:12, Val D:39, Leu A:17, Gly D:38, Leu C:17, Val D:36, Val E:36, Gly E:37, Gly E:38, Leu B:17, Gln B:15, His B:13 | Val D:39, Val D:36, Val E:36, His B:13 | Conventional hydrogen bond, carbon hydrogen bond, π -alkyl and van der Waals | | |
| | | Tau protein | PET and SPECT (combined) | T-P-2 | 4.319 | Trp A:230, Asn A:226, Leu A:174, Lys A:122, Phe A:119, Asn A:42, Ser A:45, Lys A:49, Val A:178, Leu A:229 | Trp A:230, Asn A:226, Leu A:174, Val A:178, Leu A:229 | Conventional hydrogen bond, halogen (Fluorine), alkyl, π -alkyl, water hydrogen bond and van der Waals |
| | | | | T-S-29 | 3.959 | Leu A:229, Val A:178, Leu A:174, Leu A:222, Ile A:219, Lys A:122, Asn A:175 | Leu A:229, Val A:178, Leu A:174, Leu A:222, Asn A:175 | Conventional hydrogen bonds, alkyl, π -alkyl, water hydrogen bond and van der Waals |
| | | T-P-10 | 1.311 | Lys A:122, Gly A:171, Asn A:175, Leu A:174, Glu A:182, Asn A:226, Val A:178 | Leu A:174, Val A:178 | π -alkyl, water hydrogen bond and van der Waals | | |
| | | T-P-7 | 1.957 | Leu A:174, Val A:178, Arg A:129, Asn A:175, Ser A:45, Val A:46, Asn A:42, Phe A:119, Lys A:49, Leu A:222, Asn A:226 | Leu A:174, Val A:178 | π -alkyl, water hydrogen bond and van der Waals | | |

and receptor plays a vital role in binding. This observation supports the occurrence of **TPSA(Tot)** (total polar surface area using N, S, O, and P polar contributions) and **nHDon**

(number of donor atoms) descriptors in the QSAR models. Furthermore, formation of π -sulfur bonds can be correlated with the **T(O..S)** descriptor, where the presence of sulfur

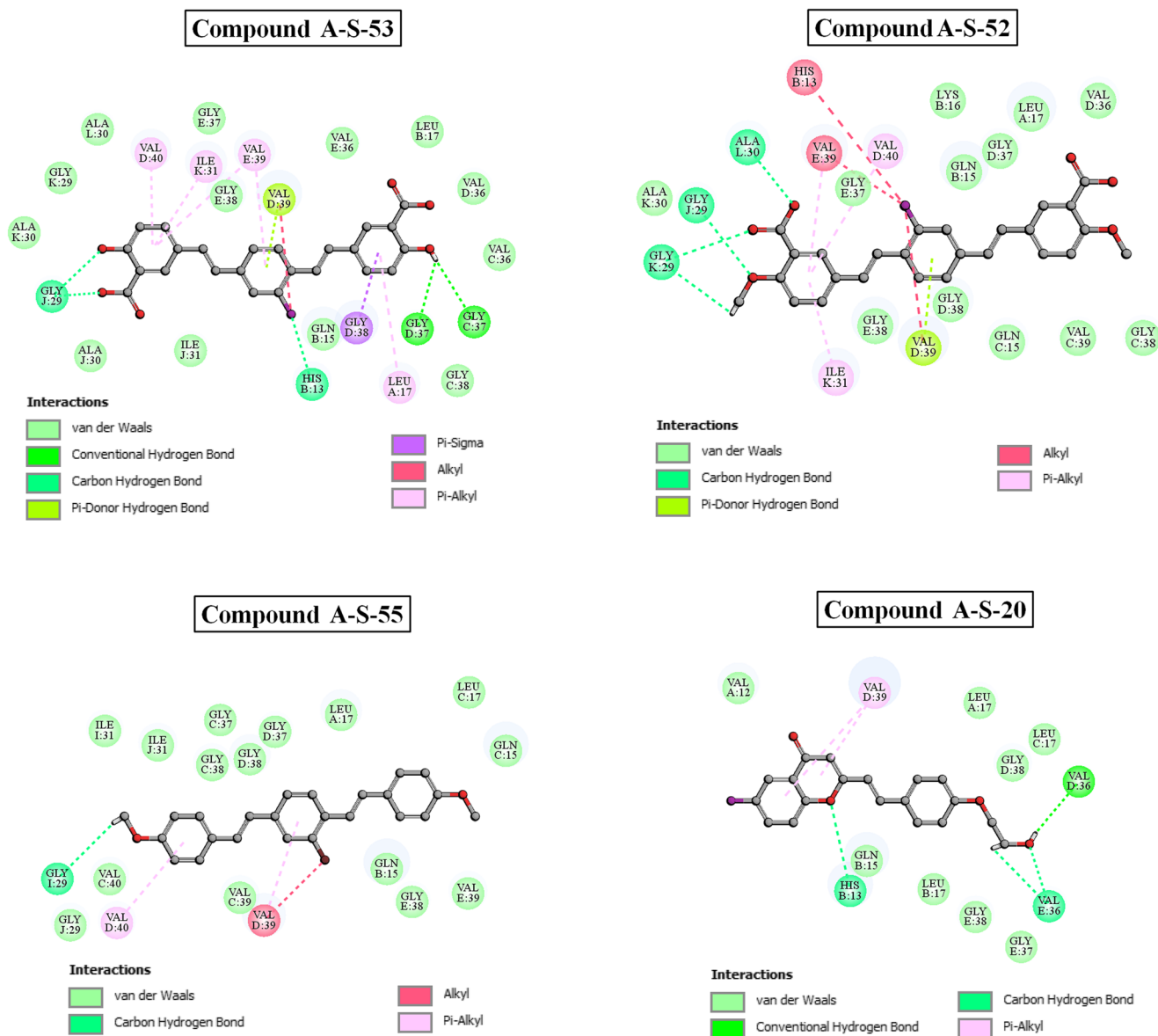


Fig. 9 Molecular interactions between high / low active SPECT imaging agents and A β protein

atoms in the molecules is essential. Val A12 residue forms hydrophobic interaction with compound **A-P-56** due to the presence of bromine (halogen) which corroborates with the **nArX** descriptor proving that the presence of halogen groups is beneficial for binding.

Molecular docking for selected SPECT imaging agents against A β plaques

In compounds like **A-S-53** and **A-S-52** having higher binding affinity ($pK_i = 6.0$ and 5.77 respectively), interaction forces include hydrogen bonding (carbon-hydrogen bonding, conventional hydrogen bonding and π -donor hydrogen bonds), π interactions (like π -alkyl, π -sigma interactions, π -lone pair

interactions and amide- π interactions), and alkyl interactions. The amino acid residues interacting with compound **A-S-53** are Gly J:29, His B:13, Gly D:38, Leu A:17, Gly D:37, Gly C:37, Val D:39, Val E:39, Ile K:31, and Val D:40. In Fig. 9, we can see the interactions for the most stable pose, where Val D:40, Ile K:31, Val D:39, and Leu A:17 make π -alkyl [77, 78] interactions with the ligand due to the presence of unsaturation in the ligand moiety. Gly D:38 makes π -sigma interaction with the ligand. Hydrogen bond interactions are observed with Gly J:29, Val D:39, Gly D:37, and Gly C:37. In compound **A-S-52**, hydrogen bond interaction such as carbon hydrogen bonds is observed with Gly K:29, Gly J:29 and Ala L:30 whereas Val D:39 makes π -donor hydrogen bond interaction. π -Alkyl interaction is observed with His B:13, Val D:40, and

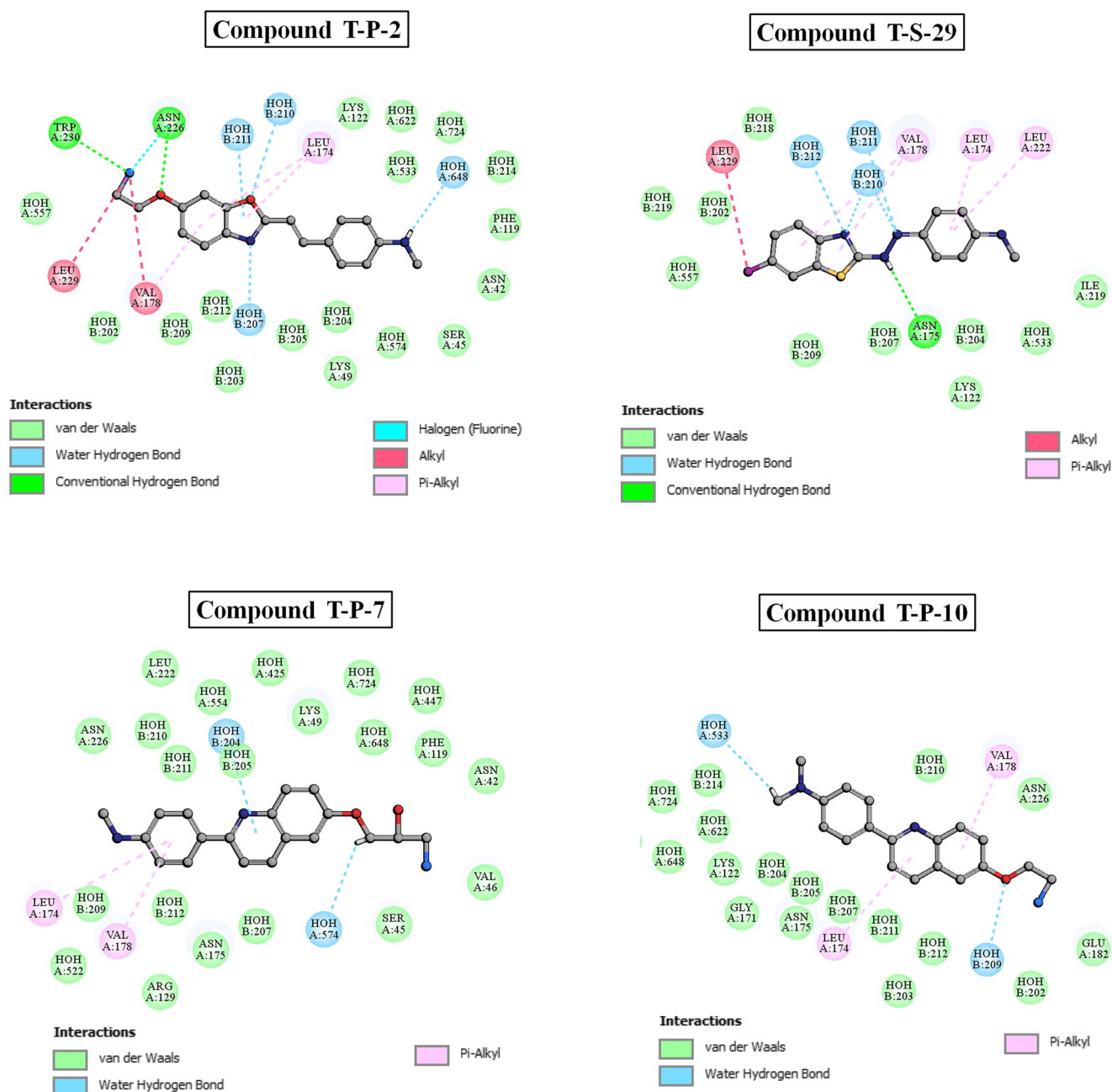


Fig. 10 Molecular interactions between high / low active PET or SPECT imaging agents and tau protein

Ile K:31. The interacting amino acid residues are Ile K:31, Val D:39, Val D:40, Val E:39, His B:13, Ala L:30, Gly J:29, and Gly K:29.

In compounds like **A-S-55** and **A-S-20** having low binding affinity ($pK_i = 2.18$ and 2.66 respectively), similar kinds of interactions are observed like hydrogen bond and π interactions but the number of interacting residues is much less (Fig. 9). The docking sites for both high and low binding affinity SPECT imaging agents targeted against $A\beta$ are given in Table 2.

Relation with QSAR models From the docking study, it is observed that hydrogen bonding formation between the protein

receptor and ligand molecule plays an important role in binding affinity of the later. This observation corroborates with the **SAacc** (denotes the surface area of acceptor atoms) descriptor occurred in the QSAR model.

Molecular docking for selected PET and SPECT imaging agents against tau protein

The tau protein (PDB ID: 6FAU) was docked with higher and lower active imaging agents in order to study their binding pattern and the molecular interactions occurring between them. In compounds like **T-P-2** and **T-S-29** with high binding

affinities ($pK_i = 4.319$ and 3.959 respectively), higher number of hydrogen bonding interactions and π interactions has been observed. Compound **T-P-2** makes interaction with Trp A:230, Asn A:226, Leu A:174, Val A:178, and Leu A:229 amino acid residues (Fig. 10). The stable pose makes π -alkyl interaction with Val A:178 and Leu A:174. The fluorine atom makes alkyl interaction with Leu A:229 and Val A:178 and halogen interaction with Asn A:226. The amino acid residues interacting with compound **T-S-29** are Leu A:229, Val A:178, Leu A:174, and Leu A:222. From Fig. 10, it is seen that π interaction is the predominant binding mode with the protein (as observed with Val A:178, Leu A:174 and Leu A:222). Other interactions noticed are alkyl interaction and various hydrogen-bonding interactions. Low affinity compounds include **T-P-7** and **T-P-10** ($pK_i = 1.311$ and 1.957) which showed less number of interactions in comparison to higher affinity compounds (in Fig. 10). Two π -alkyl interactions are observed for both the compounds, with Leu A:174 and Val A:178 in both the cases. The docking sites for both high and low binding affinity PET and SPECT imaging agents targeted against tau protein are given in Table 2.

Relation with QSAR models In the docking study, it is observed that π interactions play a vital role in ligand-receptor binding. This observation supports the occurrence of **D/Dtr09** (distance/detour ring of order 9) descriptor which is a ring descriptor. Increased number of aromatic nuclei will increase the value of this descriptor thereby increasing the binding affinity, also paving way for more π interactions. Also, π interactions corroborate with **SaaCH** descriptor, where aaCH represents $-\text{CH}$ groups in an aromatic nucleus. From the observations, it is concluded that aromaticity is a major feature regulating the binding affinity of PET and SPECT imaging agents.

QSAR modeling and molecular docking studies for newly designed PET and SPECT imaging agents

A set of 12 imaging agents (6 each for both PET and SPECT) was designed for, and their QSAR prediction and docking studies were performed to understand the binding properties towards $A\beta$ plaques. Also, another 6 imaging agents (PET and SPECT combined) targeting tau protein were designed for QSAR model prediction and molecular binding. From the QSAR analysis, it was found that all the compounds designed for both $A\beta$ and tau protein gave good predicted binding affinity (Table S1 in Supplementary Materials) and also falls under the model applicability domain as calculated by DModX method. The docking interactions as given in Supplementary Materials (Figs. S7, S8, and S9) also support the observations found for the actual dataset compounds. Similar interactions are observed in case of the newly

designed compounds, thus ensuring the validity of the new design.

Conclusion

The present research used chemometric tools for investigating the binding affinity of PET and SPECT against $A\beta$ plaques and tau protein. The three QSAR models developed through DCV method in this study give knowledge about the essential structural requirements necessary for improved binding affinity against $A\beta$ plaques and tau fibril. Many of the imaging agents used for modeling inhibit plaque formation, in addition to just binding to β -amyloid. Thus, these compounds can also be considered as multifunctional imaging agents (useful for both binding and inhibition) [81]. Double cross-validation proved its efficacy in modeling large dataset compounds previously [82, 83]. In the present study, we have utilized small size datasets (< 50 compounds in two cases) where DCV has proved its competence in searching for optimum combination of descriptors for generating models with good predictive ability. Thus, it can be concluded that DCV can not only be applied in modeling of large datasets but it also is suitable for modeling smaller dataset compounds. Furthermore, new sets of designed PET and SPECT imaging agents with better predicted binding properties are reported in the current report. Further experiments might be conducted in the future on these potential compounds.

Funding information PD received financial assistance from the Department of Atomic Energy—Board of Research in Nuclear Sciences (DAE-BRNS).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Schilling LP, Zimmer ER, Shin M, Leuzy A, Pascoal TA, Benedet AL, Borelli WV, Palmieri A, Gauthier S, Rosa-Neto P (2016) Imaging Alzheimer's disease pathophysiology with PET. *Dement Neuropsychol* 10:79–90
- Klunk WE (1998) Biological markers of Alzheimer's disease. *Neurobiol Aging* 2:145–147
- Selkoe DJ (2001) Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* 81:741–766
- Wimo A, Winblad B, Aguero-Torres H, von Strauss E (2003) The magnitude of dementia occurrence in the world. *Alzheimer Dis Assoc Disord* 17:63–67
- Alzheimer Association. Alzheimer's and dementia facts and figures <https://www.alz.org/alzheimers-dementia/facts-figures>. Accessed on 20 Nov 2018
- Okamura N, Furumoto S, Harada R, Tago T, Yoshikawa T, Foderotavoletti M, Mulligan RS, Villemagne VL, Akatsu H, Yamamoto T

- (2013) Novel ^{18}F -labeled arylquinoline derivatives for noninvasive imaging of tau pathology in Alzheimer disease. *J Nucl Med* 54:1420–1427
7. Duyckaerts C, Clavaguera F, Potier M-C (2019) The prion-like propagation hypothesis in Alzheimer's and Parkinson's disease. *Curr Opin Neurol* 32:266–271
 8. Hamley IW (2012) The amyloid beta peptide: a chemist's perspective. Role in Alzheimer's and fibrillization. *Chem Rev* 112:5147–5192
 9. Lichtenberg B, Mandelkow EM, Hagestedt T, Mandelkow E (1988) Structure and elasticity of microtubule-associated protein tau. *Nature* 334:359
 10. Barghorn S, Davies P, Mandelkow E (2004) Tau paired helical filaments from Alzheimer's disease brain and assembled in vitro are based on β -structure in the core domain. *Biochemistry* 43:1694–1703
 11. Bondareff W, Mountjoy CQ, Roth M, Hauser DL (1989) Neurofibrillary degeneration and neuronal loss in Alzheimer's disease. *Neurobiol Aging* 10:709–715
 12. Bobinski M, Wegiel J, Wisniewski HM, Tarnawski M, Bobinski M, Reisberg B, De Leon MJ, Miller DC (1996) Neurofibrillary pathology—correlation with hippocampal formation atrophy in Alzheimer disease. *Neurobiol Aging* 17:909–919
 13. Ono M, Hayashi S, Matsumura K, Kimura H, Okamoto Y, Ihara M, Takahashi R, Mori H, Saji H (2011) Rhodanine and thiohydantoin derivatives for detecting tau pathology in Alzheimer's brains. *ACS Chem Neurosci* 2:269–275
 14. Wang Y, Klunk WE, Debnath ML, Huang G-F, Holt DP, Shao L, Mathis CA (2004) Development of a PET/SPECT agent for amyloid imaging in Alzheimer's disease. *J Mol Neurosci* 24:55–62
 15. Yang Y, Cui M, Jin B, Wang X, Li Z, Yu P, Jia J, Fu H, Jia H, Liu B (2013) $^{99\text{m}}\text{Tc}$ -labeled dibenzylideneacetone derivatives as potential SPECT probes for in vivo imaging of β -amyloid plaque. *Eur J Med Chem* 64:90–98
 16. Kung HF, Choi SR, Qu W, Zhang W, Skovronsky D (2009) ^{18}F stilbenes and styrylpyridines for PET imaging of A β plaques in Alzheimer's disease: a miniperspective. *J Med Chem* 53:933–941
 17. Rojo LE, Alzate-Morales J, Saavedra IN, Davies P, Maccioni RB (2010) Selective interaction of lansoprazole and astemizole with tau polymers: potential new clinical use in diagnosis of Alzheimer's disease. *J Alzheimers Dis* 19:573–589
 18. Jensen JR, Cisek K, Funk KE, Naphade S, Schafer KN, Kuret J (2011) Research towards tau imaging. *J Alzheimers Dis* 26:147–115
 19. Fodero-Tavoletti MT, Okamura N, Furumoto S, Mulligan RS, Connor AR, McLean CA, Cao D, Rigopoulos A, Cartwright GA, O'keefe G (2011) ^{18}F -THK523: a novel in vivo tau imaging ligand for Alzheimer's disease. *Brain* 134:1089–1100
 20. Villemagne VL, Furumoto S, Fodero-Tavoletti M, Harada R, Mulligan RS, Kudo Y, Masters CL, Yanai K, Rowe CC, Okamura N (2012) The challenges of tau imaging. *Future Neurol* 7:409–421
 21. Ono M, Saji H (2011) SPECT imaging agents for detecting cerebral β -amyloid plaques. *Int J Mol Imaging* 2011. <https://doi.org/10.1155/2011/543267>
 22. Small GW, Agdeppa ED, Kepe V, Satyamurthy N, Huang S-C, Barrio JR (2002) In vivo brain imaging of tangle burden in humans. *J Mol Neurosci* 19:321–327
 23. Shoghi-Jadid K, Small GW, Agdeppa ED, Kepe V, Ercoli LM, Siddarth P, Read S, Satyamurthy N, Petric A, Huang S-C (2002) Localization of neurofibrillary tangles and beta-amyloid plaques in the brains of living patients with Alzheimer disease. *Am J Geriatr Psychiatry* 10:24–35
 24. Hansch C, Leo A, Hoekman DH (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. American Chemical Society, Washington, DC
 25. Hansch C, Leo A, Mekapati SB, Kurup A (2004) Qsar and Adme. *Bioorg Med Chem* 12:3391–3400
 26. Klein C, Kaiser D, Kopp S, Chiba P, Ecker GF (2002) Similarity based SAR (SIBAR) as tool for early ADME profiling. *J Comput Aided Mol Des* 16:785–793
 27. Toropova MA (2017) Drug metabolism as an object of computational analysis by the Monte Carlo method. *Curr Drug Metab* 18:1123–1131
 28. Toropova AP, Toropov AA (2018) CORAL: Monte Carlo method to predict endpoints for medical chemistry. *Mini Rev Med Chem* 18:382–391
 29. Toropova AP, Toropov AA, Begum S, Achary PGR (2018) Blood brain barrier and Alzheimer's disease: similarity and dissimilarity of molecular alerts. *Curr Neuropharmacol* 16:769–785
 30. Toropova MA, Toropov AA, Raška Jr I, Rašková M (2015) Searching therapeutic agents for treatment of Alzheimer disease using the Monte Carlo method. *Comput Biol Med* 64:148–154
 31. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intell Lab Syst* 159:108–126
 32. Herholz K, Ebmeier K (2011) Clinical amyloid imaging in Alzheimer's disease. *Lancet Neurol* 10:667–670
 33. Cohen AD, Rabinovici GD, Mathis CA, Jagust WJ, Klunk WE, Ikonomic MD (2012) Using Pittsburgh compound B for in vivo PET imaging of fibrillar amyloid-beta. *Adv Pharmacol* 64:27–81
 34. Zhu L, Ploessl K, Kung HF (2014) PET/SPECT imaging agents for neurodegenerative diseases. *Chem Soc Rev* 43:6683–6691
 35. Mathis CA, Wang Y, Holt DP, Huang G-F, Debnath ML, Klunk WE (2003) Synthesis and evaluation of ^{11}C -labeled 6-substituted 2-arylbenzothiazoles as amyloid imaging agents. *J Med Chem* 46:2740–2754
 36. Ono M, Kawashima H, Nonaka A, Kawai T, Haratake M, Mori H, Kung M-P, Kung HF, Saji H, Nakayama M (2006) Novel benzofuran derivatives for PET imaging of β -amyloid plaques in Alzheimer's disease brains. *J Med Chem* 49:2725–2730
 37. Qu W, Kung M-P, Hou C, Jin L-W, Kung HF (2007) Radioiodinated aza-diphenylacetylenes as potential SPECT imaging agents for β -amyloid plaque detection. *Bioorg Med Chem Lett* 17:3581–3584
 38. Ono M, Cheng Y, Kimura H, Watanabe H, Matsumura K, Yoshimura M, Iikuni S, Okamoto Y, Ihara M, Takahashi R (2013) Development of novel ^{123}I -labeled pyridyl benzofuran derivatives for SPECT imaging of β -amyloid plaques in Alzheimer's disease. *PLoS One* 8:e74104
 39. Fuchigami T, Yamashita Y, Kawasaki M, Ogawa A, Haratake M, Atarashi R, Sano K, Nakagaki T, Ubagai K, Ono M (2016) Characterisation of radioiodinated flavonoid derivatives for SPECT imaging of cerebral prion deposits. *Sci Rep* 5:18440
 40. Maya Y, Ono M, Watanabe H, Haratake M, Saji H, Nakayama M (2008) Novel radioiodinated aurones as probes for SPECT imaging of β -amyloid plaques in the brain. *Bioconjug Chem* 20:95–101
 41. Alagille D, DaCosta H, Baldwin RM, Tamagnan GD (2011) 2-Arylimidazo [2, 1-b] benzothiazoles: a new family of amyloid binding agents with potential for PET and SPECT imaging of Alzheimer's brain. *Bioorg Med Chem Lett* 21:2966–2968
 42. Maya Y, Okumura Y, Kobayashi R, Onishi T, Shoyama Y, Barret O, Alagille D, Jennings D, Marek K, Seibyl J (2015) Preclinical properties and human in vivo assessment of 123 I-ABC577 as a novel SPECT agent for imaging amyloid- β . *Brain* 139:193–203
 43. Kung M-P, Hou C, Zhuang Z-P, Skovronsky DM, Zhang B, Gur TL, Trojanowski JQ, Lee VMY, Kung HF (2002) Radioiodinated styrylbenzene derivatives as potential SPECT imaging agents for amyloid plaque detection in Alzheimer's disease. *J Mol Neurosci* 19:7–10

44. Pan J, Mason NS, Debnath ML, Mathis CA, Klunk WE, Lin K-S (2013) Design, synthesis and structure–activity relationship of rhodium 2-arylbenzothiazoles as β -amyloid plaque binding agents. *Bioorg Med Chem Lett* 23:1720–1726
45. Okamura N, Suemoto T, Furumoto S, Suzuki M, Shimadzu H, Akatsu H, Yamamoto T, Fujiwara H, Nemoto M, Maruyama M (2005) Quinoline and benzimidazole derivatives: candidate probes for in vivo imaging of tau pathology in Alzheimer's disease. *J Neurosci* 25:10857–10862
46. Declercq L, Celen S, Lecina J, Ahamed M, Tousseyn T, Moechars D, Alcazar J, Ariza M, Fierens K, Bottelbergs A (2016) Comparison of new tau PET-tracer candidates with [18 F] T808 and [18 F] T807. *Mol Imaging* 15:1536012115624920
47. Tago T, Furumoto S, Okamura N, Harada R, Adachi H, Ishikawa Y, Yanai K, Iwata R, Kudo Y (2016) Structure–activity relationship of 2-arylquinolines as PET imaging tracers for tau pathology in Alzheimer disease. *J Nucl Med* 57:608–614
48. Hashimoto H, Kawamura K, Takei M, Igarashi N, Fujishiro T, Shiomi S, Watanabe R, Muto M, Furutsuka K, Ito T (2015) Identification of a major radiometabolite of [11 C] PBB3. *Nucl Med Biol* 42:905–910
49. Tago T, Furumoto S, Okamura N, Harada R, Ishikawa Y, Arai H, Yanai K, Iwata R, Kudo Y (2014) Synthesis and preliminary evaluation of 2-arylhydroxyquinoline derivatives for tau imaging. *J Label Compd Radiopharm* 57:18–24
50. Matsumura K, Ono M, Hayashi S, Kimura H, Okamoto Y, Ihara M, Takahashi R, Mori H, Saji H (2011) Phenylidiazanyl benzothiazole derivatives as probes for in vivo imaging of neurofibrillary tangles in Alzheimer's disease brains. *MedChemComm* 2:596–600
51. MarvinSketch software, <https://www.chemaxon.com>. Accessed 28 Dec 2018
52. Dragon version 7, Kodesrl, Milan, Italy, 2016; software available at <http://www.taletemi.it/index.htm>. Accessed 03 Jan 2019
53. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* 17:241–253
54. Golmohammadi H, Dashtbozorgi Z, Acree Jr WE (2012) Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur J Pharm Sci* 47:421–429
55. Park H-S, Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. *Expert Syst Appl* 36:3336–3341
56. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opin Drug Discov* 13:1075–1089
57. Pope PT, Webster JT (1972) The use of an F-statistic in stepwise regression procedures. *Technometrics* 14:327–340
58. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58:109–130
59. Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform* 6:47
60. Roy K, Kar S, Das RN (2015) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, New York
61. Todeschini R, Ballabio D, Grisoni F (2016) Beware of unreliable Q^2 ! A comparative study of regression metrics for predictivity assessment of QSAR models. *J Chem Inf Model* 56:1905–1913
62. Chirico N, Gramatica P (2012) Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J Chem Inf Model* 52:2044–2058
63. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 52:396–408
64. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33
65. Paravastu AK, Leapman RD, Yau W-M, Tycko R (2008) Molecular structural basis for polymorphism in Alzheimer's β -amyloid fibrils. *Proc Natl Acad Sci* 105:18349–18354
66. Andrei SA, Meijer FA, Neves JF, Brunsvelde L, Landrieu I, Ottmann C, Milroy L-G (2018) Inhibition of 14-3-3/Tau by hybrid small-molecule peptides operating via two different binding modes. *ACS Chem Neurosci* 9:2639–2654
67. BIOVIA Discovery studio 2018. <http://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/requirements/technical-requirements-410.html>. Accessed 08 Feb 2019
68. Wu G, Robertson DH, Brooks Iii CL, Vieth M (2003) Detailed analysis of grid-based molecular docking: a case study of CDOCKER—a CHARMM-based MD docking algorithm. *J Comput Chem* 24:1549–1562
69. Benfenati E (2011) Quantitative structure–activity relationships (QSAR) for pesticide regulatory purposes. Elsevier, Amsterdam
70. Chartrand G, Johns GL, Tian S (1993) Detour distance in graphs. *Ann Discrete Math* 55:127–136
71. Akarachantachote N, Chadcham S, Saithanu K (2014) Cutoff threshold of variable importance in projection for variable selection. *Int J Pure Appl Math* 94:307–322
72. Jackson JE (2005) A user's guide to principal components. John Wiley & Sons, New Jersey
73. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 1:45–63
74. U. Simca-P, 10.0, info@umetrics.com, www.umetrics.com, Umea, Sweden, 2002. Accessed 22 Jan 2019
75. Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47:2345–2357
76. Alkorta I, Rozas I, Elguero J (1998) Non-conventional hydrogen bonds. *Chem Soc Rev* 27:163–170
77. Ribas J, Cubero E, Luque FJ, Orozco M (2002) Theoretical study of alkyl- π and aryl- π interactions. Reconciling theory and experiment. *J Org Chem* 67:7057–7065
78. Echeverría J (2017) Alkyl groups as electron density donors in π -hole bonding. *CrystEngComm* 19:6289–6296
79. Martínez CR, Iverson BL (2012) Rethinking the term “ π -stacking”. *Chem Sci* 3:2191–2201
80. Shiri F, Shahraki S, Baneshi S, Nejati-Yazdinejad M, Majd MH (2016) Synthesis, characterization, in vitro cytotoxicity, in silico ADMET analysis and interaction studies of 5-dithiocarbamate-1, 3, 4-thiadiazole-2-thiol and its zinc (ii) complex with human serum albumin: combined spectroscopy and molecular docking investigations. *RSC Adv* 6:106516–106526
81. Darras FH, Pang Y-P (2017) On the use of the experimentally determined enzyme inhibition constant as a measure of absolute binding affinity. *Biochem Biophys Res Commun* 489:451–454
82. De P, Roy K (2018) Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. *SAR QSAR Environ Res* 29:319–337
83. Khan K, Benfenati E, Roy K (2019) Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. *Ecotoxicol Environ Saf* 168:287–297

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: a QSAR approach

Priyanka De¹ · Joyita Roy¹ · Dhananjay Bhattacharyya² · Kunal Roy¹

Received: 6 April 2020 / Accepted: 19 May 2020 / Published online: 25 May 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Recently, adenosine A_{2A} receptor antagonists have been identified as an interesting drug target for the treatment of Parkinson's disease (PD). Radiolabelled molecular imaging technologies such as positron emission tomography (PET) have emerged in the research field of medicinal chemistry as a diagnostic tool for PD. In the current study, we have performed quantitative structure–activity relationship (QSAR) analysis of 35 xanthine ligand PET tracers as A_{2A} R (adenosine receptors) antagonists in order to determine their structural features required to have binding affinity and selectivity towards A_{2A} R. The division of the dataset into training and test sets was done using a random method, while the feature selection for the binding affinity was done using Genetic Algorithm (GA). The best model with five descriptors was obtained using the spline option in the GA run. QSAR models with four descriptors were also developed for A_{2A} R selectivity, where significant descriptors were selected from the large pool of descriptors using stepwise regression method followed by Best Subset Selection (BSS) method. Furthermore, to improve the quality of the external predictions, we used the “Intelligent Consensus Predictor” tool (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/). Both the models showed robustness in terms of statistical parameters. Molecular docking studies have been carried out to understand the molecular interactions between the ligand and receptor, and the results are then correlated with the structural features obtained from the QSAR models. Furthermore, the information derived from the newly found descriptors gives an insight for the development of new candidate PET tracers for the use in PD.

Keywords Parkinson disease (PD) · Positron emission tomography (PET) · QSAR

Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder of the central nervous system characterized by muscle rigidity, bradykinesia, and tremor. The disease affects older people, and it is known that 2–3% of the population \geq 65 years of age are more prone towards this disease [1]. It is also associated with loss of dopaminergic neurons in the

substantia nigra, lewy body generation, and abnormal clustering of α -synuclein protein, which is directly connected to expectancy of long life. Hence, effective research for neurodegenerative disease treatment is one of the vital clinical needs of today's life. The current therapy of PD includes restoration of dopamine with levodopa in the striatum of the brain. However, to maintain the therapeutic level, the dosage has to be increased which does not prevent the underlying neuronal loss [2, 3]. On the other hand, such long-term treatment in addition may cause adverse effects which include levodopa-induced dyskinesia and behavioral disturbances in the individuals [4, 5].

Adenosine enzyme inhibitors can be considered an alternative medication in the treatment of PD having less degree of adverse effects. Adenosine is an endogenous modulator of different physiological functions in the peripheral tissues in addition to the central nervous system (CNS). It is a purine nucleoside having four varieties of subtypes consisting of A_1 , A_{2A} , A_{2B} , and A_3 . A_{2A} receptors are highly expressed in striatum (dopamine-rich areas of the CNS) where it is almost co-

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11224-020-01560-6>) contains supplementary material, which is available to authorized users.

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

¹ Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

² Computational Science Division, Saha Institute of Nuclear Physics, Kolkata 700064, India

located with dopamine D_2 receptor on GABAergic striatopallidal neurons [6]. A_{2A} antagonistically interferes with the D_2 receptor and as a result decreases the affinity of the D_2 receptor for dopamine upon stimulation and show opposite effect on motor function [7]. Thus, adenosine A_{2A} receptor blockade may show motoric improvement as proven by many animal models [8–11].

Positron emission tomography (PET) [12] and single-photon emission computed tomography (SPECT) [13] are non-invasive methodologies which make use of the dynamic distribution of the radiotracers and provide 3D map of the brain quantifying the biological processes. These imaging agents help in the detection and quantification of dopamine and adenosine receptors in the brain thereby creating a path for early detection of the disease. PET studies are superior to SPECT in terms of accurate results and in determining the temporal measurements of radioactivity with their regional distributions. Agonists and antagonists containing positron-emitting radioisotopes can be introduced in vivo to get 3D image of the receptors which have been helpful in CNS diagnosis. The PET tracers can be used as in vivo-imaging agents in order to improve the pharmacokinetics, physicochemical properties, and mapping of the receptor as per interest. As search of new compounds with desired activity is time-consuming and expensive, pharmaceutical companies have a great interest upon theoretical approaches to design compounds with desired activity.

Quantitative structure–activity relationships (QSAR) have gained a lot of attention in molecular modeling field and are beneficial due to less involvement of human resource and cost-effectiveness [14, 15]. It attempts to develop a correlation between the chemical structures with a well-defined activity. It expresses chemical structures and physiological property in the numerical form and develops a mathematical correlation between them. Furthermore, this relationship can be used to predict the biological response of other existing chemical structures. QSAR-based studies have shown useful applications in drug discovery, molecular modeling, pharmaceutical toxicity modeling, pharmacokinetics/toxicokinetics, data mining, environmental toxicity (ecotoxicity), chemical or drug property modeling, food science, agricultural sciences, pesticide toxicity, fragrance, nanoscience (Nano-QSAR), and many other fields [16–24]. QSAR is also used to predict the absorption, distribution, metabolism, excretion, and toxicological (ADMET) of drug like compounds [25, 26]. QSAR has widespread applications in drug design, medicinal chemistry, and predictive toxicology. It has also become an effective tool in understanding and determining the major biochemical features associated with the Parkinson's disease [27, 28].

In the present study, we have tried to develop QSAR models with PET tracers of xanthine ligands as A_{2A} R (adenosine receptors) antagonist using only 2D descriptors to

explore the structural features required for binding affinity towards A_{2A} R and selectivity of the tracers between A_{2B} and A_{2A} receptors.

Materials and methods

Dataset

The experimental binding affinity and selectivity data of 35 xanthine ligand-based PET tracers were taken from a previously published literature [29] and applied for QSAR modeling to determine the essential structural features needed for binding affinity and explore the structural requirements necessary to be present in the antagonists for selectivity towards A_{2A} adenosine receptors. The experimental values of selectivity and binding affinity (K_i) ranged from 0.1–20 nM and 7.84–16,500 nM respectively, and the details are provided in Supplementary Material I (Table S1). The experimental values were converted into negative logarithm scale during modeling and were used as independent values. No compounds with binding affinity data were removed during modeling but some compounds (**14**, **32**, **33**, and **34**) with no experimental selectivity values were eliminated during modeling. Here, the binding affinity and selectivity were separately used as endpoints or independent variables in modeling. The compounds for both the dataset were represented in the MarvinSketch software [30] with proper aromatization and addition of hydrogen bond as necessary.

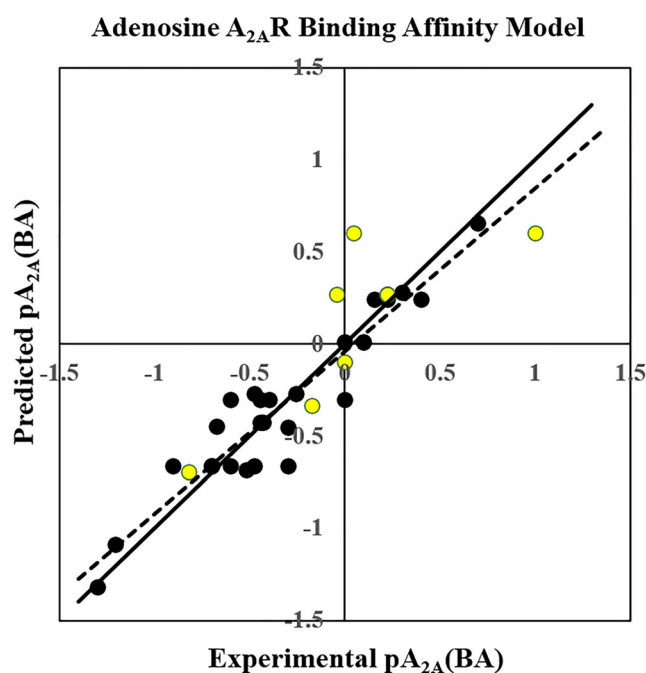


Fig. 1 Observed vs predicted A_{2A} R binding affinity scatter plot

Molecular descriptors

In the present study, QSAR models were developed using a selected class of two-dimensional molecular descriptors involving E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments, molecular property descriptors, and extended topochemical atom (ETA) indices. The ETA descriptors were calculated using the PaDel-Descriptor software [31], whereas the non-ETA descriptors were calculated using the Dragon 7 software [32]. Intercorrelated ($|r| > .95$), constant (variance < 0.0001), and other incompetent and redundant data was removed using an in-house software available at <http://dtclab.webs.com/software-tools> before model development.

Dataset division

Dataset division is a crucial part of QSAR modeling in order to develop a properly validated and robust model. Rational data division ensures an unbiased external validation along with uniform data distribution [33]. The division of the dataset into training set ($\sim 70\%$) and test set ($\sim 30\%$) was performed employing random dataset division method [34] for both binding affinity and selectivity end points. The training set was used for model development, and the test set was used for model validation.

Variable selection and model development

Prior to model development, we have performed variable selection strategies such as Genetic Algorithm (GA) [35, 36] and stepwise regression [35, 37] for binding affinity and selectivity, respectively, to extract the important and influential descriptors and created a reduced pool of descriptors. After obtaining the important descriptors, we went for model development. The best model with five descriptors was obtained using the spline option in the GA run on Discovery Studio version 4.1 for the binding affinity. On the other hand, for A_{2A} R selectivity, four models with four descriptors were selected from the Best Subset Selection (BSS) method based on MAE criteria [38]. Furthermore, to improve the quality of the external prediction via “intelligent” selection of multiple models, we have applied an “Intelligent consensus predictor” tool [39] developed in our laboratory [40].

Statistical validation metrics

The statistical quality of the models developed in the present study was rigorously examined using multiple approaches to check the robustness and predictivity of the developed models. All the models were validated both externally and internally. Various parameters like determination coefficient R^2 , explained variance R^2_a , variance ratio (F), and standard

error of estimate (s) were computed. Internal predictivity parameters such as predicted residual sum of squares (PRESS) and leave-one-out cross-validated correlation coefficient (Q^2_{LOO}) were also calculated along with external predictivity parameters like R^2_{pred} or Q^2_{F1} , Q^2_{F2} , and concordance correlation coefficient (CCC) [41]. It has been reported that consensus models are better in performance in comparison with an individual model [41]. Therefore, we have also performed “Intelligent Consensus Prediction (ICP)” using multiple models to see whether the quality of predictions can be increased through an intelligent selection.

Applicability domain

Applicability domain (AD) [42] is a theoretical region in the chemical space developed based on modeled descriptors and modeled response of the training set, where the developed model could make predictions basing on some logical reliability. Here, we have checked AD using standardization approach using the tool developed in our laboratory [40].

Molecular docking

Molecular docking analysis has been implemented in the present work that helps in understanding the intermolecular interactions taking place between the PET tracer antagonists and the A_{2A} receptor. The protein structure for adenosine A_{2A} receptor is retrieved from the protein data bank with PDB ID:3UZA [43]. The X-ray crystal structure of the protein consists of a bound ligand T4G commonly known as 6-(2,6-dimethylpyridin-4-yl)-5-phenyl-1,2,4-triazin-3-amine (formula: $C_{16}H_{15}N_5$). Before docking the target PET tracers, protein preparation was done by cleaning the protein for any missing residues, explicit hydrogen addition, and generation of the docking site. The generation of active docking site was done in the BIOVIA Discovery Studio platform from the ligand-binding domain of the bound ligand T4G by the selection of the ligand and generating the site “from current selection” program in receptor-ligand interaction module of the software. After the generation of the active ligand-binding domain, the bound ligand was removed for new molecule docking. For ligand preparation, the PET tracers were put through small molecule module in the Discovery Studio platform where a series of ligand conformers were generated. Each of these generated conformers was then used in the CDOCKER module energy for molecular docking involving CHARMM interaction [44]. The CDOCKER interaction energy parameter (kJ/mol) was checked for all the receptor ligand complexes, and the top scoring (most negative, thus favorable to binding) poses were kept.

Results and discussion

Based on the binding affinity and selectivity endpoints of 35 xanthine PET tracer antagonists of adenosine A_{2A} receptor, we have developed one model for the binding affinity ($Q^2 = 0.85$, $R^2 = 0.90$, $Q^2_{F1} = 0.80$) and 4 models ($Q^2 = 0.80$ – 0.87 , $R^2 = 0.87$ – 0.91 , $Q^2_{F1} = 0.84$ – 0.85) for selectivity. All the models were externally and internally validated which showed model robustness and good predictivity in terms of the statistical results. We have also checked the r_m^2 parameters for both internal sets ($\overline{r_{m(loo)}^2}$, $\Delta r_{m(loo)}^2$) and external sets ($r_{m(test)}^2$ and $\Delta r_{m(test)}^2$), and the statistical results were above the critical point justifying the reliability of the models. To improve the quality of the external prediction for selectivity, we also performed “*Intelligent Consensus Prediction*” of the multiple MLR models using the ICP tool [39], and found that the consensus predictions were better than the individual MLR model-derived predictions. The winner model was consensus model 0 (CM0).

Modeling binding affinity of PET tracers towards adenosine (A_{2A}) receptor

The model for binding affinity consists of five descriptors: C-025, F09 [N-O], nBnz, NRS, and nCIR which significantly influence the binding of the antagonists to the adenosine (A_{2A}) receptor. The 5 descriptor MLR

model (Eq. 1) developed using Genetic Function Algorithm (GFA) could predict 85.0% variance of the training set and 80.0% of the test set. The values of all descriptors appearing in the model for training and test set compounds are given in Supplementary Material II (Excel file) and the scatter plot of the observed vs. predicted binding affinity is shown in Fig. 1.

$$pKi(A_{2A}R) = -0.849(\pm 0.2167) \\ -0.36271(\pm 0.06190) C-025 \\ + 0.17693(\pm 0.05895) F09[N-O] \\ -0.52109(\pm 0.07616) NRS \\ + 0.81699(\pm 0.09908) nBnz \\ + 0.3024(\pm 0.03363) nCIR$$

$$n_{\text{training}} = 25, R^2 = 0.901, R^2_{\text{adj}} = 0.875, Q^2 = 0.850, S \\ = 0.170027, F = 34.62, \text{PRESS} \\ = 0.833306, \overline{r_{m(\text{LOO})}^2} = 0.790, \Delta r_{m(\text{LOO})}^2 \\ = 0.072, \text{MAE-based criteria} = \text{Moderate}$$

$$n_{\text{test}} = 10, Q^2_{F1} = 0.80, Q^2_{F2} = 0.681, \overline{r_{m(\text{test})}^2} \\ = 0.54, \Delta r_{m(\text{test})}^2 = 0.23, \text{MAE-based criteria} = \text{Good}$$

Table 1 Definition and contribution of all the descriptors obtained from the MLR models (models developed by using binding affinity)

| Sl. no. | Name of descriptors | Descriptor type | Contribution | Discussion | Probable mechanism of binding |
|---------|---------------------|-----------------------------------|--------------|--|---|
| 1 | C-025 | Atom-centered fragment descriptor | -ve | C-025 can be depicted as R-CR-R , where ‘R’ can be any group linked to carbon and ‘-’ is any aromatic bond. It is the number of fragments in which a C (sp ²) aromatic atom is bound to three carbon atoms, two of them by an “aromatic bond” and the third by a simple single bond | Flexibility which helps in accommodating the antagonist well in the receptor pocket |
| 2 | nBnz | Ring descriptor | +ve | Indicates number of benzene-like rings | π - π Stacking interaction |
| 3 | F09 [N-O] | 2D atom pair descriptor | +ve | Frequency of N-O fragment at the topological distance 9 | Hydrogen bonding |
| 4 | NRS | Ring descriptor | -ve | A ring descriptor indicates number of ring systems within a molecule | - |
| 5 | nCIR | Ring descriptor | +ve | Number of circuits, i.e., larger loops around two or more rings in a molecule | Hydrophobic interaction/ π - π stacking interaction |

Essential features required for binding and receptor interaction

The descriptors obtained in the QSAR model (Table 1) give an insight regarding the mechanism of interaction occurring during binding of the xanthine PET tracer antagonists to adenosine A_{2A} receptor. Unsaturation and aromaticity play a dominating role in regulating the receptor binding affinity which is evident from the occurrence of descriptors such as **C-025**, **nBnz**, **NRS**, and **nCIR**. Descriptors like nBnz and nCIR have positive influences on the adenosine A_{2A} receptor binding (Fig. 2). But on the other hand, descriptors like C-025 and NRS have negative effects on the binding affinity of the PET tracers (Fig. 3). The occurrence of these similar types of descriptors with opposite influence is contradictory and leads to a conclusion that aromaticity provided by benzene nucleus (as seen in compounds like **A-32** and **A-23**) is more important for binding. On the other hand, the presence of heterocyclic aromatic rings and fused-ring systems decrease the overall binding affinity of the radiotracer molecule (found in compounds **A-1**, **A-2**, and **A-20**).

The 2D atom pair descriptor **F09 [N-O]** gives information about the electronegativity of the compounds, and the positive coefficient of the descriptor suggests that higher occurrence of nitrogen and oxygen at topological distance 9 would enhance binding affinity of the compounds as seen in compounds **A-4** and **A-32**. It is found that the presence of electronegative atoms in the compounds or chemical structures can influence the binding to the receptor through hydrogen bonding [45].

Molecular docking Molecular docking helped in understanding the optimized conformation of the complex between the imaging agent and A_{2A} receptor and gave evidences related to the orientation of the imaging agents at the binding zone of the receptor. The major goal was to understand the molecular interactions taking place during radiotracer binding and correlate these findings with QSAR analysis. The docking analysis showed the predominance of different types of π bonding interactions and hydrogen-bonding interactions. In higher active compounds (Fig. 4) like **A-4**, **A-8**, and **A-25** ($pA_{2A}R(BA) = 0.699$, 1.000, and 0.398 respectively), the interaction forces include mainly hydrogen-bonding interactions (conventional hydrogen bond and carbon-hydrogen bond interaction), π interactions (π -cation, π -donor hydrogen, π - π stacked, π - π T-

shaped, and π -alkyl). Other interactions include halogen and alkyl interaction in compound **A-4** and salt bridge formation in compound **A-8**. Higher number of interacting residues supports the fact that these compounds have higher binding affinity. Compounds having binding affinity in the medium range (Fig. 5) like compound numbers **A-14** and **A-27** ($pA_{2A}R(BA) = -0.301$ and -0.255 respectively) make less number of interactions with the adenosine receptor, but the type of interactions remains similar, i.e., π interactions and hydrogen-bonding interactions. The lowest active compounds (Fig. 5) like compound numbers **A-20** and **A-35** ($pA_{2A}R(BA) = -1.301$ and -1.204 respectively) show the least number of interactions. All the details of binding including interacting residues and type of binding interactions are given in Table 2.

Relationship with QSAR models The docking study shows different types of π interactions occurring between the PET radiotracer molecules and adenosine A_{2A} receptor. This observation supports the occurrence of **nBnz** and **nCIR** descriptors obtained in the QSAR models. The presence of aromatic rings like benzene can enhance binding with the receptor through aromatic π - π stacking interaction with the phenyl/imidazole residue of the receptor [46]. The interaction of these antagonists through π - π stacking interaction eventually blocks the receptor in the indirect pathway thus blocking the activity of GABA-mediated influence in the globus pallidus pars externa (GPe). This helps the PD patients to gain the motor function again by regaining the balance between direct and indirect pathway. Nitrogen and oxygen are capable of hydrogen bond formation and various types of hydrogen bonding as observed in both higher active and lower active compounds, and this can be also correlated to the **F09[N-O]** descriptor which gives an idea about the electronegativity of the molecule.

Modeling selectivity of PET tracers towards adenosine (A_{2A}) receptor

In the current work, we have developed four MLR models to understand the selectivity of the PET tracer molecules towards adenosine A_{2A} receptor. A single QSAR model may not be efficient enough for the prediction of activity since the property of molecules cannot be understood by a limited number of

Fig. 2 Features increasing the binding affinity (pKi) value

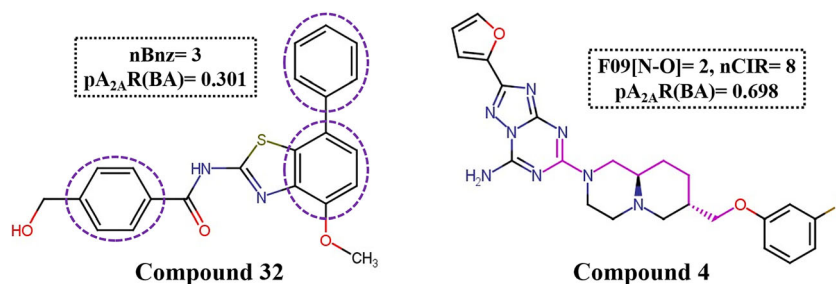
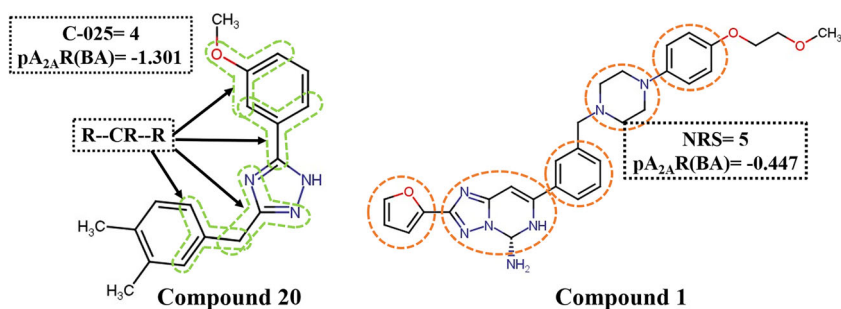


Fig. 3 Features decreasing the binding affinity (pKi) value



features. The use of multiple models for prediction using consensus approach helps in reducing model uncertainty by enhancing the prediction quality of the external set and also in reducing the prediction errors [38]. The four MLR models are given below:

Model 1

$$\log A_{2A}R(SeI) = 0.5875(\pm 0.4130) + 0.4643(\pm 0.1574)C-027 - 0.8679(\pm 0.1797)C-040 + 0.7245(\pm 0.1006)F09[N-O] + 0.8382(\pm 0.01749)ETA_Beta.s$$

$$n_{\text{training}} = 21, R^2 = 0.915, R^2_{\text{adj}} = 0.893, Q^2 = 0.867, S = 0.234982, F = 42.88,$$

$$PRESS = 1.37546, \overline{r^2}_{m(\text{LOO})} = 0.81227, \Delta r^2_{m(\text{LOO})} = 0.07373, \text{MAE-based criteria} = \text{Moderate}$$

$$n_{\text{test}} = 10, Q^2_{F1} = 0.84, Q^2_{F2} = 0.81, \overline{r^2}_{m(\text{test})} = 0.7682, \Delta r^2_{m(\text{test})} = 0.11949, \text{MAE-based criteria} = \text{Good}$$

Model 2

$$\log A_{2A}R(SeI) = 0.36359(\pm 0.43605) - 0.76227(\pm 0.18863)C-040 - 0.05224(\pm 0.02421)T(F..Cl) + 0.71046(\pm 0.11057)F09[N-O] + 0.09777(\pm 0.01808)ETA_Beta.s$$

$$n_{\text{training}} = 21, R^2 = 0.90, R^2_{\text{adj}} = 0.87, Q^2 = 0.82, S = 0.274853, F = 35.21,$$

$$PRESS = 1.05627, \overline{r^2}_{m(\text{LOO})} = 0.7526, \Delta r^2_{m(\text{LOO})} = 0.05874, \text{MAE-based criteria} = \text{Moderate}$$

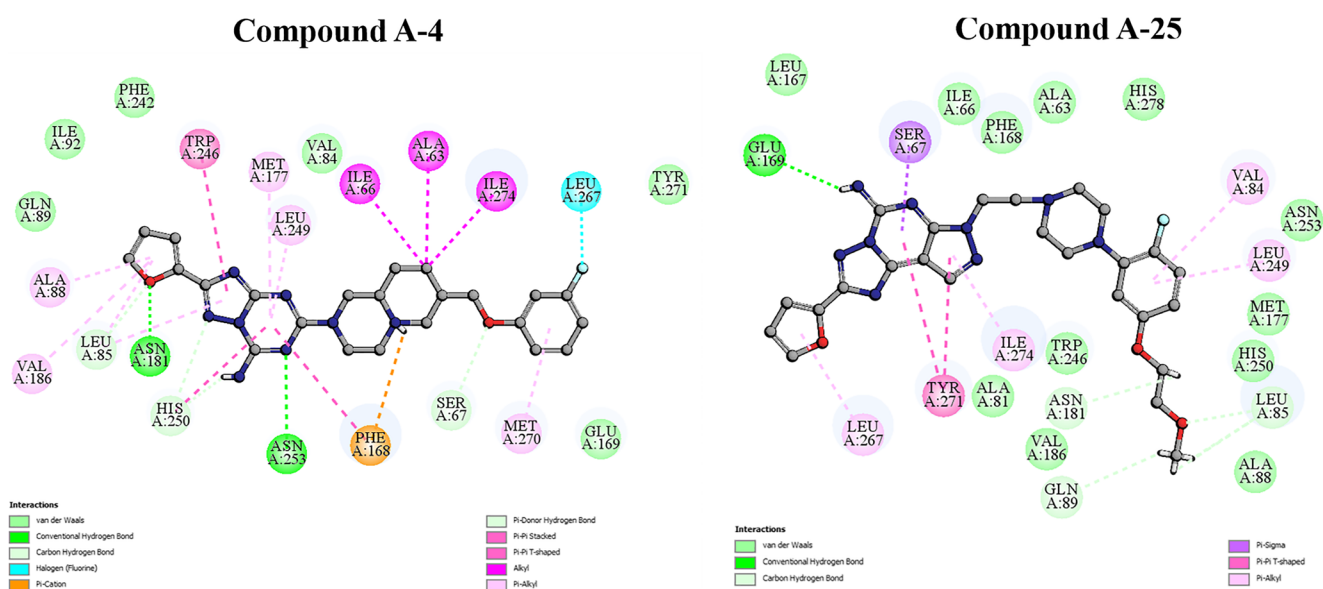


Fig. 4 Docking interactions for compounds having higher binding affinity (pKi)

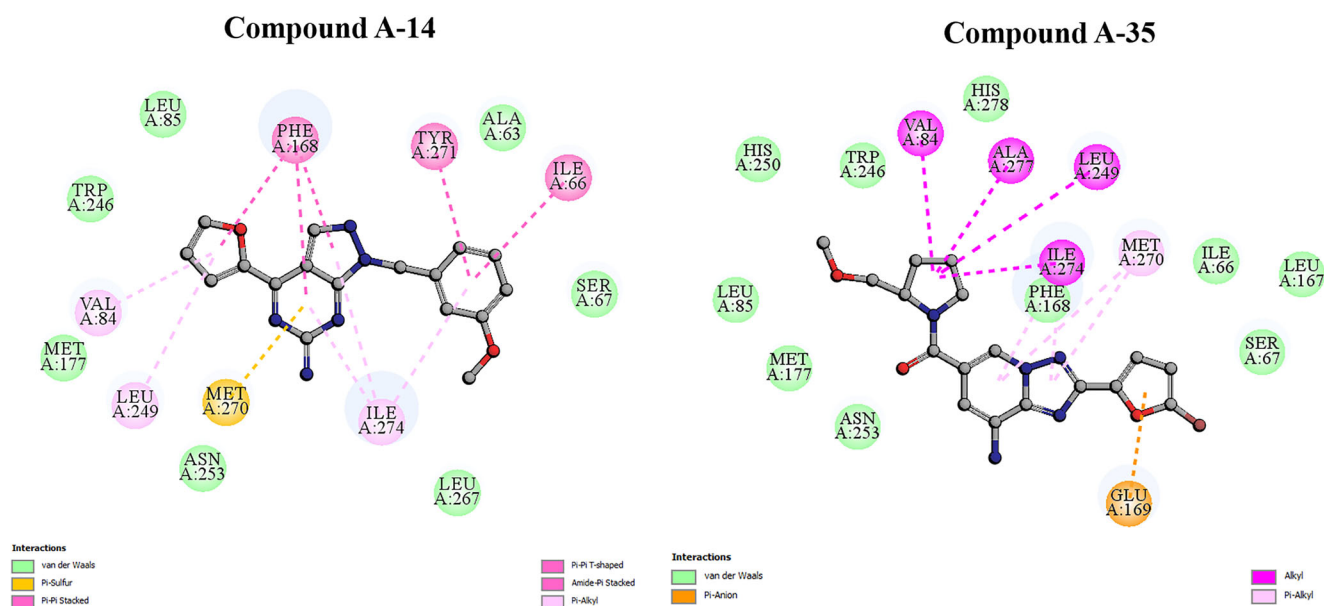


Fig. 5 Docking interactions for compounds having medium (A-14) and low (A-35) binding affinity (pKi)

$$\begin{aligned}
 n_{\text{test}} &= 10, Q_{F1}^2 = 0.84, Q_{F2}^2 = 0.82, \overline{r_{m(\text{test})}^2} \\
 &= 0.7737, \Delta r_{m(\text{test})}^2 = 0.04197, \text{MAE-based criteria} \\
 &= \text{Good}
 \end{aligned}$$

Model 3

$$\begin{aligned}
 \log A_{2A}R(\text{Sel}) &= 0.9642(\pm 0.4535) \\
 &+ 0.31245(\pm 0.08846) \text{ nCIC} \\
 &+ 0.4848(\pm 0.1856) \text{ C-027} \\
 &- 0.9394(\pm 0.2114) \text{ C-040} \\
 &+ 0.6662(\pm 0.1184) \text{ F09[N-O]} \\
 n_{\text{training}} &= 21, R^2 = 0.883, R_{\text{adj}}^2 = 0.854, S = 0.274853, F \\
 &= 30.27, \\
 \text{PRESS} &= 1.72765, Q^2 = 0.833, \overline{r_{m(\text{LOO})}^2} \\
 &= 0.76, \Delta r_{m(\text{LOO})}^2 = 0.12, \text{MAE-based criteria} \\
 &= \text{Moderate}, \\
 n_{\text{test}} &= 10, Q_{F1}^2 = 0.84, Q_{F2}^2 = 0.82, \overline{r_{m(\text{test})}^2} \\
 &= 0.77, \Delta r_{m(\text{test})}^2 = 0.13, \text{MAE-based criteria} \\
 &= \text{Good}
 \end{aligned}$$

Model 4

$$\begin{aligned}
 \log A_{2A}R(\text{Sel}) &= 1.3245(\pm 0.2988) - 0.6702(\pm 0.2119) \text{ C-040} \\
 &+ 0.10445(\pm 0.04427) \text{ SsssN} \\
 &+ 0.05519(\pm 0.01932) \text{ F07[C-C]} \\
 &+ 0.5954(\pm 0.1263) \text{ F09[N-O]} \\
 n_{\text{training}} &= 21, R^2 = 0.872, R_{\text{adj}}^2 = 0.84, S = 0.287861, F \\
 &= 27.24, \\
 \text{PRESS} &= 2.09555, Q^2 = 0.827, \overline{r_{m(\text{LOO})}^2} = 0.717, \Delta r_{m(\text{LOO})}^2 \\
 &= 0.131, \text{MAE-based criteria} = \text{Moderate}, \\
 n_{\text{test}} &= 10, Q_{F1}^2 = 0.85, Q_{F2}^2 = 0.83, \overline{r_{m(\text{test})}^2} = 0.78, \Delta r_{m(\text{test})}^2 \\
 &= 0.07, \text{MAE-based criteria} = \text{Good}
 \end{aligned}$$

The significant descriptors obtained from the four MLR models (M1–M4) contributing to A_{2A} receptor selectivity are C-040, C-027, F09 [N-O], ETA_Beta_s, nCIC, T (F..Cl), SsssN, and F07[C-C]. All the descriptors positively contribute to the A_{2A} receptor selectivity, except C-040, as identified from the regression coefficients of the descriptors and summarized in Table 3. We have also checked the applicability domain of the developed MLR models. The models showed good predictive ability as per the statistical results. The details of the descriptors, their contribution, and frequency of appearance in all the four models are explained elaborately in

Table 2 Details of interacting residues and different types of binding interaction occurring between the PET imaging agents and the target protein (adenosine A_{2A} receptor)

| Compound no. | Activity | Binding affinity [pA _{2A} R(BA)] | Interacting residues | Binding interactions |
|--------------|----------|---|---|---|
| A-4_6 | High | 0.699 | Ala A:88, Val A:186, Leu A:85, Asn A:181, His A:250, Asn A:253, Phe A:168, Ser A:67, Met A:270, Leu A:267, Ile A:274, Ala A:63, Ile A:66, Leu A:249, Met A:177, Trp A:246 | Conventional hydrogen bond, carbon hydrogen bond, halogen (fluorine), π -cation, π -donor hydrogen bond, π - π stacked, π - π T-shaped, alkyl, π -alkyl |
| A-8 | High | 1.000 | Met A:270, Asn A:253, Leu A:249, Phe A:168, Ala A:81, Ile A:66, Glu A:169 | Conventional hydrogen bond, carbon hydrogen bond, π - π stacked, π -alkyl, salt bridge |
| A-25 | High | 0.398 | Leu A:267, Tyr A:271, Ile A:274, Asn A:181, Gln A:89, Leu A:85, Leu A:249, Val A:84, Ser A:67, Glu A:169 | Conventional hydrogen bond, carbon hydrogen bond, π -sigma, π - π T-shaped, π -alkyl |
| A-14 | Medium | -0.301 | Val A:84, Leu A:249, Met A:270, Ile A:274, Ile A:66, Tyr A:271, Phe A:168 | π -sulfur, π - π T-shaped, π - π stacked, amide- π stacked, π -alkyl |
| A-27 | Medium | -0.255 | Asn A:253, Ser A:67, Ile A:274, Leu A:167, Glu A:169, Ala A:63, Ile A:66, Leu A:249, Val A:84 | Conventional hydrogen bond, carbon hydrogen bond, π -anion, π -alkyl |
| A-20 | Low | -1.301 | Val A:84, Leu A:249, Leu A:267, Tyr A:271, Ser A:67, Ile A:274, Asn A:253 | Conventional hydrogen bond, π - π T-shaped, π -sigma, π -alkyl, alkyl |
| A-35 | Low | -1.204 | Val A:84, Ala A:277, Leu A:249, Ile A:274, Met A:270, Glu A:169 | π -alkyl, alkyl, π -anion |

Table 3. The values of all descriptors appearing in the models for training and test set compounds are given in the Supplementary Material II (Excel file) and the scatter plots of the observed vs. predicted selectivity values are given in Figure 6.

Mechanistic interpretation All the descriptors obtained in the four models and their frequency give an idea about their importance in modeling the selectivity of the PET tracers towards adenosine A_{2A} receptor. The descriptors like C-027, F09[N-O], SsssN, T(F..Cl), and ETA_Beta_s appearing in the models give information about the electronic feature of the compounds and are essential when the selectivity of receptor is considered (Fig. 7). Electronegativity is a chemical property that describes the tendency of an atom to draw electron towards itself. If a compound contains higher number of electronegative atoms in its structure, then the selectivity of the A_{2A} receptor for that compound also increases.

The presence of atom-centered fragments like C-027 (R-CH-X) in compounds like A-23 and A-25 increase the antagonist selectivity of the PET compounds. Since 'X' represents any electronegative atom like O, N, S, P, Se, and halogens, the presence of heteroatoms increases the selectivity of the compounds towards A_{2A} receptor. The descriptor F09[N-O] explains the frequency of presence of nitrogen and oxygen at the topological distance 9, and its positive regression coefficient indicates its influential activity on the antagonistic behavior of the imaging agents (as seen in compounds A-4 and A-27). Another similar kind of descriptor is T (F..Cl), explaining the information about sum of topological distances

between F and Cl atoms in the chemical structure. These descriptors give information about the electronegative atoms, i.e., nitrogen and oxygen in F09[N-O] and fluorine and chlorine in T(F..Cl). ETA_Beta_s ($\Sigma\beta_s$) is an extended topochemical atom (ETA) descriptor, which can be represented as sum of β_s values of all non-hydrogen vertices divided by 2. The term ' β_s ' can be denoted as

$$\Sigma\beta_s = \Sigma x\sigma$$

Here, x represents contribution of sigma bonds and σ signifies parameters related to sigma bonds. During the computation of β values, the sigma bond value for two similar types of electronegative atoms should be considered 0.5, and dissimilar electronegative atoms should be considered 0.75. This suggests that compounds bearing dissimilar heteroatoms will have greater selectivity to A_{2A} receptor as seen in compounds A-25, A-23, and A-4. Sigma bonds connected with different heteroatoms will have higher descriptor values indicating that the presence of dissimilar heteroatoms is more favorable for selectivity than similar heteroatoms. E-state descriptor SsssN (>N—) encodes the intrinsic electronic state of the nitrogen atom as perturbed by the electronic influence of other molecules with the context of topological character within the molecule. The electronegative contribution of nitrogen is well-depicted in this descriptor, and the positive regression coefficient shows that an increase in the number of tertiary nitrogen benefits in receptor selectivity as seen in compounds A-30 and A-4.

Other descriptors which significantly contribute to A_{2A} receptor selectivity are nCIC, F07[C-C], and C-040. These

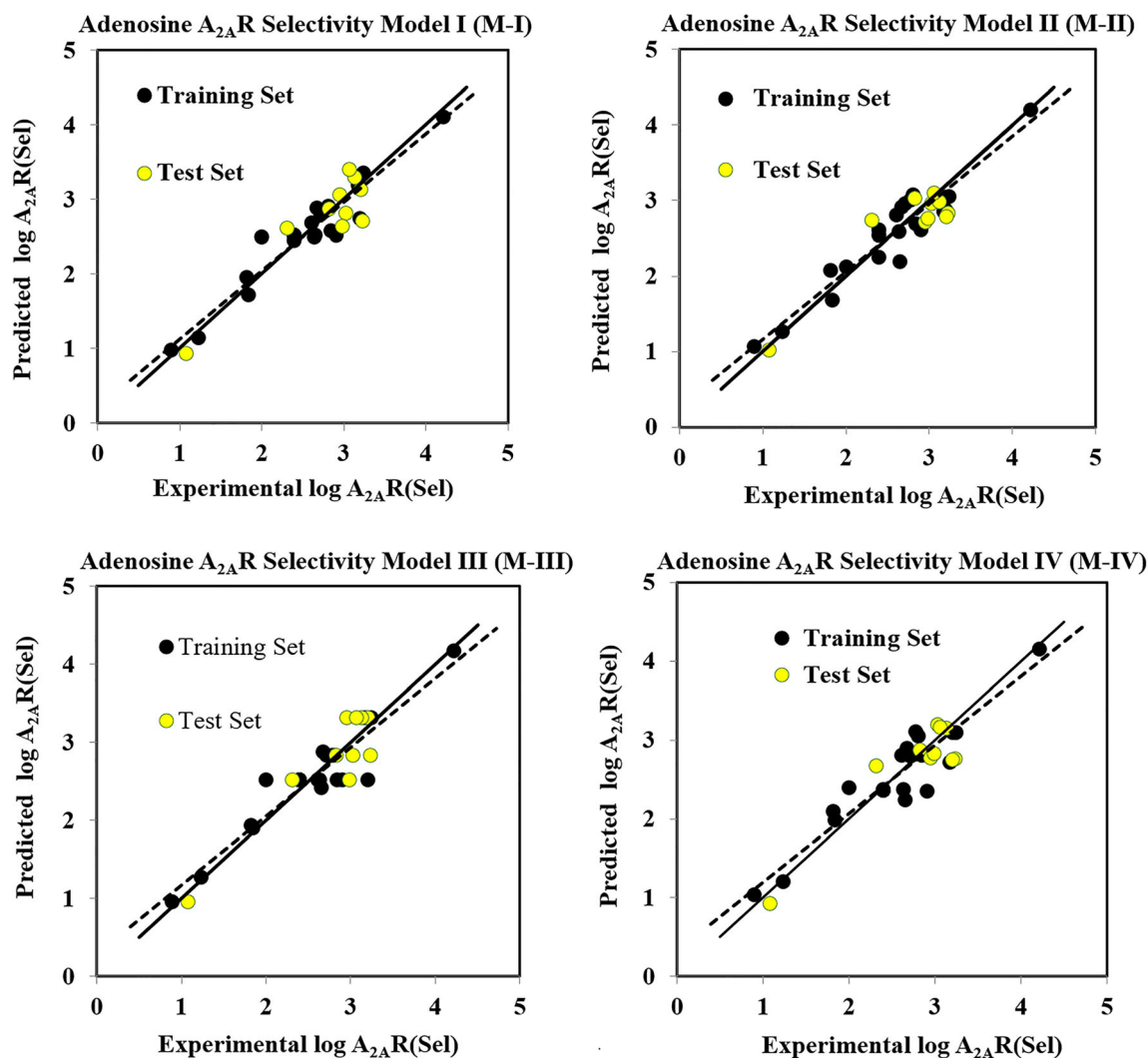


Fig. 6 Observed vs predicted $A_{2A}R$ selectivity plots for all four MLR models

descriptors give information about the number of rings present, type of bonds, and size of the antagonists showing selectivity towards the receptor. The number of rings (cyclomatic number) in the structure is indicated by **nCIC** descriptor. The positive regression coefficient of the descriptor suggests that the presence of high number of rings increases the selectivity towards the A_{2A} receptor as observed in compounds **A-25** and **A-4**. **F07[C-C]**, a 2D atom pair stands for frequency of C–C fragment at the topological distance 7. It provides information about the size (chain length) of the molecule. This means that with an increase in the number of this fragment, i.e., carbon chain, the selectivity towards the A_{2A} receptor increases (as in compounds **A-4** and **A-25**). The atom-centered fragment descriptor, **C-040** (Table 3) gives information about the number of carbon atoms that are attached to heteroatoms by single/double or triple bonds in the straight chain length. The negative regression coefficient suggests that an increase in the number of such fragments decreases the selectivity of the compound towards the A_{2A} receptor as seen in compounds

A-6, **A-7**, and **A-35**. As this fragment suggests high number of double and triple bonds attached with the carbon, it can be concluded that unsaturation in the straight chain of the antagonists is unfavorable for the receptor selectivity.

Intelligent consensus predictions For further refinement of the predictions obtained from the individual models, we have applied intelligent consensus modeling methods. Consensus modeling helps in enhancing the prediction performance of the models and also reduces the test set errors. It was observed that consensus prediction of the test set compounds (Table 4) is better in terms of both MAE-based criteria and predicted R^2 parameter. Four different consensus approaches were used employing “Intelligent Consensus Prediction” tool [39]: CM0 (simple average of predictions), CM1 (average of predictions from the ‘qualified’ individual models), CM2 (weighted average predictions (WAPs) from ‘qualified’ individual models), and CM3 (best selection of predictions (compound-wise) from ‘qualified’ individual models). From

Table 3 Definition, frequency, and contribution of all the descriptors obtained from the MLR models

| Sl. no. | Name of descriptors | Type of descriptor | Contribution | Discussion | Frequency of descriptors |
|---------|---------------------|---------------------------|--------------|---|--------------------------|
| 1 | C-027 | Atom-centered fragment | +ve | Counts for certain structural fragment (R--CH--X) in the antagonist, where 'R' can be any group linked to carbon and '--' is any aromatic bond. X can be any electronegative atom (O, N, S, P, Se, halogens) | 3 |
| 2 | ETA_Beta_s | ETA indices | +ve | Sum of all sigma bond contributions considering non-hydrogen vertices divided by 2. The descriptor deals with the presence of dissimilar heteroatoms. | 1 |
| 3 | F09 [N-O] | 2D atom pairs | +ve | Frequency of the N-O fragment at the topological distance 9 | 4 |
| 4 | SsssN | Atom-type E-state indices | +ve | E-state of ssssN which encodes the intrinsic electronic state of the nitrogen atom as perturbed by the electronic influence of other molecules with the context of topological character within the molecule. SsssN is the atom-type E-state of all tertiary nitrogen in molecules. | 1 |
| 5 | nCIC | Ring descriptors | +ve | Number of rings (cyclomatic number) present in the antagonist | 2 |
| 6 | C-040 | Atom-centered fragment | -ve | Represented as R-C(=X)-X/R-C#X/X = C = X fragments where number of carbon atoms are attached to heteroatoms by single/double or triple bonds | 4 |
| 7 | F07[C-C] | 2D atom pairs | +ve | Frequency of C-C at topological distance 7 | 1 |

the four consensus model obtained, CM0 was found to be the best.

Applicability domain Applicability domain (AD) is an important tool for reliable application of QSAR models. It can be considered a “theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable” [42, 47]. We have checked the AD of all the models using standardization approach [48] to check whether any molecule in the test set lies outside the AD of a model. From the domain of

applicability analysis, it was found that there were no test set compounds outside the AD, and no compound in the training set came as an outlier (see Supplementary II Excel file).

Comparison with a previously published model A direct comparison between the current and a previously published model [29] is infeasible due to the differences in the composition of training and test sets. However, the current model can be considered more advantageous since it has been developed using simple and easily interpretable two-dimensional descriptors

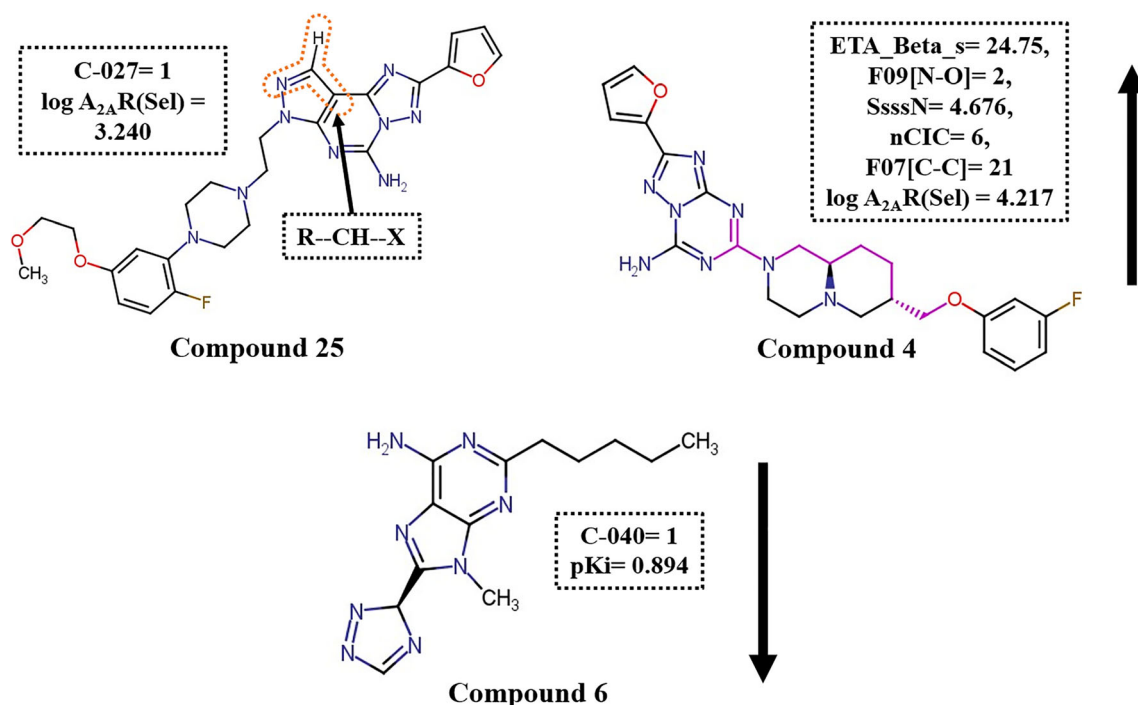
**Fig. 7** Features affecting the adenosine A_{2A} selectivity

Table 4 Detailed summary of the QSAR models and consensus models obtained for selectivity PET tracer compounds for adenosine A_{2A} selectivity (the quality of the best model CM0 is shown in italics)

| Dataset | Type of model | Training set statistics | | | | | Test set statistics | | | | | | |
|---------------------------|---------------|-------------------------|---------------------------------------|-----------|----------------------------------|-----------------------------------|--|------------------------------|-------------|-----------------------------------|------------------------------------|-------------|-------------|
| | | Model R ² | Model Q ² _(LOO) | MAE_train | r ² _{m(LOO)} | Δr ² _{m(LOO)} | R ² _{pred} or Q ² _{F1} | Q ² _{F2} | CCC | r ² _{m(test)} | Δr ² _{m(test)} | MAE (95%) | MAE |
| Individual models (N1–N5) | IM1 | 0.92 | 0.87 | Good | 0.81 | 0.07 | 0.84 | 0.81 | 0.77 | 0.12 | 0.18 | Good | |
| | IM2 | 0.90 | 0.82 | Moderate | 0.75 | 0.06 | 0.84 | 0.82 | 0.77 | 0.04 | 0.24 | Good | |
| | IM3 | 0.88 | 0.83 | Moderate | 0.76 | 0.12 | 0.84 | 0.82 | 0.77 | 0.13 | 0.18 | Good | |
| | IM4 | 0.87 | 0.83 | Moderate | 0.72 | 0.13 | 0.85 | 0.83 | 0.78 | 0.07 | 0.22 | Good | |
| Consensus models | CM0 | - | - | - | - | - | <i>0.88</i> | <i>0.86</i> | <i>0.93</i> | <i>0.82</i> | <i>0.10</i> | <i>0.16</i> | <i>Good</i> |
| | CM1 | - | - | - | - | - | 0.86 | 0.84 | 0.92 | 0.82 | 0.11 | 0.18 | Good |
| | CM2 | - | - | - | - | - | 0.85 | 0.83 | 0.92 | 0.82 | 0.11 | 0.18 | Good |
| | CM3 | - | - | - | - | - | 0.83 | 0.81 | 0.90 | 0.80 | 0.10 | 0.19 | Good |

The quality of the best model CM0 is provided in italics

which does not require any conformational analysis or energy minimization before their calculation.

Conclusion

Parkinson's disease is a neurodegenerative disease affecting the elderly person around the world. An important target for its treatment is blocking adenosine A_{2A} receptor which is co-located with the D₂ receptor and is pharmacologically opposite in motor function. Many studies hint that blocking A_{2A} receptor would be a beneficial strategy in the treatment of PD. Thus, this work endeavors exploring QSAR analysis to correlate the chemical structures with their biological activity with the aim to filter the essential chemical features of an antagonist for selectivity and binding affinity to A_{2A} receptor. The computational approach used in this work consists firstly the calculation of the molecular descriptors, and secondly, correlating these descriptors with the binding affinity and selectivity using different chemometric tools such as Genetic Function Algorithm (GFA), Best Subset Selection (BSS) method, and Intelligent consensus predictor (ICP) tools. The statistical quality of the models was checked using traditional metrics both internally and externally. We have also discussed about the contributions of the descriptors in the light of known binding mechanisms such as π-π stacking interaction, hydrophobic interaction, and hydrogen bonding with the different protein residues present in the receptor binding sites. From the insights obtained from such mechanism, we found that electronegative atoms and presence of aromatic ring like benzene are favorable for enhancing the binding affinity to the A_{2A} receptor. Furthermore, the docking studies supported the conclusions derived from the QSAR studies. In conclusion, the study highlights the pharmacophoric features mainly responsible for antagonizing adenosine receptors that can be further

modified for better binding and selectivity to A_{2A} receptor. In case of selectivity also, electronegativity and aromaticity of the compounds play essential and influential roles. The simple two-dimensional (2D) descriptors appearing in all the models are easier to compute requiring no conformation analysis or energy minimization process. Thus, this information would help in the future development and synthesis of newer PET tracer targeted towards adenosine receptor.

Funding information PD thanks Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship. JR received financial assistance from the Department of Atomic Energy—Board of Research in Nuclear Sciences (DAE-BRNS) (ref. 36(3)/14/08/2017-BRNS). KR thanks DAE-BRNS for a major research project (ref. 36(3)/14/08/2017-BRNS).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkman J, Schrag AE, Lang AE (2017) Parkinson disease. Nat Rev Dis Primers 3:1–21
- Voss T, Ravina B (2008) Neuroprotection in Parkinson's disease: myth or reality? Curr Neurol Neurosci Rep 8:304–309
- Ahmed SS, Ahameethunisa A, Santosh W (2010) QSAR and pharmacophore modeling of 4-arylthieno [3, 2-d] pyrimidine derivatives against adenosine receptor of Parkinson's disease. J Theor Comput Chem 9:975–991
- Chen JJ, Swope DM (2007) Pharmacotherapy for Parkinson's disease. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 27:161S–173S
- Jankovic J, Stacy M (2007) Medical management of levodopa-associated motor complications in patients with Parkinson's disease. CNS Drugs 21:677–692

6. Fredholm BB, IJzerman AP, Jacobson KA, Klotz KN, Linden J (2001) International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors. *Pharmacol Rev* 53:527–552
7. Fuxe K, Ferré S, Genedani S, Franco R, Agnati LF (2007) Adenosine receptor–dopamine receptor interactions in the basal ganglia and their relevance for brain function. *Physiol Behav* 92: 210–217
8. Chen JF, Xu K, Petzer JP, Staal R, Xu YH, Beilstein M, Sonsalla PK, Castagnoli K, Castagnoli N, Schwarzschild MA (2001) Neuroprotection by caffeine and A2A adenosine receptor inactivation in a model of Parkinson's disease. *J Neurosci* 21:RC143–RC143
9. Grondin R, Bedard PJ, Tahar AH, Gregoire L, Mori A, Kase H (1999) Antiparkinsonian effect of a new selective adenosine A2A receptor antagonist in MPTP-treated monkeys. *Neurology* 52: 1673–1673
10. Ongini E, Monopoli A, Impagnatiello F, Fredduzzi S, Schwarzschild M, Chen JF (2001) Dual actions of A2A adenosine receptor antagonists on motor dysfunction and neurodegenerative processes. *Drug Dev Res* 52:379–386
11. Ikeda K, Kurokawa M, Aoyama S, Kuwana Y (2002) Neuroprotection by adenosine A2A receptor blockade in experimental models of Parkinson's disease. *J Neurochem* 80:262–270
12. Pike VW (2009) PET radiotracers: crossing the blood–brain barrier and surviving metabolism. *Trends Pharmacol Sci* 30:431–440
13. Rahmim A, Zaidi H (2008) PET versus SPECT: strengths, limitations and challenges. *Nucl Med Commun* 29:193–207
14. Roy K (2018) Quantitative structure–activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. *J Indian Chem Soc* 95:1497–1502
15. Gramatica P (2020) Principles of QSAR modeling: comments and suggestions from personal experience. *IJQSPR* 5:1–37. <https://doi.org/10.4018/IJQSPR.20200701.oa1>
16. Puzyn T, Leszczynski J, Cronin MT, eds. (2010) Recent advances in QSAR studies: methods and applications, Vol. 8 Springer Science & Business Media, Berlin, Germany
17. Gao DW, Wang P, Yang L, Peng YZ, Liang H (2002) Study on the screening of molecular structure parameter in QSAR model. *J Environ Sci Heal A* 37:601–609
18. Tropsha A (2004) Application of predictive QSAR models to database mining. *Chemoinformatics Drug Discov* 23:437–455
19. Roy K (2020) Ecotoxicological QSARs. Springer, New York
20. Kar S, Roy K, Leszczynski J (2017) In: Roy K. (eds) On applications of QSARs in food and agricultural sciences: history and critical review of recent developments. *Advances in QSAR Modeling*, Springer, Cham, Switzerland
21. Ojha PK, Roy K (2018) Chemometric modeling of odor threshold property of diverse aroma components of wine. *RSC Adv* 8:4750–4760
22. Tantra R, Oksel C, Puzyn T, Wang J, Robinson KN, Wang XZ, Ma CY, Wilkins T (2015) Nano (Q) SAR: Challenges, pitfalls and perspectives. *Nanotoxicology* 9:636–642
23. Mikolajczyk A, Gajewicz A, Mulkiewicz E, Rasulev B, Marchelek M, Diak M, Hirano S, Zaleska-Medynska A, Puzyn T (2018) Nano-QSAR modeling for ecosafe design of heterogeneous TiO₂-based nano-photocatalysts. *Environ Sci Nano* 5:1150–1160
24. Mikolajczyk A, Sizochenko N, Mulkiewicz E, Malankowska A, Rasulev B, Puzyn T (2019) A chemoinformatics approach for the characterization of hybrid nanomaterials: safer and efficient design perspective. *Nanoscale* 11:11808–11818
25. Hoekman D (1996) Exploring QSAR fundamentals and applications in chemistry and biology, volume 1. hydrophobic, electronic and steric constants, Volume 2. *J. Am. Chem. Soc.* 1995, 117, 9782. *J Am Chem Soc* 118:10678–10678
26. Klein C, Kaiser D, Kopp S, Chiba P, Ecker GF (2002) Similarity based SAR (SIBAR) as tool for early ADME profiling. *J Comput Aided Mol Des* 16:785–793
27. Sebastián-Pérez V, Martínez MJ, Gil C, Campillo NE, Martínez A, Ponzoni I (2019) QSAR Modelling to identify LRRK2 inhibitors for Parkinson's disease. *J Integr Bioinform* 16
28. Khanfar MA, Al-Qtaishat S, Habash M, Taha MO (2016) Discovery of potent adenosine A2a antagonists as potential anti-Parkinson disease agents. Non-linear QSAR analyses integrated with pharmacophore modeling. *Chem Biol Interact* 254:93–101
29. Tamiji Z, Salahinejad M, Niazi A (2018) Molecular modeling of potential PET imaging agents for adenosine receptor in Parkinson's disease. *Struct Chem* 29:467–479
30. MarvinSketch software, <https://www.chemaxon.com>. Accessed on 05 Jan 2020
31. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32:1466–1474
32. Dragon version 7, Kodesrl, Milan, Italy, 2016; software available at <http://www.taletе.mi.it/index.htm>. Accessed 07 Jan 2020
33. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* 17: 241–253
34. Golbraikh A, Tropsha A (2000) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers* 5:231–243
35. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opin Drug Discovery* 13:1075–1089
36. Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, Cornwall, Great Britain
37. Pope PT, Webster JT (1972) The use of an F-statistic in stepwise regression procedures. *Technometrics* 14:327–340
38. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33
39. Roy K, Ambure P, Kar S, Ojha PK (2018) Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J Chemom* 32:e2992
40. DTC Lab QSAR Tools http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab
41. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen* 14: 450–474
42. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 1:45–63
43. Congreve M, Andrews SP, Doré AS, Hollenstein K, Hurrell E, Langmead CJ, Mason JS, Ng IW, Tehan B, Zhukov A, Weir M (2012) Discovery of 1, 2, 4-triazine derivatives as adenosine A2A antagonists using structure based drug design. *J Med Chem* 55: 1898–1903
44. Wu G, Robertson DH, Brooks Iii CL, Vieth M (2003) Detailed analysis of grid-based molecular docking: a case study of CDOCKER—a CHARMM-based MD docking algorithm. *J Comput Chem* 24:1549–1562
45. Pan AC, Borhani DW, Dror RO, Shaw DE (2013) Molecular determinants of drug–receptor binding kinetics. *Drug Discov Today* 18:667–673
46. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, IJzerman AP, Stevens RC (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* 322:1211–1217

47. Yun YH, Wu DM, Li GY, Zhang QY, Yang X, Li QF, Cao DS, Xu QS (2017) A strategy on the definition of applicability domain of model based on population analysis. *Chemom Intell Lab Syst* 170: 77–83
48. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. *Chemom Intell Lab Syst* 145:22–29

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting dopamine receptor

Priyanka De¹ · Kunal Roy¹

Received: 19 July 2020 / Accepted: 12 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Dopamine (D2) receptor has emerged as a potent drug target for the diagnosis and treatment of Parkinson's disease (PD). Radiolabelled imaging such as positron emission tomography (PET) has been recognized as an important tool in medicinal chemistry useful for the early diagnosis of PD. The present study explores quantitative structure—activity relationship analysis of 34 PET imaging agents targeted toward dopamine D2 receptor. The dataset division into training and test sets was done using Euclidean distance division method, while the feature selection was done by double cross-validation-genetic algorithm method. Finally, a five-descriptor partial least squares regression model was derived after carrying out the best subset selection applied on the significant descriptors. The developed model showed robustness in terms of statistical parameters. Finally, the structural information derived from the model descriptors gives an insight for the development of new candidate D2-PET imaging for the use in PD.

Keywords Positron emission tomography · Parkinson's disease · Quantitative structure—activity relationship

1 Introduction

Parkinson's disease (PD) is considered as the second most common progressive neurodegenerative disorder associated with a selective degeneration of the dopaminergic neurons in the substantia nigra pars compacta and loss of projecting nerve fibers in the striatum. It is estimated that more than 10 million people are living with PD worldwide and the occurrence of PD increases with age [1]. About four percent of people with PD are diagnosed before 50 years of age, and men are more prone to this disease than women (about 1.5 times more) [1]. The neurons involved in this disease control the motor movements like resting tremor, muscular

rigidity, bradykinesia, and postural imbalance [2]. Patients with this disease also experience a combination of non-motor symptoms like sleep disturbances, dementia, fatigue, anxiety, depression, apathy, cognitive impairment, olfactory dysfunction, pain, sweating and constipation [3].

Neuroimaging studies are non-invasive methods which help in providing an in vivo image of the nigrostriatal dopaminergic system and further assessment of the extent of neuronal loss associated with PD. Radioactive tracers that selectively bind with dopamine receptors are involved in positron emission tomography (PET) imaging and lately single photon emission computed tomography (SPECT) imaging for research and clinical purposes [4]. PET imaging is a powerful analytical tool which is able to detect in vivo changes in the brain function [5]. PET imaging involves quantification of brain metabolism, abundance of a receptor and its binding in different neurotransmitter systems, and alterations in blood flow in specific region in the brain [5]. PET imaging is considered better than SPECT imaging in terms of accuracy and its regional distributions [6]. Heiss and Hilker (2003) [7] studied that the radiotracer ¹⁸F-fluorodopa (FDOPA) is capable of measuring dopamine deficiency, both its synthesis and storage at the pre-synaptic striatal nerve endings, thus allowing FDOPA-PET in the diagnosis of PD in early disease stages. Wu et al. [8] characterized the clinical

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00214-020-02687-9>) contains supplementary material, which is available to authorized users.

Published as part of the special collection of articles "Festschrift in honour of Prof. Ramon Carbó-Dorca".

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

¹ Drug Theoretics and Cheminformatics Laboratory,
Department of Pharmaceutical Technology, Jadavpur
University, Kolkata 700032, India

features and associated cerebral glucose metabolism pattern of cognitive impairments in (PD) using ^{18}F -fluorodeoxyglucose (^{18}F -FDG) PET imaging. Glaab et al. integrated blood metabolomics data with PET imaging information which gave better diagnostic discrimination power in understanding cellular processes, including oxidative stress response and inflammation [9].

There is a continuous search of new compounds with improved properties and lowered toxicity which takes enormous human resource and cost into its requirement. Thus, theoretical approaches are gaining more importance among the pharmaceutical and chemical industries enabling logical design of pharmaceutical agents. Currently, quantitative structure-activity relationship (QSAR) has gained great interest in the process of modern drug discovery and design [10, 11]. The study attempts to build a relationship between the chemical properties with a well-defined endpoint as the compounds' activity (QSAR) or property (QSPR) or toxicity (QSTR). QSAR acts as an effective tool in the prediction of biological response (activity/property/toxicity) of existing chemical compounds.

In the present study, we have developed a QSAR model with two-dimensional (2D) molecular descriptors to explore the correlations of the molecular structure of a series of PET tracers against the binding affinity of dopamine (D2) receptor.

2 Materials and methods

2.1 Dataset

Dopamine (D2) receptor binding affinity (K_i) data of 34 PET imaging agents were taken from different literature as mentioned in Table 1. The experimental binding affinity for all the compounds was measured using the same assay protocol, i.e., rat striatal homogenate (RSH) assay method. This datum was applied in the development of a 2D-QSAR model to determine the essential structural features required for good binding to the D2 receptor. The binding affinity (K_i) values for the PET imaging agents were converted to their negative logarithm ($\text{p}K_i$) form and then used for modeling. The compounds were represented using the MarvinSketch software [12] with proper aromatization and addition of hydrogen bond as necessary.

2.2 Molecular descriptors

QSAR models were developed using a selected class of two-dimensional molecular descriptors. The descriptors were *E*-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments, and molecular property descriptors. These descriptors were calculated

Table 1 Dataset compounds with their observed binding affinity (in $\text{p}K_i$)

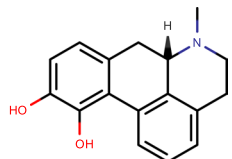
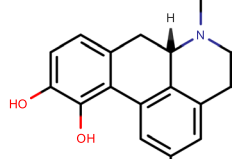
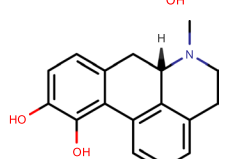
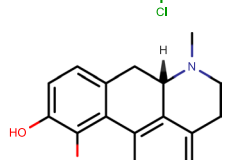
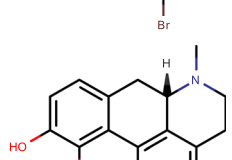
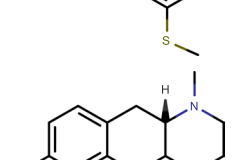
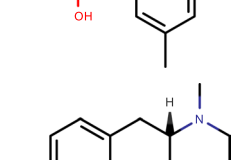
| Compound no | Structure | $\text{p}K_i$ | Reference |
|-------------|--|---------------|-----------|
| 1 |  | 2.321 | [13] |
| 2 |  | 4.420 | [14] |
| 3* |  | 2.652 | [14] |
| 4 |  | 2.752 | [14] |
| 5 |  | 2.262 | [15] |
| 6 |  | 2.684 | [13] |
| 7* |  | 3.951 | [14] |

Table 1 (continued)

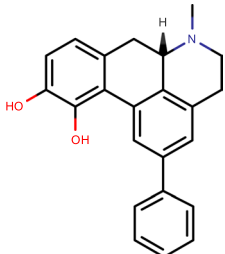
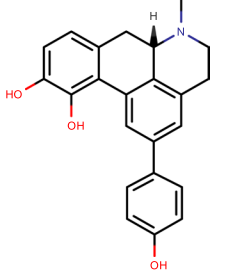
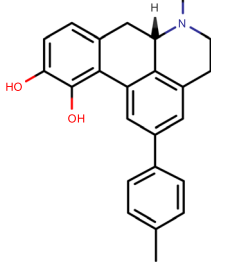
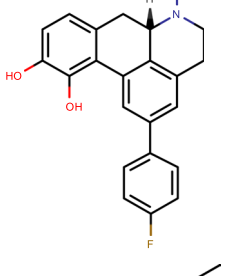
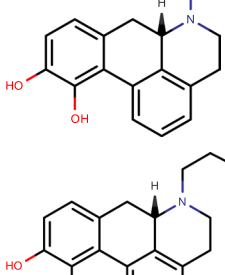
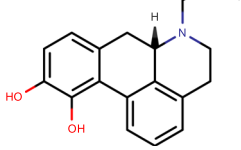
| Compound no | Structure | pKi | Reference |
|-------------|---|-------|-----------|
| 8 |  | 2.932 | [13] |
| 9 |  | 3.401 | [13] |
| 10 |  | 1.460 | [16] |
| 11 |  | 1.839 | [16] |
| 12 |  | 4.658 | [17] |
| 13 |  | 4.097 | [17] |

Table 1 (continued)

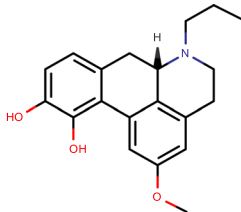
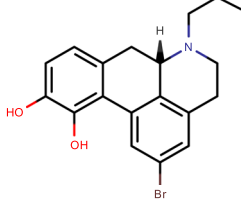
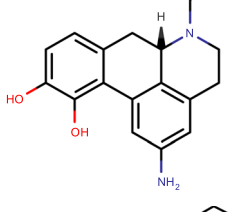
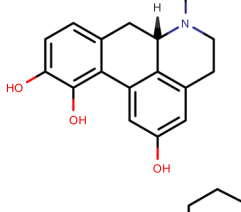
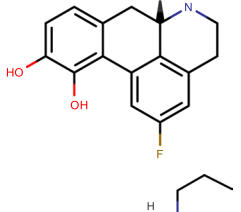
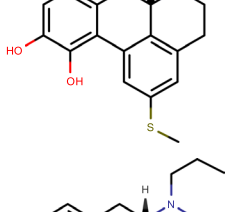
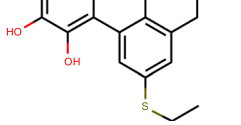
| Compound no | Structure | pKi | Reference |
|-------------|--|-------|-----------|
| 14 |  | 4.770 | [14] |
| 15 |  | 4.824 | [18] |
| 16 |  | 4.036 | [18] |
| 17 |  | 5.276 | [14] |
| 18 |  | 5.921 | [18] |
| 19* |  | 3.428 | [15] |
| 20 |  | 3.108 | [15] |

Table 1 (continued)

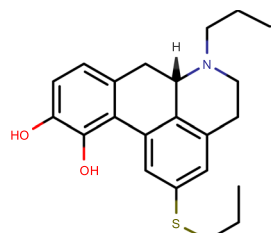
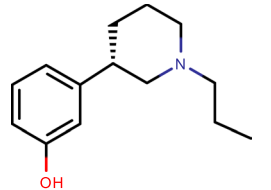
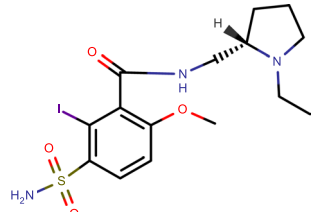
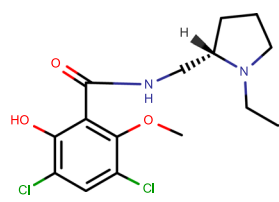
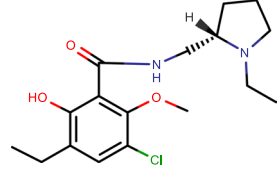
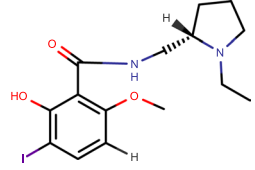
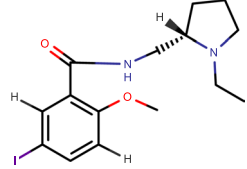
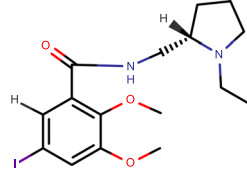
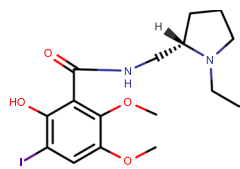
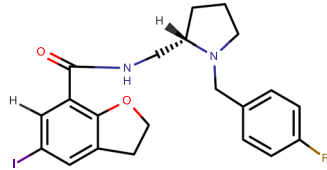
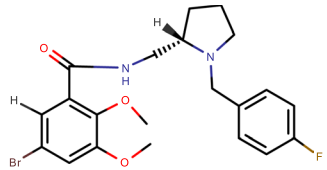
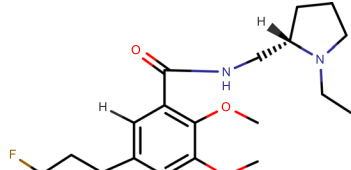
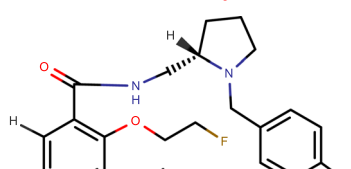
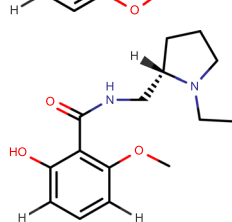
| Compound no | Structure | pKi | Reference |
|-------------|---|-------|-----------|
| 21 |  | 2.807 | [15] |
| 22* |  | 3.114 | [19] |
| 23 |  | 3.824 | [20] |
| 24 |  | 3.959 | [20] |
| 25* |  | 5.046 | [20] |
| 26* |  | 4.367 | [20] |
| 27 |  | 3.523 | [20] |
| 28* |  | 5.602 | [20] |

Table 1 (continued)

| Compound no | Structure | pKi | Reference |
|-------------|--|-------|-----------|
| 29 |  | 5.721 | [20] |
| 30 |  | 4.975 | [20] |
| 31 |  | 4.833 | [20] |
| 32 |  | 5.699 | [20] |
| 33 |  | 2.886 | [20] |
| 34 |  | 2.507 | [21] |

Compounds marked with "*" are test set compounds

using Dragon 7 [22] descriptor calculator. A total of 403 Dragon descriptors were calculated. Before the development of the QSAR model, the data were curated [23] by removing intercorrelated ($|r| > 0.95$), constant (variance < 0.0001), and other noisy and redundant data by using data pretreatment software developed in our laboratory and available from <https://dtclab.webs.com/software-tools>. After data pretreatment, the number of descriptors was reduced to 179.

2.3 Dataset splitting

Splitting of the dataset into training and test sets is a vital step in QSAR modeling, and it enables the development of a robust and well-validated model. Data division must be done in such a way that the points representing both training and test set are well scattered within the whole descriptor space defined by the entire dataset. The training set is used for model development and the test set for model validation. The division of the dataset was executed by one of the most extensively used methods, Euclidean distance division method, where the Euclidean distances for all of the compounds in the dataset are calculated and the compounds are then sorted, based on the Euclidean distance [24].

2.4 Variable selection and model development

The main aim of the present study is to develop a well-validated QSAR model to understand the binding of PET imaging agents toward dopamine (D2) receptor for the diagnosis of Parkinson's disease. Critical selection of statistically significant descriptors ensures improvement in the quality of the model. Prior to development of the QSAR model, we have extracted a number of significant descriptors using double cross-validation-genetic algorithm (DCV-GA) approach applied on the training set compounds [25–27]. Finally, a partial least squares (PLS) [28] regression model was generated using descriptors selected from the best subset selection (BSS).

Double cross-validation (DCV) is an attractive statistical design which combines both model generation and model assessment with the aim to produce better models [25, 29]. Sometimes the fixed composition of a training set can lead to biased descriptor selection. DCV method helps in better descriptor selection by dividing the training set into 'n' calibration and validation sets. This results in diverse compositions of the modeling set, thus removing any bias in descriptor selection. DCV technique consists of two nested cross-validation loops commonly known as internal and external cross-validation loops. In the external loop, the data objects are split randomly into disjoint subsets known as training set compounds and test set compounds. The training set compounds are involved in the internal loop for the purpose of model development and model selection, and the test set is used solely for the intention of checking model predictivity. Further, in the internal loop, the training set compounds are repetitively split into calibration (construction) and validation sets by employing the *k*-fold cross-validation technique (here, *k* = 10) [29] and producing *k* iterations to construct calibration and validation sets.

The calibration objects are used to derive different models by altering the tuning parameter(s) of the model (i.e., the descriptors), whereas the validation objects are used to guess the models' error. The model with the lowest cross-validated error is selected. The test compounds in the outer loop are employed to assess the predictive performance of the selected model.

In the current study, descriptor selection in the DCV platform was done using genetic algorithm (GA) approach. GA is a model optimization approach with an algorithm inspired by the theory of evolution [26]. GA has five basic steps: (i) coding of variables; (ii) initiation of population; (iii) evaluation of the response; (iv) reproduction; and (v) mutation. Steps (iii) to (v) are repeated until a termination criterion is reached. The criterion can be based on a lack of improvement in the response or simply on a maximum number of generations or on the total time allowed for the elaboration.

2.5 Statistical validation metrics

Validation of the robustness and predictive ability of the developed models is a very crucial step in a QSAR study. A meticulous examination of the statistical quality of the developed model has been done to judge the robustness in terms of reliability and predictivity measures using various internal and external validation parameters. For determining the quality of the developed model, statistical parameters like determination coefficient R^2 and explained variance R_a^2 were calculated. Other parameters including internal predictivity parameters such as predicted residual sum of squares (PRESS) and leave-one-out cross-validated correlation coefficient (Q_{LOO}^2) were also calculated along with external predictivity parameters like R_{pred}^2 or Q_{F1}^2 , Q_{F2}^2 , and concordance correlation coefficient (CCC) [30]. Further, we have also calculated r_m^2 metrics (i.e., r_m^2 and Δr_m^2) for both training and test set compounds [31]. Validation using mean absolute error (MAE)-based criteria for both external and internal validation was done [32]. The Q_{ext}^2 -based criteria do not always interpret the correct prediction quality because of the impact of the response range as well as the distribution of the values of the response in both the training and test set compounds; so MAE was calculated to check the average error [32]. Figure 1 shows the flowchart of the present work methodology.

3 Results and discussion

3.1 Modeling binding affinity of PET tracers toward dopamine (D2) receptor

The final PLS model of three latent variables (LVs) consisted of five descriptors that explains the binding properties of the

PET radioligands toward dopamine receptor. The final model is given below:

$$\text{pKi} = 4.512 - 0.184 \times \text{SaaCH} - 1.554 \times \text{B08}[C-S] + 0.060 \times \text{SsF} - 2.350 \times \text{B10}[N-F] + 1.425 \times \text{B10}[C-O]$$

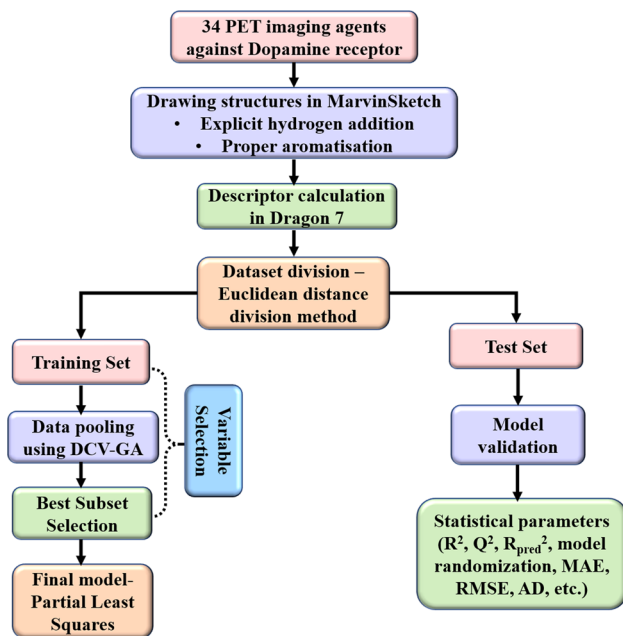


Fig. 1 Flowchart of the present work methodology

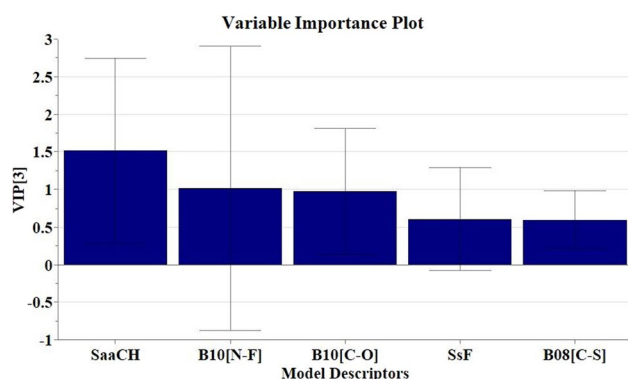


Fig. 2 Variable importance plot of the PLS model

Table 2 Descriptor meaning and their contribution

| Serial no | Descriptor | Descriptor type | Contribution | Discussion |
|-----------|------------|-------------------|--------------|---|
| 1 | SaaCH | Atom-type E-state | -ve | Sum of the atom-type E-state values for aromatic -CH groups |
| 2 | B08[C-S] | 2D atom pairs | -ve | Presence or absence of carbon and sulfur at the topological distance 8 |
| 3 | SsF | Atom-type E-state | +ve | Sum of the atom-type E-state values for -F fragments |
| 4 | B10[N-F] | 2D atom pairs | -ve | Presence or absence of nitrogen and fluorine at the topological distance 10 |
| 5 | B10[C-O] | 2D atom pairs | +ve | Presence or absence of carbon and oxygen at the topological distance 10 |

$$\begin{aligned} n_{\text{training}} &= 27, R^2 = 0.731, R_{\text{adj}}^2 = 0.696, Q^2 = 0.623, \overline{r_{m(\text{LOO})}^2} \\ &= 0.507, \Delta r_{m(\text{LOO})}^2 = 0.159, \text{MAE}(\text{train}) \\ &= 0.528, \text{SD}(\text{train}) = 0.550, \text{PRESS} = 15.392 \end{aligned}$$

$$\begin{aligned} n_{\text{test}} &= 7, Q_{F1}^2 = 0.687, Q_{F2}^2 = 0.664, \overline{r_{m(\text{test})}^2} = 0.742, \Delta r_{m(\text{test})}^2 = 0.116, \\ \text{MAE}(\text{test}) &= 0.505, \text{SD}(\text{test}) = 0.280, \text{CCC}(\text{Test}) = 0.812 \end{aligned}$$

3.2 Mechanistic interpretation

The variable importance plot (VIP) (Fig. 2) gives an idea about the influence of the individual descriptors on the model and thereby on the binding affinity [33]. The order of importance of the descriptors was found as follows: SaaCH, B10[N-F], B10[C-O], SsF, and B08[C-S]. The VIP gives an understanding that descriptors SaaCH and B10[N-F] are highly influential due to their VIP scores being more than one. The regression coefficient plot (not shown) provides a basic understanding about the contribution of the individual descriptor on the model [28]. It is seen that the descriptors SaaCH, B08[C-S], and B10[N-F] negatively contribute to the response, while the descriptors SsF and B10[C-O] positively contribute to the response. The details of the descriptors and their contributions are given in Table 2 and also explained below in detail. The observed vs predicted scatter plot is shown in Fig. 3.

The E-state indices descriptor SaaCH gives idea on the sum of the atom-type E-state values for aromatic -CH groups. From the regression coefficient of the descriptor, it can be inferred that aromaticity hinders the binding of the PET compounds to the D2 receptor as in compounds **8** (SaaCH = 18.392) (Fig. 4), **10** (SaaCH = 16.63), and **11** (SaaCH = 14.214). These compounds are aromatic and have high SaaCH values, and they have lower binding affinity values (pKi = 2.931, 1.460, and 1.839). Further, in compounds like **29** and **32**, aromaticity is less as compared to the previously mentioned compounds, thus having lower values for the descriptor (SaaCH = 3.583 and 1.640, respectively). These compounds have better binding affinity (compound **29** (pKi = 5.700) and compound **32** (pKi = 5.721)) toward dopamine receptor.

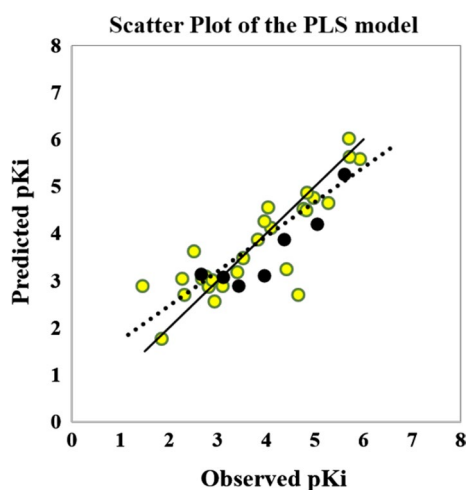


Fig. 3 Observed versus predicted pKi plot

The next important descriptor is B10[N-F] (2D atom pair type), and the negative contribution implies that the presence of nitrogen and fluorine at the topological distance 10 will hinder the binding affinity seen in compounds **11** (B10[N-F] = 1; pKi = 1.838) (Fig. 4) and **33** (B10[N-F] = 1; pKi = 2.886). Further, the absence of this fragment will increase the binding affinity as observed in compounds

29 (B10[N-F] = 0; pKi = 5.700) and **32** (B10[N-F] = 0; pKi = 5.721). The effect of the electronegativity of fluorine atom on nitrogen is a determining factor for the good binding which is latter explained while studying the descriptor SsF. The closeness between nitrogen and fluorine atom explains how the binding will occur.

B10[C-O] is another 2D atom pair descriptor representing the presence or absence of C-O fragment at the topological distance 10. The descriptor positively influences the binding affinity of the PET tracers toward dopamine receptor as seen in compounds **18** (B10[C-O] = 1; pKi = 5.921) (Fig. 4), **29** (B10[C-O] = 1; pKi = 5.721), and **32** (B10[C-O] = 1; pKi = 5.700). The presence of this kind of fragment affects the electronegativity of the compounds essential for binding. The absence of this fragment on the other hand decreases the dopamine binding affinity observed in compounds like **1** (pKi = 2.321) and **5** (pKi = 2.262).

The *E*-state values for the descriptor SsF depend on the number of fluorine atoms present in a PET tracer molecule. From the regression coefficient, it can be understood that with increasing fluorine atoms the binding affinity also increases as observed in **18** (SsF = 14.107; pKi = 5.921), **32** (SsF = 12.490; pKi = 5.698) (Fig. 4), and **31** (SsF = 13.108; pKi = 4.833). The electronegative fluorine atom is presumed to decrease electron charge density on nitrogen atoms. This

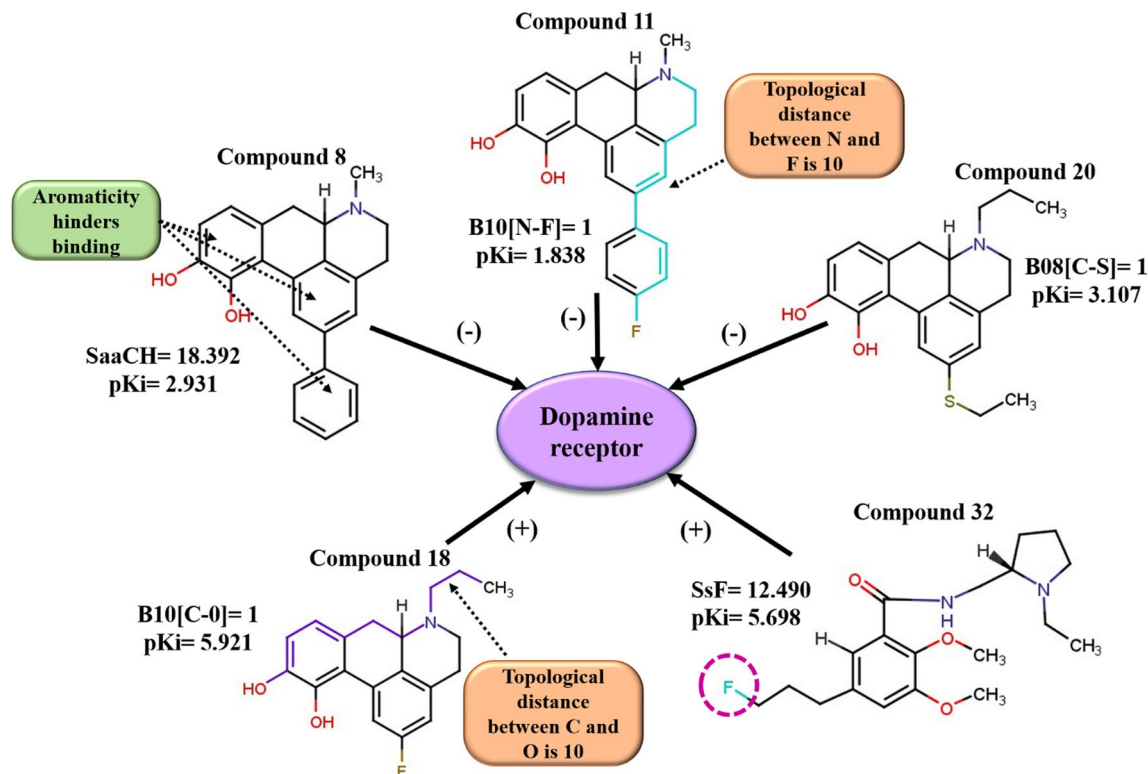


Fig. 4 Descriptors appearing in the PLS model and their contribution

reduces nitrogen basicity and its prospect to get protonated at physiological pH which is a basic requirement for good binding to dopamine receptors [34].

The least important descriptor is B08[C-S], which is also a 2D atom pair descriptor and gives an idea of the presence or absence of C-S fragment at a topological distance 8. The negative contribution suggests that the presence of this fragment will result in a decreased binding affinity toward the dopamine receptor which is observed in compounds **21** ($pK_i = 2.807$) and **20** ($pK_i = 3.107$) (Fig. 4). Alternatively, compounds like **18** ($pK_i = 5.921$), **29** ($pK_i = 5.721$) and **32** ($pK_i = 5.698$) have no such fragment, thus having higher binding affinity.

From the descriptors and their contributions, we can draw an inference that the oxygen for B10[C-O] and fluorine for SsF impart an electronegative character to the PET ligands which plays an essential role for the good dopamine (D2) binding.

3.3 Plot Interpretation

1 *Loading Plot*— This plot gives a relationship between the X-variables (i.e., the descriptors) and Y-variable (i.e., response) [35]. In Fig. 5, five X-variables and one Y-variable are shown. Generally, the plot is developed with the first and second components. A loading plot provides an insight about how much a variable contributes to a model and which variable provides the maximum footprint. For interpretation, the distance from the origin is taken under consideration. Descriptors which are similar in nature and providing similar contribution are correlated and grouped together. Descriptors which are situ-

ated far away from the plot origin are supposed to have greater impact on the Y-response. From the loading plot it, is seen that descriptors SaaCH and B10[N-F] are far away from the plot origin supporting their higher influence also explained by the VIP. The positive or negative algebraic symbol is also taken under consideration in a PLS plot. Features explained by descriptors SsF and B10[C-O] are beneficial for binding because of their closeness to pK_i in the plot. On the other hand, SaaCH, B10[N-F] and B08[C-S] are present in the negative side of the plot origin and are detrimental for good binding.

2 *Score Plot*— Figure 6 shows the distribution of the compounds in the latent variable space as defined by the scores. We have plotted the scores of the first two components t_1 and t_2 . The applicability domain of the model is designated by the ellipse, as defined by Hotelling's t^2 . Hotelling's t^2 defines multivariate generalization of Student's t test. The method offers a check for compounds adhering to multivariate normality [36]. Compounds which are situated near each other in the plot have similar properties, whereas compounds which are far from each other have dissimilar properties with respect to their binding affinity toward dopamine receptor. As an example, we can take compounds **14**, **15**, **16**, and **17** which are clubbed together as a group on the plot space and can be considered to be with similar properties. On the other hand, compounds **18** and **12** are completely located on the opposite side of the origin and far from each other and they represent heterogeneity in their properties. Since there are no compounds out of the ellipse, we can conclude that there are no outliers according to this method.

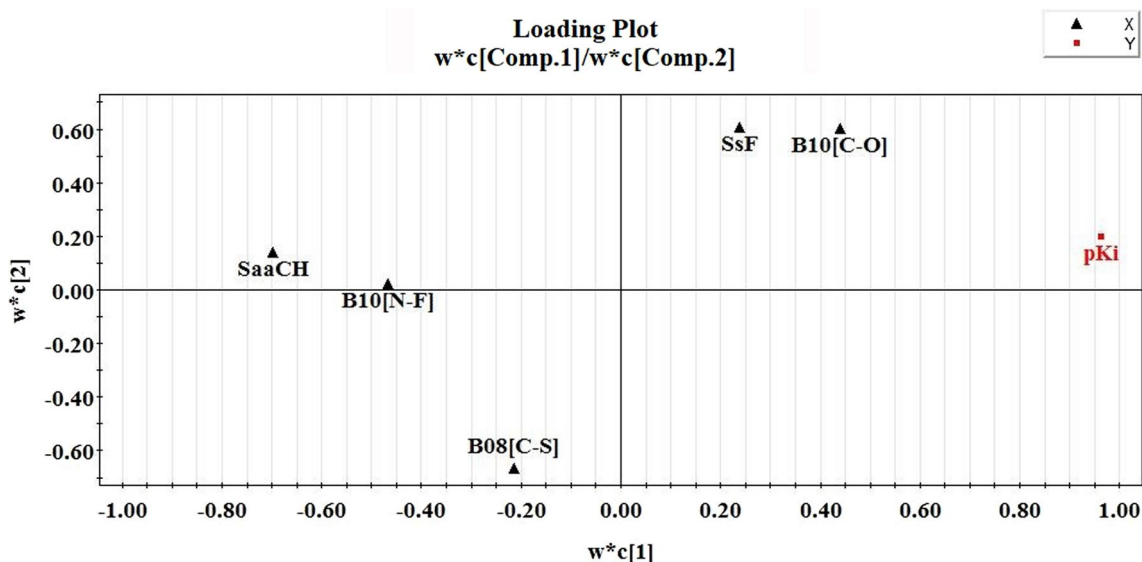
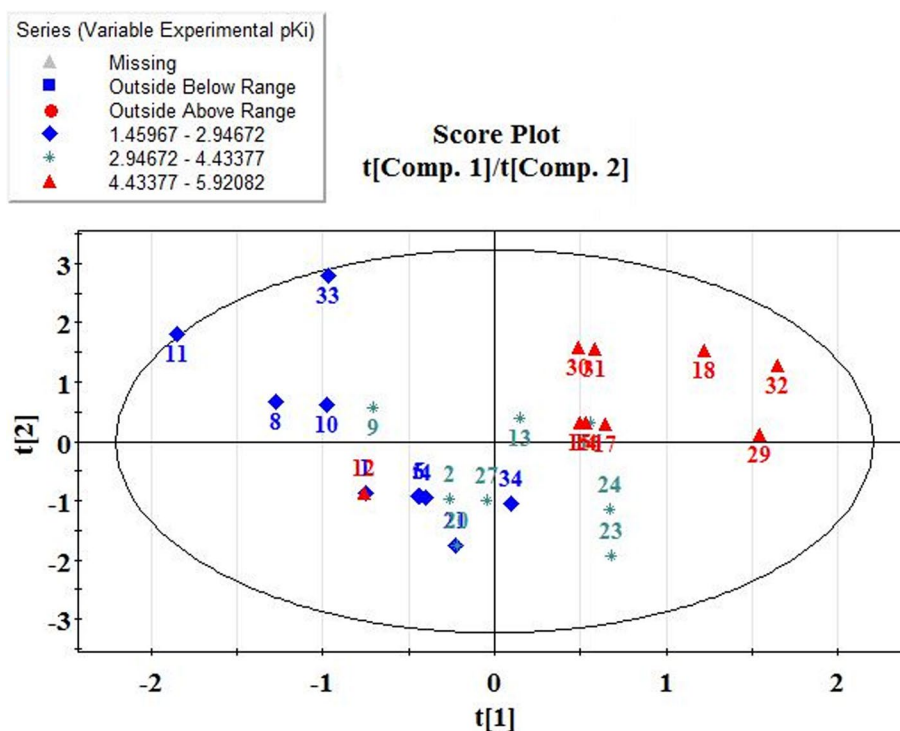


Fig. 5 Loading plot of the PLS model

Fig. 6 Score plot of the PLS model

Y-Randomization Plot— Model randomization gives a notion about the model significance and ensures that the model is not an outcome of a chance correlation [37]. A randomized model is generated by the development of multiple models by shuffling or reordering different combinations of *X*-or *Y*-variables (here *Y*-variable only) and based on the fit of the reordered model. In the present study, we have used 100 permutations which can be changed according to the choice of the user. A randomized model should have very poor statistics. The R^2

and Q^2 values for the random models (*Y*-axis) are plotted against correlation coefficient between the original *Y* values and the permuted *Y* values (*X*-axis); the R_y^2 intercept should not exceed 0.3, and the Q_y^2 intercept should not exceed 0.05. Figure 7 shows the correlation between original *Y*-vector and permuted *Y*-vector versus cumulative R_y^2 , cumulative Q_y^2 plot where R_y^2 intercept = 0.09 and Q_y^2 intercept = -0.393 proving the model is robust and non-random.

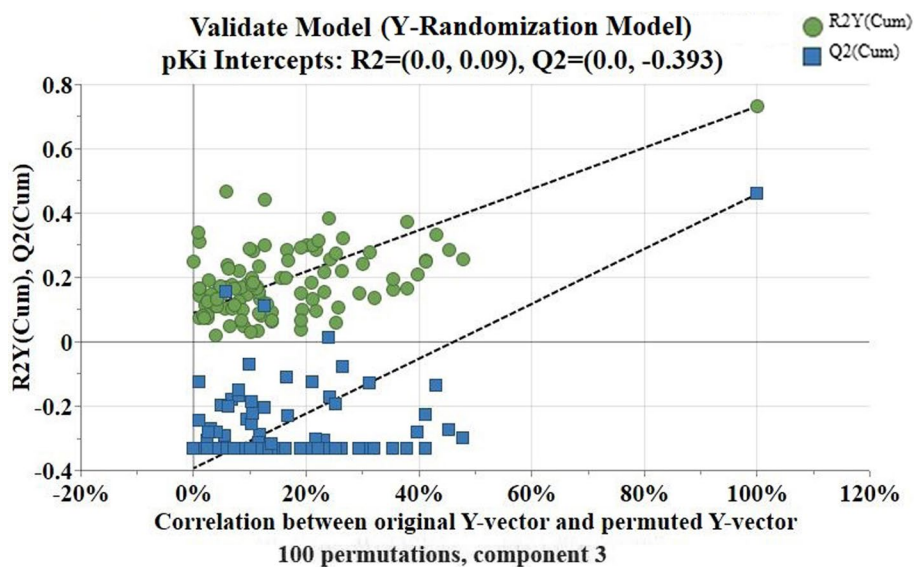
Fig. 7 *Y*-randomization plot of the PLS model

Fig. 8 DModX applicability domain of the training set

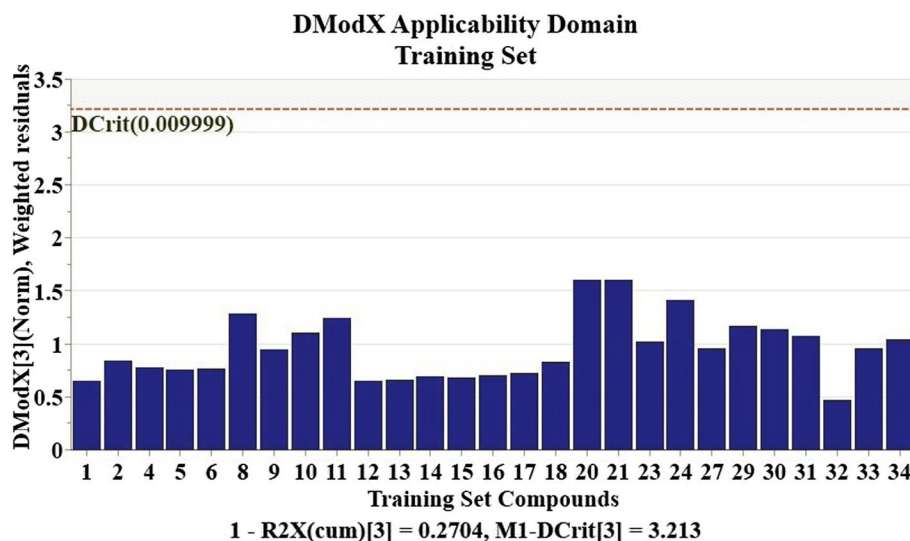
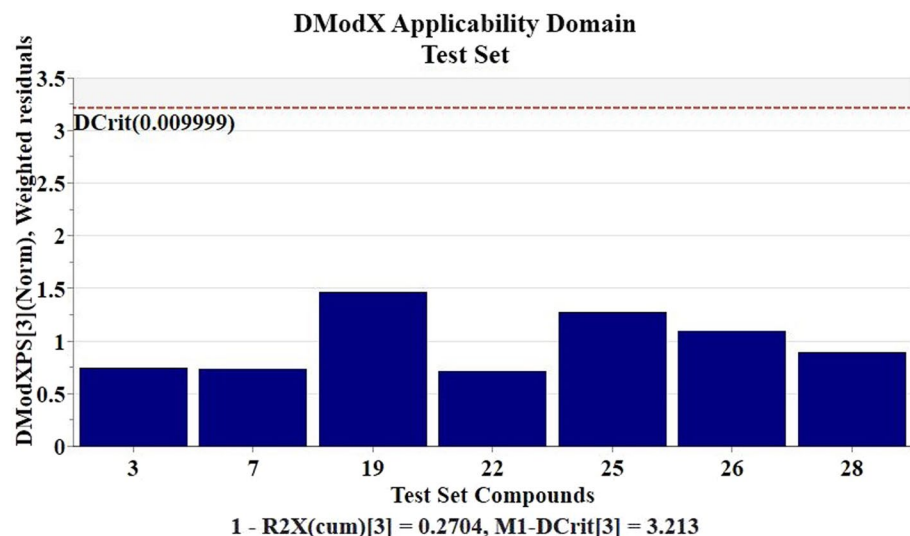


Fig. 9 DModX applicability domain of the test set



3 *Applicability Domain (AD)*— The prediction reliability of a particular model is dependent on its applicability domain (AD) assessment. Applicability domain (AD) “represents a chemical space from which a model is derived and where a prediction is considered to be reliable” [38]. The AD evaluation was done using the DModX (distance to model) in the *X*-space using SIMCA 16.0.2 software available at <https://landi.ng.umetrics.com/downloads-simca>. The AD plots are given in Figs. 8 and 9 and for training and test sets, respectively, and it is found that there are no outliers in case of training set, and none of the compounds are outside AD in case of the test set at 99% confidence level (D -crit = 0.009999, M -Dcrit [3] = 3.213).

4 Conclusion

In vivo imaging targeting dopamine receptor is a subject of extensive studies nowadays. Dopamine plays a vital role in controlling the pathophysiology of Parkinson’s disease. Hence, it can be treated as a suitable target in controlling the disease. The present study aims in the development of a 2D QSAR model of a group of 34 PET imaging agents having affinity toward dopamine D2 receptor. The 2D QSAR model developed is simple and interpretable and provides knowledge about the basic structural features required for good dopamine binding. The use of simple two-dimensional descriptors reduces the need of time-consuming computational approaches of conformational analysis or energy

minimization; thus, the developed model may be suitable for the quick screening purposes.

Acknowledgements Special issue to Celebrate 80th Birthday of Prof Ramon Carbó-Dorca

Funding PD thanks Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship. KR thanks Science and Engineering Research Board (SERB), New Delhi, for financial assistance under the MATRICS scheme (File number MTR/2019/000008). Financial assistance from DAE-BRNS under the scheme 36 (3)/14/08/2017-BRNS is also thankfully acknowledged.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Parkinson's Foundation (2020) Understanding Parkinson's, Statistics. <https://www.parkinson.org/Understanding-Parkinsons/Statistics>. Accessed on 02 July 2020
2. Jankovic J (2008) Parkinson's disease: clinical features and diagnosis. *J Neurol Neurosurg Psychiatry* 79(4):368–376
3. Barone P (2010) Neurotransmission in Parkinson's disease: beyond dopamine. *Eur J Neurol* 17(3):364–376
4. Antonini A, Moresco R, Gobbo C, De Notaris R, Panzacchi A, Barone P, Calzetti S, Negrotti A, Pezzoli G, Fazio F (2001) The status of dopamine nerve terminals in Parkinson's disease and essential tremor: a PET study with the tracer [11-C] FE-CIT. *Neurol Sci* 22(1):47–48
5. Politis M, Piccini P (2012) Positron emission tomography imaging in neurological disorders. *J Neurol* 259(9):1769–1780
6. De P, Roy J, Bhattacharyya D, Roy K (2020) Chemometric modeling of PET imaging agents for diagnosis of Parkinson's disease: a QSAR approach. *Struct Chem*. <https://doi.org/10.1007/s11224-020-01560-6>
7. Heiss WD, Hilker R (2004) The sensitivity of 18-fluorodopa positron emission tomography and magnetic resonance imaging in Parkinson's disease. *Eur J Neurol* 11(1):5–12
8. Wu L, Liu FT, Ge JJ, Zhao J, Tang YL, Yu WB, Yu H, Anderson T, Zuo CT, Chen L (2018) Clinical characteristics of cognitive impairment in patients with Parkinson's disease and its related pattern in 18F-FDG PET imaging. *Hum Brain Mapp* 39(12):4652–4662
9. Glaab E, Trezzi JP, Greuel A, Jäger C, Hodak Z, Drzezga A, Timmermann L, Tittgemeyer M, Diederich NJ, Eggers C (2019) Integrative analysis of blood metabolomics and PET brain neuroimaging data for Parkinson's disease. *Neurobiol Dis* 124:555–556
10. Roy K (2018) Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. *J Indian Chem Soc* 95(12):1497–2150
11. Gramatica P (2020) Principles of QSAR modeling: comments and suggestions from personal experience. *IJQSPR* 5(3):61–97
12. MarvinSketch software (2020). <https://www.chemaxon.com> Accessed on 25 May 2020
13. Sipos A, Kiss B, Schmidt É, Greiner I, Berényi S (2008) Synthesis and neuropharmacological evaluation of 2-aryl-and alkylapomorphines. *Bioorg Med Chem* 16(7):3773–3779
14. Gao Y, Baldessarini RJ, Kula NS, Neumeyer JL (1990) Synthesis and dopamine receptor affinities of enantiomers of 2-substituted apomorphines and their N-n-propyl analogs. *J Med Chem* 33(6):1800–1805
15. Tóth M, Berényi S, Csutorás C, Kula NS, Zhang K, Baldessarini RJ, Neumeyer JL (2006) Synthesis and dopamine receptor binding of sulfur-containing apomorphines. *Bioorg Med Chem* 14(6):1918–1923
16. Søndergaard K, Kristensen JL, Palner M, Gillings N, Knudsen GM, Roth BL, Begtrup M (2005) Synthesis and binding studies of 2-arylapomorphines. *Org Biomol Chem* 3(22):4077–4081
17. Gao Y, Ram VJ, Campbell A, Kula NS, Baldessarini RJ, Neumeyer JL (1990) Synthesis and structural requirements of N-substituted norapomorphines for affinity and activity at dopamine D-1, D-2, and agonist receptor sites in rat brain. *J Med Chem* 33(1):39–44
18. Baldessarini R, Kula N, Gao Y, Campbell A, Neumeyer J (1991) R (–) 2-fluoro-nn-propylnorapomorphine: a very potent and D2-selective dopamine agonist. *Neuropharmacology* 30(1):97–99
19. Vasdev N, Natesan S, Galineau L, Garcia A, Stableford WT, McCormick P, Seeman P, Houle S, Wilson AA (2006) Radiosynthesis, ex vivo and in vivo evaluation of [11C] preclamol as a partial dopamine D2 agonist radioligand for positron emission tomography. *Synapse* 60(4):314–331
20. Chumpradit S, Kung M, Billings J, Mach R, Kung H (1993) Fluorinated and iodinated dopamine agents: D2 imaging agents for PET and SPECT. *J Med Chem* 36(2):221–228
21. Murphy RA, Kung HF, Kung MP, Billings J (1990) Synthesis and characterization of iodobenzamide analogs: potential D-2 dopamine receptor imaging agents. *J Med Chem* 33(1):171–178
22. Dragon version 7 (2016) Kodesrl, Milan, Italy. <https://www.talet.e.mi.it/index.htm>. Accessed on 26 May 2020
23. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29(6–7):476–488
24. Golmohammadi H, Dashtbozorgi Z, Acree WE Jr (2012) Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur J Pharm Sci* 47(2):421–429
25. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intell Lab Syst* 159:108–126
26. Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, Cornwall, Great Britain
27. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opin Drug Discov* 13(12):1075–1089
28. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58(2):109–130
29. Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform* 6(1):47
30. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen* 14(6):450–474
31. Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm2 metrics for validation of QSPR models. *Chemom Intell Lab Syst* 107(1):194–205
32. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33
33. Akarachantachote N, Chadcham S, Saithanu K (2014) Cutoff threshold of variable importance in projection for variable selection. *Int J Pure Appl Math* 94(3):307–322

34. Finnema SJ, Bang-Andersen B, Wikstrom HV, Halldin C (2010) Current state of agonist radioligands for imaging of brain dopamine D2/D3 receptors in vivo with positron emission tomography. *Curr Top Med Chem* 10(15):1477–1498
35. De P, Aher RB, Roy K (2018) Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector *Aedes aegypti*: application of ETA indices. *RSC Adv* 8(9):4662–5467
36. Jackson JE (2005) *A user's guide to principal components*, vol 587. Wiley, United States of America
37. Topliss JG, Edwards RP (1979) Chance factors in studies of quantitative structure-activity relationships. *J Med Chem* 22(10):1238–1244
38. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 1(1):45–63

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Computational modeling of PET imaging agents for vesicular acetylcholine transporter (VACHT) protein binding affinity: application of 2D-QSAR modeling and molecular docking techniques

Priyanka De¹ · Kunal Roy¹

Received: 27 December 2022 / Accepted: 31 March 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

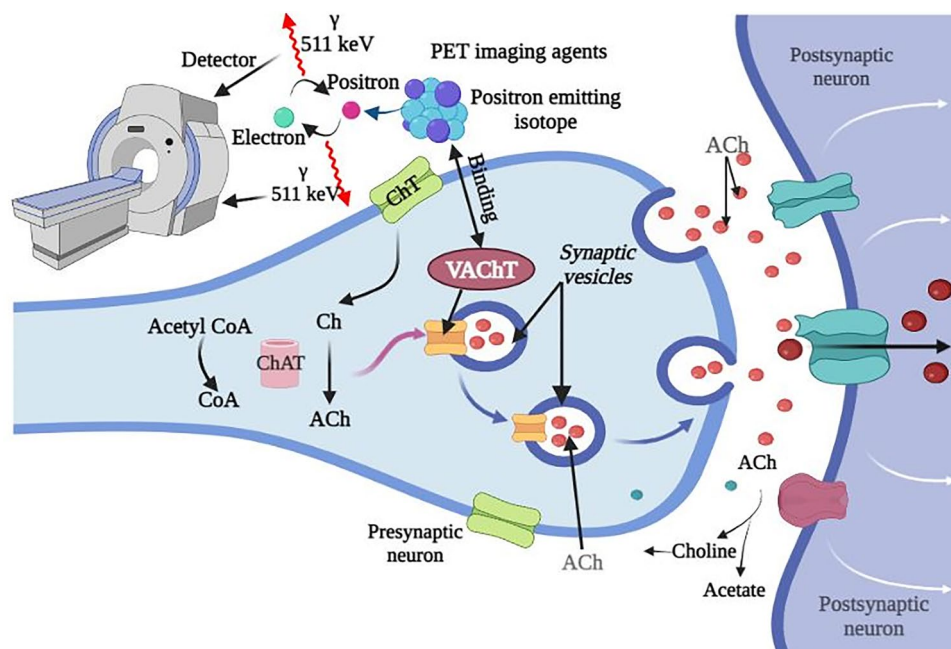
Abstract

The neurotransmitter acetylcholine (ACh) plays a ubiquitous role in cognitive functions including learning and memory with widespread innervation in the cortex, subcortical structures, and the cerebellum. Cholinergic receptors, transporters, or enzymes associated with many neurodegenerative diseases, including Alzheimer's disease (AD) and Parkinson's disease (PD), are potential imaging targets. In the present study, we have developed 2D-quantitative structure–activity relationship (2D-QSAR) models for 19 positron emission tomography (PET) imaging agents targeted against presynaptic vesicular acetylcholine transporter (VACHT). VACHT assists in the transport of ACh into the presynaptic storage vesicles, and it becomes one of the main targets for the diagnosis of various neurodegenerative diseases. In our work, we aimed to understand the important structural features of the PET imaging agents required for their binding with VACHT. This was done by feature selection using a Genetic Algorithm followed by the Best Subset Selection method and developing a Partial Least Squares-based 2D-QSAR model using the best feature combination. The developed QSAR model showed significant statistical performance and reliability. Using the features selected in the 2D-QSAR analysis, we have also performed similarity-based chemical read-across predictions and obtained encouraging external validation statistics. Further, we have also performed molecular docking analysis to understand the molecular interactions occurring between the PET imaging agents and the VACHT receptor. The molecular docking results were correlated with the QSAR features for a better understanding of the molecular interactions. This research serves to fulfill the experimental data gap, highlighting the applicability of computational methods in the PET imaging agents' binding affinity prediction.

✉ Kunal Roy
kunal.roy@jadavpuruniversity.in

¹ Drug Theoretics and Cheminformatics Laboratory,
Department of Pharmaceutical Technology, Jadavpur
University, Kolkata 700032, India

Graphical abstract



Keywords QSAR · PET imaging agents · Read-across · Alzheimer's drugs

Introduction

According to World Health Organisation (WHO), currently, more than 55 million people live with dementia worldwide, and there are more than 10 million cases new cases every year. Dementia is characterized by the loss or decline in memory or other cognitive impairment commonly observed in neurodegenerative disorders like Alzheimer's disease (AD), Parkinson's disease (PD), schizophrenia, and Down's syndrome. The severity of dementia-associated cognitive dysfunction is connected with the loss of cholinergic synaptic elements in the cortex and subcortical regions of the brain (Bohnen and Albin 2011; Hampel et al. 2018). Cholinergic neurons are accountable for synaptic transmission as well as neuronal modulation in various regions of the central and peripheral nervous systems. Cholinergic neurotransmission controls cognitive functions including learning and memory. Acetylcholine (ACh) is one of the main neurotransmitters secreted by cholinergic neurons to perform a plethora of physiological functions (Prado et al. 2013). ACh is produced at the nerve terminals from acetyl coenzyme A (acetyl CoA) and choline by Choline acetyltransferase (ChAT) enzyme. The neurotransmitters are then transported and stored in synaptic vesicles by transporters called vesicular ACh transporters (VACHTs), before being released in the synaptic cleft (Amenta and Tayebati 2008). Neurodegenerative diseases have common events of cholinergic impairment.

Thus, radiolabeling of these vesicular transporters would provide a presynaptic marker of cholinergic innervation. The depletion in ChAT and AChE levels, occurring in several neurodegenerative diseases, are potential measuring targets for these imaging agents (Bergmann et al. 1978; Mountjoy 1986; Mountjoy et al. 1984) (Fig. 1). Imaging cholinergic neurotransmission in vivo with positron emission tomography (PET) provides noteworthy information about disease progression.

Radio imaging of presynaptic VACHT was first done using ^{18}F -fluoroethoxybenzovesamicol (^{18}F -FEOBV), a PET ligand, which was later successfully rendered into clinical application. Vesamicol (2-(4-phenylpiperidino)cyclohexanol) was reported to bind to VACHT and is considered to be a useful lead for developing new PET imaging agents for mapping cholinergic signaling in vivo (Giboureau et al. 2012). Kitamura et al. (2016) found that o-methyl-trans-decalinvesamicol (OMDV) demonstrated a high binding affinity and selectivity for VACHT and can be used in the early diagnosis of Alzheimer's disease (AD). ^{11}C JOMDV was synthesized and investigated as a new PET ligand for VACHT imaging through in vivo evaluation. Kilbourn et al. (2009) used (2R,3R)-5- ^{18}F fluoroethoxybenzovesamicol in micro PET imaging to determine the regional brain pharmacokinetics of rat and rhesus monkey brains. Horsager et al. (2022) evaluated human in vivo VACHT distribution in 13 peripheral organs using a 70 min dynamic ^{18}F

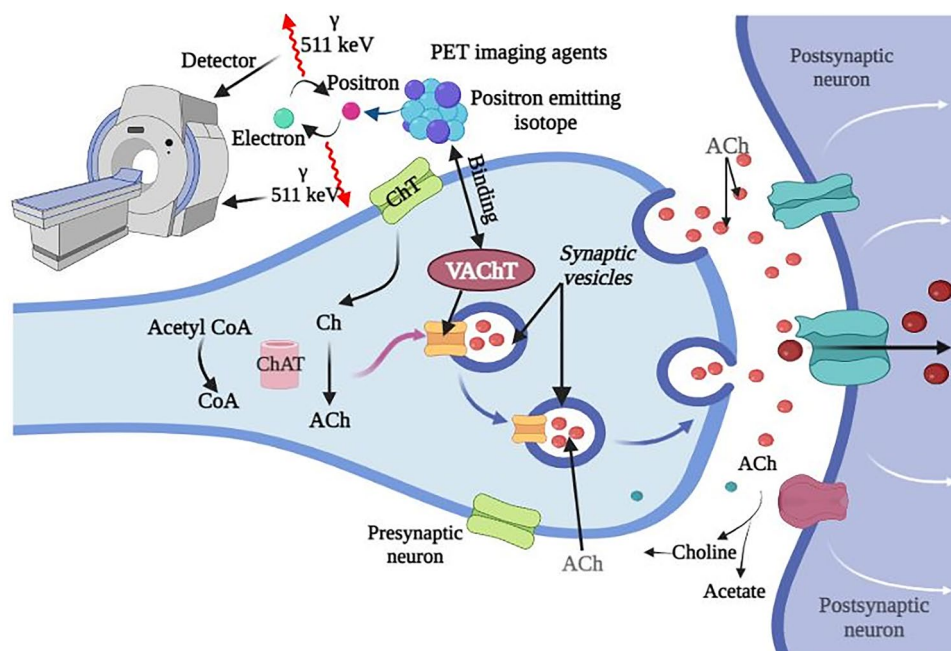


Fig. 1 PET imaging of vesicular acetylcholine transporter to study its role in presynaptic cholinergic innervations. VACHT helps in the transport of Acetylcholine (ACh), the essential neurotransmitter regulating AD and PD, through the synaptic vesicles. PET imaging technology helps in the diagnosis of the increase or decrease in the number of VACHT receptors. Mechanism of VACHT: In the cytoplasm of

nerve endings, ACh is synthesized by the enzyme ChAT, and then it is loaded into synaptic vesicles by VACHT. Upon any nerve impulse, vesicles fuse to the plasma membrane and release the neurotransmitter ACh. (AD: Alzheimer's disease; PD: Parkinson's disease; CoA: Coenzyme A; ChAT: Choline acetyltransferase)

fluoroethoxybenzovesamicol ($[^{18}\text{F}]\text{FEOBV}$) PET/CT protocol. Tu et al. (2009) synthesised nine fluorine-containing VACHT inhibitors and screened them as potential PET tracers for imaging the VACHT. Earlier, imaging of AD was possible through the amyloid beta detection; however, these methods were not useful to evaluate the therapeutic efficacy of the AD treatment. The dysfunction of presynaptic cholinergic neurons is associated with loss of choline acetyltransferase (ChAT) (enzyme synthesizing ACh) and the vesicular acetylcholine transporter (VACHT) (Reinikainen et al. 1990). Thus, these internal molecules function as novel cranial molecular targets for developing new imaging agents for their detection.

Driven by the continuous search for new entities with improved properties and considerably lower toxicities, theoretical approaches are of high priority within the chemical and pharmaceutical industries. This provides a logical design of chemicals or pharmaceuticals with reduced time and cost. Quantitative structure–activity relationship (QSAR) has gained immense importance in the pharmaceutical industries as an effective tool for the predictions when experimental data is limited (Gramatica 2020). QSAR has enormous applications in medicinal chemistry, drug designing, and toxicity prediction. Another chemometric approach, similarity-based quantitative read-across (Chatterjee et al. 2022), can also be used for data gap filling. This method uses a weighted average approach to

quantitatively predict similar query compounds. Read-across approach, due to its transparency, has a strong potential for providing confident predictions.

In the present research, we have strived to develop a two-dimensional QSAR model with 19 PET imaging agents acting against vesicular acetyl choline transporter. The selection of a small dataset was due to the non-availability of a larger number of experimental data. Here, QSAR modeling plays a pivotal role for providing promising predictions when data is scarce. To revalidate our predictions, we have performed leave-one-out and leave-many-out cross-validation tests. We have also performed read-across based predictions to analyze the predictive ability of the features obtained from QSAR analysis. Besides these, we also have performed molecular docking analysis to corroborate its results with QSAR analysis. Further, we have used two external datasets of PET imaging agents for their VACHT binding predictions (vide infra) using our developed 2D-QSAR model.

Materials and methods

In the present study, 2D-quantitative structure–activity relationship (2D-QSAR) models were developed for 19 positron emission tomography (PET) imaging agents targeted against presynaptic vesicular acetylcholine

transporter (VACHT). VACHT assists in the transport of ACh into the presynaptic storage vesicles, and it is one of the main targets for the diagnosis of various neurodegenerative diseases.

The dataset

According to the OECD principle, dataset selection with a defined endpoint is the first essential step while developing a QSAR model. For our present work, the binding affinity (K_i) values of 19 PET imaging agents acting against vesicular acetylcholine transporter were procured from different previously published literature (Kovac et al. 2010; Tu et al. 2015, 2009). The VACHT binding affinity of the dataset compounds was assayed by the same experimental protocol of the competitive displacement of 5 nM [^3H] vesamicol on homogenates of PC12 cells (Zea-Ponce et al. 2005). The binding affinity data which was expressed as K_i were converted to its negative logarithmic form (pK_i). The structures obtained from different sources were then represented in MarvinSketch version 15.12.7.0 software with proper explicit hydrogen addition and aromatization. The 19 PET imaging agents used for the present study is given in Table 1. The dataset compounds obtained from three sources had some common compounds. The compound IDs in Table 1 are given in such a way so that the common compounds are not repeated.

Molecular descriptors

The molecular descriptor is a fundamental component of QSAR and other in-silico models since it formally represents a molecule's structure numerically. Descriptors provide a mathematically meaningful relationship between the molecular structure and biological activities, physico-chemical and toxicological properties of chemicals (Mauri et al. 2017). Descriptors can be classified into different categories depending on the process of calculation or scheme of experimental determination or concept of the origin. For ease of interpretation, the present work involved the use of eight main types of two-dimensional (2D) descriptors, viz., E-state indices, extended topochemical atom (ETA), connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments and molecular property descriptors. The descriptors were calculated using alvaDesc descriptor calculator (Alvascience, alvaDesc version 2.0.6, 2021, <https://www.alvascience.com>). With the intention to minimize the redundant and incompetent data, inter-correlated descriptors (correlation greater than 0.95) were removed from the original descriptor pool. This

resulted in a final pool of 188 descriptors which was used as input variables for QSAR modeling.

Feature selection and model development

In general, a QSAR model development involves a training set and a test for model development and validation purposes respectively. However, owing to the small number of compounds in our dataset, we did not apply the general method of data division (Király et al. 2022; Kovács et al. 2021; Rácz et al. 2021). It is natural that all the descriptors calculated through AlvaDesc will not be able describe the binding properties of the PET imaging agents. Therefore, to further reduce the data pool, we have applied the Genetic Algorithm (Sukumar et al. 2014) feature selection method to choose essential features required for binding. Further, we have executed the Best Subset Selection (available from <http://dtclab.webs.com/software-tools>) on the reduced pool of 12 descriptors obtained from the GA. Finally, the acquired pool of descriptors was applied to develop the final model using the partial least squares (PLS) regression (Wold et al. 2001). PLS converts the original descriptors into the new latent variable space thus lowering the dimensionality and obviating the inter-correlation among the original descriptors.

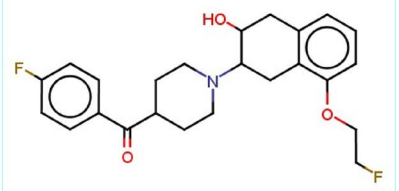
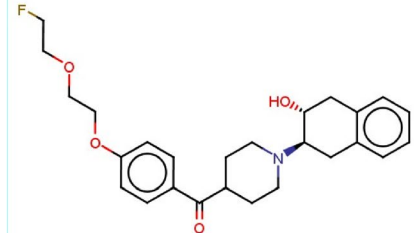
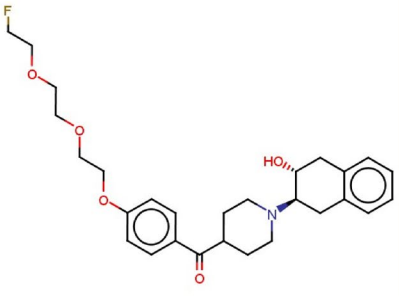
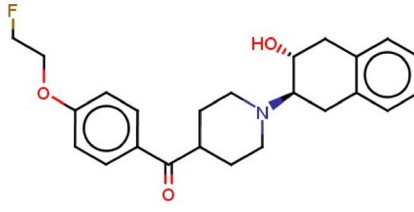
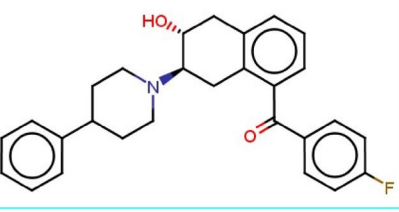
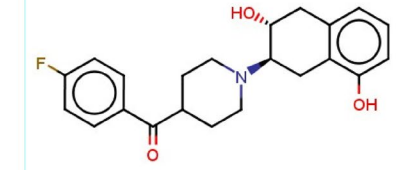
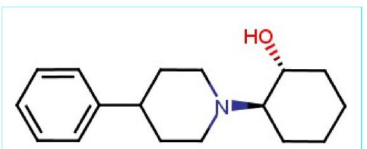
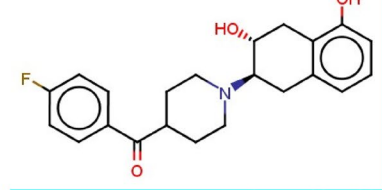
Machine learning-based read across predictions

In the current work, we have employed a machine learning-based Read-across prediction tool which relies on similarity approaches. The predictions were made using the tool Quantitative Read Across v4.0 developed by Chatterjee et al. (2022) available from <https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>. The main similarity approaches involved in this tool are Euclidean distance-based similarity, Gaussian kernel function, and Laplacian kernel function-based similarity estimation. Please note that read-across does not develop any statistical model like QSAR and make predictions only based on the similarity values. Thus, this approach may be good when a limited number of source compounds is available (Banerjee and Roy 2022). For read-across predictions, we have divided the dataset into training and test sets. The prediction scheme starts with the initial optimization of hyperparameters (sigma and gamma values; distance and similarity thresholds) which requires division of the training into sub-training and sub-test sets into different combinations. This step is followed by the selection of the best setting of hyperparameters which is then applied to the original training and test sets.

Molecular docking

In this study, molecular docking was performed using the most and least active compounds from the initial dataset to

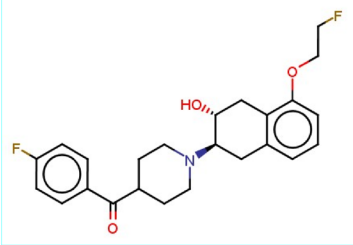
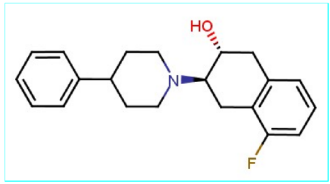
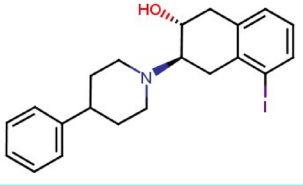
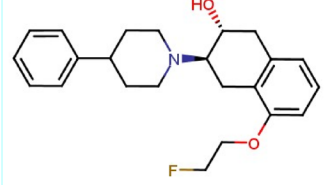

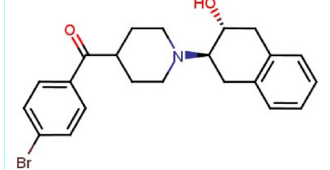
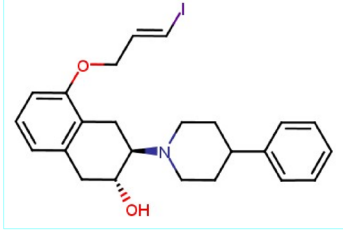
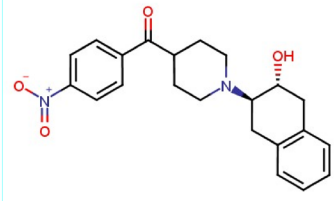
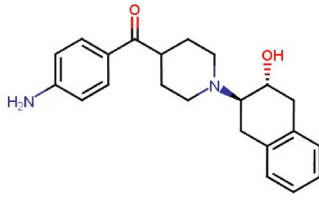
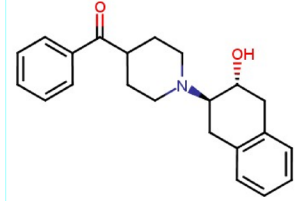
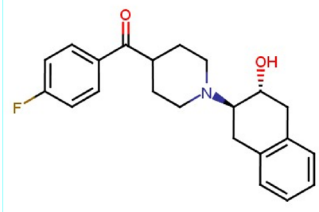
Table 1 PET radiotracers target vesicular acetylcholine transporters (VACHT)

| Structures with compound IDs | <i>pK_i</i> | Structures with compound IDs | <i>pK_i</i> |
|--|-----------------------|--|-----------------------|
|  1 | 2.239 |  2 | 3.060 |
|  3 | 2.921 |  5 | 2.770 |
|  6 | 2.569 |  7 | 2.337 |
|  9 | 2.261 |  10 | 1.252 |

identify the interaction pattern with the target. Owing to the non-availability of any protein structure for VACHT in the protein data bank, we have retrieved the predicted protein structure from the AlphaFold Protein Structure Database (available from <https://alphafold.ebi.ac.uk/entry/Q16572>) with the UniProt: Q16572, Source organism: Homo sapiens (Human), and AlphaFold id: AF-Q16572-F1-model_v2. AlphaFold is an artificial intelligence (AI) system established by DeepMind that predicts a protein's three-dimensional (3D) structure from its amino acid sequence (Jumper et al. 2021; Varadi et al. 2022). We have then validated the reliability of the predicted structure using the Ramachandran

plot server embedded in Biovia Discovery Studio 4.1 which represents the good quality of the model (see Fig. 2). In this study, multiple active sites at the surface of the protein were predicted using the Biovia discovery studio 4.1 client platform from the "define and edit binding site" using the module "generate active site from receptor cavities", and the ligand was docked into each site to identify the favorable binding site (identified most favorable active site coordinate x: 16.478, y: 6.38307, Z: -15.9527, the radius of the sphere: 26). Initially, a total of sixteen binding sites were identified where the standard compound "vesamicol" was docked. It was found that vesamicol binds at core of site 1

Table 1 (continued)

| | | | |
|---|-------|--|-------|
|  <p>11</p> | 1.032 |  <p>12</p> | 2.185 |
|  <p>15</p> | 1.801 |  <p>16</p> | 1.476 |
|  <p>20</p> | 3.658 |  <p>21</p> | 3.602 |
|  <p>22</p> | 3.347 |  <p>23</p> | 3.319 |
|  <p>25</p> | 2.770 |  <p>27</p> | 2.367 |
|  <p>29</p> | 0.967 | | |

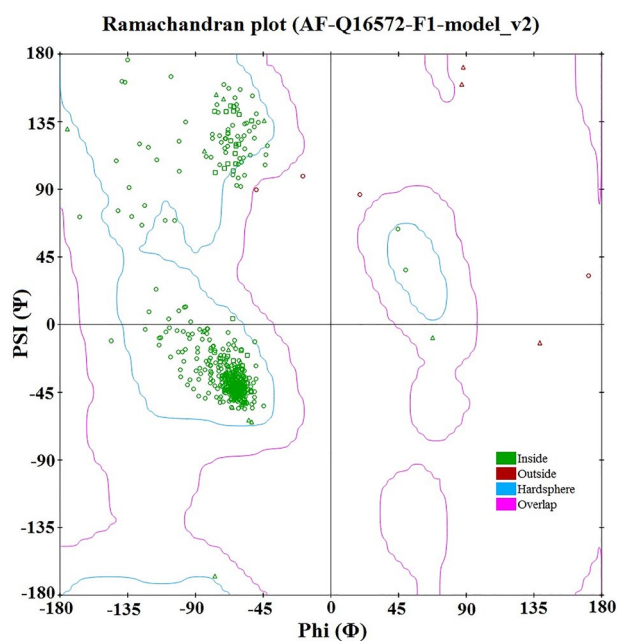


Fig. 2 Ramachandran plot for Vesicular acetylcholine transporter model (UniProt: Q16572, Source organism: Homo sapiens (Human), and AlphaFold id: AF-Q16572-F1-model_v2). Ramachandran plot shows 435 residues (97.098%) reside in the most favored region, 10 (2.232%) residues reside in the preferable region and only 3 (0.670%) reside in the unfavorable region

of the protein with a good binding energy (27.572 kcal/mol) and interactions (shown later in the “Results and discussion” section). Out of other 15 sites, molecular docking failed in five docking sites and in case of the rest ten sites vesamicol did not bind at the docking site (outside the grid). Thus, site 1 was chosen for further docking analysis. Ligand preparation was performed using selected high and low active compounds by running them through the Discovery Studio platform’s ‘small-molecule module’, where several ligand conformers were formed. Each of these generated conformers was subsequently employed in the CDOCKER module for molecular docking using a CHARMM-based molecular dynamic scheme (Wu et al. 2003). The CDOCKER interaction energy parameter (kcal/mol) was examined for all receptor-ligand complexes, and the highest-scoring (more negative; hence favorable to binding) poses with only non-covalent interactions (ionic bonds, hydrophobic interactions, hydrogen bonds, etc.) were kept for future investigation.

Results and discussions

The present work demonstrates the contribution of different structural attributes of PET imaging agents required for binding to and quantifying the presence of vesicular acetylcholine transporter. The main work is focused on the development of a

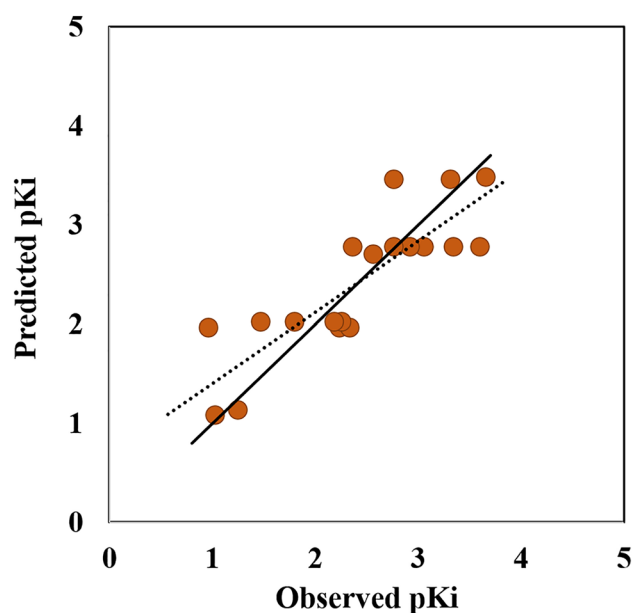


Fig. 3 Observed versus predicted scatter plot of the PLS model

simple 2D-QSAR model to obtain the major structural features responsible for binding. These features were further validated using the structural similarity-based read-across approach as well as molecular docking techniques.

QSAR modeling of binding affinity of PET imaging agents towards VACHT

The dataset procured for this study consisted of 19 compounds. A three-descriptor partial least squares (PLS) regression model with two latent variables (LVs) was developed which could explain 71.77% of the variance. The leave-one-out cross-validated determination coefficient (i.e., $Q_{LOO}^2 = 0.523$) is above the critical threshold value fulfilling the statistical reliability of the model. We have also calculated the leave-many-out squared correlation coefficient ($Q_{LMO(25\%)}^2$), and the result obtained was above the threshold value (Roy et al. 2015). The observed versus predicted pKi (Supplementary S1) scatter plot is shown in Fig. 3. In cases, where residuals are high (Fig. 3), clearly some contributing features important for the response have remained unidentified and not included in the model. This is usual for models developed from a small data set, as due to limited variability of a particular (important) feature in the data set, the feature is not captured by the modeling algorithm. As more and more data become available, the model can be refined subsequently. However, with the available data, the presently developed model may be a good start as a tool for future predictions.

$$pK_i = 2.018 - 0.831 \times B06[N-O] + 0.757 \times F08[C-N] - 0.812 \times F09[N-F]$$

$$N = 19, R^2 = 0.718, Q_{(LOO)}^2 = 0.523,$$

$$Q_{LMO(25\%)}^2 = 0.598, r_{m(LOO)}^2 = 0.439,$$

$$\Delta r_{m(LOO)}^2 = 0.027, MAE = 0.335, SD = 0.273$$

The descriptors appearing in the final PLS model are all 2D atom pair descriptors suggesting the importance of the presence of a particular atom pair in the PET tracer molecule. The 2D atom pair descriptors are mainly dependent on the topological distance between two atoms pairs. Thus, the value of the descriptors can be similar for many compounds resulting in same predicted pKi values for many dataset compounds (Fig. 3). The residual might be high in such cases, but all the compounds are inside the applicability domain of the model (vide infra). The variable importance plot (Akarachantachote et al. 2014) given in Fig. 4 shows the significance level of each descriptor toward VACHT binding affinity. The descriptor **F09[N-F]** was the most significant descriptor with VIP Score > 1 (VIP = 1.289) followed by **F08[C-N]** (VIP = 1.043) and **B06[N-O]** (VIP = 0.502). F09[N-F] which contributes negatively to the binding affinity, is the frequency of the

N-F fragment at a topological distance 9. Compounds like **10** and **11** (Fig. 4) have nitrogen and fluorine at the topological distance 9, thereby decreasing the binding affinity towards VACHT, whereas in compounds like **21** and **23**, the N-F fragment at 9 distance is absent, and the pKi values are high.

The next important 2D atom pair descriptor is F08[C-N] which denotes the frequency of C-N fragment at the topological distance 8. The positive regression coefficient indicates that with an increase in the frequency of C-N at the 8 distance, the binding affinity will increase as observed in compounds like **20** (Fig. 4), **23**, and **25**. These compounds have three such fragments and have high pKi values of 3.658, 3.319, and 2.700 respectively.

The least important among all the descriptors is B06[N-O] which implies the presence or absence of an N-O fragment at a topological distance 6. The negative contribution indicates that the presence of such a fragment will decrease the VACHT binding of the PET imaging agents as seen in compounds like **10** and **11** (Fig. 4). These compounds have a very low binding affinity towards (1.251 and 1.032 respectively) VACHT receptor.

The significance and validity of the developed model were further analyzed using some important PLS plots, namely, the loading plot, randomization plot, and applicability domain (AD) which are described below.

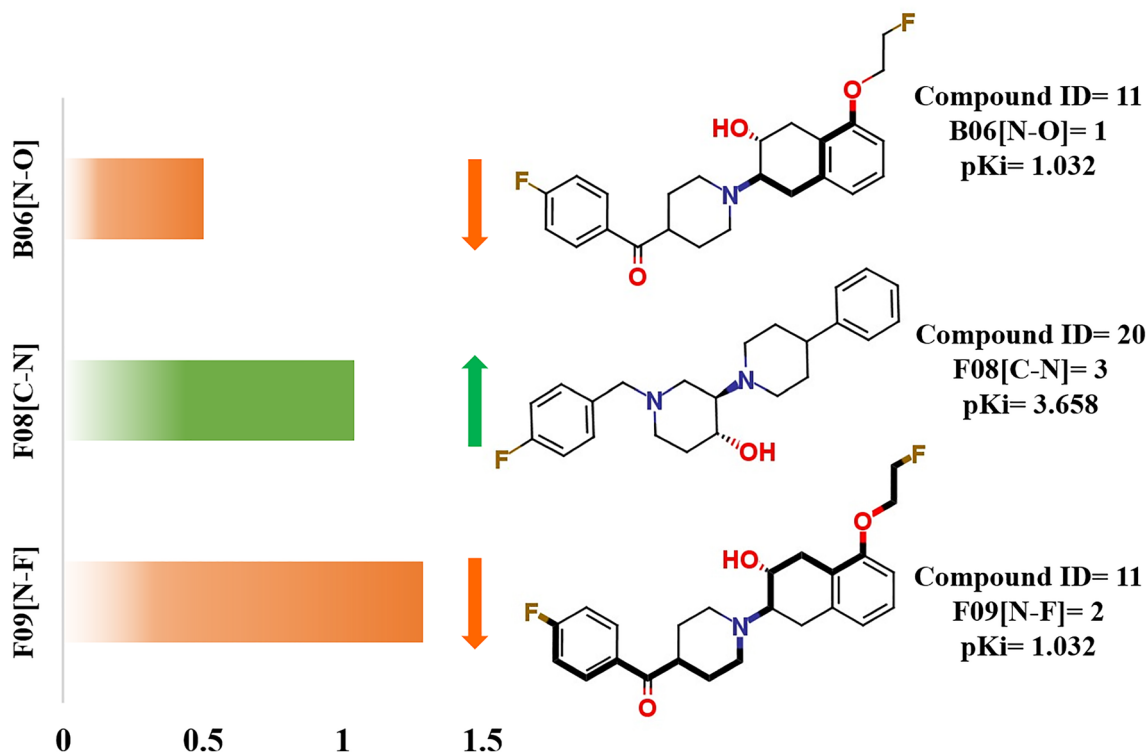


Fig. 4 Variable importance plot and significance of the descriptors appearing in the PLS model

A **loading plot** (Fig. 5) explains the relationship between the independent variables or descriptors (X-variables) with the dependent variable or pKi values (Y-variable). The influence of the descriptors on the developed model can be recognized from the loading plot. Descriptors that are far from the plot origin (like F08[C-N] and F09[N-F]) contribute significantly more toward the binding affinity. Descriptors with different meanings appear distantly from each other in the loading plot.

Model randomization confirms that the model is not the outcome of any chance correlation (Topliss and Edwards 1979). The **randomization plot** determines the statistical significance and robustness of the model. Multiple models are generated during a randomization plot development by shuffling different combinations of either X-variables (X-randomization) or Y-response (Y-randomization). The Y-randomization was performed in the present study with 100 permutations for each model for random model generation. For a non-random model R_y^2 intercept should not exceed 0.3 and Q_y^2 intercept should not exceed 0.05. The randomization plot given in Fig. S1 (Supplementary File S2) shows that the developed model is non-random and robust and is suitable for prediction.

According to OECD guideline 3, a developed QSAR model should possess a defined chemical domain of applicability. AD can be interpreted as a chemical space defined by the structural information or molecular properties of the chemicals used in the model development (Gadaleta et al. 2016). Compounds present within this chemical space can only be properly predicted. In this study, the DModX (distance to model in X-space) method of AD determination (Kar et al. 2018; Vargas et al. 2018) at a 99% confidence interval (D-crit = 0.009999) was applied using SIMCA

16.0.2 software (Wu et al. 2010). DModX represents the unexplained variation (residuals), and it can be explained as the distance to the model X space corresponding to the X residuals standard deviation (Vargas et al. 2018). The DModX value of an observation i can be calculated using the formula $S_i = \sqrt{\frac{\sum e_{ik}^2}{(K-A)}} / \sqrt{\frac{\sum e_k^2}{(N-A-A0)(K-A)}}$, where e_{ik} is the X-residual of the observation i and variable k , $\sum e_k^2$ is the squared sum of the residuals, N is the number of observations, K is the number of x-variables, and A is the number of latent variables, $A0$ is 1 if the model is centered and 0 otherwise. The DModX is asserted to be F-distributed, and thus, can be used to analyse if the observation is significantly far away from the PLS model presuming the data is normally distributed. The AD plot (Fig. 6) shows none of the compounds was an outlier. The PET compounds selected for the VACHT binding and imaging contained a basic core structure of 2-(piperidin-1-yl)cyclohexan-1-ol which is also the main core moiety of standard compound vesamicol. The QSAR model developed in the present research contains 2D-atom pair features which can predict compounds with or without the core structures as evident from the external set predictions (vide infra).

Although we developed our QSAR model from the whole set due to the limited availability of the experimental data, to further check its validity for external predictions, we have additionally split the dataset into training and test sets, and redeveloped three models with the same combination of descriptors (given in the Supplementary Section S1). The models were found to be robust and predictive.

Fig. 5 Loading plot of the PLS model

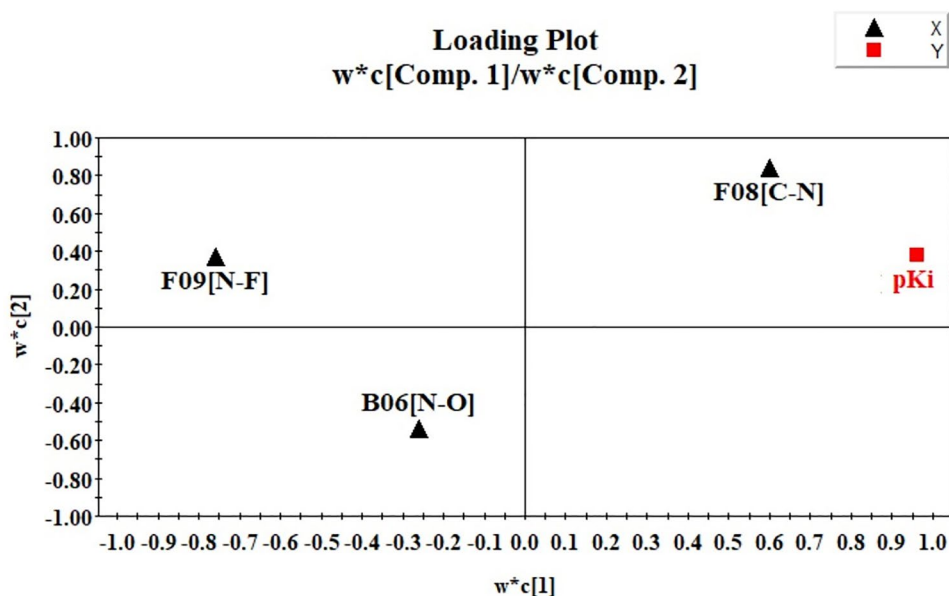
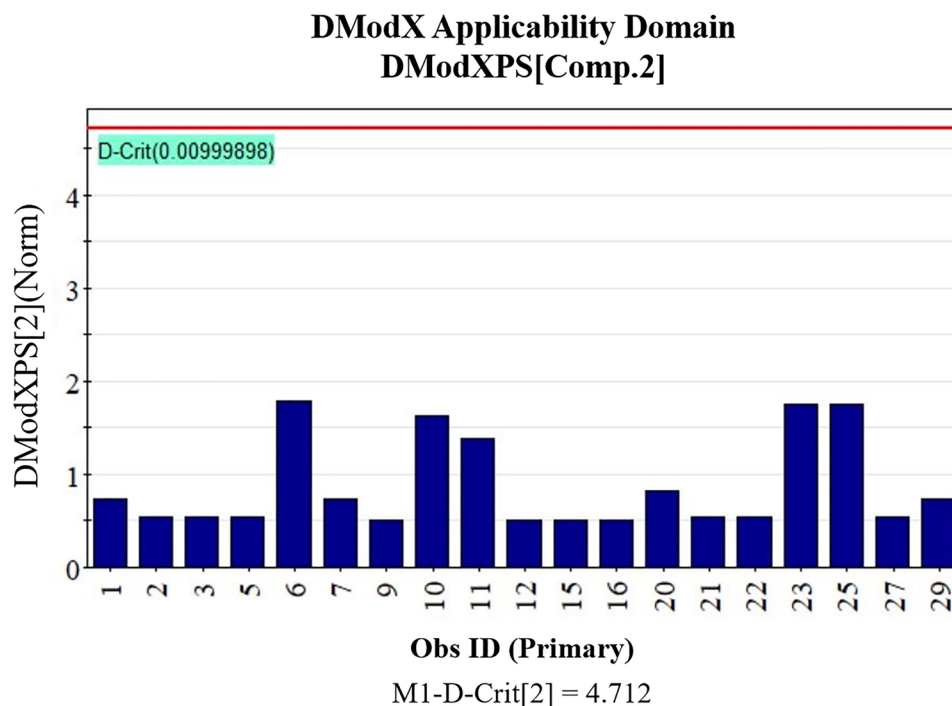


Fig. 6 DModX AD plot of the PLS model

Read-across based predictions

To explore the predictivity of the selected features used for QSAR modeling, a similarity-based read-across prediction was performed by using a group of five compounds (compound ID: **3**, **11**, **12**, **21**, and **27**) as the test set (Chatterjee et al. 2022). Read-across was also previously performed on small datasets (< 20 compounds) successfully (Gajewicz et al. 2014, 2017; Gajewicz 2017a, b). In the current work, three types of similarity were measured: the Euclidean Distance-based, the Gaussian Kernel Similarity-based, and the Laplacean Kernel Similarity based predictions using Read-Across-v4.1 (<https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>) tool and after hyperparameter optimization using Auto_RA_Optimizer-v1.0 tool (<https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>), it was found that the external validation results obtained from quantitative Read-Across algorithm using Gaussian Kernel Similarity-based functions

($Q_{F1}^2 = 0.763$, $Q_{F2}^2 = 0.763$, $RMSE = 0.414$, $MAE = 0.331$) was better compared to the results obtained with the other two read-across approaches (Table 2).

Molecular docking

Molecular docking must include a reasonably accurate model of energy and should be able to deal with the combinatorial complexity experienced by the molecular flexibility of the docking partners. In the present research, molecular docking studies were performed to understand the individual molecular interactions and orientation of the imaging agents occurring at the binding zone of the VACHT receptor (Fig. 7). In the present work, the protein structure for VACHT was not available in PDB, hence, we have procured the predicted protein structure from AlphaFold Protein Structure Database. The selected protein structure was further validated using the famous Ramachandran plot to improve the accuracy of prediction. From the Ramachandran

Table 2 Comparison between three types of read-across predictions

| Method | N_{train} | R^2 | $Q_{(LOO)}^2$ | MAE | N_{test} | Q_{F1}^2 | Q_{F2}^2 | MAE |
|--------------------|-------------|-------|---------------|-------|------------|--------------|--------------|--------------|
| QSAR | 19 | 0.718 | 0.523 | 0.335 | – | – | – | – |
| Read-Across | | | | | | | | |
| Euclidean distance | 14 | – | – | – | 5 | 0.189 | 0.189 | 0.596 |
| Gaussian Kernel | – | – | – | – | – | 0.763 | 0.763 | 0.331 |
| Laplacian Kernel | – | – | – | – | – | 0.719 | 0.719 | 0.380 |

Bold values indicate the best predictions

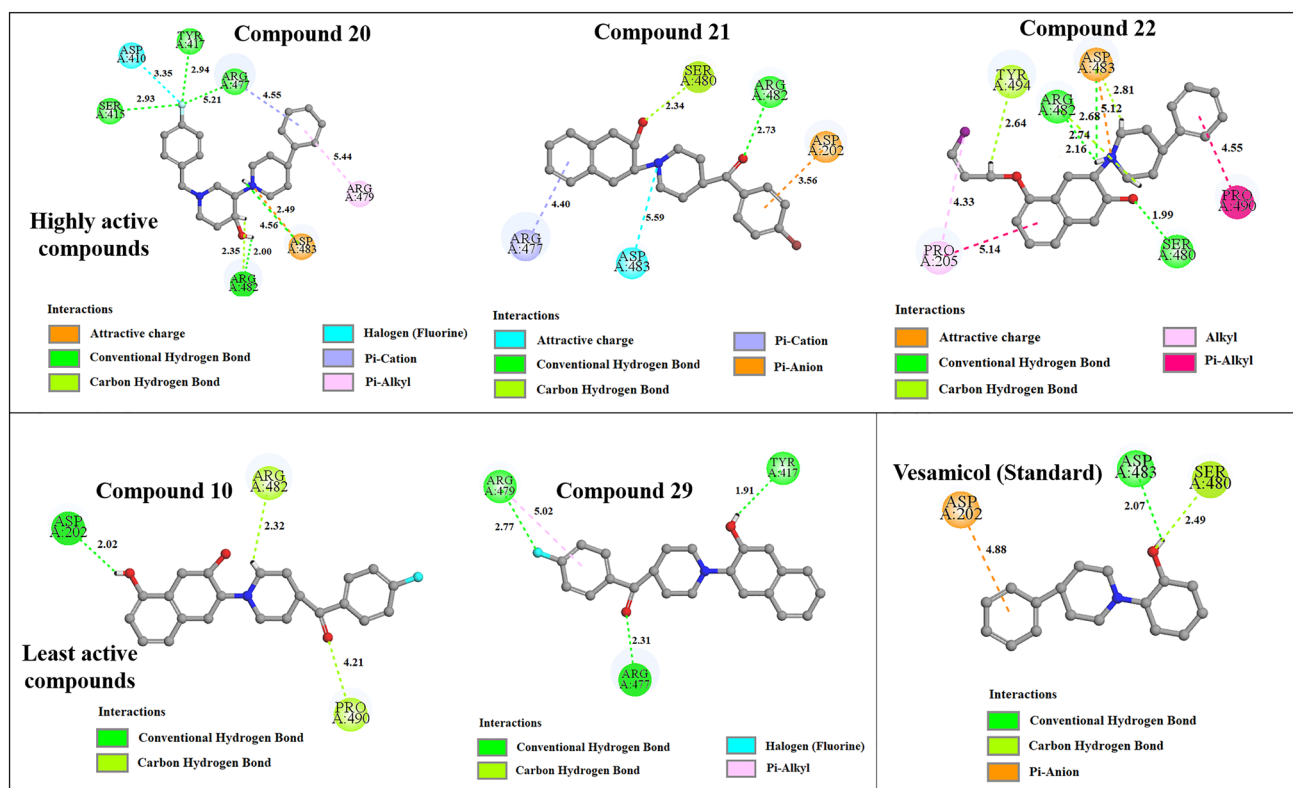


Fig. 7 Molecular docking interactions of highly active, least active and standard compounds against VACHT binding

plot, we found that 435 residues (97.098%) reside in the most favored region, 10 (2.232%) residues reside in the preferable region and only 3 (0.670%) reside in the unfavorable region. During docking, the physiological pH of brain was considered at which the piperidine N is protonated. The tautomers were also considered. Further, validation using the standard compound, i.e., vesamicol was performed by docking at the binding site to understand its nature of interactions. Again, both high and low-active compounds were also used for the docking study. In the case of vesamicol (compound **9**), which has a moderate binding affinity ($pK_i = 2.261$), the interaction forces include hydrogen bond interactions (both conventional and carbon-hydrogen bond interactions) and π -anion interactions. The amino acid residues engaged in vesamicol binding are Asp A:202, Asp A:483, and Ser A:480. Comparing vesamicol-VACHT binding interactions with highly active compounds like compound ID **20** ($pK_i = 3.658$), **21** ($pK_i = 3.602$), and **22** ($pK_i = 3.347$), it was observed that similar interactions were also involved in their binding (Fig. 7). However, it was found that these highly active compounds were docked with higher number of interactions at their binding site with far better binding (Table 3). For compound **20**, halogen (fluorine) interactions, attractive charge, π -cation, and π -alkyl interactions were active along with hydrogen bond interactions. In the case of compound

21, additional interactions include attractive charge, π -anion, and π -cation interactions. Similarly, in the case of compound **22**, attractive charges, alkyl, and π -alkyl interactions were active along with conventional hydrogen bond and carbon-hydrogen bond interactions. The attractive charge interaction of Asp A:483 amino acid with the nitrogen of piperidine moiety of all three high active compounds was a noteworthy finding inferring the importance of the fragment in VACHT binding.

In the case of lower active compounds like compound **10** ($pK_i = 1.032$) and compound **29** ($pK_i = 0.967$) (Fig. 7), the number of molecular interactions was much less than the higher active ones (Table 3). Conventional hydrogen bond and carbon-hydrogen bond interactions were prevalent, with additional halogen and π -alkyl in the case of compound **29**.

Relationship with QSAR features

From QSAR modeling, it was found that F08[C-N] is the only positively contributing descriptor. Therefore, the presence of nitrogen in the PET imaging agent is very essential for good VACHT binding. In the case of highly active compounds (compounds **20**, **21**, and **22**) used for molecular docking, it was found that attractive charge interaction was prevalent in all three compounds which occurred between

Table 3 The interacting residues and different types of binding interaction occurring between the PET imaging agents and VACHT

| Compound | Category | pKi | Binding amino acids | Types of interactions |
|---------------|---------------|-------|---|---|
| 9 (Vesamicol) | Standard | 2.261 | Asp A:202, Asp A:483, Ser A:480 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, and π -anion interaction |
| 20 | Highly active | 3.658 | Ser A:415, Asp A:410, Tyr A:417, Arg A:477, Arg A:479, Asp A:483, Arg A:482 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, attractive charge, halogen (fluorine) interaction, π -cation, and π -alkyl interaction |
| 21 | | 3.602 | Arg A:477, Asp A:483, Ser A:480, Arg A:482, Asp A:202 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, attractive charge, π -cation, and π -anion interactions |
| 22 | | 3.347 | Pro A:205, Tyr A:494, Arg A:482, Asp A:483, Ser A:480, Pro A:490 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, attractive charge, alkyl, and π -alkyl interactions |
| 10 | Least active | 1.032 | Asp A:202, Arg A:482, Pro A:490 | Conventional hydrogen bond interactions and carbon-hydrogen interactions |
| 29 | | 0.967 | Arg A:479, Arg A:477, Tyr A:417 | Conventional hydrogen bond interactions, carbon-hydrogen interactions, halogen (fluorine), and π -alkyl interactions |

Asp A:483 amino acid with the nitrogen of piperidine moiety of the PET tracer. These two observations correlate with each other and thus it can be inferred that nitrogen (as piperidine moiety) is essential for good VACHT binding.

True external set predictions

For the analysis of the predictivity of the developed model, we have considered two PET datasets previously used by our group (De et al. 2019; De and Roy 2020) for their VACHT binding predictions (Table S1). Dataset D1 was initially used for amyloid beta imaging and dataset D2 was used for Dopamine (D2) imaging. The prediction quality was further verified using by the application of “Prediction Reliability Indicator” tool (Roy et al. 2018) available from <https://dtclab.webs.com/software-tools>. The prediction tool reported “Good” quality prediction for all the compounds and they were all inside the AD of the model (Supplementary Files S1 and S2). Thus, these compounds can also be considered as potential PET imaging agents for VACHT subject to experimental validation.

Conclusions

The neurotransmitter acetylcholine (ACh) plays a ubiquitous role in cognitive functions including learning and memory with widespread innervation in the cortex, subcortical structures and the cerebellum. Cholinergic receptors, transporters, or enzymes associated with many neurodegenerative diseases, including Alzheimer’s disease (AD) and Parkinson’s disease (PD), are potential imaging targets. In the present study, we have developed a

2D-QSAR model for 19 positron emission tomography (PET) imaging agents targeted against presynaptic vesicular acetylcholine transporter (VACHT). In our work, we aimed to understand the important structural features of the PET imaging agents required for their binding with VACHT. This was done by the feature selection using a Genetic Algorithm followed by the Best Subset Selection method and developing a Partial Least Squares- based 2D QSAR model using the best feature combination. The developed QSAR model showed significant statistical performance and reliability ($R^2 = 0.718$, $Q^2_{(LOO)} = 0.523$, $Q^2_{LMO(25\%)} = 0.598$). Using the features selected in the 2D-QSAR analysis, we have also performed similarity-based chemical read-across predictions and obtained encouraging external validation statistics. From the developed QSAR model, it was found that the presence of nitrogen in the PET tracer molecule potentiates the binding affinity towards the VACHT receptor. This was further confirmed by molecular docking studies where nitrogen in the piperidine moiety produced attractive charge interaction with **Asp A:483** amino acid of VACHT. In the future, this study will help in the prediction of newly developed compounds within the applicability domain of the model targeted toward VACHT.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40203-023-00146-4>.

Acknowledgements PD thanks Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship (file no. ISRM/11(61)/2019). Financial assistance for SERB, New Delhi under the MATRICS scheme (MTR/2019/000008) is thankfully acknowledged.

Author contributions PD: computation, analysis, and initial draft. KR: conceptualization, editing, supervision, and funding.

Funding Funding is provided by ICMR, New Delhi (PD), SERB, New Delhi (KR).

Availability of data and material Some of the data and materials are available in Supplementary Materials of this paper. Additional data are available from the authors on request.

Code availability DTC Lab software tools are available from http://teqip.jdvu.ac.in/QSAR_Tools/.

Declarations

Conflict of interest None.

References

- Akarachantachote N, Saithanu K, Chadcham S, Akarachantachote N, Chadcham S, Saithanu K (2014) Cutoff threshold of variable importance in projection for variable selection. *Int J Pure Appl Math* 94(3):307–322
- Amenta F, Tayebati SK (2008) Pathways of acetylcholine synthesis, transport and release as targets for treatment of adult-onset cognitive dysfunction. *Curr Med Chem* 15(5):488–498
- Banerjee A, Roy K (2022) First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. *Mol Divers* 26(5):2847–2862
- Bergmann K, Tomlinson BE, Blessed G, Gibson PH, Perry RH (1978) Correlation of cholinergic abnormalities with senile plaques and mental test scores in senile dementia. *Br Med J* 2(6150):1457–1459
- Bohnen NI, Albin RL (2011) The cholinergic system and Parkinson disease. *Behav Brain Res* 221(2):564–573
- Chatterjee M, Banerjee A, De P, Gajewicz-Skretna A, Roy K (2022) A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environ Sci Nano* 9(1):189–203
- De P, Roy K (2020) QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting dopamine receptor. *Theor Chem Acc* 139(12):176
- De P, Bhattacharyya D, Roy K (2019) Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease. *Struct Chem* 30(6):2429–2445
- Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 1(1):45–63
- Gajewicz A (2017a) What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. *Nanoscale* 9(24):8435–8448. <https://doi.org/10.1039/C7NR02211E>
- Gajewicz A (2017b) Development of valuable predictive read-across models based on “real-life” (sparse) nanotoxicity data. *Environ Sci Nano* 4(6):1389–1403
- Gajewicz A, Cronin MT, Rasulev B, Leszczynski J, Puzyn T (2014) Novel approach for efficient predictions properties of large pool of nanomaterials based on limited set of species: nano-read-across. *Nanotechnology* 26(1):015701. <https://doi.org/10.1088/0957-4484/26/1/015701>
- Gajewicz A, Jagiello K, Cronin MTD, Leszczynski J, Puzyn T (2017) Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. *Environ Sci Nano* 4(2):346–358. <https://doi.org/10.1039/C6EN00399K>
- Giboureau N, Mat Som I, Boucher-Arnold A, Guilloteau D, Kas-siou M (2012) PET radioligands for the vesicular acetylcholine transporter (VACHT). *Curr Top Med Chem* 10(15):1569–1583
- Gramatica P (2020) Principles of QSAR modeling: comments and suggestions from personal experience. *IJQSPR* 5(3):61–97
- Hampel H, Mesulam MM, Cuellar AC, Farlow MR, Giacobini E, Grossberg GT et al (2018) The cholinergic system in the pathophysiology and treatment of Alzheimer's disease. *Brain* 141(7):1917–1933
- Horsager J, Okkels N, Van Den Berge N, Jacobsen J, Schact A, Munk OL et al (2022) In vivo vesicular acetylcholine transporter density in human peripheral organs: an [18F]FEOBV PET/CT study. *EJNMMI Res* 12(1):1–11
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O et al (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589
- Kar S, Roy K, Leszczynski J (2018) Applicability domain: a step toward confident predictions and decidability for QSAR modeling. In: *Computational toxicology: methods and protocols*, pp 141–169
- Kilbourn MR, Hockley B, Lee L, Sherman P, Quesada C, Frey KA et al (2009) Positron emission tomography imaging of (2R,3R)-5-[18F]fluoroethoxybenzovesamicol in rat and monkey brain: a radioligand for the vesicular acetylcholine transporter. *Nucl Med Biol* 36(5):489–493
- Király P, Kiss R, Kovács D, Ballaj A, Tóth G (2022) The relevance of goodness-of-fit, robustness and prediction validation categories of OECD-QSAR principles with respect to sample size and model type. *Mol Inform* 41(11):2200072
- Kitamura Y, Kozaka T, Miwa D, Uno I, Azim ul MA, Ogawa K et al (2016) Synthesis and evaluation of a new vesamicol analog o-[11C]methyl-trans-decalinvesamicol as a PET ligand for the vesicular acetylcholine transporter. *Ann Nucl Med* 30(2):122–129
- Kovac M, Mavel S, Deuther-Conrad W, Méheux N, Glöckner J, Wenzel B et al (2010) 3D QSAR study, synthesis, and in vitro evaluation of (+)-5-FBVM as potential PET radioligand for the vesicular acetylcholine transporter (VACHT). *Bioorg Med Chem* 18(21):7659–7667
- Kovács D, Király P, Tóth G (2021) Sample-size dependence of validation parameters in linear regression models and in QSAR. *SAR QSAR Environ Res* 32(4):247–268
- Mauri A, Consonni V, Todeschini R (2017) Molecular descriptors. In: *Handb Comput Chem*. pp 2065–93
- Mountjoy CQ (1986) Correlations between neuropathological and neurochemical changes. *Br Med Bull* 42(1):81–85
- Mountjoy CQ, Rossor MN, Iversen LL, Roth M (1984) Correlation of cortical cholinergic and GABA deficits with quantitative neuropathological findings in senile dementia. *Brain* 107(2):507–518
- Prado VF, Roy A, Kolisnyk B, Gros R, Prado MAM (2013) Regulation of cholinergic activity by the vesicular acetylcholine transporter. *Biochem J* 450(2):265–274
- Rác A, Bajusz D, Héberger K (2021) Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules* 26(4):1111
- Reinikainen KJ, Soininen H, Riekkinen PJ (1990) Neurotransmitter changes in alzheimer's disease: implications to diagnostics and therapy. *J Neurosci Res* 27(4):576–586
- Roy K, Kar S, Das RN (2015) *Statistical Methods in QSAR/QSPR. A Prim. QSAR/QSPR Model*. Springer, Cham pp 37–59
- Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure-activity relationship derived predictions for new query chemicals? *ACS Omega* 3(9):11392–11406

- Sukumar N, Prabhu G, Saha P (2014) Applications of genetic algorithms in QSAR/QSPR modeling. In: *Appl Metaheuristics Process Engg.* pp 315–24
- Topliss JG, Edwards RP (1979) Chance factors in studies of quantitative structure-activity relationships. *J Med Chem* 22(10):1238–1244
- Tu Z, Efange SMN, Xu J, Li S, Jones LA, Parsons SM et al (2009) Synthesis and in vitro and in vivo evaluation of 18F-labeled positron emission tomography (PET) ligands for imaging the vesicular acetylcholine transporter. *J Med Chem* 52(5):1358–1369
- Tu Z, Zhang X, Jin H, Yue X, Padakanti PK, Yu L et al (2015) Synthesis and biological characterization of a promising F-18 PET tracer for vesicular acetylcholine transporter. *Bioorg Med Chem* 23(15):4699–4709
- Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G et al (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 50(D1):D439–D444
- Vargas JM, Nielsen S, Cárdenas V, Gonzalez A, Aymat EY et al (2018) Process analytical technology in continuous manufacturing of a commercial pharmaceutical product. *Int J Pharm* 538(1–2):167–178
- Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58(2):109–130
- Wu G, Robertson DH, Brooks CL, Vieth M (2003) Detailed analysis of grid-based molecular docking: a case study of CDOCKER—A CHARMM-based MD docking algorithm. *J Comput Chem* 24(13):1549–1562
- Wu Z, Li D, Meng J, Wang J (2010) Introduction to SIMCA-P and its application. In: *Handbook of Partial Least Squares: Concepts, Methods and Applications.* pp 757–774
- Zea-Ponce Y, Mavel S, Assaad T, Kruse SE, Parsons SM, Emond P et al (2005) Synthesis and in vitro evaluation of new benzovesamicol analogues as potential imaging probes for the vesicular acetylcholine transporter. *Bioorganic Med Chem* 13(3):745–753

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling

Priyanka De¹ · Dhananjay Bhattacharyya² · Kunal Roy¹

Received: 25 November 2019 / Accepted: 19 December 2019 / Published online: 10 January 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Radiosensitizers are aimed to augment tumor cell killing by radiation while having much less effect on normal tissues. Nitroimidazoles and related analogues are efficient radiation sensitivity enhancers, and they particularly work on hypoxic tumor cells. In the current study, we have developed two partial least squares (PLS) regression-based two-dimensional quantitative structure-activity relationship (2D-QSAR) models using a novel class of 84 nitroimidazole compounds to understand their radiosensitization effectiveness ($pC_{1.6}$). Feature selection was done by genetic algorithm along with stepwise regression, while model validation was performed using various stringent validation criteria following the strict rules of OECD guidelines of QSAR validation. The variables included in the models were obtained from Dragon (version 7.0) and simplex representation of molecular structures (SiRMS) (version 4.1.2.270) software. The developed models were robust, externally predictive, and useful tools to predict the radiosensitization effectiveness of nitroimidazole compounds. True external prediction was carried out using a group of six nitroimidazole derivatives and the model reliability was checked using the *Prediction Reliability Indicator* tool (<http://dtclab.webs.com/software-tools>). Furthermore, the developed models will give an insight for development of new radiosensitizers with enhanced radiation sensitivity.

Keywords Radiosensitizers · Radiosensitization effectiveness · QSAR · SiRMS

Introduction

Radiation, surgery, and chemotherapy have been the major approaches of treatment for cancer and malignancies for more than 40 years. Combination therapy including radiation and chemotherapy often termed as chemoradiation has provided promising results in targeting, diagnosis, and treatment of human malignancy. With recent discoveries, newer molecules targeting specific pathophysiology or molecular pathways have come into the forefront. The use of antibodies or

hormones labeled with radionuclides to deliver radiation in the systemic circulation has enlarged the concept of radiosensitizers [1]. Nitroimidazoles have proven to be efficient radiation sensitivity enhancer particularly in hypoxic tumor cells [2]. Hypoxia is a particular pathophysiological condition arising due to inefficient vascularization of tumors, causing an alteration in tumor metabolism [3], and metastasis [4], and is associated with poor diagnosis and resistance to therapeutic agents [5]. Nitroimidazole radiosensitizers are relatively non-toxic molecules, and they replace oxygen in oxidizing radiation-induced DNA free radicals to generate cytotoxic DNA strand breakage [6].

A number of studies performed previously have elaborately explained the role of nitroimidazole derivatives in radiation sensitivity enhancement. 1-Methyl-5-sulfonamide-4-nitroimidazole (MJL-1–191-VII) sensitizes hypoxic cells with its electron affinity, but does not affect the radiosensitivity of aerated cells when added to cells 5 min prior to irradiation [7]. 2-nitroimidazoles like misonidazole and etanidazole has ability to kill hypoxic cell by increasing the cells' radiation sensitivity via radiochemical and biochemical means known as "preincubation effect" [8].

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11224-019-01481-z>) contains supplementary material, which is available to authorized users.

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

¹ Department of Pharmaceutical Technology, Drug Theoretics and Cheminformatics Laboratory, Jadavpur University, Kolkata 700032, India

² Computational Science Division, Saha Institute of Nuclear Physics, Kolkata 700064, India

Molecular modeling studies such as quantitative structure-activity relationships (QSAR) [9] are effective tools in prediction of radiosensitization effectiveness due to lack of data and proper experimental facilities. QSAR studies have found immense applications in the prediction of absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties of drug and other organic biologicals [10–12]. Computational ADMET in combination with *in vivo* and *in vitro* predictions helps in reducing the chances of safety related issues [13]. Many pharmaceutical and chemical industries, commercial software developers, and research groups are developing new QSAR models for ADMET properties utilizing large databases or compilation of published data. A wide number of computational research work describing oral absorption and bioavailability [14, 15], metabolism [16], volume of distribution [17], and enzyme inhibition and induction [18, 19] have been carried out in recent years. The theory of QSAR is applied not only to model activity and toxicity, but also properties of materials in the form of quantitative structure-property relationships (QSPR). Radiosensitization effectiveness can be considered as a property of the nitroimidazole compounds and can thus be subjected to QSAR analysis. Many such property based QSAR models for radiopharmaceuticals have been developed previously by different groups of researchers [20–24]. A properly validated QSAR model could generate radiosensitization data for groups of such related chemicals, and such predictions have the ability to substitute experimental evaluation to an extent.

Feature selection is an essential step for unbiased development of QSAR models. The selection of a reduced pool of descriptors by using multilayered variable selection strategy has proven to be an effective method in QSAR model development and easier data handling. Furthermore, feature selection can reduce the chances of intercorrelation among the descriptors [25]. The current study presents QSAR models for predicting the radiosensitization effectiveness of a dataset of 84 nitroimidazole derivatives. Two-dimensional descriptors calculated from Dragon and SiRMS software were capable enough in developing well-validated and predictive models. Simplex representation of molecular structures (SiRMS) descriptors helped in providing a comprehensive understanding of the basic fragments contributing towards the improvement of radiosensitization effectiveness of the nitroimidazole derivatives. The 2D-QSAR models were developed with an intention of producing statistically robust predictions for radiosensitization effectiveness of nitroimidazole derivatives. Furthermore, we have also predicted some related nitroimidazole compounds to prove the validity of the developed models.

Materials and methods

A data of 86 nitroimidazoles possessing radiosensitizing properties are used for two-dimensional QSAR (2D-QSAR) study

[26]. Radiosensitization capacities of the compounds can be understood by radiosensitization effectiveness, expressed as $C_{1.6}$, which can be represented as the corresponding concentration of a given compound when its sensitization enhancement ratio (SER) accomplishes 1.6. Higher value of $C_{1.6}$ indicates lower bioactivity of radiosensitization effectiveness. For analysis purpose, the source literature had converted the endpoint $C_{1.6}$ to its negative logarithmic scale ($pC_{1.6}$, where $pC_{1.6} = -\log(C_{1.6})$). Two compounds (one radical and one salt) were removed, and the final dataset of 84 compounds is used for model development. The structures of the compounds were drawn in MarvinSketch software (version 14.10.27) [27] with proper aromatization and hydrogen bond addition and saved as MDL.mol, a recommended format for further descriptor calculation.

Descriptor calculation

For developing the first 2D-QSAR model, a pool of 270 descriptors was calculated using Dragon version 7 [28] software. This model was developed using specific classes of descriptors including E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments and molecular property descriptors. Additionally, SiRMS descriptors were calculated using SiRMS (version 4.1.2.270) [29] tool. Simplex representations of molecular structure (SiRMS) descriptors symbolize a class of diverse molecular features developed from 1D to 4D molecular structures. These are tetratomic fragments of different simplex descriptors having predefined chirality, composition, and symmetry [29]. SiRMS descriptors consider both connected and unconnected fragments and also take into account not only the nature of atoms but also their different chemical and physical properties like charge, lipophilicity, electronegativity, atomic refraction, donor/acceptor of hydrogen in the potential Hbond, etc. In our study, we have used 2D SiRMS descriptors only in order to avoid conformational complexity and energy minimization requirements for higher dimensional descriptors and to derive reproducible models. The constant (variance < 0.0001), intercorrelated ($|r| > 0.95$) variables and other incompetent data were removed using an in house software available at <http://dtclab.webs.com/software-tools> before model development.

Dataset splitting

A well-validated QSAR model is the main objective of any QSAR study which can be obtained through proper division of the dataset into training (used for model development) and test (used for model validation) sets. An unbiased external validation with uniform distribution of compounds into training and test sets can be obtained through rational dataset division [30]. For 2D-QSAR modeling, the whole dataset utilized for modeling was divided into training (75%) and test (25%)

sets using modified k -Medoids (Modified k -medoid GUI 1.3) [31, 32] method of dataset division.

Variable selection and QSAR model development

Development of well-validated QSAR models in order to understand the radiosensitization effectiveness of the dataset compounds was the main aim of the present study. Critical evaluation process helped in the selection of statistically significant models. In this study, we have built two QSAR models; a 2D-QSAR model to deduce a relationship between the molecular properties of the nitroimidazoles and their radiosensitization properties. For the model with Dragon descriptors, a pool of 32 descriptors were selected using Genetic Algorithm (GA) [33, 34] modeling implemented in double cross-validation (DCV) [35] tool (version 1.2). Then, the final model was generated using Partial Least Squares (PLS) regression [33, 36] method using descriptors selected from best subset selection (BSS). In case of SiRMS, the number of descriptors generated was large, i.e., about more than ten thousand. Handling of this large data is very much complicated, and so we have applied stepwise regression on the large pool of SiRMS descriptors to find out the essential descriptors contributing to the radiosensitization properties of the dataset. After descriptor thinning, the obtained pool of 300 descriptors was further subjected to multilayered stepwise regression to obtain a manageable number of descriptors and run best subset selection for development of five descriptors models. From the developed models obtained after best subset selection, we have selected one model based on different validation parameters for the test set. Finally, we have run a partial least squares regression (PLS) using SIMCA-P software [37] and developed a PLS model.

Statistical validation metrics

We have rigorously examined the statistical quality of the derived models to judge the robustness in terms of reliability and predictivity measures using various internal and external validation parameters. In the present work we have computed various statistical parameters like determination coefficient R^2 , explained variance R_a^2 , variance ratio (F), and standard error of estimate (s). Since these quality parameters are not sufficient to assess the predictive ability of the model, we have further used additional parameters that could properly validate our predictions. For internal predictions, leave-one-out cross-validation ($Q_{(LOO)}^2$) was reported, and for external predictions, parameters like R_{pred}^2 or Q_{F1}^2, Q_{F2}^2 and concordance correlation coefficient (CCC), were calculated [38]. We have also calculated r_m^2 metrics (i.e., $\overline{r_m^2}$ and Δr_m^2) for both training and test set compounds [39]. We have also validated the models using mean absolute error (MAE) based criteria for

both external and internal validation [40]. This was done since the Q_{ext}^2 based criteria do not always offer the correct indication of the prediction quality because of the influence of the response range as well as the distribution of the values of response in both the training and test set compounds [40].

Results and discussion

Statistically significant 2D-QSAR models using Dragon and simplex (SiRMS) descriptors explaining the chemical features required for good radiosensitization are presented in the following section. The observed versus predicted $pC_{1.6}$ values are plotted for both the models is shown in Fig. 1.

2D-QSAR model using dragon descriptors

$$pC_{1.6} = 3.612 + 0.613 \times (C-035) - 0.285 \times nCp - 1.129 \\ \times (C-043) + 0.068 \times (H-052) - 1.630 \times (C-042) \\ + 0.295 \times nRNHR.$$

$$N_{train} = 63, R^2 = 0.773, R_{adj}^2 = 0.757, Q_{(LOO)}^2 \\ = 0.746, r_{m(Train)}^2 = 0.647, \Delta r_{m(Train)}^2 \\ = 0.173, MAE(Train) = 0.246, SD(Train) \\ = 0.195, RMSEC = 0.30, Quality = Good N_{test} \\ = 21, Q_{F1}^2 = 0.752, Q_{F2}^2 = 0.724, r_{m(Test)}^2 \\ = 0.608, \Delta r_{m(Test)}^2 = 0.216, CCC (Test) \\ : 0.831, MAE(Test) = 0.240, SD(Test) \\ = 0.204, RMSEP = 0.31, Quality = Moderate$$

Model 1

The PLS model with 4 latent variables (LVs) could predict 74.6% variance of the training set and 75.2% of the test set. Important internal and external metrics used to determine the quality of the QSAR model are listed in eq. 1. Mechanistic interpretation of the six descriptors obtained in the model would give us an insight about the structural features of the nitroimidazoles which are likely to influence their radiosensitization effectiveness. The obtained descriptors are C-035, nCp, C-043, H-052, C-042, and nRNHR. The model contains four atom-centered fragments **C-035** (R–CX..X; positive contribution), **C-043** (X–CR..X, negative contribution), **H-052** (hydrogen (H^c) attached to sp³ carbon (C⁰) with one X

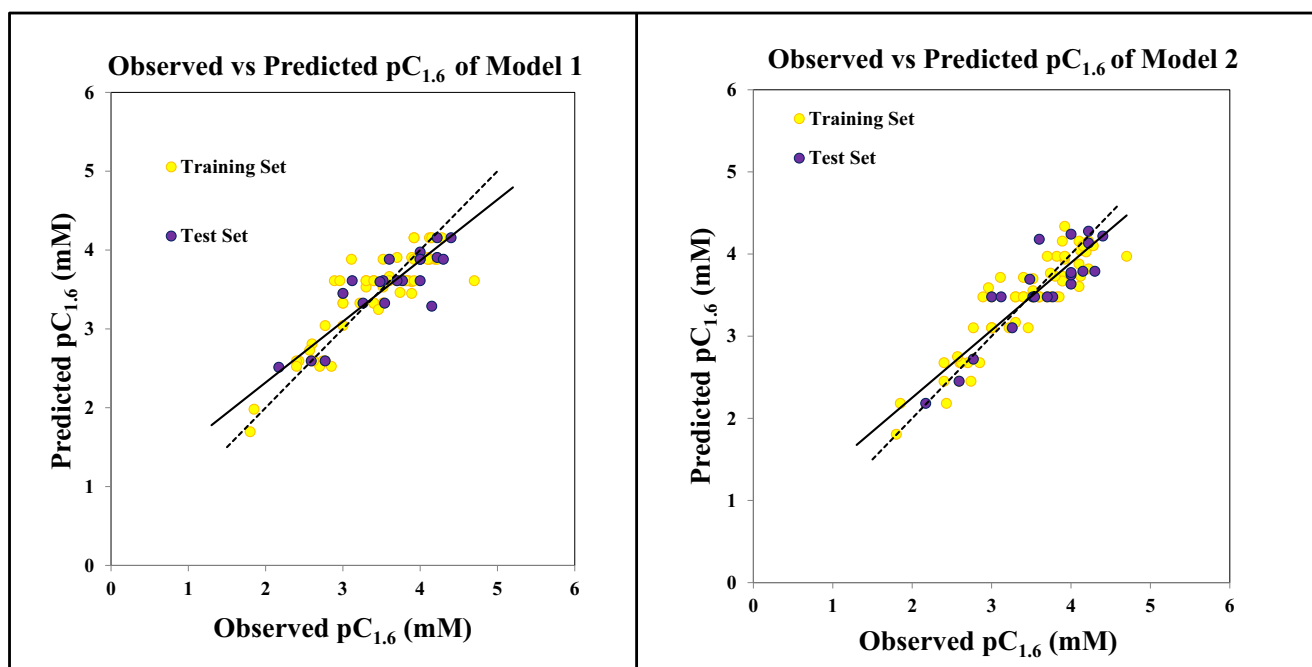


Fig. 1 Scatter plots for observed vs predicted $pC_{1.6}$ values for Model 1 and Model 2

attached to next carbon, “e” represents the formal oxidation number; positive contribution) and **C-042** ($X-CH..X$; negative contribution). These descriptors are further explained with molecular structures from the dataset in Fig. 2. The other two descriptors belonging to functional group counts are nCp (number of terminal primary C (sp^3); negative contribution) and nRNHR (number of secondary amines (aliphatic); positive contribution). The descriptors obtained in the model gives us an idea regarding the vital features essential for better radiosensitization which includes the position of nitro group in the imidazole moiety. Atom-centered fragment-based descriptors like C-042 and C-043 could explain that presence of nitro group at position 4 and position 5 would decrease the $pC_{1.6}$.

The variable importance plot (VIP) [41] analysis gives us a premonition that C-042 and C-035 are the most important descriptors ($VIP > 1$) and contributing mostly towards the radiation enhancement of the compounds. The loading plot gives the relationship between the Y variable ($pC_{1.6}$) and the X variables (descriptors). For interpretation of the loading, the distance from the plot origin is considered, where similar types of descriptors with similar properties are located together. The variables which are far away from the plot origin are considered to have stronger impact on the model. This statement is verified by descriptors C-042 and C-035 which are proved to have higher impact from the VIP values also. The closeness of any descriptor to the Y variable signifies its higher influence on the response. The VIP and loading plot are shown in Fig. 3.

The 2D-QSAR model with Dragon descriptors gives an insight about the importance of the position of nitro group in the nitroimidazole compounds. Also it is found that the

presence of secondary aliphatic amine has significant importance on radiosensitization.

2D-QSAR model using SiRMS descriptors

We have further tried to improve the quality of the model by the use of SiRMS descriptors. The obtained 2D-QSAR model using SiRMS descriptors for radiosensitization effectiveness of nitroimidazoles was highly robust in terms of the statistical parameters as the values of quality metrics were above the recommended threshold as currently practiced [39].

$$pC_{1.6} = 1.381 + 0.802 \times Fr3(elm)/CNN/12s, 13a/ \\ + 0.494 \times SA(chg)/ACDD/12s, 14a, 34s/6 \\ + 0.004 \times SA(chg)/BCCC/14s, 34s/4 \\ + 0.377 \times Fr5(type)/C.3C.ARC.ARC.ARN.AR/12s, 23a, 25a, 45a/ \\ + 0.269 \times Fr(en)/CCCD/15s, 23s, 25s, 34a/$$

$$N_{train} = 63, R^2 = 0.82, R_{adj}^2 = 0.81, Q_{(L00)}^2 = 0.79, r_{m(100)}^2 = 0.70, \Delta r_{m(100)}^2 = 0.14, \\ MAE_{train} = 0.22, SD_{train} = 0.18, RMSEC = 0.26, Quality_{(Train)} = Moderate$$

$$N_{test} = 21, Q_{F1}^2 (or R_{pred}^2) = 0.80, Q_{F2}^2 = 0.77, r_{m(Test)}^2 = 0.70, \Delta r_{m(Test)}^2 = 0.05, \\ CCC(Test) = 0.88, MAE_{test} = 0.23, SD_{test} = 0.16, RMSEP = 0.28, Quality_{(Test)}$$

= Moderate

Model 2

The PLS equation with 3 LVs is able to predict 79% variance of the training set (Q^2) and 80% of the test set (R_{pred}^2). The various internal and external metric values obtained are given in eq. 2. The observed and predicted radiosensitization effectiveness values of the nitroimidazoles are listed in Table S1 in the Supplementary Section.

From VIP (Fig. 4) the descriptors from highest to lowest order of significance are as follows: Fr3(elm)/C_N_N/N/1_2s,1_3a/, S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6,

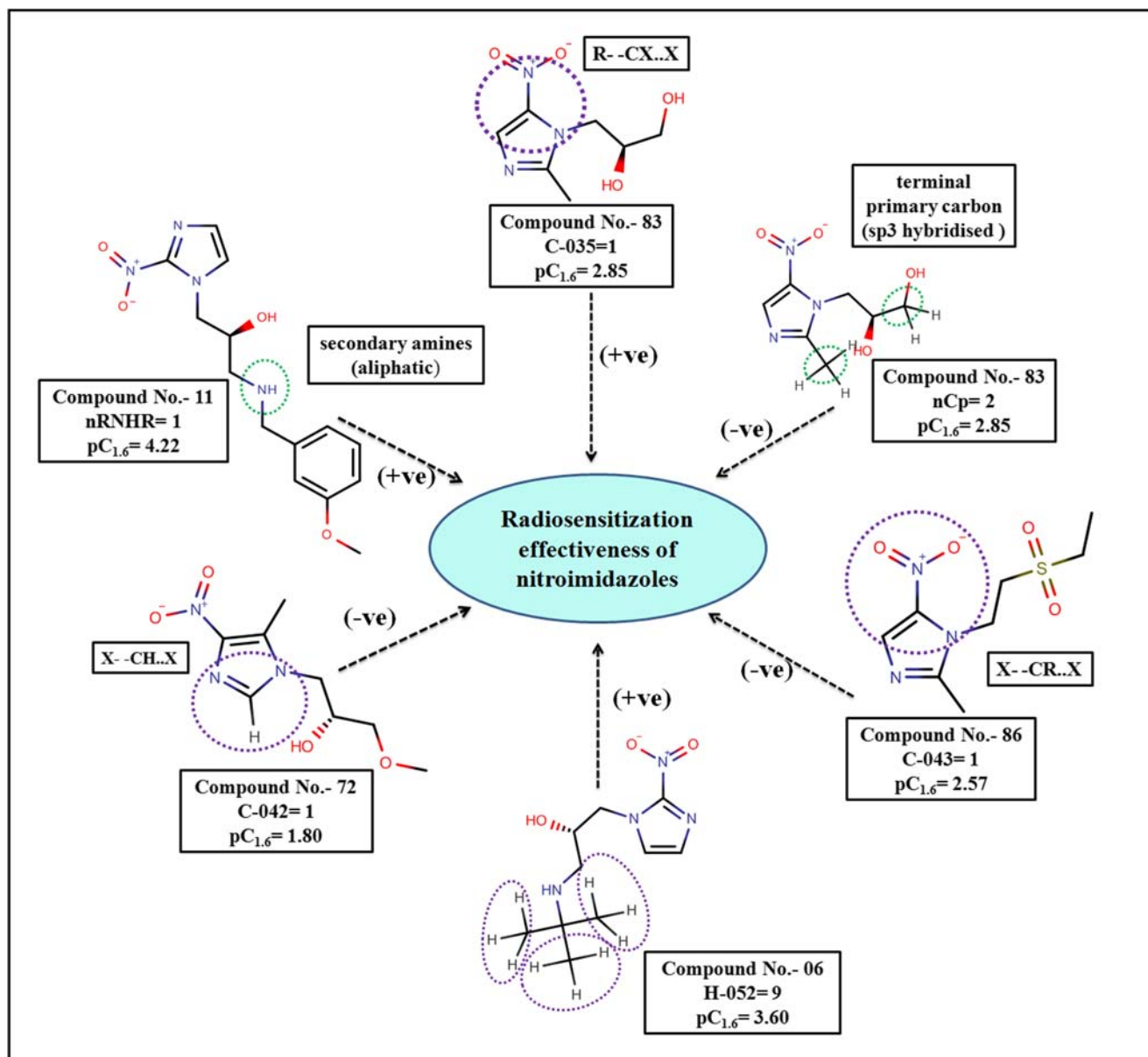


Fig. 2 Descriptor features obtained from Dragon controlling the radiosensitization effectiveness of nitroimidazoles

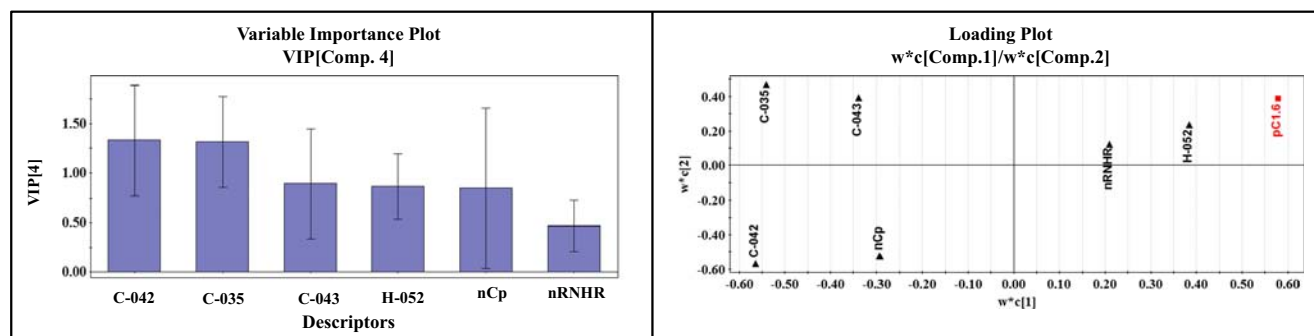


Fig. 3 VIP and loading plot of Model 1

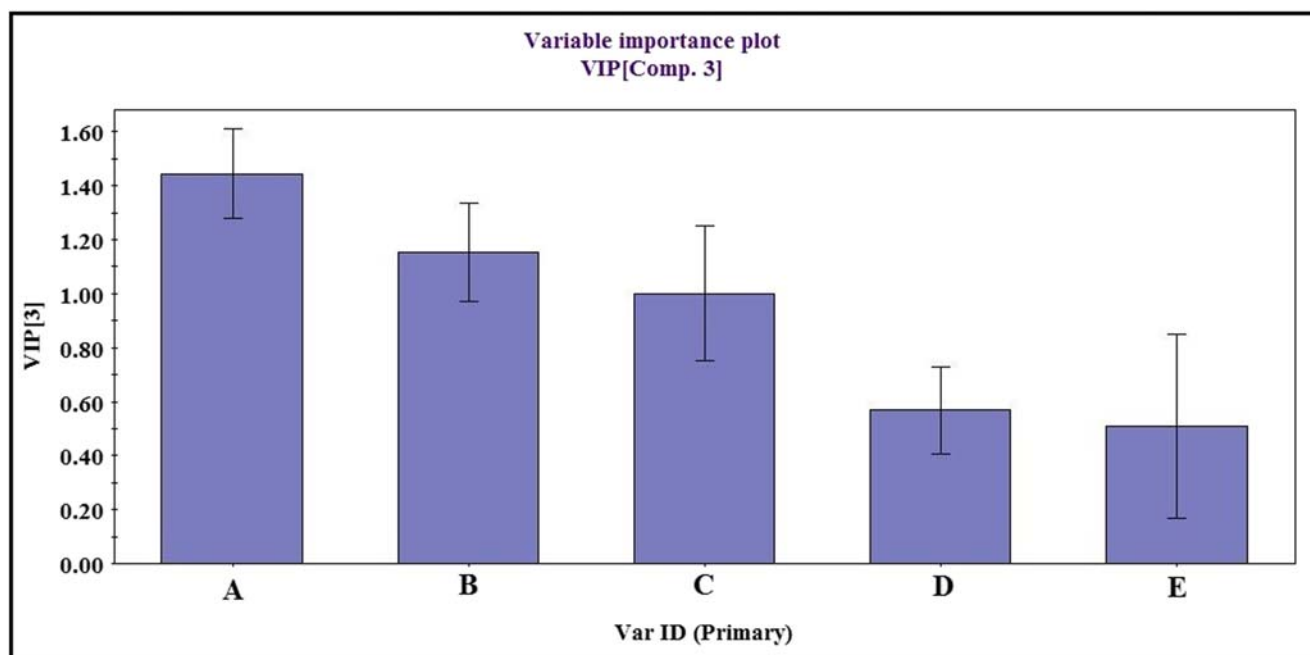


Fig. 4 Variable importance plot of SiRMS model. (A- Fr3(elm)/C_N_N/1_2s,1_3a/, B- S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, C- S_A(chg)/B_C_C_C/1_4s,3_4s/4, D- Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/, E- Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/)

S_A(chg)/B_C_C_C/1_4s,3_4s/4, Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/ and Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/. The loading plot developed using first two components describe the relationship between the X variables and Y variable is shown in Fig. 5.

The highest contributing descriptor is **Fr3(elm)/C_N_N/1_2s,1_3a/** which is a three atomic fragment depicted by N-C=N (**Box 1**). Here, the unsaturation between carbon and nitrogen takes place within the imidazole moiety and the other

nitrogen is from the nitro group. This descriptor has a positive impact on the radiosensitization of the nitroimidazoles thus with higher number of such fragments increases the $pC_{1.6}$ value. All the compounds in the dataset have this particular group once or twice. Compounds with two fragments of this kind has higher $pC_{1.6}$ values as prominently seen in compounds like **63, 47, 11, 53, 46, 51, 43, 45, 10, 22, 54**, etc. Compounds with only one fragment have considerably lower $pC_{1.6}$ values as observed in **72, 71, 82, 78, 75, 86, 80, 81, 85**,

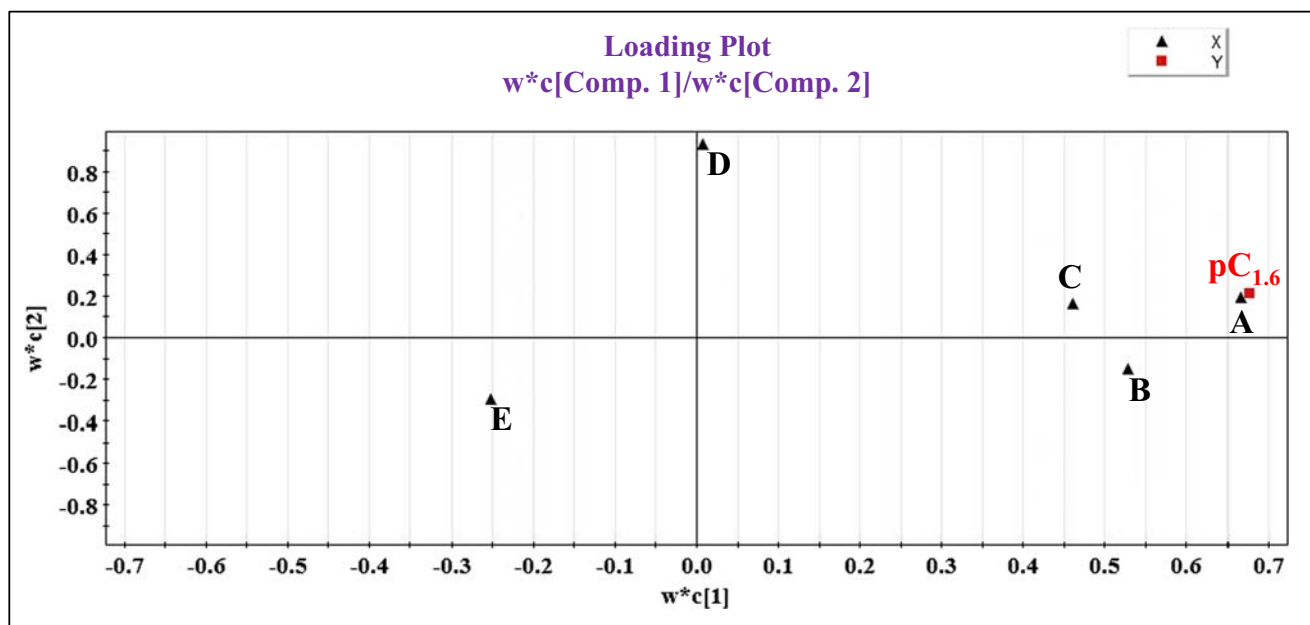


Fig. 5 Loading plot of the SiRMS model. (A - Fr3(elm)/C_N_N/1_2s,1_3a/, B - S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, C - S_A(chg)/B_C_C_C/1_4s,3_4s/4, D- Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/, E- Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/)

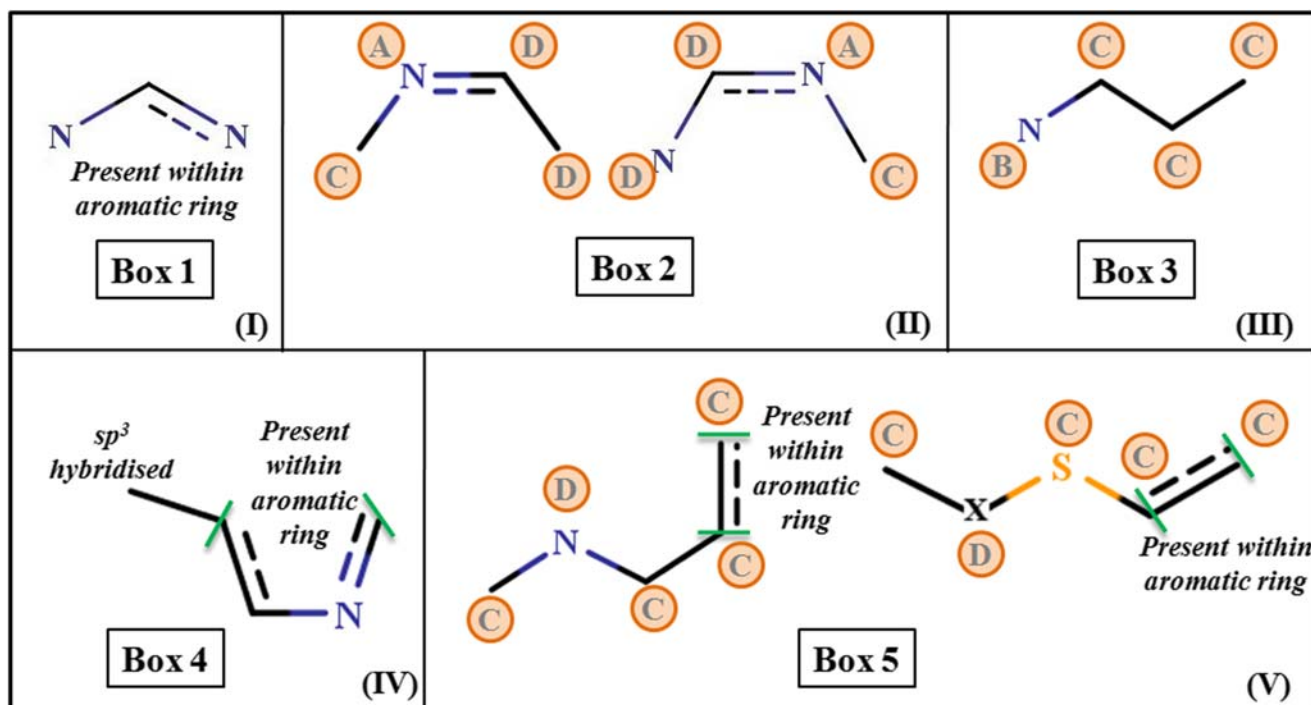


Fig. 6 Simplex representation of molecular structures (SiRMS) fragments appearing in the nitroimidazole dataset. **I**- Fr3(elm)/C_N_N/1_2s,1_3a/, **II**- S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, **III**- S_A(chg)/B_

C_C_C/1_4s,3_4s/4, **IV**- Fr5(type)/C.3_C_AR_C_AR_C_AR_N_AR/1_2s,2_3a,2_5a,4_5a/, **V**- Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/

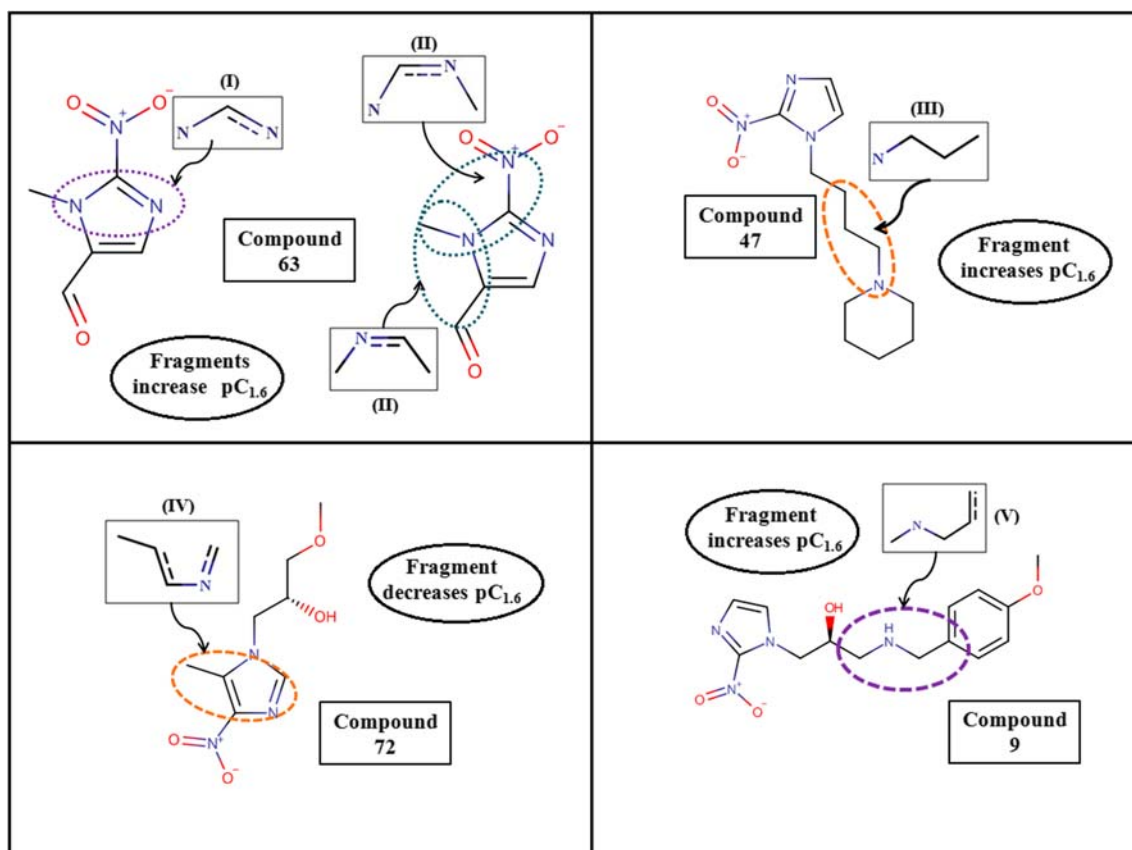


Fig. 7 SiRMS features controlling the increase or decrease in $pC_{1.6}$

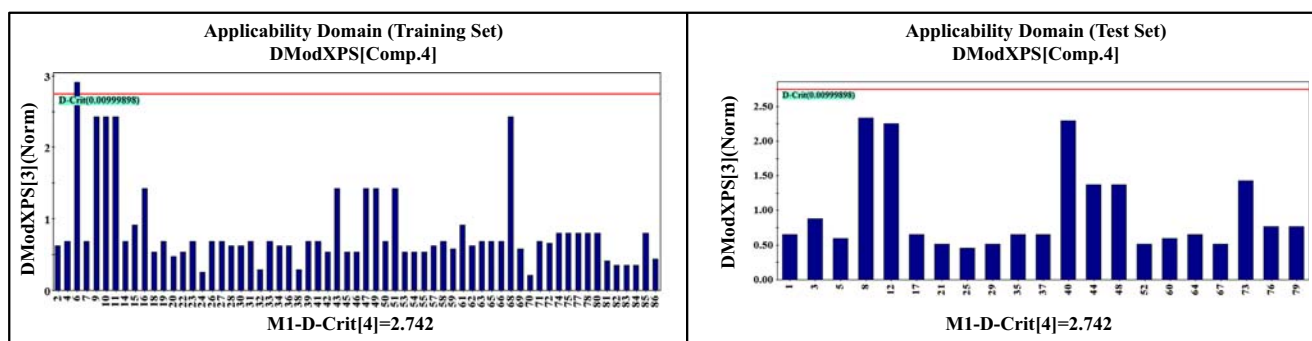


Fig. 8 Applicability Domain of training and test set of Model 1 (with Dragon descriptors) at 99% confidence level

84, etc. Thus, the importance of this fragment leads us to a conclusion that the presence of nitro groups in nitroimidazole should be between N1 and N3 positions of imidazole moiety so as to show better radiosensitization property.

The second important descriptor is **S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6** that represents the partial charge of any of the four atom fragment as given in **Box 2**. The fragment here has two possibilities, one with single nitrogen present within the imidazole moiety and another with two nitrogens (one from the imidazole moiety and another from the nitro group) (given in **Box 2**). Most of the compounds having this fragment have a nitro group attached at position 2 of the imidazole ring. Thus, the position of nitro group plays a vital role in controlling the $pC_{1.6}$ value. This fragment has a positive influence on the radiosensitization effectiveness observed in compounds like **63, 66, 65, 68, 47, 11, and 53**. Compounds which are devoid of these kind of fragments have considerably low $pC_{1.6}$ value (such as in **74, 77, 80, 75, 78, 71, and 72**) (Figs. 6 and 7).

The next important descriptor is **S_A(chg)/B_C_C_C/1_4s,3_4s/4** which represents the partial charge of a four atom fragments as given in **Box 3**. The presence of the mentioned fragment (i.e., three carbon chain attached to nitrogen from a cyclic nucleus) would increase the radiosensitization effectiveness due to the positive influence of the descriptor. Compounds like **47, 51, 43, 46, 55, 49, 54, and 53** have higher

partial charges due to the presence of the mentioned fragments thereby increasing the radiosensitization effectiveness whereas in compounds with no such fragments (like in **71, 72, 82, 78, 75, 80, and 81**) the effect of such charges is not observed thereby the $pC_{1.6}$ value is less.

The next important descriptor **Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/** is a five atomic fragment signifying the following formula: C (sp^3)-C (aromatic)-C (aromatic)-C (aromatic)-N (aromatic). The structure of the possible fragment is given in **Box 4**. The presence of this type of fragment reduces the radiosensitization effectiveness as indicated by the negative influence of the descriptor on $pC_{1.6}$ value. This is well observed in compounds like **72, 59, 57, 61, 69, 62, 41, and 70**. On the other hand, absence of this fragment increases the radiosensitization property as seen in compounds such as **43, 45, 51, 46, 11, 53, 47, and 63**.

The descriptor with the least significance is **Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/** which denotes the electronegativity of the compound due to the presence of a four atomic fragment given in **Box 5**. The positive contribution suggested that the presence of any of the given fragments will influence the electronegativity of the compound thereby increasing the $pC_{1.6}$ value. Compounds **9, 10, and 11** have been reported to have two such fragments and thereby increase the radiosensitization effectiveness.

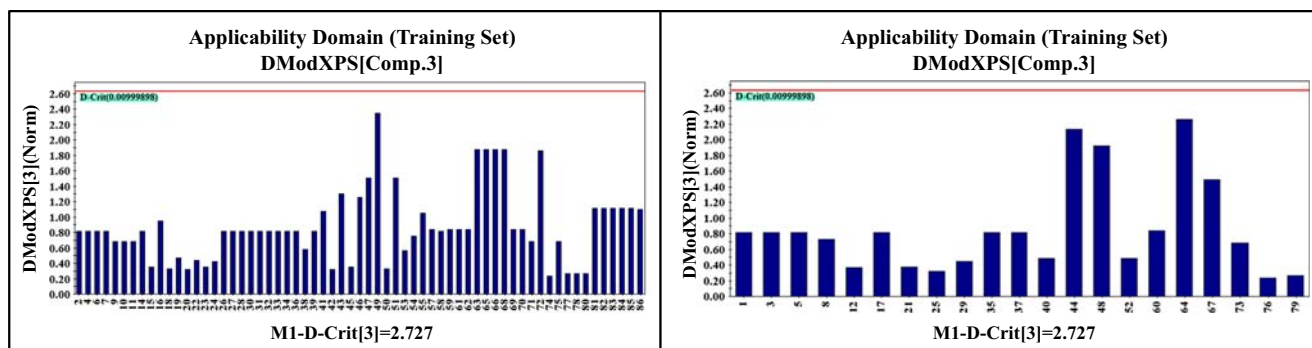


Fig. 9 Applicability Domain of training and test set of Model 2 (with SiRMS descriptor) at 99% confidence level

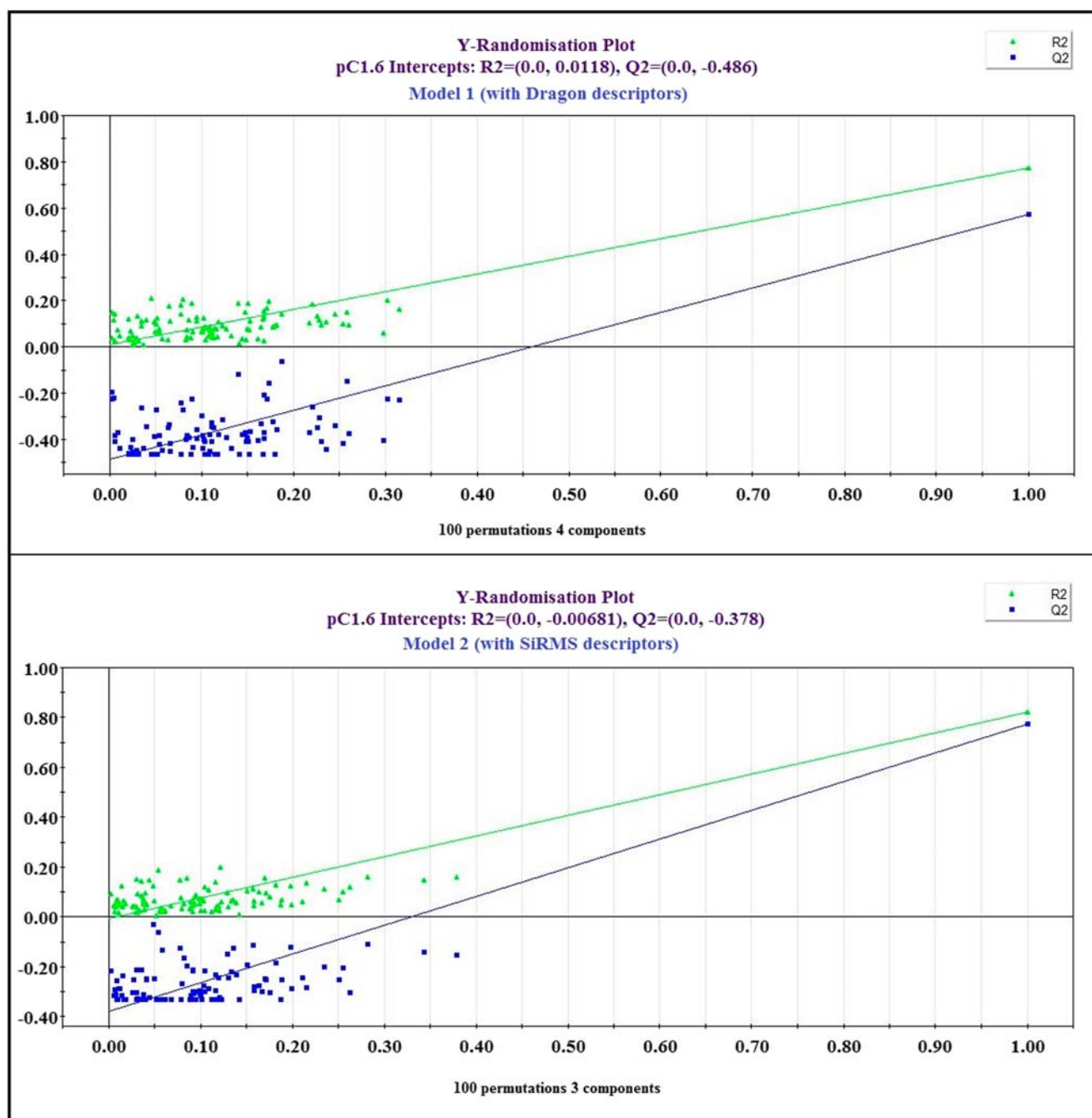


Fig. 10 Y-randomization plots for Model 1 and Model 2

Applicability domain assessment

The prediction reliability of both the 2D-QSAR models is determined by the applicability domain (AD) assessment. AD gives a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable [42]. AD assessment for both the models was performed using DModX (distance to model in the X-space) approach at 99% confidence level (Figs. 8 and 9). Both the models displayed good coverage of domain of

applicability showing maximum number of compounds in the AD (only compound 6 is outside the AD in case of Model 1, i.e., 2D-QSAR model with Dragon descriptors). There were no outliers obtained from the test set for both the models. We have also performed AD assessment at 95% confidence level for both the models as given in the Supplementary Materials (Figures S1 and S2) and found that in this case three compounds in the test set were outside AD for the model with Dragon descriptors and two compounds in the test set for the model with SiRMS descriptors.

Table 1 External dataset and their predicted $pC_{1.6}$ values

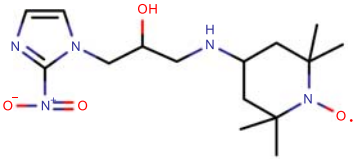
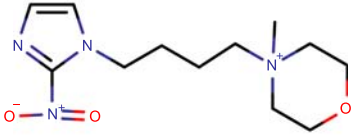
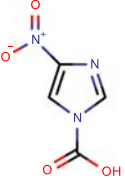
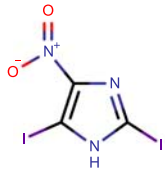
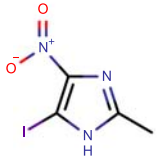
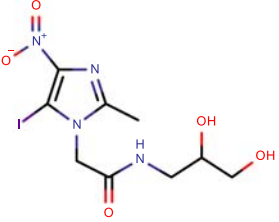
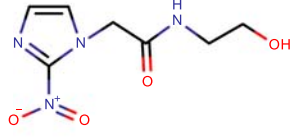
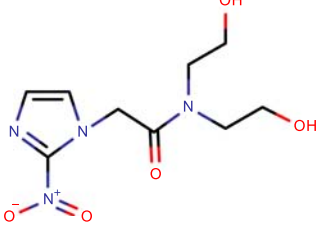
| Compound Number | Structure | Observed $pC_{1.6}$ | Predicted $pC_{1.6}$ using model 1 | Predicted $pC_{1.6}$ using model 2 | Reference |
|-----------------|---|---------------------|------------------------------------|------------------------------------|-----------|
| P-1 |  | 4.05 | 3.58 | 3.67 | [26] |
| P-2 |  | 2.89 | 3.88 | 3.82 | [26] |
| P-3 |  | - | 1.98 | 2.18 | [44] |
| P-4 |  | - | 4.22 | 2.18 | [44] |
| P-5 |  | - | 2.81 | 2.18 | [44] |
| P-6 |  | - | 2.53 | 2.18 | [44] |
| P-7 |  | - | 3.33 | 3.48 | [45] |
| P-8 |  | - | 3.04 | 3.48 | [45] |

Table 2 Prediction quality [46] for the true external dataset

| Compound number | Prediction status of model with Dragon descriptors | | | Prediction status of model with SiRMS descriptors | | |
|-----------------|--|--------------------|--|---|--------------------|--|
| | Composite score | Prediction quality | AD status (using standardization approach) | Composite score | Prediction quality | AD status (using standardization approach) |
| P-1 | 3 | Good | Outside AD | 3 | Good | In |
| P-2 | 3 | Good | In | 3 | Good | In |
| P-3 | 2 | Moderate | In | 3 | Good | In |
| P-4 | 3 | Good | In | 3 | Good | In |
| P-5 | 3 | Good | In | 3 | Good | In |
| P-6 | 3 | Good | Outside AD | 3 | Good | In |
| P-7 | 3 | Good | In | 3 | Good | In |
| P-8 | 3 | Good | In | 3 | Good | In |

Y-randomization

Y-randomization plot analysis helps to understand the statistical significance of the model. The randomization plot confirms that the model is not the result of any chance correlation [43]. In this process, a number of models are generated by shuffling different combinations of X or Y variables (here Y variable only) based on the fit of the reordered model. In our work, we have used 100 permutations for random model generation. A model with no chance correlation would show very poor statistics for the randomized models, i.e., R_Y^2 intercept should not exceed 0.3 and Q_Y^2 intercept should not exceed 0.05 [43]. The randomization plots given in Fig. S8 show that the developed models are non-random and robust (as understood from their R_Y^2 and Q_Y^2 values) and are suitable for prediction of the radiosensitization effectiveness within the AD of the model (Fig. 10).

True external predictions

Prediction of responses for external compounds based on their molecular features using chemometric methods can reduce the experiment costs and animal handling. To verify the predictive power of both the models, we have used a set of eight nitroimidazole derivatives (Table 1) as an external prediction set [26, 44, 45]. The original dataset in the source literature

contain 86 nitroimidazoles but we have removed two of them and used the rest 84 for modeling. These two compounds are now used for prediction purpose. In addition to this, the domain of applicability and their predictive reliability are analyzed using **Prediction Reliability Indicator** tool [46]. The prediction quality and domain of applicability are given in Table 2. From the prediction status, it can be inferred that model with fragment-based SiRMS descriptors provides better prediction than model with dragon descriptors.

Comparison with the previously published research

In the previously published research by Long and Liu (2010) [26], the authors developed MLR and projection pursuit regression (PPR) [47–49] models using complex descriptors such as geometrical, electrostatic, and quantum chemical descriptors. The models developed by us cannot be critically compared to the previously published since the calibration and validation set compositions are different. However, it can be found that our MLR model developed using SiRMS descriptor is better in terms of both training and test set validation metrics if we consider their MLR model (Table 3). Also the current model comes with an added advantage of presence of lower number of simple descriptors and non-requirement of conformation analysis or energy minimization prior to their calculation. Furthermore, the PPR based model reported in the

Table 3 Comparison of the current SiRMS model with previously developed MLR model

| Model | Total no. of compounds used | No. of compounds in the training set | No. of compounds in the test set | Descriptor type | No. of descriptors in final model | Training set | | | Test set | |
|--------------------|-----------------------------|--------------------------------------|----------------------------------|---------------------------|-----------------------------------|--------------|-------|-------|------------|-------|
| | | | | | | R^2 | Q^2 | RMSEC | Q_{F1}^2 | RMSEP |
| Current study | 84 | 63 | 21 | 2D (fragment-based SiRMS) | 5 (3 LVs) | 0.82 | 0.79 | 0.26 | 0.80 | 0.28 |
| Long and Liu, 2010 | 86 | 68 | 18 | 3D | 6 | 0.80 | 0.76 | 0.28 | 0.76 | 0.28 |

previous study is derived from a more complicated process which uses projection based approach to convert high dimensional data to lower dimension. Moreover, 3D descriptors were used in the previous work. MLR or PLS models are more straight-forward and reproducible as used in the current work. In addition, 2D descriptors used in the present work are easy to compute and do not need any conformation analysis or energy minimization process.

Conclusion

This study targets for the development of fragment-based 2D-QSAR models for predicting radiosensitization of nitroimidazole derivatives. The simplex descriptors give an insight about the fragments and their proper position in the nitroimidazole ring that enhance or decline the radiosensitization effectiveness. Also reduction in the large data pool by using multilayered variable selection is shown for better handling of a large pool of descriptors and removing chances of intercorrelation among them. Further, the newly developed models were used for prediction of eight external compounds and their prediction reliability was checked.

Funding information PD thanks Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship. KR thanks BRNS, Department of Atomic Energy, Govt. of India, for a Major Research Project (36(3)/14/08/2017-BRNS).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Kvols LK (2005) Radiation sensitizers: a selective review of molecules targeting DNA and non-DNA targets. *J Nucl Med* 46:187S
- Bonnet M, Hong CR, Gu Y, Anderson RF, Wilson WR, Pruijn FB, Wang J, Hicks KO, Hay MP (2014) Novel nitroimidazolealkylsulfonamides as hypoxic cell radiosensitizers. *Bioorg Med Chem* 22:2123–2132
- Caims RA, Harris IS, Mak TW (2011) Regulation of cancer cell metabolism. *Nat Rev Cancer* 11:85
- Chang Q, Jurisica I, Do T, Hedley DW (2011) Hypoxia predicts aggressive growth and spontaneous metastasis formation from orthotopically grown primary xenografts of human pancreatic cancer. *Cancer Res* 71:3110–3120
- Rohwer N, Cramer T (2011) Hypoxia-mediated drug resistance: novel insights on the functional interaction of HIFs and cell death pathways. *Drug Resist Updat* 14:191–201
- Wilson WR, Hay MP (2011) Targeting hypoxia in cancer therapy. *Nat Rev Cancer* 11:393
- Astor M, Hall EJ, Martin J, Flynn M, Biaglow J, Parham JC (1982) Radiosensitizing and cytotoxic properties of ortho-substituted 4- and 5-nitroimidazoles: role of NPSH reactivity. *Int J Radiat Oncol Biol Phys* 8:409–413
- Koch CJ, Skov KA (1994) Enhanced radiation-sensitivity by preincubation with nitroimidazoles: effect of glutathione depletion. *Int J Radiat Oncol Biol Phys* 29:345–349
- Roy K (2018) Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. *J Indian Chem Soc* 95:1497–1502
- Hansch C, Leo A, Hoekman DH (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. American Chemical Society Washington, DC
- Hansch C, Leo A, Mekapati SB, Kurup A (2004) Qsar and Adme. *Bioorg Med Chem* 12:3391–3400
- Klein C, Kaiser D, Kopp S, Chiba P, Ecker GF (2002) Similarity based SAR (SIBAR) as tool for early ADME profiling. *J Comput Aided Mol Des* 16:785–793
- Merlot C (2010) Computational toxicology—a tool for early safety evaluation. *Drug Discov Today* 15:16–22
- Xu X, Zhang W, Huang C, Li Y, Yu H, Wang Y, Duan J, Ling Y (2012) A novel chemometric method for the prediction of human oral bioavailability. *Int J Mol Sci* 13:6964–6982
- Yoshida F, Topliss JG (2000) QSAR model for drug human oral bioavailability. *J Med Chem* 43:2575–2585
- Roy H, Nandi S (2019) In silico modeling in drug metabolism and interaction: current strategies of lead discovery. Bentham Science Publishers, Sharjah
- Simeon S, Montanari D, Gleeson MP (2019) Investigation of factors affecting the performance of in silico volume distribution QSAR models for human, rat, mouse, dog & monkey. *Mol Inform* 38:1900059
- Halder AK, Cordeiro M (2019) Development of multi-target chemometric models for the inhibition of class I PI3K enzyme isoforms: a case study using QSAR-Co tool. *Int J Mol Sci* 20:4191
- Dmitriev AV, Lagunin AA, Karasev DI, Rudik AV, Pogodin PV, Filimonov DA, Porokov VV (2019) Prediction of drug-drug interactions related to inhibition or induction of drug-metabolizing enzymes. *Curr Top Med Chem* 19:319–336
- Salahinejad M (2015) Quantitative structure property relationships on formation constants of radiometals for radiopharmaceuticals applications. *J Radioanal Nucl Chem* 303:671–680
- Singh S, Ojha H, Tiwari AK, Kumar N, Singh B, Mishra AK (2010) Design, synthesis, and in vitro antiproliferative activity of benzimidazole analogues for radiopharmaceutical efficacy. *Cancer Biother Radiopharm* 25:245–250
- Yoshizuka K, Pietzsch H-J, Seifert S, Stephan H (2013) Quantitative structure property relationship of logP for radiopharmaceutical technetium and rhenium complexes by using molecular dynamics calculations. *Solvent Extr Res Dev, Jpn* 20:15–27
- Santos L, Pilar Cornago M, Izquierdo MC, Consuelo Lopez-Zumel M, Smeyers YG (1989) Electron affinity/radiosensitizing activity relationship for quaternary 5-nitroimidazole derivatives. *Quantum chemical QSAR. Quant Struct-Act Rel* 8:214–217
- Wardman P, Clarke ED (1987) Redox properties and rate constants in free-radical mediated damage. *Br J Cancer Suppl* 8:172
- De P, Bhattacharyya D, Roy K (2019) Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease. *Struct Chem* 30:2429–2445
- Long W, Liu P (2010) Quantitative structure activity relationship modeling for predicting radiosensitization effectiveness of nitroimidazole compounds. *J Radiat Res* 51:563–572
- MarvinSketch software, <https://www.chemaxon.com>. Accessed 26 Aug 2019
- Dragon version 7, Kodesrl, Milan, Italy, 2016; software available at <http://www.taletе.mi.it/index.htm>. Accessed 26Aug 2019
- Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) Hierarchic

- system of QSAR models (1D–4D) on the base of simplex representation of molecular structure. *J Mol Model* 11:457–467
30. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* 17:241–253
 31. Park H-S, Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. *Expert Syst Appl* 36:3336–3341
 32. Drug Theoretics and Cheminformatics (DTC) laboratory software tools <https://dtclab.webs.com/software-tools> Accessed 28 Aug 2019
 33. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opin Drug Discov* 13:1075–1089
 34. Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, Cornwall, Great Britain
 35. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intell Lab Syst* 159:108–126
 36. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58:109–130
 37. U. Simca-P, 10.0, info@umetrics.com, www.umetrics.com, Umea, Sweden, 2002. Accessed 30 Aug 2019
 38. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen* 14:450–474
 39. Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm^2 metrics for validation of QSPR models. *Chemom Intell Lab Syst* 107:194–205
 40. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33
 41. Akarachantachote N, Chadcham S, Saithanu K (2014) Cutoff threshold of variable importance in projection for variable selection. *Int J Pure Appl Math* 94:307–322
 42. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 1:45–63
 43. Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47:2345–2357
 44. Krause W, Jordan A, Scholz R, Jimenez J-LM (2005) Iodinated nitroimidazoles as radiosensitizers. *Anticancer Res* 25:2145–2151
 45. Brown JM, Ning YY, Brown DM, Lee WW (1981) SR-2508: a 2-nitroimidazole amide which should be superior to misonidazole as a radiosensitizer for clinical use. *Int J Radiat Oncol Biol Phys* 7:695–703
 46. Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? *ACS Omega* 3:11392–11406
 47. Friedman JH, Stuetzle W (1981) Projection pursuit regression. *J Am Stat Assoc* 76:817–823
 48. Du Y, Liang Y, Yun D (2002) Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *J Chem Inf Comput Sci* 42:1283–1292
 49. Liu H, Yao X, Liu M, Hu Z, Fan B (2007) Prediction of gas-phase reduced ion mobility constants (K_0) based on the multiple linear regression and projection pursuit regression. *Talanta* 71:258–263

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: application of small dataset modeling

Priyanka De¹ · Kunal Roy¹

Received: 1 December 2020 / Accepted: 15 January 2021 / Published online: 25 January 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

In recent years, hypoxic cell radiosensitizers have evolved as potential molecules in the diagnosis of cancer and in clinical radiotherapy. Nitroimidazole and its sulfonamide analogues are effective radiosensitizers working on hypoxic tumor cells. The application of QSAR modeling technique has paved an easier way for the prediction of newly developed compounds. In the present study, we have used 21 nitroimidazole sulfonamide analogues to develop 2D quantitative structure-activity relationship (QSAR) models and determine their structural features essential for two radiosensitization properties, viz., sensitizer enhancement ratio and survival ratio. The models were developed using the small dataset modeler software (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/), and model validation was performed using various stringent validation criteria. The developed models are robust, predictive, and should be useful tools to predict the radiosensitization of nitroimidazole sulfonamides. Furthermore, we have used the “prediction reliability indicator” tool to check the predictive ability of the developed models using 14 external nitroimidazole sulfonamide derivatives. We have also developed quantitative structure-activity-activity relationship (QSAAR) models for the two endpoints.

Keywords QSAR · QSAAR · Nitroimidazole sulfonamide · Radiosensitizer · Small dataset modeler

Introduction

Hypoxia is a principal component of the tumor microenvironment, which is considered to be the pivotal cause of clinical radioresistance and local failure. Oxygen is considered as the best radiosensitizer by far; however, metabolic consumption of oxygen limits its diffusion into hypoxic tumor cells [1]. Hypoxia has a chief role in cancer progression manipulating angiogenesis [2], vasculogenesis [3], and activation of a glycolytic shift in metabolism [4], invasion enhancement, and metastasis [5]. Radiation therapy is an anchoring treatment for many types of cancer; however, there is a great challenge to augment radiation damage to the tumor tissues and reduce side effects to healthy tissues. Radiosensitizers are promising agents in controlling hypoxia by enhancing tumor tissue injury through accelerating DNA damage and producing free radicals [6].

Oxygen-mimetic radiosensitizers are potential agents in controlling radiation damage in hypoxic tumor cells. Nitroheterocyclic compounds such as nitroimidazoles have been evaluated as oxygen-mimetic agents where electron-rich nitro group is intended to react with DNA radicals produced by ionizing radiation in a similar fashion like oxygen does [6, 7]. DNA and nitro group adduct leads to DNA strand breaks and subsequent cellular apoptosis or lysis. Enhanced radiosensitization after prolonged exposure of cells to misonidazole was identified by Hall et al. [8]. However, this was restricted by delayed peripheral neuropathies when combined with fractionated radiotherapy [9]. Ro 03-8799 (pimonidazole) and SR 2508 (etanidazole) were used in combination as cell radiosensitizers in the treatment of high-grade gliomas. It was found that Ro 03-8799 is distributed extensively in the central nervous system, and SR 2508 could achieve high tumor concentrations when the blood-brain barrier is compromised [10]. Yahiro et al. studied effects of the radiosensitizer doranidazole (PR-350) on the radioresponse of murine and human tumor cells in vitro and in vivo and observed that the amount of radiosensitization of tumors induced by doranidazole is dependent on the oxygenation status of the tumors [11]. A 5-nitroimidazole derivative, nimorazole, has shown similar radiosensitization properties to misonidazole at clinically

✉ Kunal Roy
kunal.roy@jadavpuruniversity.in; kunalroy_in@yahoo.com

¹ Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

acceptable dose levels. It is clinically used in head and neck cancer along with fractionated radiotherapy (FRT) [12].

Recently, a wide range of nitroimidazole sulfonamides has been identified as potential radiosensitizers against hypoxic cancer cells [13, 14]. These sulfonamides have been considered as hypoxia-selective cytotoxins and radiosensitizers, and their variation in side chains noticeably influence the physico-chemical properties of the analogues. The compounds might have lowered aqueous solubility and raised the electron affinity of the nitroimidazole group.

Computational approaches such as quantitative structure-activity/property relationships (QSAR/QSPR) [15] are effective tools in prediction of radiosensitization properties when experimental data is scarce. The method allows virtual screening of drug libraries to find suitable drug-target for a particular disease. QSAR finds an immense application in the prediction of ADMET (absorption, distribution, metabolism, elimination, and toxicity) properties of drugs and other biologicals [16, 17]. A large number of researches have been carried out with the hope to do some predictions of the ADMET properties using the structural features of the molecules. QSAR/QSPR modeling is one such important approach where data derived from their activity profiles and their different structural features (quantitative molecular descriptors) are used [18]. Radiosensitization is a property of nitroimidazole and nitroimidazole sulfonamide derivatives and can thus be subjected to QSAR analysis. A well-validated QSAR model could evaluate and generate radiosensitization data for such related compounds when experimental data is not available.

The present study explores the features essential to show radiosensitization properties by nitroimidazole sulfonamide derivatives using QSAR and quantitative structure activity-activity relationship (QSAAR) modeling [19]. Two dimensional (2D) descriptors obtained from Dragon and SiRMS software were utilized during the development of well-validated models. A small dataset of nitroimidazole sulfonamides is used for modeling in the current study where splitting of the dataset into training and test sets would cause loss of chemical information leading to unreliable models. Thus, a “small dataset modeling” approach has been adopted using the whole dataset [20], and the developed models were subjected to leave-many-out cross-validation. Furthermore, a group of nitroimidazole sulfonamides has been predicted to prove the validity of the developed models.

Materials and methods

Dataset

In vitro radiosensitization data of selected compounds involving sensitizer enhancement ratio (drug SER) and survival ratio (drug SR) was obtained from a previously published research

work [21]. A dataset of 21 compounds given in Table 1 was selected for 2D QSAR modeling. Sensitizer enhancement ratio (SER) can be defined as the ratio of radiation dose for 1% survival without or with the drug in a condition where HCT116 cells (human colorectal carcinoma cell line) were exposed to the drug at 6–29 Gy radiation for 1 h. Survival ratio can be explained using the following expression: “SR = (cell survival with radiation)/(cell survival with drug and with radiation) interpolated from the radiation dose response curves at 15 Gy.” During modeling, the drug SER values were used as provided in the original article but drug SR values were converted into their logarithmic form (logSR) for analysis. The compounds were drawn in MarvinSketch software (version 14.10.27) [22] with hydrogen bond addition and proper aromatization and saved as MDL.mol, a suggested format for further descriptor calculation.

Molecular descriptors

The molecular descriptor is the “final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment” [23]. A selected class of 356 2D molecular descriptors was calculated from Dragon version 7 [24] software. These comprised E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments, and molecular property descriptors. Intercorrelated ($|r| > 0.95$) and constant (variance < 0.0001) variables and other incompetent data were removed using a software available at <http://dtclab.webs.com/software-tools> prior to model development. This resulted in 224 Dragon descriptors which were used for modeling. Furthermore, SiRMS descriptors were calculated using SiRMS (version 4.1.2.270) [25] tool and used along with Dragon descriptors during modeling. Simplex representations of molecular structure (SiRMS) descriptors are a class of molecular descriptors developed from 1D to 4D molecular structures involving tetratomic fragments of different simplex descriptors having predefined chirality, composition, and symmetry [25].

Model development: application of small dataset modeler

Before development of a QSAR model, the dataset is generally divided into a training set (calibration) and a test set (validation). Furthermore, a double cross-validation method [26] of model development involves two nested cross-validation loops: internal (inner) and external (outer) cross-validation loops. In the outer loop, the data points are segregated into two subsets, i.e., training and test sets. The training set is further employed in the inner loop for model building and selection purpose. The test set has the sole purpose of model

Table 1 Dataset of 21 compounds used for modeling

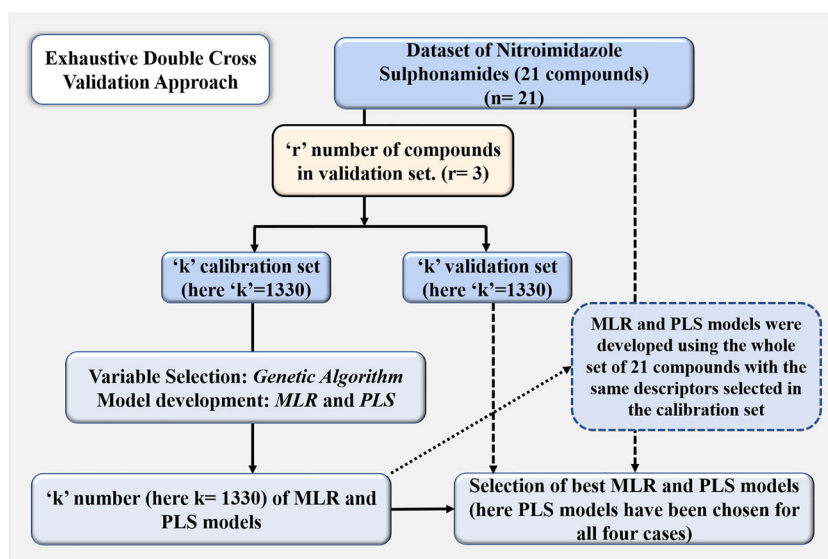
| Serial number | Compound number | Structure (SMILES) | Drug SER | Log drug SR |
|---------------|-----------------|--|----------|-------------|
| 1 | 1 | <chem>c1(n(ccn1)CC(COC)O)[N+](=O)[O-]</chem> | 1.4 | 0.833 |
| 2 | 2 | <chem>c1(n(ccn1)CC(=O)NCCO)[N+](=O)[O-]</chem> | 1.339 | 0.663 |
| 3 | 4 | <chem>c1n(c(cn1)[N+](=O)[O-])CCN1CCOCC1</chem> | 1.8 | 1.652 |
| 4 | 6 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCOC)[N+](=O)[O-]</chem> | 1.2 | 0.462 |
| 5 | 7 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCO)[N+](=O)[O-]</chem> | 1.11 | 0.255 |
| 6 | 8 | <chem>c1(n(ccn1)CS(=O)(=O)NCCCN1CCOCC1)[N+](=O)[O-]</chem> | 1.28 | 0.591 |
| 7 | 12 | <chem>c1(n(ccn1)CS(=O)(=O)NN1CCOCC1)[N+](=O)[O-]</chem> | 1.11 | 0.301 |
| 8 | 14 | <chem>c1(n(ccn1)CCS(=O)(=O)NCCCO)[N+](=O)[O-]</chem> | 1.27 | 0.623 |
| 9 | 15 | <chem>c1(n(ccn1)CCS(=O)(=O)NCCCN1CCOCC1)[N+](=O)[O-]</chem> | 1.357 | 0.699 |
| 10 | 16 | <chem>c1n(cc(n1)[N+](=O)[O-])CS(=O)(=O)NCCCOC</chem> | 1.105 | 0.114 |
| 11 | 19 | <chem>c1n(c(cn1)[N+](=O)[O-])CS(=O)(=O)NCCCO</chem> | 1.81 | 2.057 |
| 12 | 21 | <chem>c1n(c(cn1)[N+](=O)[O-])CS(=O)(=O)NCCCN1CCOCC1</chem> | 1.43 | 0.914 |
| 13 | 22 | <chem>c1n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC</chem> | 1.56 | 1.415 |
| 14 | 24 | <chem>c1n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCO</chem> | 1.81 | 2.212 |
| 15 | 26 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCOC)C</chem> | 1.34 | 0.681 |
| 16 | 28 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCO)C</chem> | 1.176 | 0.208 |
| 17 | 30 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN1CCOCC1)C</chem> | 1.68 | 1.447 |
| 18 | 31 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN(C)C)C</chem> | 1.57 | 1.173 |
| 19 | 34 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN1CCCC1)C</chem> | 1.54 | 1.134 |
| 20 | 35 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NCCCN1CCCC1)C</chem> | 1.71 | 1.380 |
| 21 | 38 | <chem>c1(n(c(cn1)[N+](=O)[O-])CCS(=O)(=O)NN1CCC(CC1)N(C)C)C</chem> | 1.67 | 1.398 |

validation. However, the present study deals with a small dataset containing a limited number of data points (21 compounds), and splitting of this dataset into training and test sets is not desirable. Small dataset modeling (http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/) involves the DCV method of modeling for small datasets without dividing the dataset into training and test sets [20]. Here, the “modeling set” in the inner loop is not generated. However, deriving all possible combinations (k) of the validation set (containing n compounds) and the calibration set (containing $n - r$ compounds) is followed. The tool has an option for the user to define the number of compounds to be kept in the validation set (r) depending on which the calibration and validation sets are defined. Calibration set compounds are used for the generation of genetic algorithm-multiple linear regression (GA-MLR) [27, 28] models, and the validation sets are utilized for model prediction purpose. A number of internal and external validation metrics are calculated in the exhaustive double cross-validation technique for all the selected models. Additionally, the software also derives partial least squares (PLS) [29] regression models corresponding to each MLR model. Furthermore, the selection of best/top model can be done in any of the five following methods mentioned:

- (i) Model (MLR/PLS) with the lowest mean absolute error or MAE (95%) in the validation set is selected.
- (ii) Model (MLR/PLS) with the lowest MAE (95%) in the modeling set is selected.
- (iii) Model (MLR/PLS) with the highest $Q_{\text{Leave-many-out}}^2$ (modeling set).
- (iv) Application of consensus modeling by using top ranking models selected based on the MAE (95%) values in the respective validation sets. Two types of consensus approaches include (a) simple arithmetic average of predictions from all the selected top models, and (b) weighted average of predictions by assigning appropriate weights to the selected top models based on the mean absolute error obtained from leave-one-out cross-validation, $MAE_{CV(95\%)}$.
- (v) A pool of unique descriptors from the top 3 models with lowest MAE (95%) of the validation set is used. These descriptors are used for further model development purpose. In case of MLR, the best subset selection (BSS) method is used which finds the best combinations of descriptors out of all the possible combinations of unique descriptors present in the selected models. In case of PLS models, the models are formed by all descriptors selected in the top models through a PLS run.

The approach proposed in small dataset modeler (Fig. 1) thus ensures the division of small dataset internally within the DCV algorithm without the actual need of a test set. Thus,

Fig. 1 The approach adopted to develop QSAR models for small-sized dataset using small dataset modeler



there is no requirement of the dataset division. The small dataset modeling approach combines data curation, exhaustive double cross-validation, and optimal model approaches including consensus predictions for model development, particularly for small datasets.

Statistical validation metrics

A rigorous analysis using multiple approaches of assessment of the model quality for measurement of the fitness, stability, robustness, and predictivity of the developed models was carried out. In the present work, we have computed various statistical parameters like determination coefficient (R^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) for internal validation. We have also calculated the leave-many-out squared correlation coefficient ($Q_{LMO(20\%)}^2$) for the final PLS models [30]. Furthermore, r_m^2 metrics [31], root mean square error (RMSE), and mean absolute error (MAE) were also calculated [32].

Results and discussion

2D QSAR models using Dragon and SiRMS descriptors explaining chemical features required for good drug radiosensitization (both SER and logSR) are shown in the following section. There are 4 models developed of which two are QSAR models and the rest two are QSAAR models. All the models are three-descriptor PLS models with 2 latent variables (LVs) showing acceptable values for all validation metrics as shown in Table 2. The validation metrics included R^2 , Q^2 , $Q_{LMO(20\%)}^2$, $\overline{r_m^2(LOO)}$, $\Delta r_m^2(LOO)$, SD (95% data; training), MAE (95% data; training), and RMSE. Furthermore, we

have calculated the Q_{F1}^2 metric for the validation set in each iteration cycle for each model during the calculation of $Q_{LMO(20\%)}^2$ (Supplementary Section). The experimental and predicted values for all the models are given in Supplementary files (S1) and the observed versus predicted plots for all the developed QSAR and QSAAR models are shown in Fig. 2. The different PLS plots including variable importance plot [33], loading plot [29], regression coefficient plot [29], and randomization plot [34] discussed later are shown in Supplementary files (SM2).

Model 1: modeling drug sensitizer enhancement ratio

$$\text{SER} = 0.931 + 0.452 \times \mathbf{H-049} - 0.238 \times \mathbf{B05[O-S]} + 0.09 \times \mathbf{F05[C-S]}$$

The first descriptor **H-049** belongs to atom-centered fragment type, which indicates H atom attached to C^3 (sp^3)/ C^2 (sp^2)/ C^3 (sp^2)/ C^3 (sp). The descriptor symbolizes the hydrogen of a CH group with the carbon bonded to varying numbers of heteroatoms in a variety of hybridizations. The descriptor has a positive contribution towards the response (Fig. 3) which is well understood from certain higher active compounds in the dataset like compounds **19** (SER = 1.81) and **24** (SER = 1.81), each of which has two H-049 fragments. On the other hand, compounds like **12** (SER = 1.11) and **16** (SER = 1.105) having only one such fragments have low SER values.

The next descriptor is **B05[O-S]**, which is a 2D atom pair descriptor demonstrating the presence or absence of oxygen and sulfur atoms at the topological distance 5. The negative contribution explains that presence of oxygen and sulfur

Table 2 Validation metrics of the four models developed using the small dataset modeler

| Model number | Endpoint | Number of descriptors | LV | R ² | Q ² | Q _{LMO} ² (20%) | $\overline{r^2}_{m(L00)}$ | $\Delta^2_{m(L00)}$ | SD (95% data; TRAIN) | MAE (95% data; TRAIN) | RMSE |
|--------------|-------------|-----------------------|----|----------------|----------------|-------------------------------------|---------------------------|---------------------|----------------------|-----------------------|-------|
| 1 | SER | 3 | 2 | 0.834 | 0.746 | 0.712 | 0.660 | 0.134 | 0.066 | 0.073 | 0.096 |
| 2 | logSR | 3 | 2 | 0.798 | 0.660 | 0.665 | 0.563 | 0.109 | 0.189 | 0.216 | 0.261 |
| 3 | QSAAR_SER | 3 | 2 | 0.993 | 0.985 | 0.982 | 0.972 | 0.012 | 0.013 | 0.016 | 0.027 |
| 4 | QSAAR_logSR | 3 | 2 | 0.991 | 0.983 | 0.983 | 0.968 | 0.014 | 0.037 | 0.046 | 0.055 |

atoms at the topological distance 5 will lower the SER values (Fig. 3) as observed in compounds **7** (SER = 1.11) and **16** (SER = 1.105). On the other hand, in compounds like **4** (SER = 1.835) and **30** (SER = 1.687), the absence of such fragment does not lower the SER value.

The descriptor **F05[C-S]**, another 2D atom pair descriptor, denotes the frequency of C-S at the topological distance 5. The positive contribution of the descriptor indicates that higher frequency of the C-S fragment at the topological distance 5 will increase the SER value (Fig. 3) as seen in compounds **30** (F05[C-S] = 3, SER = 1.68) and **38** (F05[C-S] = 3, SER = 1.67).

Model 2: modeling drug survival ratio (logSR)

$$\log\text{SR} = 1.965 - 1.08$$

$$\times S_A(\text{chg})/A_B_B_D/1_4s, 3_4s/4 - 1.073$$

$$\times C-033 - 0.108 \times F07[C-C]$$

$S_A(\text{chg})/A_B_B_D/1_4s, 3_4s/4$ represents a four atomic fragment labeled by partial charges, and its negative regression coefficient indicates that it reduces the radiosensitization property with the presence of such fragment (shown in Fig. 4). In compounds like **26** and **28**, presence of such fragment reduces the radiosensitization (logSR = 0.681 and 0.208).

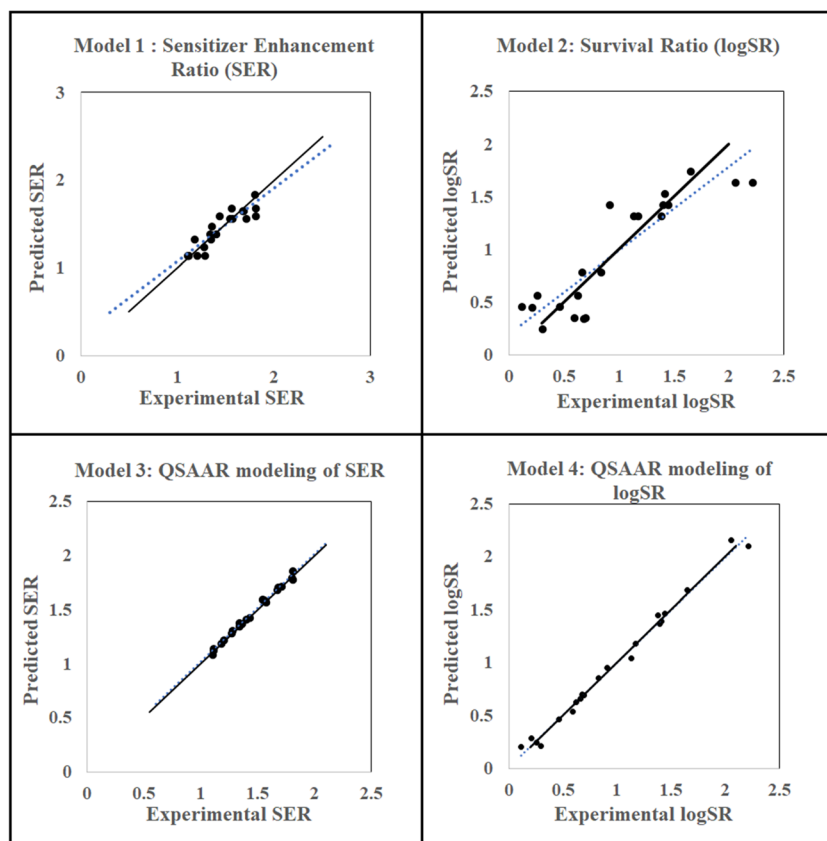
C-033 is an atom-centered fragment descriptor represented by R-CH..X fragment. “R” denotes any group linked through carbon, “-” represents an aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group, “..” represents aromatic single bonds as the C-N bond in pyrrole, and “X” is any electronegative atom (O, N, S, P, Se, halogens) [35]. The negative coefficient indicates that presence of this type of fragment lowers logSR (Fig. 4) values as observed in compounds **6** (C-033 = 1, logSR = 0.462) and **7** (C-033 = 1, logSR = 0.255).

F07[C-C] is a 2D atom pair descriptor, which signifies the frequency of the C-C fragment at the topological distance 7. The negative coefficient indicates that a higher value of the descriptor may decrease the radiosensitization (logSR value) (Fig. 4). This is observed in compounds like **12** and **8** where F07[C-C] are high (6 and 5 respectively) and their logSR values are low (0.301 and 0.591 respectively).

Quantitative structure activity-activity relationship models

Quantitative structure activity-activity relationship (QSAAR) models are mathematical expressions correlating two biological endpoints, here SER and logSR, with the aim to extrapolate any one explicit activity endpoint

Fig. 2 Scatter plots for QSAR and QSAAR models



when the experimental data is not available. This advanced technique can overcome the additional cost of manifold experimental procedures. In the present study, we have developed two QSAAR models, one taking SER as the endpoint and logSR as an independent variable and another taking logSR as the endpoint and SER as an independent variable. It was found that these two endpoints had positive correlation between themselves explaining that increase in experimental values of any of the endpoints would increase the other endpoint values and vice versa.

Model 3: QSAAR modeling of SER

$$\text{SER} = 1.084 + 0.018 \times F03[\text{C-C}] + 0.363 \times \log\text{SR} - 0.001 \times T(\text{N..O})$$

Model 3 is a PLS model with 2 latent variables and shows acceptable values of the validation metrics. Here, logSR has been used as an independent variable to produce a QSAAR model for drug SER. Thus, for any compound, if survival ratio

Fig. 3 Features increasing or decreasing SER values as explained in model 1

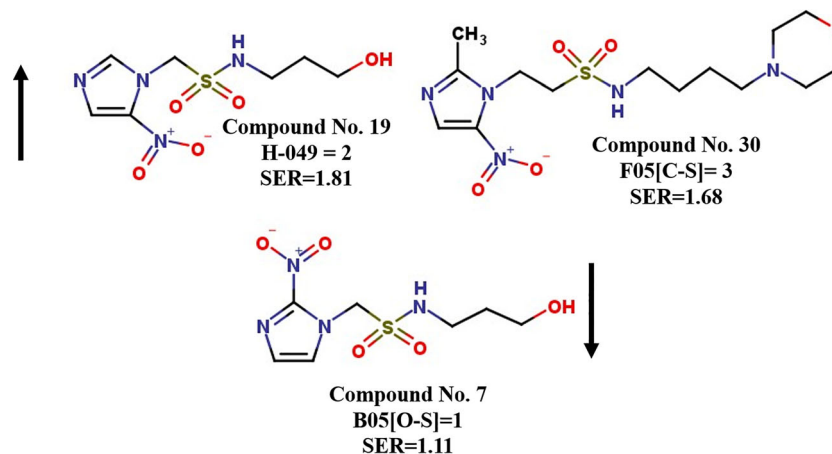
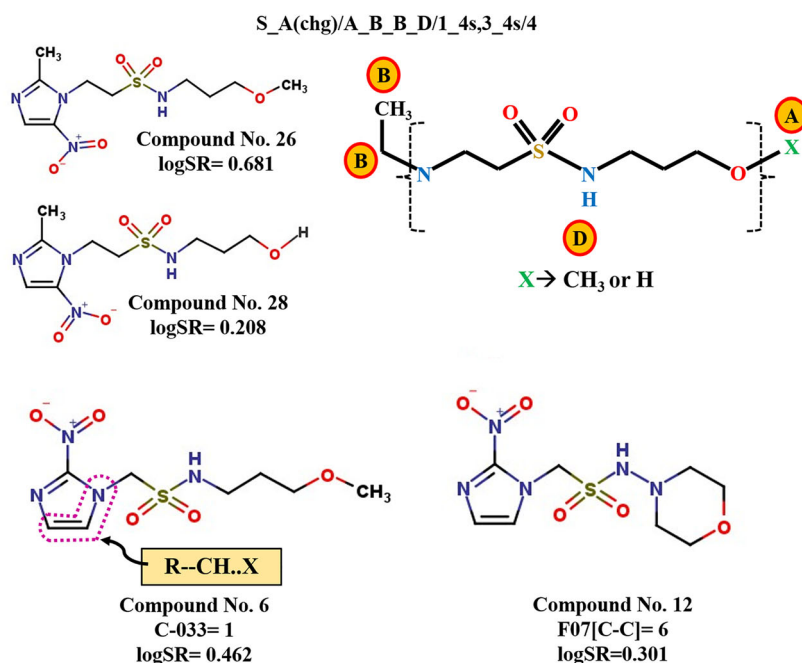


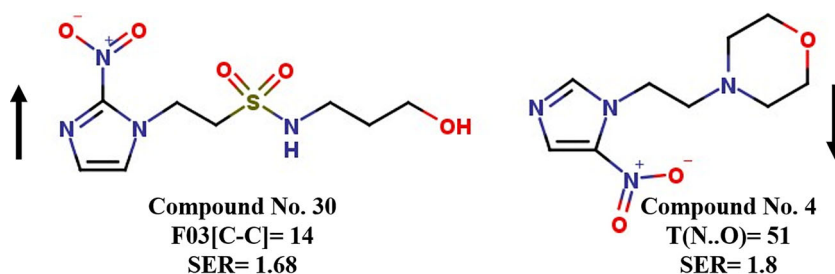
Fig. 4 Factors decreasing logSR values as explained in model 2



(SR) value is known, the SER value can be extrapolated using model 3. This reduces time and experimental expenses. In the model, logSR shows a positive regression coefficient; hence, a higher value of logSR will increase SER values as observed in compounds like **19** (logSR = 2.212, SER = 1.81) and **24** (logSR = 2.057, SER = 1.81).

The descriptor **F03[C-C]** is a 2D atom pair descriptor signifying the frequency of C-C fragments at the topological distance 3. This makes a positive contribution to the endpoint, thus indicating that with an increase in the F03[C-C] descriptor value, SER value will also increase as seen in compounds **30** (F03[C-C] = 14, SER = 1.68) and **35** (F03[C-C] = 13, SER = 1.71). Another 2D atom pair descriptor **T(N..O)** appears in the model signifying the sum of topological distances between N..O. This descriptor has a negative influence on the SER values indicating that the total distance between nitrogen and oxygen should be low for higher SER values as in compound **4** (T(N..O) = 51, SER = 1.8). Compounds with higher T(N..O) values will have lower SER values as observed in compounds **8** (T(N..O) = 130, SER = 1.28) and **12** (T(N..O) = 106, SER = 1.11). Features increasing and decreasing SER values are shown in Fig. 5.

Fig. 5 Features increasing or decreasing SER value as explained in model 3



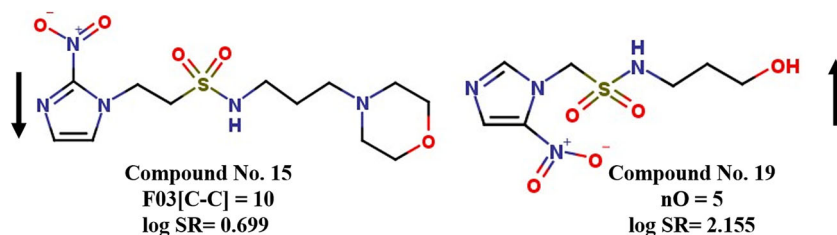
Model 4: QSAAR modeling of logSR

$$\log\text{SR} = -3.364 + 2.735 \times \text{SER} - 0.028 \times \text{F03}[\text{C-C}] + 0.125 \times n\text{O}$$

In model 4, SER has been used as an independent variable for modeling logSR. SER makes a positive contribution to logSR, proving the authenticity of the previously developed model 3 and this can be explained by the same compounds **19** and **24**.

F03[C-C] is a 2D atom pair descriptor symbolizing the frequency of the C-C fragment at the topological distance 3. The descriptor shows a negative regression coefficient, thus signifying that with an increase in F03[C-C] values, logSR value will decrease and vice versa. It is observed that in compounds **15** and **34**, the F03[C-C] values are high (10 and 11 respectively) and their logSR values are low (log SR = 0.699 and 1.134 respectively). The opposite is observed in compounds **19** (F03[C-C] = 2, logSR = 2.057) and **24** (F03[C-C] = 4, logSR = 2.212) having lower values for F03[C-C]. Descriptor **nO** is a constitutional descriptor meaning the

Fig. 6 Features increasing or decreasing logSR value as explained in model 4



number of oxygen atoms present in a molecule. The positive regression coefficient indicates that presence of oxygen atoms is beneficial for the in vitro radiosensitization (logSR). In compounds like **19** ($\log SR = 2.057$) and **24** ($\log SR = 2.212$), higher number of oxygen ($nO = 5$) contributes to a higher value of logSR. Features increasing and decreasing logSR value are shown in Fig. 6.

Plot interpretation

(i) Variable importance plot (VIP)—A VIP can provide with a better knowledge about the descriptors and their contribution in controlling the radiosensitization properties of nitroimidazole sulfonamides. The plot signifies the order of contribution of each descriptor appearing in the model. The most and least important descriptors can be identified using this plot. A variable with VIP score > 1 indicates the descriptor has higher statistical significance as compared to the one with a lower VIP value [33]. The VIP plot showing

the descriptors from higher to lower significance is given in the Supplementary Section S2 (Figs. S1–S4).

- (ii) Loading plot—The loading plot defines the relationship between X variables and Y variables [29]. The plot was developed using the two latent variables for all the four models. The plot describes the impact of the different variables. Descriptors that are grouped together have similar meanings and similar effects on the response, whereas descriptors with different meanings are situated at a considerable distance from each other. Descriptors which are situated far from the plot origin have greater impact on the response. The loading plots of the four models are given in the Supplementary Section S2 (Figs. S5–S8).
- (iii) Regression coefficient plot—The regression coefficient plot [29] gives knowledge about the positive or negative contribution of the descriptors towards the activity (SER or logSR) of the compounds. Descriptors having a positive regression coefficient indicate that with an increase

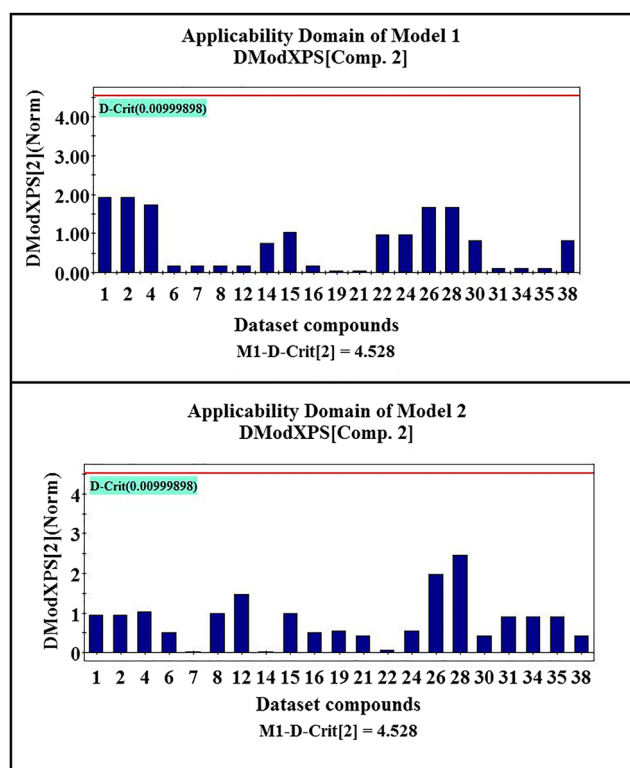


Fig. 7 DModX applicability domain plot of model 1 and model 2

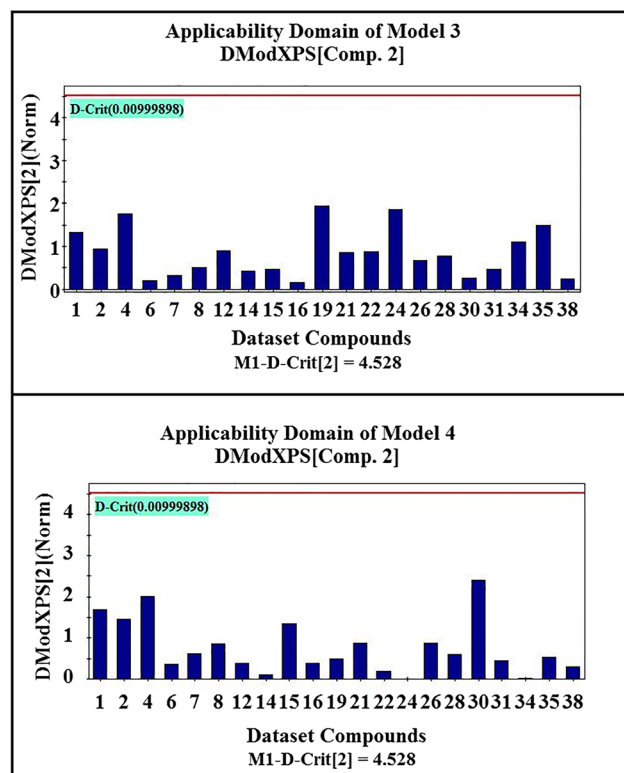


Fig. 8 DModX applicability domain of model 3 and model 4

Table 3 Prediction dataset and their predicted SER and logSR values along with prediction quality and AD status obtained from the “prediction reliability indicator” tool

| Serial no. | Compound no. | Structure (SMILES) | M1 (SER) | | | M2 (logSR) | | | M3 (QSAAR-SER) | | |
|------------|--------------|---|----------|--------------------|-----------|------------|--------------------|------------|----------------|--------------------|-----------|
| | | | Pred_SER | Prediction quality | AD status | Pred_logSR | Prediction quality | AD status | Pred_SER | Prediction quality | AD status |
| 1 | 9 | <chem>c1n(ccn1)CS(=O)(=O)NCCCNC(=O)[O-]</chem> | 0.694 | Bad/Unreliable | In | 0.355 | Moderate | In | 1.174 | Moderate | In |
| 2 | 10 | <chem>c1n(ccn1)CS(=O)(=O)NCCC(=O)O[N+](=O)[O-]</chem> | 0.694 | Bad/Unreliable | In | 0.570 | Moderate | In | 1.207 | Moderate | In |
| 3 | 11 | <chem>c1n(ccn1)CS(=O)(=O)NCCCC(=O)O[N+](=O)[O-]</chem> | 0.784 | Moderate | In | 0.570 | Moderate | In | 1.106 | Moderate | In |
| 4 | 13 | <chem>c1n(ccn1)CCS(=O)(=O)NCCCOC(=O)[O-]</chem> | 0.784 | Moderate | In | 0.462 | Moderate | In | 1.356 | Moderate | In |
| 5 | 18 | <chem>c1n(ccn1)[N+](=O)[O-]CS(=O)(=O)NCCCCO</chem> | 0.694 | Bad/Unreliable | In | 0.570 | Moderate | In | – | – | – |
| 6 | 20 | <chem>c1n(ccn1)[N+](=O)[O-]CS(=O)(=O)NCCCNC(=O)C(=O)C</chem> | 0.931 | Moderate | In | 0.355 | Moderate | In | 1.371 | Moderate | In |
| 7 | 23 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCCOC</chem> | 0.784 | Moderate | In | 0.462 | Moderate | In | 1.401 | Moderate | In |
| 8 | 25 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCCOC</chem> | 0.874 | Moderate | In | 0.456 | Moderate | Outside AD | 1.387 | Moderate | In |
| 9 | 27 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCCOC</chem> | 1.201 | Moderate | In | 1.428 | Moderate | In | 1.382 | Moderate | In |
| 10 | 29 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCCOC</chem> | 1.111 | Moderate | In | 1.320 | Moderate | In | 1.304 | Moderate | In |
| 11 | 32 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCNC(=O)C(=O)C</chem> | 1.201 | Moderate | In | 1.428 | Moderate | In | 1.471 | Moderate | In |
| 12 | 33 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCNC(=O)C(=O)C</chem> | 1.111 | Moderate | In | 1.320 | Moderate | In | 1.548 | Moderate | In |
| 13 | 36 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCNC(=O)C(=O)C</chem> | 0.874 | Moderate | In | 1.535 | Moderate | In | 1.130 | Moderate | In |
| 14 | 37 | <chem>c1n(ccn1)[N+](=O)[O-]CCS(=O)(=O)NCCCNC(=O)C(=O)C</chem> | 1.111 | Moderate | In | 1.428 | Moderate | In | 1.215 | Moderate | In |

in the descriptor values, the SER and logSR increase. On the other hand, a negative regression coefficient indicates that with an increase in the descriptor value, the SER and logSR decrease. The regression coefficient plots are given in Supplementary Section S2 (Figs. S9–S12).

- (iv) Randomization plot—Model randomization is done to ensure that the model is not the result of any chance correlation [34]. The statistical significance of the model is determined by a randomization model. During the model randomization, multiple models are generated by shuffling different combinations of X or Y variables (here Y variable) based on the fit of the reordered model. Here, we have used 100 permutations for each model for random model generation. A model not generated out of chance correlation should have poor statistics (R_y^2 intercept should not exceed 0.3 and Q_y^2 intercept should not exceed 0.05). The randomization plots given in Figs. S12–S16 show that the developed models are non-random and robust and are suitable for prediction.

Applicability domain

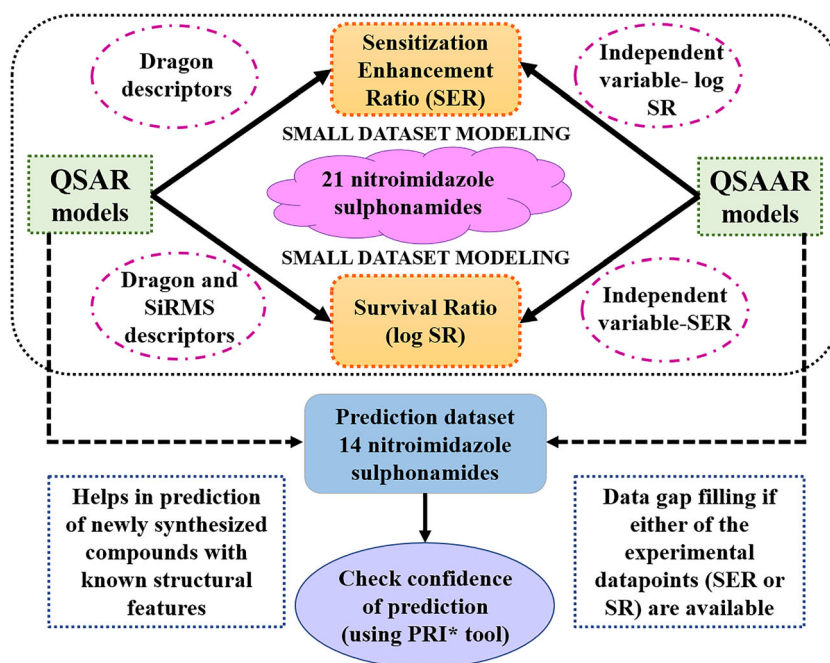
Applicability domain (AD) explains the prediction reliability of a particular model. It is the “chemical space from which a model is derived and where a prediction is considered to be reliable” [36]. AD evaluation was done using DModX (distance to model) in the X-space using SIMCA 16.0.2 software (<https://landing.umetrics.com/downloads-simca>). The AD

plots are given in Figs. 7 and 8. It is found that there is no outlier in any of the four models developed at 95% confidence level ($D\text{-crit} = 0.009999$).

Prediction dataset

A QSAR model helps in the prediction of external datasets based on their molecular features, thereby reducing the experiment costs and animal handling. To study the predictive power of the developed models, we have used 14 compounds whose SER and logSR values have been predicted. These 14 compounds were selected from Table 1 of the source article [21]. This table contained about 36 nitroimidazole sulfonamides out of which 21 compounds were used for QSAR and QSAAR modeling and rest 14 compounds were used as an external set for prediction. Furthermore, we have analyzed the prediction quality and domain of applicability using the prediction reliability indicator tool [37]. The prediction status and domain of applicability are given in Table 3. Prediction was possible for model 1 (M1), model 2 (M2), and model 3 (M3). In M1 and M2, the predicted SER and predicted logSR values were calculated for 14 compounds. In case of M3 (QSAAR-SER), SR_{15} values were obtained from source article [21] and the values were converted to logarithmic form and used as an independent variable for the calculation of predicted SER values. Prediction for model M4 was not possible since experimental SER values for the prediction compounds are not available. During prediction with model M1, three compounds had bad/unreliable predictions. This is due to the difference between the mean of the training set response and predicted value of the query compound being considerably higher. However, these compounds fall inside the AD of the

Fig. 9 Overview of the present work involving the development of QSAR and QSAAR model using small dataset modeler



model. In case of M2, one compound (compound no. 25) is outside AD; however, it shows moderate prediction quality. During prediction with model M3, all the compounds are found to have “moderate” prediction quality and are inside the model AD.

Conclusion

This study aims at developing 2D QSAR models with the notion to investigate the essential features in nitroimidazole sulfonamide analogues to show radiosensitization properties with respect to sensitizer enhancement ratio and survival ratio endpoints. The different descriptors obtained give an idea about the position of the features and type of chemical groups required to enhance or hinder these properties. Moreover, QSAAR modeling helps in correlating two endpoints (SER and logSR) and suggests how to extrapolate an endpoint if the experimental information is unavailable. The current study emphasizes on the application of the “small dataset modeler” software when the dataset is small and splitting of dataset is not worthy. Furthermore, the newly developed models were used for prediction of 14 compounds and their prediction reliability was checked. These developed QSAR and QSAAR models are able to predict newly developed nitroimidazole sulfonamide derivatives with known structural features. The complete overview of the work is explained in Fig. 9.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11224-021-01734-w>.

Code availability The DTC Lab software tools are available at http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. Contact kunal.roy@jadavpuruniversity.in for further details.

Authors' contributions PD: Model development and validation, original draft. KR: Supervision, concept, and editing.

Funding PD thanks the Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship.

Data availability Raw data of the modeling study are available in Supplementary materials.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Ethics approval This is a computational paper not requiring any ethics approval.

Consent to participate All co-authors have consent to participate in this project.

Consent for publication All co-authors have consent to publish the results.

References

- De Ridder M, Verellen D, Verovski V, Storme G (2008) Hypoxic tumor cell radiosensitization through nitric oxide. *Nitric Oxide* 19: 164–169
- Muz B, de la Puente P, Azab F, Azab AK (2015) The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* 3:83
- Kioi M, Vogel H, Schultz G, Hoffman RM, Harsh GR, Brown JM (2010) Inhibition of vasculogenesis, but not angiogenesis, prevents the recurrence of glioblastoma after irradiation in mice. *J Clin Invest* 120:694–705
- Chiche J, Brahimi-Horn MC, Pouyssegur J (2010) Tumour hypoxia induces a metabolic shift causing acidosis: a common feature in cancer. *J Cell Mol Med* 14:771–794
- Hill RP, Marie-Egyptienne DT, Hedley DW (2009) Cancer stem cells, hypoxia and metastasis. *Semin Radiat Oncol* 19:106–111
- Wardman PJCO (2007) Chemical radiosensitizers for use in radiotherapy. *Clin Oncol* 19:397–417
- Suto MJ (1991) Radiosensitizers. *Annu Rep Med Chem* 26:151–160
- Hall EJ, Astor M, Biaglow J, Parham JC (1982) The enhanced sensitivity of mammalian cells to killing by X rays after prolonged exposure to several nitroimidazoles. *IJROBP* 8:447–451
- Saunders M, Dische S (1996) Clinical results of hypoxic cell radiosensitisation from hyperbaric oxygen to accelerated radiotherapy, carbogen and nicotinamide. *Br J Cancer* 27:S271
- Newman HFV, Bleehen NM, Ward R, Workman P (1988) Hypoxic cell radiosensitizers in the treatment of high grade gliomas: a new direction using combined Ro 03-8799 (pimonidazole) and SR 2508 (etanidazole). *IJROBP* 15:677–684
- Yahiro T, Masui S, Kubota N, Yamada K, Kobayashi A, Kishii K (2005) Effects of hypoxic cell radiosensitizer doranidazole (PR-350) on the radioresponse of murine and human tumor cells in vitro and in vivo. *J Radiat Res* 46:363–372
- Metwally MAH, Frederiksen KD, Overgaard J (2014) Compliance and toxicity of the hypoxic radiosensitizer nimorazole in the treatment of patients with head and neck squamous cell carcinoma (HNSCC). *Acta Oncol* 53:654–661
- Hong CR, Wang J, Hicks KO, Hay MP (2016) Efficient protocol for the identification of hypoxic cell radiosensitizers. *Tumor Microenviron* pp. 269–290
- Bonnet M, Hong CR, Gu Y, Anderson RF, Wilson WR, Pruijn FB, Wang J, Hicks KO, Hay MP (2014) Novel nitroimidazole alkylsulfonamides as hypoxic cell radiosensitizers. *Bioorg Med Chem* 22:2123–2132
- Roy K (2015) Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment. *IGI Global*
- Hansch C, Leo A, Mekapati SB, Kurup A (2004) Qsar and Adme. *Bioorg Med Chem* 12:3391–3400
- Klein C, Kaiser D, Kopp S, Chiba P, Ecker GF (2002) Similarity based SAR (SIBAR) as tool for early ADME profiling. *J Comput Aided Mol Des* 16:785–793
- Tareq Hassan Khan M (2010) Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr Drug Metab* 11:285–295
- Lessigiarska I, Worth AP, Netzeva TI, Dearden JC, Cronin MT (2006) Quantitative structure-activity-activity and quantitative structure-activity investigations of human and rodent toxicity. *Chemosphere* 65:1878–1887
- Ambure P, Gajewicz-Skretna A, Cordeiro MND, Roy K (2019) New workflow for QSAR model development from small data sets: Small Dataset Curator and Small Dataset Modeler. Integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. *J Chem Inf Model* 59:4070–4076

21. Bonnet M, Hong CR, Wong WW, Liew LP, Shome A, Wang J, Gu Y, Stevenson RJ, Qi W, Anderson RF, Pruijn FB (2018) Next-generation hypoxic cell radiosensitizers: nitroimidazole alkylsulfonamides. *J Med Chem* 61:1241–1254
22. MarvinSketch software, <https://www.chemaxon.com>. Accessed on 26 Nov 2020
23. Todeschini R, Consonni V (2009) *Molecular descriptors for chemoinformatics*. Wiley-VCH, Weinheim
24. Dragon version 7, Kodesrl, Milan, Italy, 2016; software available at <http://www.talete.mi.it/index.htm>. Accessed on 28 Nov 2020
25. Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) Hierarchic system of QSAR models (1D–4D) on the base of simple representation of molecular structure. *J Mol Model* 11:457–467
26. Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intell Lab Syst* 159:108–126
27. Devillers J (1996) *Genetic algorithms in molecular modeling*. Academic Press, Cornwall, Great Britain
28. Venkatasubramanian V, Sundaram A (2002) Genetic algorithms: introduction and applications. *Encycl Comput Chem* 2
29. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58:109–130
30. Roy K, Kar S, Das RN (2015) *Statistical methods in QSAR/QSPR in a primer on QSAR/QSPR modeling: fundamental concepts*. Springer, Cham
31. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 52:396–408
32. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33
33. Akarachantachote N, Chadcham S, Saithanu K (2014) Cutoff threshold of variable importance in projection for variable selection. *Int J Pure Appl Math* 94:307–322
34. Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47:2345–2357
35. Todeschini R, Consonni V (2009) *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references (Vol. 41)*. John Wiley & Sons
36. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 1:45–63
37. Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? *ACS Omega* 3:11392–11406

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Nitroaromatics as hypoxic cell radiosensitizers: A 2D-QSAR approach to explore structural features contributing to radiosensitization effectiveness



Priyanka De, Kunal Roy*

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India

ARTICLE INFO

Keywords:

Hypoxia
Nitroaromatics
Radiosensitization
Radiosensitization effectiveness
QSAR

ABSTRACT

Hypoxia is the prime component of tumor microenvironment that plays a pivotal role in cancer progression. Nitroaromatic compounds are known to enhance the sensitivity of hypoxic cells to ionizing radiation. The application of computational tools like Quantitative Structure-Activity Relationship (QSAR) can be used to predict newly developed nitroaromatics or compounds with missing data. In the present work, three datasets consisting of 18 nitrofurans, 11 nitrothiophenes and 84 nitroimidazoles were utilised for two-dimensional QSAR modeling to retrieve their structural features essential to elicit radiosensitivity. The work comprises two parts: (i) local modeling using individual datasets; and (ii) global modeling by clubbing the three datasets. The two-dimensional descriptors were calculated using Dragon (version 7.0) software. The developed models were obtained using various feature selection techniques applied in “Small Dataset Modeling” and “Double Cross Validation” tools available from <https://dtclab.webs.com/software-tools>. Finally, the models were validated using stringent metrics following the Organisation for Economic Co-operation and Development (OECD) guidelines. The developed models are robust, predictive, and are useful tools to predict the radiosensitization of newly developed nitroaromatics. Furthermore, the global model was used to predict two external sets comprising 10 and 47 compounds, and the prediction ability was validated using the “Prediction Reliability Indicator” tool.

1. Introduction

Nitroaromatic drugs have been applied to radiation therapy owing to their effectiveness in enhancing radiation damages selectively in hypoxic mammalian cells at nontoxic concentration. These drugs are known to cause a specific group of mutagenesis that require cell hypoxia as their metabolic activation and expression [1]. Nitroaromatic sensitizers are able to control the rate of local tumor growth by traditional radiotherapy while hypoxia plays a limiting role. A second discrete function of these agents involves selective cytotoxicity of the drug to hypoxic tumor cell. Moreover, these drugs have the ability to identify and locate hypoxic cells making it suitable for tumor diagnosis [2]. DNA damaging potential of nitroaromatics is an effective knowledge for the development of newer antineoplastic drugs. Bioreduction ability of nitroaromatics allows generation of free radicals in intracellular environments with a low oxygen concentration; a typical circumstance occurring in solid tumors encompassing areas of hypoxia resulting from inadequate blood supply [3,4]. The radiosensitizing activity of these compounds is dependent on two important properties: electron affinity and reduction potential. The ionizing radiation generates free radicals which has marked ability to

cause DNA damage and cellular death [5]. Nitroheterocyclic compounds like nitroimidazoles, nitrofurans and nitrothiophenes are recognized oxygen-mimetic agents in which electron rich reactive nitro group reacts with DNA, after which the DNA and nitro group adduct causes DNA strand breakage and subsequent cell death. The mechanism for DNA damage by aromatic nitro compounds is shown in Fig. 1. Radiosensitizers are aimed to augment tumor killing while having minimal effect on normal tissues. An ideal radiosensitizer should be able to selectively target hypoxic cells and reach them faster in adequate concentrations. It should have least toxicity itself and minimum or controllable augmentation of radiation induced toxicity [6].

Oxygen is the prime hypoxic cell radiosensitizer and compounds which mimic oxygen, such as nitroimidazoles, nitrofurans and nitrothiophenes, are potential targets to combat hypoxia-associated radioresistance. In a study by Chapman et al. [7], analysis with various analogues of nitrofurans and nitroheterocyclics on Chinese hamster cells grown *in vitro* helped in understanding their toxicity and radiosensitivity. The nitrofurans exhibited good radiosensitization ability with nifuraldezone being the most potential compound. Langenbacher et al. characterised the *in vitro* hypoxic cytotoxicity and radiosensitizing efficacy of

* Corresponding author.

E-mail address: kunal.roy@jadavpuruniversity.in (K. Roy).

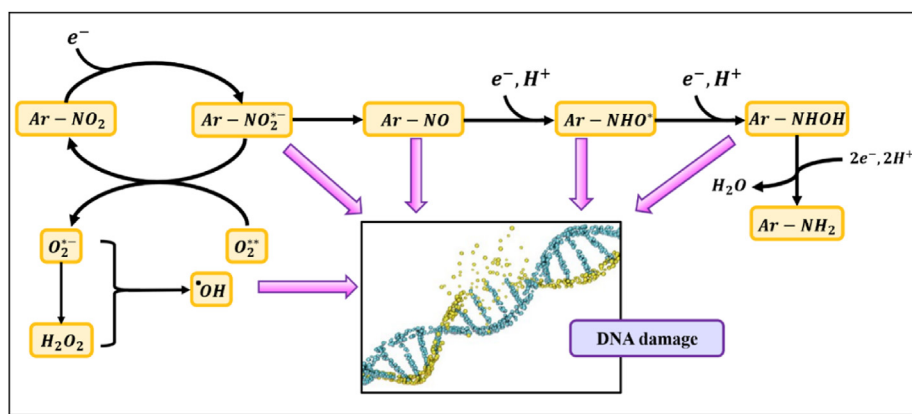


Fig. 1. Mechanism of nitroaromatic radiosensitizers in DNA damage.

N, N, N-tris[2-(2-nitro-1H-imidazole-1-yl)ethyl]amine (PRC), as a novel hypoxia-selective cytotoxic agent. The study performed on hypoxia-sensitive lymphoma and hypoxia-resistant glioblastoma cell line by colony formation assay and flow cytometry showed that PRC exerted high hypoxic cytotoxic and radiosensitizing action on both cell lines at almost absent toxicity under normoxic conditions [8]. Breccia et al. [9] synthesized a series of nitrothiophene analogues and studied their physicochemical parameters which influence their radiosensitization potential and toxicity. Compounds like NTMA (2-(5-nitrothiophen-3-yl) propanedioic acid), NTM (4-(5-nitrothiophene-2-carbonyl)morpholine), 4 and 5-NTCA (4-nitrothiophene-2-carboxylic acid and 5-nitrothiophene-2-carboxylic acid respectively) were found promising agents exhibiting good radiosensitization. Bioreductive prodrug SN38023, a nitroimidazole analogue elicits radiosensitivity on hypoxic tumor cells selectively in the absence of oxygen [10]. This compound undergoes reduction (bioreduction) within the tumor cells before exhibiting radiosensitizing property and provides suitable trigger in the inactivation of DNA protein kinase inhibitors (DNA-PKi). Misonidazole is a hypoxic radiosensitizer which can augment the antitumor role of cyclophosphamide as observed in preclinical studies [11]. It is anticipated to be an ideal radiation therapy sensitizer concerning the control of radiation-resistant tumor cells and p53 mutant tumor cells [12].

In silico tools such as quantitative structure-activity relationship (QSAR) modeling has become an effective practice for the prediction of radiosensitization properties when there is lack of data. Further, in comparison with the animal testing, *in silico* approaches are faster and less expensive, thus an *in-silico* method is very helpful towards data gap filling of a new query compound [13]. Computational methods like QSAR have significantly impacted the paradigm of drug discovery. This method allows for the calculation of physicochemical properties (e.g., lipophilicity) [14,15], the estimation of biological activity (or toxicity) [16, 17], lead optimization [18], as well as the evaluation of absorption, distribution, metabolism, and excretion (ADME) [19,20]. Radiosensitivity in terms of radiosensitization effectiveness is considered as a property of nitroaromatics which can be subjected to *in silico* QSAR analysis. Experimental data for radiosensitization of nitroaromatic compounds is very scarce; hence, well validated QSAR models assist in the prediction of such compounds and reduce experimental evaluation to a certain degree.

The current study aims to explore the maximum possible structural features expressed by nitroaromatics which are responsible for their radiosensitization effectiveness. We have collected three datasets of nitroimidazole, nitrofurans and nitrothiophene analogues which were modelled individually as local datasets as well as collectively put into global QSAR modeling. For the calculation of a large pool of over 800 two-dimensional molecular features, we have applied Dragon 7 software [21]. The individual local datasets (mainly nitrofurans and nitrothiophene data) comprise a small number of compounds (<20 compounds) and therefore

they are not suitable for usual data division into training and test sets. To avoid loss of chemical information for these kinds of datasets, “small dataset modeling” was adopted using the whole dataset [22]. For datasets with larger number of data points as in case of local nitroimidazole dataset, feature selection was carried out using the Genetic Algorithm [23,24] method. In case of global data set modeling, variable selection was performed using Genetic Algorithm employing the Double Cross Validation [25] tool. The developed 2D-QSAR models were rigorously validated by applying internationally accepted stringent validation parameters. Additionally, two different sets of external compounds were used for the prediction using the global model to check its predictive ability.

2. Materials and methods

2.1. Datasets

The radiosensitization effectiveness ($pC_{1.6}$) data for three nitroaromatics datasets (nitrofurans, nitrothiophenes and nitroimidazoles) were obtained from the previously published literature [26–28]. The datasets comprised 18 nitrofurans analogues, 11 nitrothiophenes and 84 nitroimidazole derivatives in the composite set. ‘ $C_{1.6}$ ’ is a term used to explain the radiosensitization capacities; this is the molar concentration of the compound required to give a sensitizer enhancement ratio (SER) of 1.6. Thus, a lower value for $C_{1.6}$ will give greater sensitizing efficiency. For an efficient analysis, the $C_{1.6}$ values were converted into their negative logarithmic scale ($pC_{1.6}$). The structures in the datasets were drawn in MarvinSketch software (version 14.10.27) [29] with proper aromatization and hydrogen bond addition and saved as MDL mol format.

2.2. Descriptor calculation

Before a QSAR model is developed, the structural information is converted into numerical values known as descriptors [30]. The three curated datasets were used for the calculation of descriptors using Dragon version 7 [31] software. Specific classes of descriptors were used for model development including connectivity, constitutional, topological, E-state indices, functional, 2D atom pairs, 2D autocorrelation, ring, atom-centred fragments and molecular property descriptors. The descriptors were pre-treated to reduce redundant and noisy data; constant (variance <0.0001) and intercorrelated ($|r| > 0.95$) variables were removed using an in-house software available at <http://dtclab.webs.com/software-tools> before model development.

2.3. Data set splitting and model development

Rational splitting of a dataset into training and test sets is a crucial step before a QSAR model development leading to the establishment of

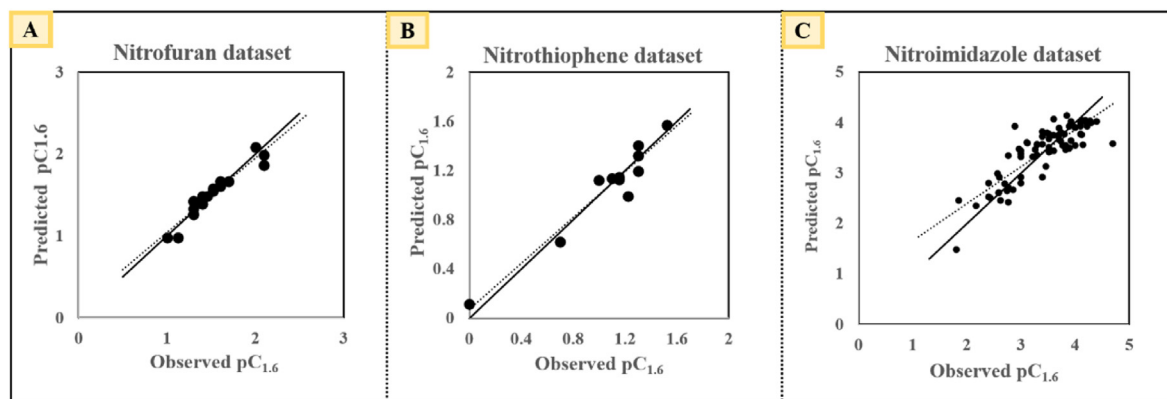


Fig. 2. Observed vs predicted $pC_{1.6}$ scatter plot for the local nitro datasets.

the models' predictive power. However, a general problem faced by *in silico* researchers during the development of ideal QSAR models is the non-availability of sufficient data suitable for data set splitting. Datasets with 25–50 datapoints or even less are difficult to divide into training and test sets and there is less chance of getting robust and predictive models. Ambure et al. [22] proposed a method for small datasets which does not require the step of data set division. “Small dataset modeling” as proposed by these authors involves the Double Cross Validation (DCV) method [25,32]. In this method, the entire dataset of n compounds is taken under consideration. The process involves the generation of all possible combinations (k) of the validation set (each containing r compounds) and the calibration set (containing $n - r$ compounds). Here, the user is allowed to set the ‘ r ’ value, i.e., the number to compounds to be retained in the validation set and depending on that all probable combinations of calibration and validation sets are generated. The models are generated using Multiple Linear Regression (MLR) [33] method using Genetic Algorithm (GA) method of feature of selection. In this scheme of exhaustive DCV, several important validation metrics are calculated for all the elected models. The selection of the best models is dependent on a set of criteria discussed in the source literature [22]. In the current study, the number of data points for nitrofurans and nitrothiophenes is relatively very small (18 and 11 respectively) for dataset division. Hence, we have utilised the “small dataset modeling” technique for efficient model development for these datasets. For the nitrofurans dataset, we have chosen the best Multiple Linear Regression (MLR) [33] model developed using the MLR plus Validation 1.3 tool available from <https://dtclab.webs.com/software-tools>. However, for the nitrothiophene dataset, the descriptors of the best MLR model were subjected to Partial Least Squares (PLS) regression [34] using the Partial Least Squares tool (available from <https://dtclab.webs.com/software-tools>). Note that PLS is a robust and generalized version of MLR which converts the original sets of descriptors into new latent variables which are lower in number in comparison to the descriptors appearing in the corresponding MLR model [34]. PLS can handle numerous and noisy variables and do not suffer from the inter-correlation problem.

In case of the nitroimidazole dataset with 84 datapoints, we have applied the Genetic Algorithm Multiple Linear Regression (GA-MLR) [23, 24] method for feature selection on the whole dataset. A pool of ten descriptors (features) was selected after this process which were further subjected to the Best Subset Selection (BSS) method which finds the best combinations of descriptors out of all the possible combinations of unique descriptors present in the selected models. The best descriptor combination obtained in this process were further subjected to PLS regression using the Partial Least Squares tool (available from <https://dtclab.webs.com/software-tools>) to obtain better quality model. In this work, we have not divided the nitroimidazole dataset though it has

sufficient amount of data points because division of the data set was earlier performed by our group in a previous work [35]. Thus, we have developed three local models from undivided data sets: nitrofurans model, nitrothiophene model and nitroimidazole model. These data sets were further clubbed to form a global dataset which was then modelled.

During modeling the global dataset, the compounds were split into training and test sets using Kennard-Stone method [36] in Dataset Division GUI 1.2 software tool available from <https://dtclab.webs.com/software-tools>. The dataset was divided into training and test sets in 7:3 ratio. Here, Genetic Algorithm method was used in the Double Cross Validation tool [25] for variable selection. A pool of 16 descriptors was selected and the final model was generated using PLS regression method using the Partial Least Squares tool (available from <https://dtclab.webs.com/software-tools>) using descriptors selected from best subset selection (BSS).

2.4. Statistical validation metrics

During the course of the present work, we have performed rigorous analysis using multiple approaches of assessment of the model quality for measurement of the stability, robustness, fitness, and predictivity of the developed models. We have computed various statistical metrics like determination coefficient (R^2), adjusted determination coefficient (R_{adj}^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) for internal validation [37]. We have also computed the leave-many-out squared-correlation coefficient ($Q_{LMO(20\%)}^2$) [38]. For external validation, in case of the global model, parameters like R_{pred}^2 or Q_{F1}^2 , Q_{F2}^2 and concordance correlation coefficient (CCC) were calculated [39]. Furthermore, we have also calculated the r_m^2 metrics (both Δr_m^2 and \bar{r}_m^2) [40] and validated the models using root mean squared error (RMSE) and mean absolute error (MAE) based criteria [41].

3. Result and discussion

3.1. Modeling local nitro datasets

2D-QSAR models for explaining radiosensitization effectiveness ($pC_{1.6}$) are discussed in this section. The QSAR models from individual class of nitro compounds were found to have good and acceptable values for all validation metrics. The validation metrics included R^2 , Q^2 , Q_{LMO}^2 , $\bar{r}_{m(LOO)}^2$, $\Delta r_{m(LOO)}^2$, MAE and RMSE. The current work proposes statistically robust and acceptable local models employing simple 2D descriptors. The observed versus predicted $pC_{1.6}$ plots for the local models are given in Fig. 2.

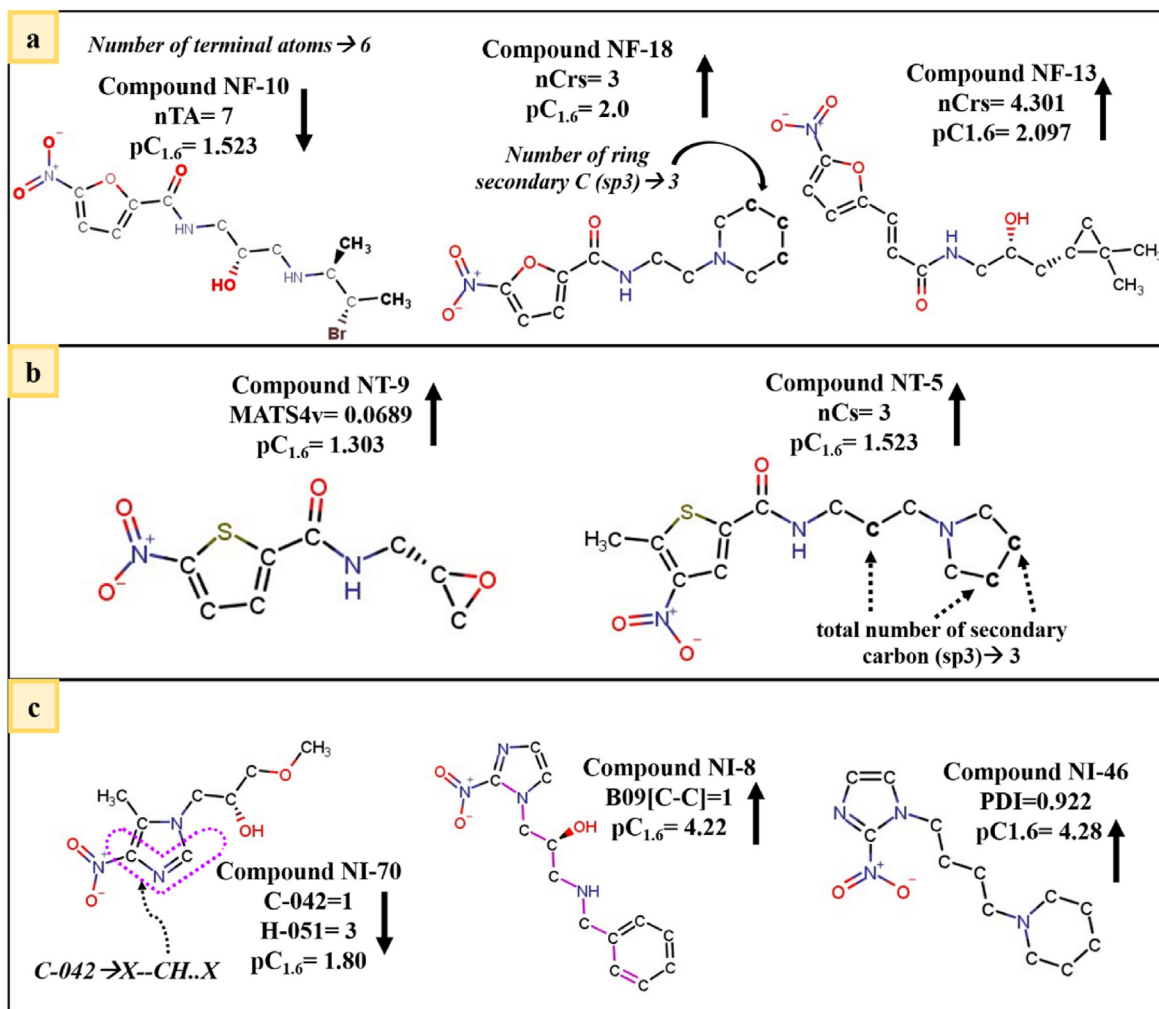


Fig. 3. Contribution of the descriptors obtained in local nitro dataset modeling towards radiosensitization effectiveness ($pC_{1.6}$ values).

3.1.1. QSAR model studying radiosensitization effectiveness of nitrofurans

$$pC_{1.6} = -0.617(\pm 0.249) - 0.361(\pm 0.039) \times nTA + 0.127(\pm 0.028) \times nCrS + 1.050(\pm 0.110) \times DBI$$

$$N = 18, R^2 = 0.911, R_{adj}^2 = 0.892, Q_{LOO}^2 = 0.842, Q_{LMO(20\%)}^2 = 0.780, \overline{r_{m(LOO)}^2} = 0.786, \Delta r_{m(LOO)}^2 = 0.037, MAE(95\%) = 0.078, RMSE = 0.090, Prediction\ Quality = Moderate$$

(1)

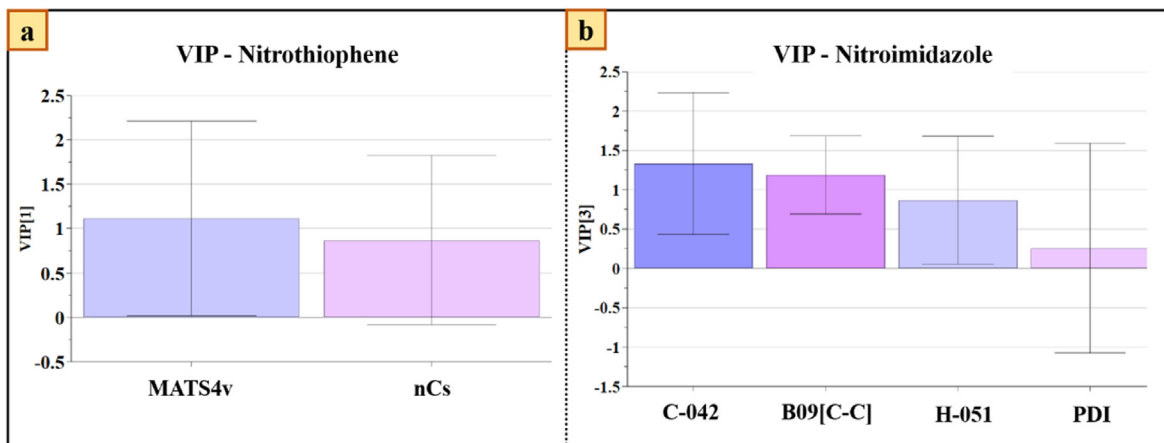


Fig. 4. Variable importance plot of local nitrothiophene and nitroimidazole datasets.

The number of data points in case of nitrofurans was very less and not suitable for data set division into training and test sets. Thus, small dataset modeling was used for robust model development where data set division is not worthy. The MLR model developed showed good determination coefficient (R^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) for internal validation. The leave-many-out predicted variance (Q_{LMO}^2) was also calculated. The descriptors appearing in the model are: **nTA** (number of terminal atoms), **nCrS** (number of ring secondary C(sp³)) and **DBI** (Dragon branching index).

The descriptor **nTA** belonging to the constitutional type has a negative contribution towards radiosensitization effectiveness; thus, compounds having higher number of terminal atoms will have lower $pC_{1.6}$ value and vice versa. This can be explained with compounds **NF-10** and **NF-12**. In compound **NF-10**, which has a lower value for $pC_{1.6}$ ($pC_{1.6} = 1.523$), the number of terminal atoms is 7 (in higher side) (Fig. 3a). Again, compound **NF-12** which has a lower number of terminal atoms ($nTA = 3$) shows higher $pC_{1.6}$ value ($pC_{1.6} = 2.097$).

nCrS represents the number of sp³ hybridised secondary carbon present in a ring system. The positive contribution implicates that with an increase in **nCrS** values, radiosensitization effectiveness will increase. This has been observed in compounds like **NF-18** (**nCrS** = 3) (Fig. 3a) and **NF-12** (**nCrS** = 1) where presence of such secondary carbon has increased the $pC_{1.6}$ value ($pC_{1.6} = 2$ and 2.097 respectively).

Another positively correlated descriptor, **DBI**, represents the branching nature of the compound. With an increase in the branching index, the radiosensitization will increase as observed in compound **NF-13** (**DBI** = 4.301, $pC_{1.6} = 2.097$) (Fig. 3a).

3.1.2. QSAR model studying radiosensitization effectiveness of nitrothiophenes

$$pC_{1.6} = 0.816 + 0.191 \times nCs + 4.555 \times MATS4v$$

$$N = 11, R^2 = 0.933, Q_{LOO}^2 = 0.807, Q_{LMO(20\%)}^2 = 0.896, \overline{r_{m(LOO)}^2} = 0.660, \Delta r_{m(LOO)}^2 = 0.178, MAE_{Fitted} = 0.081, MAE_{LOO} = 0.124, RMSE = 0.101, Prediction\ Quality = Moderate$$

(2)

Small dataset modeling was utilised again owing to the limited number of compounds in the dataset. The developed PLS model has two descriptors and one latent variable: **nCs** (total number of secondary carbon (sp³)) and **MATS4v** (Moran autocorrelation of lag 4 weighted by van der Waals volume). From the VIP plot [42] (Fig. 4a), it was found that **MATS4v** has higher VIP score than **nCs** denoting that **MATS4v** is of higher significance than **nCs**. The predicted variance explained by specific features for each latent variable is given in the Supplementary section.

MATS4v is a 2D autocorrelation descriptor, which represents the distribution mode of the atomic van der Waals volumes along the topological structure of nitrothiophenes. Here, the path connecting a pair of atoms has length 4 and applies the atomic van der Waals volumes as weighting scheme. The positive regression coefficient advocates that a higher positive value of the descriptor enhances the radiosensitivity as observed in compound **NT-9** (**MATS4v** = 0.068914, $pC_{1.6} = 1.301$) (Fig. 3b).

nCs is a functional group count descriptor and has a positive correlation with radiosensitization effectiveness. A secondary carbon is one which is bound by two other carbon atoms. Increase in the number of such fragments in nitrothiophenes will increase their radiosensitivity. This is observed in compounds like **NT-5** (**nCs** = 3, $pC_{1.6} = 1.523$) (Fig. 3b) and **NT-6** (**nCs** = 4, $pC_{1.6} = 1.301$).

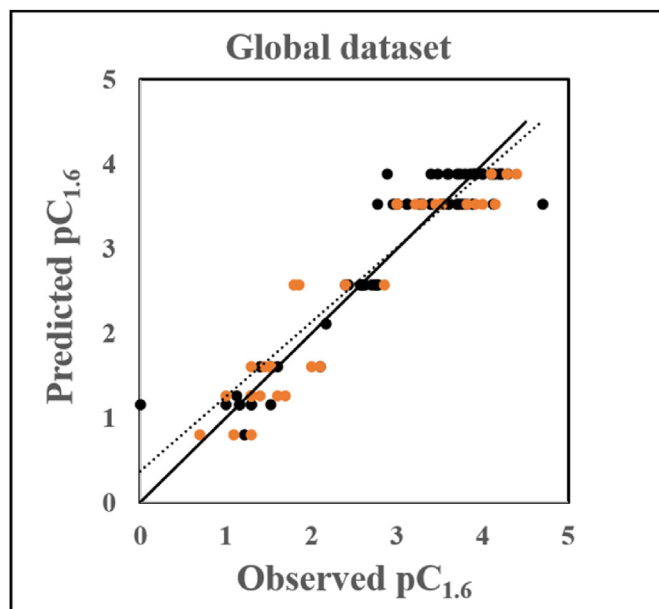


Fig. 5. Scatter plot of the global model.

3.1.3. QSAR model studying radiosensitization effectiveness of nitroimidazoles

$$pC_{1.6} = 0.873 - 1.267 \times C - 0.42 - 0.227 \times H - 0.51 + 0.287 \times B09[C - C] + 3.115 \times PDI$$

$$N = 84, R^2 = 0.733, R_{adj}^2 = 0.723, Q_{LOO}^2 = 0.701, Q_{LMO(20\%)}^2 = 0.696, \overline{r_{m(LOO)}^2} = 0.588, \Delta r_{m(LOO)}^2 = 0.193, MAE = 0.251, RMSE = 0.321, Prediction\ Quality = Moderate$$

(3)

A four descriptor PLS model with 3 latent variables (LVs) was developed for the nitroimidazole dataset. Here, the number of compounds in the dataset was relatively higher and could be divided into training and test sets for model development. However, this dataset was earlier used by our group for model development in a previously published literature [37] where division of this dataset provided acceptable results. Here, we have tried modeling for the whole dataset using GA-MLR method of variable selection followed by BSS method of model development. The descriptors selected in the best MLR model was further subjected to PLS regression with 3 LVs which showed good determination coefficient (R^2) and leave-one-out squared correlation coefficient (Q_{LOO}^2) as given in model 3. From the VIP plot (Fig. 4b), the significance of the descriptors are as follows: C-042, B09[C-C], H-051 and PDI.

The descriptor **C-042** is an atom-centred fragment descriptor representing the fragment X-CH.X (Figure), where X is any electronegative atom (O, N, S, P, Se, halogens); '-' is an aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group; and '.' is an aromatic single bond as the C-N bond in pyrrole. The negative regression coefficient implicates that increase in the number of such type of fragments in the nitroimidazole analogues will hinder its radiosensitivity. This has been observed in compounds like **NI-69** (C-042 = 1, $pC_{1.6} = 1.85$) and **NI-70** (C-042 = 1, $pC_{1.6} = 1.80$) (Fig. 3c) where the presence of C-042 fragment caused a lowering in $pC_{1.6}$ values.

The next descriptor is **B09[C-C]**, a 2D atom pair descriptor denoting the presence or absence of C-C fragment at the topological distance 9. The positive coefficient of this descriptor implies that the value of B09

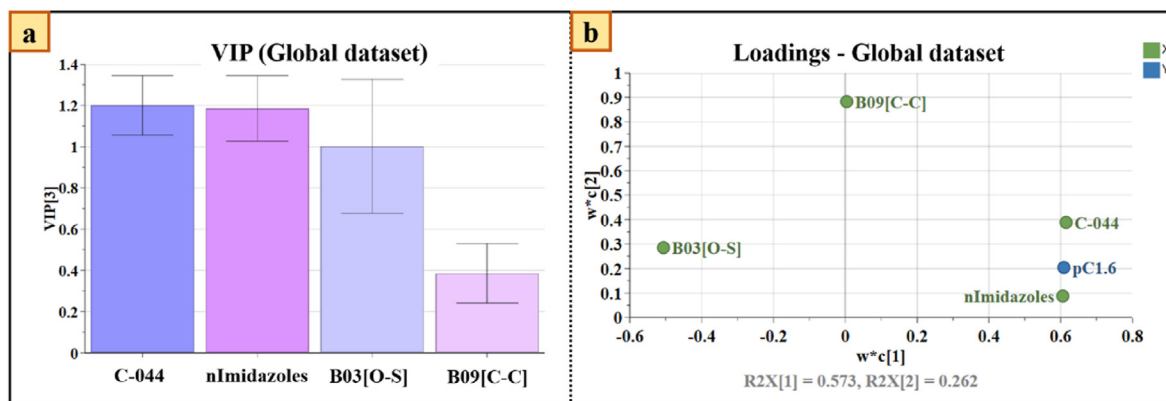


Fig. 6. Variable importance plot and loading plot of the global model.

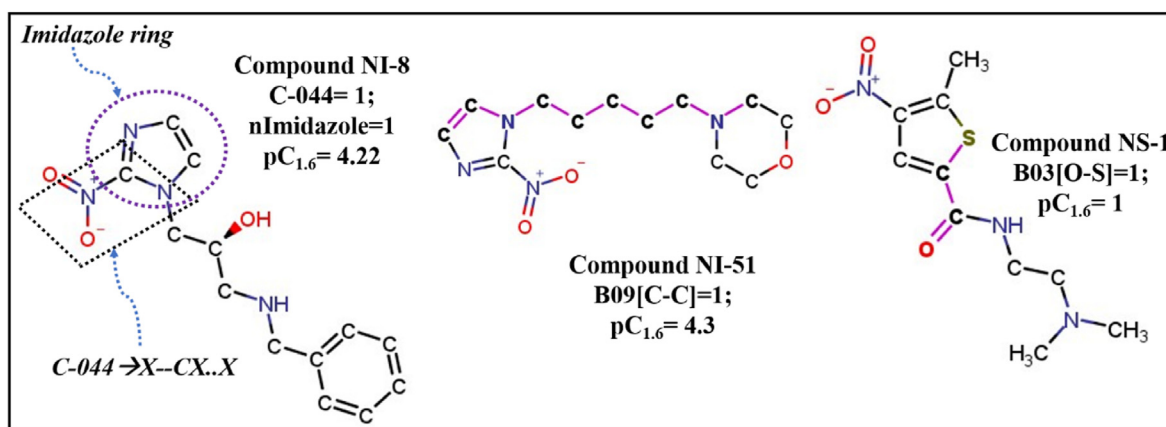


Fig. 7. Contribution of the descriptors appearing in the global model.

[C-C] is directly proportional to the radiosensitization effectiveness, which is established from the presence of such fragments in most of the active compounds (e.g., compounds NI-8 and NI-11) (Fig. 3c).

Another atom-centred fragment descriptor, H-051 corresponds to H attached to alpha carbon (where alpha carbon is any carbon attached through a single bond with $-C = X$, $-C\#X$, $-C-X$). This descriptor also contributes negatively towards the radiosensitive effectiveness; thus with an increase in the descriptor value, $pC_{1.6}$ value will decrease. This can be explained with compound number NI-70 where there are three such H-051 (H-051 = 3) fragments and $pC_{1.6}$ is low ($pC_{1.6} = 1.80$).

The last descriptor for this model is PDI or packing density index is a molecular property descriptor. PDI is described as the ratio between the McGowan volume and the total surface area [43]. The descriptor has a positive correlation with $pC_{1.6}$ thereby implicating an enhancing effect on radiosensitivity. This is observed in compounds NI-46 (PDI = 0.922; $pC_{1.6} = 4.28$) and NI-44 (PDI = 0.927; $pC_{1.6} = 4.12$) (Fig. 3c).

The loading plot [44] of the two local PLS models are given in the Supplementary Section (Figs. S1 and S2).

3.2. Modeling the global nitroaromatics dataset

The global dataset, i.e., the dataset containing all the compounds from the individual local datasets was subjected to modeling using Dragon descriptors. The dataset was divided into training and test sets by the Kennard-Stone method of data division, and then the DCV-GA method was utilised for feature selection. The final model was developed using the Best Subset Selection (BSS) method followed by PLS regression. The PLS model with 3 LVs derived exhibited 88.1% variance for the training set (86.5% in terms of leave one out variance) and 92.5%

for the test set variance (in terms of Q_{F1}^2 or R_{pred}^2). The observed versus predicted $pC_{1.6}$ plot for the global model is given in Fig. 5. The residuals of the observed and predicted $pC_{1.6}$ values for some compounds were on the higher side as evident from the scatter plot. However, it was found that all the training set and test set compounds were inside the domain of applicability [47] which will be discussed in later section.

$$pC_{1.6} = 1.256 + 1.318 \times nImidazole + 0.951 \times C - 044 + 0.354 \times B09[C - C] - 0.459 \times B03[O - S]$$

$$N_{train} = 79, R^2 = 0.881, R_{adj}^2 = 0.876, Q_{LOO}^2 = 0.865, Q_{LMO(20\%)}^2 = 0.866, \overline{r_{m(LOO)}^2} = 0.806, \Delta r_{m(LOO)}^2 = 0.101, MAE(train) = 0.262, SD(train) = 0.256, RMSEC = 0.344, Quality_{train} = Good$$

$$N_{test} = 34, Q_{F1}^2 = 0.925, Q_{F2}^2 = 0.899, \overline{r_{m(LOO)}^2} = 0.863, \Delta r_{m(LOO)}^2 = 0.006, CCC = 0.948, MAE(Test) = 0.301, SD(Test) = 0.211, RMSEP = 0.366, Quality_{test} = Good$$

The model is constituted of four descriptors, viz., nImidazole, C-044, B09[C-C] and B03[O-S]. From the VIP plot (Fig. 6a), the descriptors are in the following order of significance: C-044, nImidazole, B03[O-S] and B09[C-C].

The first descriptor is C-044, which is an atom centre fragment descriptor and represented as X-CX.X, where X is any electronegative atom (O, N, S, P, Se, halogens); '-' is an aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group; and '.' is an aromatic

Table 1
Golbraikh and Tropsha's criteria for all local and global models.

| Metrics | Acceptable range | Local Nitrofurans | Local Nitrothiophene | Local Nitroimidazole | Global dataset |
|-----------------------|-----------------------------|-------------------|----------------------|----------------------|----------------|
| r^2 | >0.6 | 0.911 | 0.933 | 0.733 | 0.881 |
| Q^2 | >0.5 | 0.842 | 0.896 | 0.701 | 0.865 |
| $ r_0 - r_0^2 $ | <0.3 | 0.008 | 0.067 | 0.094 | 0.016 |
| k | $0.85 < k < 1.15$ | 1 | 1 | 1 | 1 |
| $[(r^2 - r_0^2)/r^2]$ | $[(r^2 - r_0^2)/r^2] < 0.1$ | 0 | 0 | 0 | 0 |
| k' | $0.85 < k' < 1.15$ | 0.997 | 0.992 | 0.992 | 1 |
| $[(r^2 - r_0^2)/r^2]$ | $[(r^2 - r_0^2)/r^2] < 0.1$ | 0.009 | 0.0045 | 0.129 | 0.018 |

single bond as the C–N bond in pyrrole. Compounds showing positive values for the C-044 descriptor were found to have a specific fragment in their structure, i.e., O=NC–N. Here the O=N fragment represents the delocalized bonds in the nitro group and C–N is an aromatic single bond in pyrrole giving an idea of the 2-nitroimidazole fragment (Fig. 7). Hence, the descriptor C-044 provides a knowledge that nitroimidazoles are better

radiosensitizers having higher radiosensitization effectiveness. Next, **nImidazole** is a functional group descriptor indicating the number of imidazole present in the compound. The positive correlation gives an idea that imidazole group will increase the compounds' radiosensitivity, leading to a conclusion that nitroimidazoles are better radiosensitizers than nitrofurans or nitrothiophenes (Fig. 7).

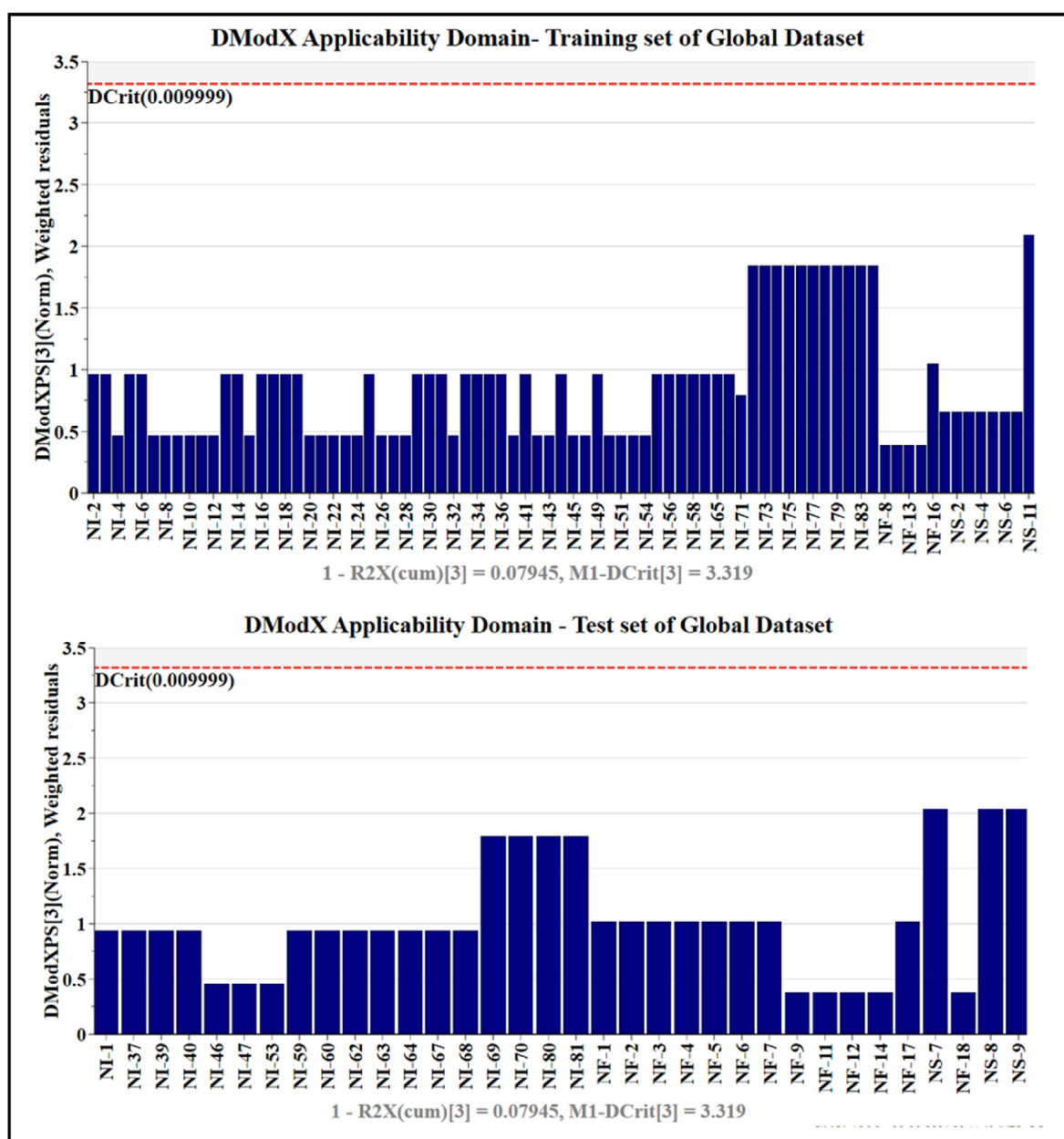


Fig. 8. Applicability domain plot for the global dataset.

Table 2
Y-Randomization model metrics for the developed local and global models.

| Models | | R_y^2 | $Q_{(LOO)y}^2$ |
|--------|----------------|---------|----------------|
| Local | Nitrofurantoin | 0.1727 | -0.3979 |
| | Nitrothiophene | -0.0311 | -0.262 |
| | Nitroimidazole | -0.0132 | -0.248 |
| Global | | -0.0305 | -0.246 |

Another 2D atom pair descriptor **B03 [O-S]** describes the presence or absence of O-S fragment at a topological distance 3. It has a negative correlation with radiosensitization effectiveness denoting that with the presence of such fragment $pC_{1,6}$ value decreases as in compounds **NS-1** ($pC_{1,6} = 1.0$) (Fig. 7) and **NS-2** ($pC_{1,6} = 0.0$).

The next descriptor is **B09 [C-C]**, a 2D atom pair descriptor, which denotes the presence or absence of C-C fragment at the topological distance 9. The positive coefficient indicates that presence of C-C fragment at distance 9 will enhance $pC_{1,6}$ values as seen in compounds like **NI-51** ($pC_{1,6} = 4.3$) (Fig. 7) and **NI-8** ($pC_{1,6} = 4.22$).

From, the descriptors obtained in the global dataset, it can be inferred that nitroimidazoles are better radiosensitizers than nitrofurantoin or

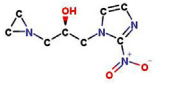
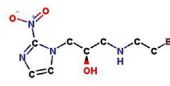
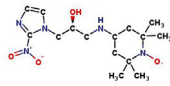
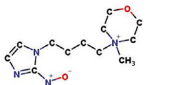
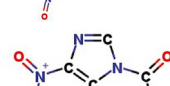
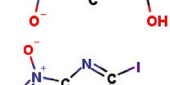
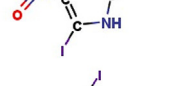
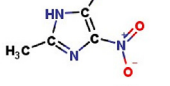
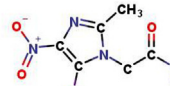
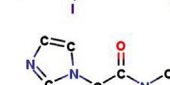
nitrothiophene analogues. Although the global model gives any idea regarding the superiority of nitroimidazoles giving better radiosensitization, the model actually takes into account a diverse group of chemicals. Also, division of dataset gives us more reliance regarding the predictivity of the model.

The loading plot explains the relationship between the descriptors (or the X-variables) with the response (or the Y-variable) [34]. The first two latent variables were utilised for the development of the plot. Through a loading plot, the impact of the descriptors on the response can be understood. Descriptors having similar meaning are grouped together close to one another. This can be explained by Fig. 6b where descriptors C-044 and nImidazole are grouped together and they almost impart the same meaning (contribution of imidazole group). Descriptors with high impact on the model are situated far from the plot origin (e.g., C-044 and nImidazole).

3.2.1. Golbraikh and Tropsha's criteria

We have calculated the Golbraikh-Tropsha's criteria [45] for all the local models as well as for the global model and reported in Table 1. All the models developed in the present study passed the criteria.

Table 3
Predicting $pC_{1,6}$ values of a true external dataset using the global model.

| Compound ID | Structure | Observed $pC_{1,6}$ | Predicted $pC_{1,6}$ (Global model) | Composite Score | Prediction Quality | AD status | Reference |
|-------------|---|---------------------|-------------------------------------|-----------------|--------------------|-----------|-----------|
| 1 |  | - | 3.879 | 3 | Good | In | [26] |
| 2 |  | - | 3.879 | 3 | Good | In | [26] |
| 3 |  | 4.05 | 3.879 | 3 | Good | In | [28] |
| 4 |  | 2.89 | 3.879 | 3 | Good | In | [28] |
| 5 |  | - | 2.574 | 3 | Good | In | [50] |
| 6 |  | - | 3.525 | 3 | Good | In | [50] |
| 7 |  | - | 2.574 | 3 | Good | In | [50] |
| 8 |  | - | 2.928 | 3 | Good | In | [50] |
| 9 |  | - | 3.879 | 3 | Good | In | [51] |
| 10 |  | - | 3.879 | 3 | Good | In | [51] |

3.2.2. Applicability domain (AD) assessment

In accordance with OECD guideline 3, any QSAR model should hold a defined domain of applicability. AD can be interpreted as a chemical space defined by the structural information or molecular properties of the chemicals used in the model development [46]. Any compound which is present within the chemical space can only be properly predicted. In the present study, for the nitrofurans data set, we have used the standardization approach [47]. There was no outlier found for the nitrofurans dataset. In case of local nitrothiophenes, local nitroimidazoles and global datasets, we have applied the DModX (distance to model in X-space) method of AD determination at 99% confidence interval ($D\text{-crit} = 0.009999$) using SIMCA 16.0.2 software (<https://landing.umatrics.com/downloads-simca>). The AD plots for the two local datasets given in Figs. S3 and S4 (in Supplementary Section) show that there was no outlier. In case of the global dataset as shown in Fig. 8, it was observed that there was neither any outlier in the training set nor any compound was outside the AD in the test set.

3.2.3. Y-randomization test

The significance of a developed QSAR model is understood by a model randomization test, and it ensures that the model is not an outcome of a chance correlation [48]. During the development of a randomized model, many models are generated by reordering or shuffling different combination of X- or Y-variables (Y-variable here) and accordingly are called X-randomization or Y-randomization. In the present work, we have used 100 permutations for all the developed models; however, this can be changed according to the choice of the user. Models which are randomly developed with y-variable shuffling should have very poor statistics. The R^2 intercept should not exceed 0.3 and the Q^2 intercept should not exceed 0.05. The metrics for the randomized models given in Table 2 and Supplementary Figs. S5 and S6 and S7 indicate that the local and global models developed are not out of chance correlation and are robust for suitable predictions.

3.2.4. True external prediction using the global model

The global model can be considered the best model here, owing to the diversity of the nitro compounds used for modeling. Further, to analyse the predictivity of the developed global model, we have considered a set of external compounds for prediction (Table 2). Predictions for these compounds were further verified by the application of "Prediction Reliability Indicator" tool [49] available from <https://dtclab.webs.com/software-tools>. The PRI results showed that predictions for all the 10 compounds were 'Good' (with Composite Score 3) and all the compounds were inside the AD of the model (Table 3). Based on the insights obtained, it can be inferred that developed global model can be used for the prediction of radiosensitization effectiveness in nitro compounds, especially for nitroimidazole derivatives. We have further computed predictions using the global model for another external dataset retrieved from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>) and checked the quality of predictions using the PRI tool. We have reported the results in the Supplementary Section (Table S2). Of note, the prediction quality was found to be good for all the external compounds. It will be interesting to verify the predictions experimentally in the future.

4. Conclusion

The present study targets for the development of 2D-QSAR models for three nitroaromatics datasets both locally and globally predicting radiosensitization effectiveness. The local models gave us an idea about the structural features required for effective radiosensitization within their own group while the global model imparted an insight regarding which type of nitroaromatic compounds are more efficient to produce better radiosensitization. The descriptors obtained in the global model clearly implicated that nitroimidazoles are better radiosensitizers as compared with nitrofurans or nitrothiazole derivatives. Moreover, all the developed

local and global models were statistically sound and well validated. The global model was further used for the prediction of two sets of external compounds, and their prediction reliability was analysed using the PRI tool.

Funding information

PD thanks Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship (File No.- ISRM/11(61)/2019).

Declaration of competing interest

The authors declare that they have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejmcr.2022.100035>.

References

- [1] F.J. Roe, Toxicologic evaluation of metronidazole with particular reference to carcinogenic, mutagenic, and teratogenic potential, *Surgery* 93 (1983) 158–164. [https://www.surgjournal.com/article/0039-6060\(83\)90294-5/pdf](https://www.surgjournal.com/article/0039-6060(83)90294-5/pdf).
- [2] A. Nunn, K. Linder, H.W. Strauss, Nitroimidazoles and imaging hypoxia, *Eur. J. Nucl. Med.* 22 (1995) 265–280. <https://doi.org/10.1007/BF01081524>.
- [3] M. Chin Chung, P. Longhin Bosquesi, J. Leandro dos Santos, A prodrug approach to improve the physico-chemical properties and decrease the genotoxicity of nitro compounds, *Curr. Pharmaceut. Des.* 17 (2011) 3515–3526. <https://doi.org/10.2174/138161211798194512>.
- [4] P.J.C.O. Wardman, Chemical radiosensitizers for use in radiotherapy, *Clin. Oncol.* 19 (2007) 397–417. <https://doi.org/10.1016/j.clon.2007.03.010>.
- [5] W.R. Wilson, M.P. Hay, Targeting hypoxia in cancer therapy, *Nat. Rev. Cancer* 11 (2011) 393–410. <https://doi.org/10.1038/nrc3064>.
- [6] L.K. Kvols, Radiation sensitizers: a selective review of molecules targeting DNA and non-DNA targets, *J. Nucl. Med.* 46 (2005) 187S–190S.
- [7] J.D. Chapman, A.P. Reuvers, J. Borsa, Effectiveness of nitrofurans derivatives in sensitizing hypoxic mammalian cells to X rays, *Br. J. Radiol. Suppl.* 46 (1973) 623–630. <https://doi.org/10.1259/0007-1285-46-548-623>.
- [8] M. Langenbacher, R.J. Abdel-Jalil, W. Voelter, M. Weinmann, S.M. Huber, In vitro hypoxic cytotoxicity and hypoxic radiosensitization, *Strahlenther. Onkol.* 189 (2013) 246–255. <https://doi.org/10.1007/s00066-012-0273-2>.
- [9] A. Breccia, F. Busi, E. Gattavecchia, M. Tamba, Reactivity of nitro-thiophene derivatives with electron and oxygen radicals studied by pulse radiolysis and polarographic techniques, *Radiat. Environ. Biophys.* 29 (1990) 153–160. <https://doi.org/10.1007/BF01210519>.
- [10] W.W. Wong, R.K. Jackson, L.P. Liew, B.D. Dickson, G.J. Cheng, B. Lipert, Y. Gu, F.W. Hunter, W.R. Wilson, M.P. Hay, Hypoxia-selective radiosensitization by SN38023, a bioreductive prodrug of DNA-dependent protein kinase inhibitor IC87361, *Biochem. Pharmacol.* 169 (2019) 113641. <https://doi.org/10.1016/j.bcp.2019.113641>.
- [11] E. Davila, L. Klein, C.L. Vogel, R. Johnson, F. Ostroy, S. Browning, E. Gorowski, R.L. Furner, C.A. Presant, Phase I trial of misonidazole (NSC# 261037) plus cyclophosphamide in solid tumor, *Jpn. J. Clin. Oncol.* 3 (1985) 121–127. <https://doi.org/10.1200/JCO.1985.3.1.121>.
- [12] S.I. Masunaga, Y. Uto, H. Nagasawa, H. Hori, K. Nagata, M. Suzuki, Y. Kinashi, K. Ono, Evaluation of hypoxic cell radio-sensitizers in terms of radio-sensitizing and repair-inhibiting potential. Dependency on p53 status of tumor cells and the effects on intratumor quiescent cells, *Anticancer Res.* 26 (2006) 1261–1270.
- [13] P. Gramatica, E. Papa, QSAR modeling of bioconcentration factor by theoretical molecular descriptors, *QSAR Comb. Sci.* 22 (2003) 374–385. <https://doi.org/10.1002/qsar.200390027>.
- [14] T. Ginex, J. Vazquez, E. Gilbert, E. Herrero, F.J. Luque, Lipophilicity in drug design: an overview of lipophilicity descriptors in 3D-QSAR studies, *Future Med. Chem.* 11 (2019) 1177–1193. <https://doi.org/10.4155/fmc-2018-0435>.
- [15] K. Mansouri, C.M. Grulke, R.S. Judson, A.J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, *J. Cheminf.* 10 (2018) 1–19. <https://doi.org/10.1186/s13321-018-0263-1>.
- [16] V. Kumar, K. Roy, Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel 3C-like protease (3CLpro) enzyme inhibitors against SARS-CoV diseases, *SAR QSAR Environ. Res.* 31 (2020) 511–526. <https://doi.org/10.1080/1062936X.2020.1776388>.
- [17] A. Seth, P.K. Ojha, K. Roy, QSAR modeling with ETA indices for cytotoxicity and enzymatic activity of diverse chemicals, *J. Hazard Mater.* 394 (2020) 122498. <https://doi.org/10.1016/j.jhazmat.2020.122498>.

- [18] M. Borrotti, D. De March, D. Slanzi, I. Poli, Designing lead optimisation of MMP-12 inhibitors, 2014, *Comput. Math Methods Med.* (2014), <https://doi.org/10.1155/2014/258627>.
- [19] C. Hansch, A. Leo, S.B. Mekapati, A. Kurup, Qsar and adme, *Bioorg. Med. Chem.* 12 (2004) 3391–3400, <https://doi.org/10.1016/j.bmc.2003.11.037>.
- [20] C. Klein, D. Kaiser, S. Kopp, P. Chiba, G.F. Ecker, Similarity based SAR (SIBAR) as tool for early ADME profiling, *J. Comput. Aided Mol. Des.* 16 (2002) 785–793, <https://doi.org/10.1023/A:1023828527638>.
- [21] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: an easy approach to molecular descriptor calculations, *Match* 56 (2006) 237–248.
- [22] P. Ambure, A. Gajewicz-Skretna, M.N.D. Cordeiro, K. Roy, New workflow for QSAR model development from small data sets: small dataset curator and small dataset modeler. integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques, *J. Chem. Inf. Model.* 59 (2019) 4070–4076, <https://doi.org/10.1021/acs.jcim.9b00476>.
- [23] J. Devillers, *Genetic Algorithms in Molecular Modeling*, Academic Press, Cornwall, Great Britain, 1996.
- [24] V. Venkatasubramanian, A. Sundaram, Genetic algorithms: introduction and applications, *Enc. Comput. Chem.* (2002) 2, <https://doi.org/10.1002/0470845015.cga003>.
- [25] K. Roy, P. Ambure, The “double cross-validation” software tool for MLR QSAR model development, *Chemometr. Intell. Lab. Syst.* 159 (2016) 108–126, <https://doi.org/10.1016/j.chemolab.2016.10.009>.
- [26] M.A. Naylor, M.A. Stephens, S. Cole, M.D. Threadgill, I.J. Stratford, P. O'Neill, E.M. Fielden, G.E. Adams, Synthesis and evaluation of novel electrophilic nitrofurans carboxamides and carboxylates as radiosensitizers and bioreductively activated cytotoxins, *J. Med. Chem.* 33 (1990) 2508–2513, <https://doi.org/10.1021/jm00171a027>.
- [27] M.D. Threadgill, P. Webb, P. O'Neill, M.A. Naylor, M.A. Stephens, I.J. Stratford, S. Cole, G.E. Adams, E.M. Fielden, Synthesis of a series of nitrothiophenes with basic or electrophilic substituents and evaluation as radiosensitizers and as bioreductively activated cytotoxins, *J. Med. Chem.* 34 (1991) 2112–2120, <https://doi.org/10.1021/jm00111a029>.
- [28] W. Long, P. Liu, Quantitative structure activity relationship modeling for predicting radiosensitization effectiveness of nitroimidazole compounds, *J. Radiat. Res.* 51 (2010) 563–572, <https://doi.org/10.1269/jrr.10053>.
- [29] MarvinSketch software, <https://www.chemaxon.com> (accessed 29 March 2021).
- [30] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, 11, John Wiley & Sons, New Jersey, United States, 2008.
- [31] Kodesrl, Milan, Italy, Dragon Version 7, 2016, <http://www.taletе.mi.it/index.htm>. (Accessed 31 March 2021).
- [32] K.D. Baumann, K. Baumann, Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation, *J. Cheminf.* 6 (2014) 1–19, <https://doi.org/10.1186/s13321-014-0047-1>.
- [33] L.S. Aiken, S.G. West, S.C. Pitts, Multiple Linear Regression, *Handbook of psychology*, 2003, pp. 481–507, <https://doi.org/10.1002/0471264385.wei0219>.
- [34] S. Wold, M. Sjöstöm, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometr. Intell. Lab. Syst.* 58 (2001) 109–130, [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1).
- [35] P. De, D. Bhattacharyya, K. Roy, Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modelling, *Struct. Chem.* 31 (2020) 1043–1055, <https://doi.org/10.1007/s11224-019-01481-z>.
- [36] A. Saptoro, M.O. Tade, H. Vuthaluru, A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models, *Chem. Prod. Process Model.* 7 (2012), <https://doi.org/10.1515/1934-2659.1645>.
- [37] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, *Exp. Opin. Drug Discov.* 2 (2007) 1567–1577, <https://doi.org/10.1517/17460441.2.12.1567>.
- [38] K. Roy, S. Kar, R.N. Das, *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*, Springer, New York, United States, 2015.
- [39] K. Roy, I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, *Comb. Chem. High Throughput Screen.* 14 (2011) 450–474, <https://doi.org/10.2174/138620711795767893>.
- [40] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm2 metrics for validation of QSPR models, *Chemometr. Intell. Lab. Syst.* 107 (2011) 194–205, <https://doi.org/10.1016/j.chemolab.2011.03.011>.
- [41] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, *Chemometr. Intell. Lab. Syst.* 152 (2016) 18–33, <https://doi.org/10.1016/j.chemolab.2016.01.008>.
- [42] N. Akarachantachote, S. Chadcham, K. Saithanu, Cutoff threshold of variable importance in projection for variable selection, *Int. J. Pure Appl. Math.* 94 (2014) 307–322, <https://doi.org/10.12732/ijpam.v94i3.2>.
- [43] A.J. A. Pirovano, S. Brandmaier, M.A. Huijbregts, A.M. Ragas, K. Veltman, A.J. Hendriks, The utilisation of structural descriptors to predict metabolic constants of xenobiotics in mammals *Environ. Toxicol. Pharmacol.* 39 (2015) 247–258, <https://doi.org/10.1016/j.etap.2014.11.025>.
- [44] P. De, R.B. Aher, K. Roy, Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector *Aedes aegypti*: application of ETA indices, *RSC Adv.* 8 (2018) 4662–4670, <https://doi.org/10.1039/C7RA13159C>.
- [45] A. Golbraikh, A. Tropsha, Beware of q²_L, *J. Mol. Graph. Model.* 20 (2002) 269–276.
- [46] D. Gadaleta, G.F. Mangiatordi, M. Catto, A. Carotti, O. Nicolotti, Applicability domain for QSAR models: where theory meets reality, *IJQSPR* 1 (2016) 45–63, <https://doi.org/10.4018/IJQSPR.2016010102>.
- [47] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, *Chemometr. Intell. Lab. Syst.* 145 (2015) 22–29, <https://doi.org/10.1016/j.chemolab.2015.04.013>.
- [48] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure–activity relationships, *J. Med. Chem.* 22 (1979) 1238–1244, <https://doi.org/10.1021/jm00196a017>.
- [49] K. Roy, P. Ambure, S. Kar, How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? *ACS Omega* 3 (2018) 11392–11406, <https://doi.org/10.1021/acsomega.8b01647>.
- [50] W. Krause, A. Jordan, R. Scholz, J.L.M. Jimenez, Iodinated nitroimidazoles as radiosensitizers, *Anticancer Res.* 25 (2005) 2145–2151.
- [51] J.M. Brown, Y.Y. Ning, D.M. Brown, W.W. Lee, SR-2508: a 2-nitroimidazole amide which should be superior to misonidazole as a radiosensitizer for clinical use, *Int. J. Radiat. Oncol. Biol. Phys.* 7 (1981) 695–703, [https://doi.org/10.1016/0360-3016\(81\)90460-0](https://doi.org/10.1016/0360-3016(81)90460-0).



Prediction reliability of QSAR models: an overview of various validation tools

Priyanka De¹ · Supratik Kar² · Pravin Ambure³ · Kunal Roy¹

Received: 28 December 2021 / Accepted: 14 February 2022 / Published online: 10 March 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

The reliability of any quantitative structure–activity relationship (QSAR) model depends on multiple aspects such as the accuracy of the input dataset, selection of significant descriptors, the appropriate splitting process of the dataset, statistical tools used, and most notably on the measures of validation. Validation, the most crucial step in QSAR model development, confirms the reliability of the developed QSAR models and the acceptability of each step in the model development. The present review deals with various validation tools that involve multiple techniques that improve the model quality and robustness. The double cross-validation tool helps in building improved quality models using different combinations of the same training set in an inner cross-validation loop. This exhaustive method is also integrated for small datasets (< 40 compounds) in another tool, namely the small dataset modeler tool. The main aim of QSAR researchers is to improve prediction quality by lowering the prediction errors for the query compounds. ‘Intelligent’ selection of multiple models and consensus predictions integrated in the intelligent consensus predictor tool were found to be more externally predictive than individual models. Furthermore, another tool called Prediction Reliability Indicator was explained to understand the quality of predictions for a true external set. This tool uses a composite scoring technique to identify query compounds as ‘good’ or ‘moderate’ or ‘bad’ predictions. We have also discussed a quantitative read-across tool which predicts a chemical response based on the similarity with structural analogues. The discussed tools are freely available from <https://dtclab.webs.com/software-tools> or http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/ and <https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home> (for read-across).

Keywords QSAR · Validation · Double cross-validation · Small dataset modeling · Intelligent consensus prediction · Read across

Introduction

A growing number of research have been conducted in recent years, wherein computational methods have been used to predict the physicochemical properties and biological activities of chemical compounds. Quantitative structure–activity relationship (QSAR) (Dearden 2016) modeling

is a popular in silico technique performed to find out a quantitative correlation between the structural features (known as descriptors) and a known response (activity/property/toxicity) for a set of molecules using various chemometric methodologies. QSAR evolves at the crossroads of chemistry, statistics, biology, and toxicological studies. The main aim is to identify and optimize new leads to shorten the time and reduce expenditure for drug discovery (Hsu et al. 2017). The fundamental assumption regarding QSAR modeling is that a chemical structure possesses unique features (geometric, steric, and electronic properties) responsible for its physical, chemical, and biological properties.

The European Union (EU) envisaged that QSAR models would increasingly be used for hazard and risk assessments of chemicals (Commission of the European Communities 2001). It is also necessary to create and apply QSARs to address animal welfare concerns by replacing, reducing,

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

¹ Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

² Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS 39217, USA

³ ProtoQSAR S.L., Valencia, Spain

and refining animal testing in toxicological assessments. In November 2004, the European Commission and the OECD (Organisation for Economic Co-operation and Development) member countries adopted principles for validation of QSAR models for use in regulatory assessment of chemical safety (Organisation for Economic Co-operation and Development (OECD 2004). According to the agreed guidelines of OECD, a QSAR model should be developed with

- (a) A defined endpoint,
- (b) An unambiguous algorithm to guarantee model transparency,
- (c) A defined domain of applicability,
- (d) Proper measures of validation including internal performance (as determined by goodness-of-fit and robustness) and predictivity (as represented by external validation), and
- (e) Possible mechanistic interpretation.

Validation is crucial for the development and application of any QSAR model. It confirms the reliability of the developed model and the acceptability of each step through model development. The debate between internal versus external validation prevails predominantly among QSAR practitioners (Roy 2007). Some QSAR studies reported an inconsistency between internal and external predictivity (Novellino et al. 1995; Norinder 1996). According to researchers, there might be an inconsistency between internal and external predictability, i.e., high internal predictivity may result in low external predictivity and vice versa (Kubinyi 1998). However, external validation is considered the ‘gold standard’ of checking predictive potential of QSAR models. Some researchers consider cross-validation to be more appropriate for checking the predictive ability of QSAR models to circumvent the loss of information from splitting the dataset into training and test sets (Héberger 2017). Several validation metrics (as discussed later) are used to check the quality of predictions generated by regression-based and classification-based QSAR models (Gramatica and Sangion 2016; Todeschini et al. 2016).

The present review has discussed several prediction reliability tools exploring various strategies to determine model reliability and predictivity. We have discussed the tools that engage in the model-building through a double cross-validation approach on large and small datasets. Furthermore, we have explained the utility of intelligent selection of multiple models and various forms of consensus prediction. We have also mentioned a tool that explains a similarity-based reliability scoring approach to understand the quality of predictions for a new query compound and ensure the developed models’ reliability. We have further reported a

similarity-based quantitative read-across tool addressing the quality of predictions both quantitatively and qualitatively.

Predictive QSAR model development approaches

Modern QSAR methods use multiple descriptors combined with the application of both linear and non-linear modeling approaches with a strong emphasis on rigorous model validation to afford robust and predictive QSAR models. Several types of research along with our understanding of QSAR model development and validation led us to establish a general outline of QSAR model workflow as described in Fig. 1. This figure illustrates the classical QSAR model development algorithm which includes: (a) collection of pertinent data with a defined endpoint, (b) descriptor calculation and data pre-treatment, (c) model development through analysis of the correlation between input data and descriptors calculated, (d) validation of the model, and (e) design and prediction of the activity of new query molecules. The QSAR modeling scheme is further described briefly in the following section.

- (i) **Dataset preparation and data curation:** One of the most challenging parts of QSAR is dataset collection with a “defined endpoint” as explained in OECD principle 1. The intent is to confirm the transparency of the endpoint aimed for prediction models, considering that a given endpoint could be dependent on the experimental protocol and the experimental conditions. Data curation is an essential and time-consuming step in the QSAR model development process. Erroneous data (both in chemical structures and biological data) retrieved from online sources require strict curation to avoid false or non-predictive models (Ambure and Cordeiro 2020).
- (ii) **Calculation of molecular descriptors:** The molecular structures applied for QSAR modeling need to be translated into numbers, i.e., molecular descriptors. The molecular descriptor is an encoded representation of the information about a chemical compound in the form of numerical values based on its chemical constitution, allowing the correlation of chemical structure with physical properties, chemical reactions, or biological activity (Consonni and Todeschini 2010). In a QSAR model, descriptors of a molecule, which describe specific aspects of a molecule, are predictors (X) of the dependent variable (Y). A QSAR study uses a variety of descriptors that can be classified into different dimensions or categories, as shown in Table 1.

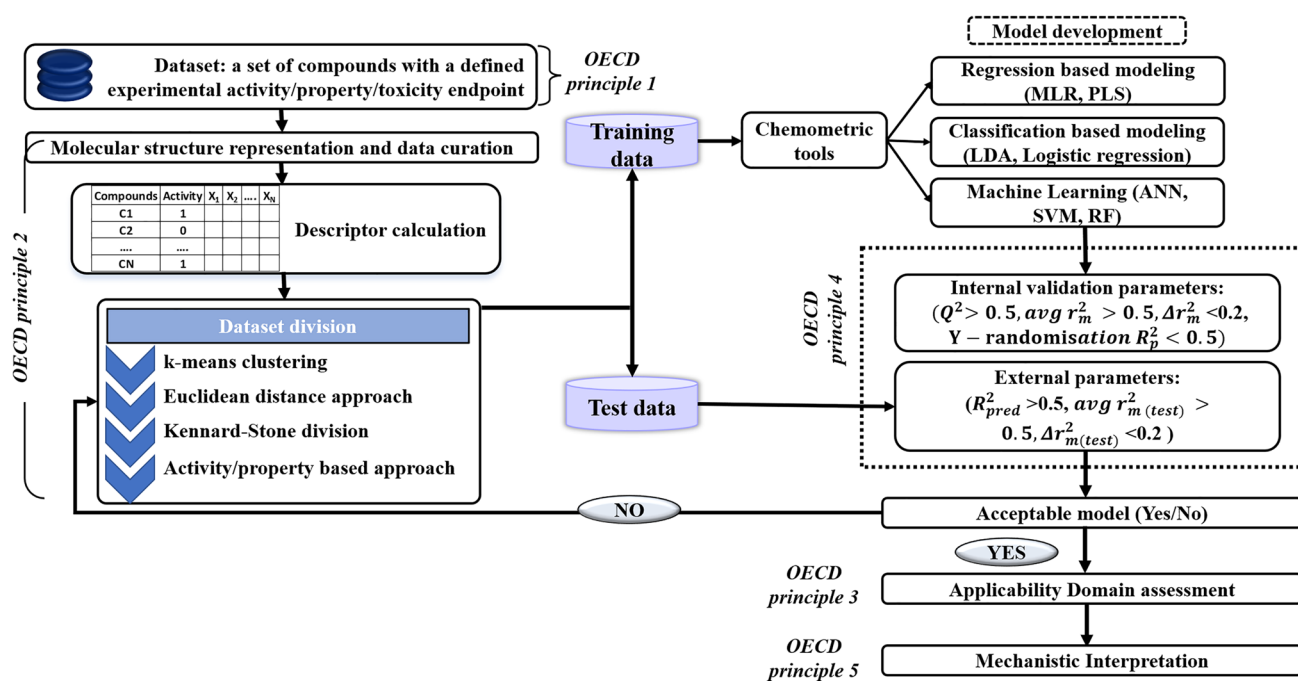


Fig. 1 Schematic representation of QSAR methodology according to OECD guidelines

- (iii) **Dataset division:** A predictive model's performance must be determined by dividing the dataset into a training set and a test set. Among all chemicals, only the training set molecules are used for developing QSAR models, and the external predictivity of the models is examined through the use of test set compounds. In developing the QSAR model, it is necessary to select a training set in a way, such that it encompasses a wide chemical domain. The test set compounds must lie within the chemical space of the training set. Dataset division involves different methods including (a) Euclidean distance (diversity-based) (Golmohammadi et al. 2012), (b) Kennard-Stone (Kennard and Stone 1969), (c) k-means clustering (Likas et al. 2003), (d) sorted response (Roy 2018), etc.
- (iv) **Feature selection:** A feature selection process is a vital step that involves identifying important predictor variables to develop correlations with the response variable. Feature selection helps decrease the model complexity, decreases the risk of overfitting or overtraining, and helps select the most critical descriptors among a pool of hundreds or thousands. In this way, the dimensionality of input descriptors is minimized without the loss of essential information (Goodarzi et al. 2012). Finally, these selected descriptors are used to build a mathematical model linking to the biological activity

of the corresponding compounds. According to the OECD guidelines, several feature selection techniques have been applied using a mechanistic basis including, genetic algorithms, genetic function approximation (GFA), forward selection, backward elimination, stepwise regression, simulated annealing, etc.

- (v) **Model development algorithms:** The OECD guideline 2 explains that a QSAR model should be developed using an “unambiguous algorithm” (Directorate 2007). The rule focuses on bringing transparency in model-building, rendering it reproducible to others and making it possible to achieve the endpoint estimates. This embraces the methods implemented during data pre-treatment, division of data, feature selection, and model development. Linear modeling techniques involve multiple linear regression (MLR) (Pope and Webster 1972; De and Roy 2018), ordinary least squares (OLS), partial least squares (PLS) (Wold et al. 2001), principal component analysis (PCA) (Abdi and Williams 2010), principal component regression (PCR), etc.

In QSAR, model-building tools can be grouped into two major categories: regression-based approach and classification-based approach. Regression-based approaches are effective when both dependent (response variable) and independent (molecular descriptors) variables are quantitative (Roy et al. 2015a; b). In the case of classification-based modeling, a relationship between the descriptors and the graded values

Table 1 Types of 0D-3D descriptors used in the QSAR study

| Dimension of descriptors | Parameters | Examples | |
|-----------------------------------|---|--|---|
| 0D | Constitutional indices | Number of atoms, number of non-H atoms, number of bonds, number of aromatic bonds, sum of atomic van der Waals volumes (scaled on carbon atom), etc. | |
| | Molecular property Atom and bond counts | Unsaturation count, unsaturation index, hydrophilic factor, unsaturation index | |
| 1D | Fragment counts, fingerprints | Atom centered fragments (C-001, H-046, O-056, etc.) | |
| 2D | Topological | Wiener index (W), Zagreb group indices, Balaban <i>J</i> index, Randic branching index (χ), Molecular connectivity index, subgraph count, Chi indices, etc. | |
| | Structural | Chiral centers, rotatable bonds, HBond donor, HBond acceptor | |
| | Physicochemical parameters (thermodynamic parameters) | Heat of formation (Hf), Log of the partition coefficient using Ghose and Crippen's method (AlogP), Desolvation free energy (Fh2o, Foct) | |
| | Connectivity indices | Average connectivity index, valence connectivity index, solvation connectivity index, modified Randic index, connectivity topochemical index, perturbation connectivity index | |
| | Functional group counts | Number of terminal primary C(sp ³), number of total secondary C(sp ³), number of ring quaternary C(sp ³), number of carboxylic acids, number of hydroxyl groups, etc. | |
| | 2D matrix based | Balaban-like index from adjacency matrix, logarithmic spectral positive sum from adjacency matrix, spectral absolute deviation from adjacency matrix, etc. | |
| | 2D atom pairs | Presence or absence of any two atoms at a particular topological distance (B01[C-C], B09[C-F], etc.), frequency of two atoms at a particular topological distance (F01[C-F], F05[O-N]), sum of occurrence of two atoms at a particular topological distance (T(N..I), T(O..N)) | |
| | 3D | Electronic | Dipole moment, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), superdelocalizability |
| | | Spatial | The radius of gyration, Jurs descriptors, area, density, volume, etc. |
| | | Receptor surface analysis parameters | Hydrophobicity, partial charge, electrostatic (ELE) potential, van der Waals (VDW) potential, and hydrogen bonding propensity |
| Molecular shape analysis | | Difference volume (DIFFV), Common overlap steric volume (COSV), Common overlap volume ratio (Fo), Noncommon overlap steric volume (NCOSV), Root mean square to shape reference (ShapeRMS) | |
| Geometric Other 3D descriptors | | Molecular eccentricity, sphericity, asphericity, aromaticity index, gravitational index 3D matrix based (Wiener like index, Randic like index, Balaban-like index, etc. all from geometric matrix, spectral moment,), 3D autocorrelations (3D Topological distance-based descriptors: unweighted; weighted by mass, polarizability, van der Waals volume, Sanderson electronegativity, ionization potential), 3D Morse descriptors, WHIM descriptors, GETAWAY descriptors, quantum-chemical descriptors | |

0D, 1D, and 2D descriptors may be collectively grouped under the broad class of 2D descriptors in general

of the response variable(s) is established. Here, the response is offered in a Boolean form like active/inactive and positive/negative or categorical (as observed in linear discriminant analysis, logistic regression, and cluster analysis).

- (vi) **Determination of domain of applicability:** One of the most essential checkpoints in QSAR modeling is determining the applicability domain (AD) of a model as explained in OECD principle 3. The applicability domain denotes a physicochemical space (both the response and chemical structure space) within which a QSAR model can predict with a certain degree of reliability (Roy et al. 2015a, b). This space is defined by the features explained by the

compounds in the training set and is mandatory to examine whether the prediction of test set molecules is reliable or not. The concept of AD was used to avoid an unjustified extrapolation of property predictions.

- (vii) **QSAR model validation:** Before interpreting and predicting biological responses of untested compounds, any QSAR model needs to be validated. Here, the model's predictive power is established, and the ability to reproduce the biological activities of the untested compounds is measured. In consonance with the fourth principle of OECD guidelines, statistical validation of models in terms of goodness-of-fit, robustness, and predictivity is an extremely impor-

tant step during QSAR model development. The validation of QSAR models is crucial if these models are used for virtual screening. Each validation parameter aims to judge the accuracy of prediction, i.e., determining whether the experimental value is close to the model-derived value. The model fitness determined using the coefficient of determination or correlation coefficient from the training set measures the degree of achieved correlation between the experimental (Y_{exp}) and calculated (Y_{calc}) response values. Data fitting does not confirm the predictability of a model but instead demonstrates the model's statistical quality. Different internal and external validation metrics for both regression and classification modeling are utilized to check model prediction quality which is discussed later in the following section.

- (viii) **Mechanistic interpretation:** The fifth OECD principle focuses on identifying the features of the variables that may contribute to a more thorough understanding of the response being modeled. Chemicals that act specifically using a specific mechanism can only be designed and developed with absolute certainty using the structural analogues. However, it is evident that furnishing mechanistic information may not always be feasible. The rule suggests that the modeler should report if any such information is available, facilitating future research on that endpoint. A mechanistic interpretation from the literature can be added, and therefore, the fifth OECD principle encourages the reporting of such information to enrich the physicochemical understanding of response being modeled.

Regression and classification validation metrics

The reliability of a developed QSAR model is confirmed through the validation process. The quality of input data, dataset diversity, predictability on an external set, applicability domain determination, and mechanistic interpretability are also confirmed through various validation metrics. QSAR model validation can be classified into two major types: (a) internal validation and (b) external validation. Internal validation in QSAR modeling involves activity prediction of the molecules/compounds used for generating the model. This is followed by estimating metrics for detecting the precision of predictions. Internal validation is useful in the case of cross-validation approaches (Konovalov et al. 2008) where the internal data are partitioned into calibration (training) and validation (test) subsets. The calibration set is used for model-building purposes, and the validation set is utilized for model predictivity assessment. Assessment of

prediction capability and applicability of a QSAR model to predict newly designed or untested molecules is done using external validation metrics. In most cases, some compounds from the original datasets are used for validation purpose when true external data points are limited or not available.

Regression-based validation metrics

One of the main quality metrics to check the goodness-of-fit of a regression model is the determination coefficient (R^2) which measures the variation of observed data with the fitted ones. The maximum possible value for R^2 is 1, which defines a perfect correlation.

Adjusted R^2 (R^2_{adj}) is a modified version of the determination coefficient and is also known as the explained variance. The R^2_{adj} parameter incorporates the information of the number of samples and the independent variables used in the model.

Considering the internal validation for a regression-based QSAR model, the leave-one-out cross-validation (Q^2_{LOO}) metric is calculated. Here, a model is developed by modifying the original training set of n compounds by removing one compound. The activity of the omitted compound is then predicted using the model developed with $n-1$ compounds. This cycle is repeated until all the training set compounds have been eliminated once and the predicted activity data are obtained for all the training set compounds. The model predictivity is thus measured using the predicted residual sum of squares (PRESS) and cross-validated R^2 (Q^2) (Table 2). The PRESS value is defined as the sum of squared differences between the experimental and leave-one-out predicted data. The standard deviation of error of predictions (SDEP) is calculated from the PRESS value (Table 2). A model is considered satisfactory if the value of Q^2 is higher than the predetermined value of 0.6. However, numerous evidences suggested that leave-one-out prediction should neither be considered as the ultimate standard for judging the predictive power of models nor for model selection (Konovalov et al. 2007; Veerasamy et al. 2011). There is a chance of overfitting and overestimation in LOO due to structural redundancy (Höltje and Sippl 2001). Leave-many-out (LMO) or leave-some-out (LSO) might be a better alternative where a part of the training data is held out ($(1 \leq m < n$, where n is a sample size) and the remaining data are modeled. The model is developed using the remaining compounds in each cycle, and the hold-out compounds are predicted. This cycle continues till all the compounds are predicted, and the predicted values are used for the calculation of Q^2_{LMO} . Therefore, the LMO technique partly reflects external validation in the context of internal validation.

Although, Q^2_{LOO} provides a measure of model robustness, it may not be sufficient to characterize the performance of the model during prediction of new query/test compounds. Furthermore, Q^2_{LOO} can provide an overestimation of model

Table 2 Validation metrics for regression modeling

| Parameters | Equation | Description |
|---|---|--|
| Determination coefficient (R^2) | $R^2 = 1 - \frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum (Y_{\text{obs}} - \bar{Y}_{\text{training}})^2}$ | Metric to check the goodness-of-fit of a regression model. It measures the variation of observed data with the predicted ones. The maximum possible value for R^2 is 1, which defines a perfect correlation. Y_{obs} denotes the observed response values for the training set, and Y_{pred} denotes the calculated response values for the training set of compounds. $\bar{Y}_{\text{training}}$ is the mean observed response of the training set compounds |
| Explained variance or adjusted R^2 (R^2_{adj}) | $R^2_{\text{adj}} = \frac{\{(n-1)R^2\} - p}{n-p-1}$ | Modified version of the determination coefficient. The R^2_{adj} parameter incorporates the information of the number of samples and the independent variables used in the model. n is the number of training set compounds and p is the number of predictor variables |
| Leave-one-out cross-validation (Q^2_{LOO}) | $Q^2_{\text{LOO}} = 1 - \frac{\sum (Y_{\text{obs}(\text{training})} - Y_{\text{pred}(\text{training})})^2}{\sum (Y_{\text{obs}(\text{training})} - \bar{Y}_{\text{training}})^2}$ | Cross-validated $R^2(Q^2)$ is checked for internal validation. $Y_{\text{obs}(\text{training})}$ is the observed response, and $Y_{\text{pred}(\text{training})}$ is the predicted response of the training set molecules based on the leave-one-out (LOO) technique |
| Predictive R^2 or R^2_{pred} or $Q^2_{\text{ext}(F1)}$ | $Q^2_{\text{ext}(F1)} = 1 - \frac{\sum (Y_{\text{obs}(\text{test})} - Y_{\text{pred}(\text{test})})^2}{\sum (Y_{\text{obs}(\text{test})} - \bar{Y}_{\text{training}})^2}$ | This metric employed for judging external predictivity. It is a measure of correlation between the observed and predicted data of test set. $Y_{\text{obs}(\text{test})}$ is the observed response, and $Y_{\text{pred}(\text{test})}$ is the predicted response of the test set molecules. $\bar{Y}_{\text{training}}$ denotes the mean observed response of the training set |
| $Q^2_{\text{ext}(F2)}$ | $Q^2_{\text{ext}(F2)} = 1 - \frac{\sum (Y_{\text{obs}(\text{test})} - Y_{\text{pred}(\text{test})})^2}{\sum (Y_{\text{obs}(\text{test})} - \bar{Y}_{\text{test}})^2}$ | It helps in the judgment of predictivity of a model using the test set mean (\bar{Y}_{test}). |
| $Q^2_{\text{ext}(F3)}$ | $Q^2_{\text{ext}(F3)} = 1 - \frac{\left[\sum (Y_{\text{obs}(\text{test})} - Y_{\text{pred}(\text{test})})^2 \right] / n_{\text{test}}}{\left[\sum (Y_{\text{obs}(\text{train})} - \bar{Y}_{\text{training}})^2 \right] / n_{\text{train}}}$ | $Q^2_{\text{ext}(F3)}$ is measured to determine external predictivity employing both training and test set features. $Y_{\text{obs}(\text{test})}$ is the observed response, and $Y_{\text{pred}(\text{test})}$ is the predicted response of the test set molecules. $Y_{\text{obs}(\text{training})}$ is the observed response and $\bar{Y}_{\text{training}}$ denotes the mean observed response of the training set molecules. The threshold for $Q^2_{\text{ext}(F3)}$ is 0.5 |
| Concordance correlation coefficient (CCC) | $CCC = \bar{p}_c = \frac{2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$ | The concordance correlation coefficient (CCC) measures both precision and accuracy detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. 'n' denotes the number of compounds, and x_i and y_i signify the mean of observed and predicted values, respectively |
| Root mean square error in predictions ($RMSE_p$) | $RMSE_p = \sqrt{\frac{\sum (Y_{\text{obs}(\text{test})} - Y_{\text{pred}(\text{test})})^2}{n_{\text{test}}}}$ | It gives a measure of model external validation. A lower value of this parameter is desirable for good external predictivity |
| r^2_m metrics | $\bar{r}^2_m = \frac{r^2 + r'^2_m}{2} \text{ and } \Delta r^2_m = r^2 - r'^2_m $ where $r^2_m = r^2 \times (1 - \sqrt{r^2 - r'^2_m})$ $r'^2_m = r^2 \times \left(1 - \sqrt{r^2 - r'^2_m} \right)$ | r^2 is the squared correlation coefficient value between observed and predicted response values, and r^2_0 and r'^2_m are the respective squared correlation coefficients when the regression line is passed through the origin by interchanging the axes. For the acceptable prediction, the value of all Δr^2_m metrics should preferably be lower than 0.2 provided that the value of r^2_m is more than 0.5 (Ojha et al. 2011) |
| Predicted residual sum of squares (PRESS) | $\text{PRESS} = \sum (Y_{\text{obs}} - Y_{\text{pred}})^2$ | Sum of squared differences between experimental and predicted data. Y_{obs} and Y_{pred} correspond to the observed and LOO predicted values |

Table 2 (continued)

| Parameters | Equation | Description |
|--|--|---|
| Standard deviation of error of prediction (SDEP) | $SDEP = \sqrt{\frac{PRESS}{n}}$ | The value of standard deviation of error of prediction (SDEP) is calculated from PRESS. n refers to the number of observations |
| Mean absolute error (MAE) | $MAE = \frac{1}{n} \times \sum Y_{obs} - Y_{pred} $ | This is also known as average absolute error (AAE) and is considered a better index of errors in the context of predictive modeling studies |

quality as a result of structural redundancy in the training set data. Thus, the performance of a model on an external dataset is considered mandatory for the judgment of predictivity. The metric employed for judging external predictivity is termed as predictive R^2 or R^2_{pred} or $Q^2_{ext(F1)}$. The $Q^2_{ext(F1)}$ metric is characterized by a minimum threshold value of 0.6, i.e., models showing a value more than 0.6 are considered to be externally predictive with the ideal value being 1.0. Schüürmann and co-workers (Schüürmann et al. 2008) defined another external validation metric $Q^2_{ext(F2)}$ for the judgment of the predictivity of a model using the test set. Consonni et al. (2009) introduced another external validation metric $Q^2_{ext(F3)}$. This metric measures the model predictability and is sensitive to the selection of training dataset and tends to penalize models fitted to a very homogeneous data set even if predictions are close to the truth, with a threshold value being 0.6.

Another metric that checks the model reliability is the concordance correlation coefficient (CCC) metric (Chirico and Gramatica 2011). It measures both precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively. Any deviation of the regression line from the concordance line (line passing through the origin) gives a value of CCC smaller than 1. The desirable threshold value for CCC is 0.85.

The root-mean-square error in predictions ($RMSE_p$) gives a measure of model external validation. This metric is comparatively simpler and directly depicts the prediction errors for the test set observations concerning the total number of test set samples. A lower value of this metric is desirable for good external predictivity.

The r^2_m metrics: the training set mean value and the distance of the mean from the response values of each compound play a decisive role in computing the Q^2 values. The Q^2 value increases with the rise in the value of the denominator in the expression of Q^2 . Thus, even for a considerable deviation between the predicted and observed response values, satisfactory Q^2 values may be obtained, if the molecules exhibit a considerably broad range of response data. Using the concept of regression through origin approach, Roy et al. (2012) introduced a new metric r^2_m or modified r^2 that penalizes the r^2 value of

a model when there is large deviation between r^2 (squared correlation coefficient values between the observed (Y axis) and predicted (X axis) values of the compounds with intercept) and r_0^2 (squared correlation coefficient values between the observed (Y axis) and predicted (X axis) values of the compounds without intercept) values (Table 1).

MAE-based criteria: in a study, Roy et al. (2016) have shown that the conventional correlation-based external validation metrics ($Q^2_{ext(F1)}$, $Q^2_{ext(F2)}$) often provide biased judgment of model predictivity, since such metrics are influenced by factors such as response range and distribution of data. Here, the authors have defined a set of criteria using simple ‘mean absolute error’ (MAE) and the corresponding standard deviation (σ) measure of the predicted residuals to judge the external predictivity of the models. Note that $MAE = \frac{1}{n} \times \sum |Y_{obs} - Y_{pred}|$, where Y_{obs} and Y_{pred} are the respective observed and predicted response values of the test set comprising n number of compounds. The response range of training set compounds has been employed here to define the threshold values. Furthermore, the authors have proposed the application of the ‘MAE based criteria’ on 95% of the test set data by removing 5% data with high predicted residual values precluding the possibility of biased prediction quality due to any outlier prediction. The following criteria for MAE prediction are followed:

- i. Good predictions: in easier terms, an error of 10% of the training set range should be acceptable, while an error more than 20% of the training set range should be a very high error. Thus, the criterion for good predictions is as follows:

$$MAE \leq 0.1 \times \text{training set range and } (MAE + 3\sigma) \leq 0.2 \times \text{training set range.}$$

Here, σ value indicates the standard deviation of absolute errors for the test data. For a normal distribution pattern, $\text{mean} \pm 3\sigma$ covers 99.7% of the data points.

- ii. Bad predictions: a value of MAE more than 15% of the training set range is considered high, while an error higher than 25% of the training set range is judged as very high. Thus, prediction is considered bad when

$$\text{MAE} > 0.15 \times \text{training set range or } (\text{MAE} + 3\sigma) > 0.25 \times \text{training set range.}$$

Predictions which do not fall under either of the above two conditions may be considered as of moderate quality. This criterion is applied for judging the quality of test set prediction when there are at least 10 data points signifying statistical reliability and there is no systemic error in model predictions.

Randomisation of response (Y-scrambling)–Randomisation is an assessment to ensure the developed QSAR model is not due to chance, thereby giving an idea of model robustness (Rücker et al. 2007). In this technique, validation metrics are checked by repetitive permutation of the response data (Y) of n compounds in the training set with respect to the X (descriptor) matrix which is kept unchanged. The calculations are repeated with randomized activities, followed by a probabilistic examination of the results. Every run will yield approximations of R^2 and Q^2 , which are recorded. For an acceptable QSAR model, the average correlation coefficient (R_r) of randomized models should be less than the correlation coefficient (R) of a non-random model. The difference between mean-squared correlation coefficients of the randomized (R_r^2) and that of the non-random (R^2) models

can be obtained through R_p^2 calculation ($R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$). A robust QSAR model should have R_p^2 value less than 0.5. At the ideal condition, the average value of R^2 for the randomized models should be zero, i.e., R_r^2 should be zero. Consequently, in such a case, the value of R_p^2 should be equal to the value of R^2 for the developed QSAR model. Thus, as proposed by Todeschini, the corrected formula of $R_p^2(c_{R_p^2})$ is $c_{R_p^2} = R \times \sqrt{R^2 - R_r^2}$ (Todeschini 2010).

Classification-based QSAR validation metrics

In a binary classification model, several validation metrics are utilized to evaluate the model's performance in terms of accurate qualitative prediction of the dependent variable. Classification models are generally assessed using a statistical method that is based on the Bayesian approach (Ghosh et al. 2020). A binary classification model is typically a two-class model, i.e., positive and negative, or active and inactive. The results obtained can be arranged in a contingency table (also known as confusion matrix) (Table 3). The

Table 3 Contingency table or confusion matrix for classification modeling

| | | Experimental | | Total |
|-----------|----------|---------------------|---------------------|--------------------|
| | | Active | Inactive | |
| Predicted | Active | True positive (TP) | False positive (FP) | TP+FP |
| | Inactive | False negative (FN) | True negative (TN) | FN+TN |
| Total | | TP+FN | FP+TN | N = TP+FP+FN+TN |

Table 4 Validation metrics for classification modeling

| Sl No. | Classification metric | Equation |
|--------|---------------------------------------|---|
| 1 | Sensitivity | $\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$ |
| 2 | Specificity | $\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$ |
| 3 | Precision | $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$ |
| 4 | Accuracy | $\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{TN}+\text{FP}}$ |
| 5 | F-measure | $\text{F-measure}(\%) = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Sensitivity}}}$ |
| 6 | G-means | $\text{G-means} = \sqrt{\text{Specificity} \times \text{Sensitivity}}$ |
| 7 | Cohen's Kappa (κ) | $P_r(a) = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})}$ $P_r(e) = \frac{[(\text{TP}+\text{FP}) \times (\text{TP}+\text{FN})] + [(\text{TN}+\text{FP}) \times (\text{TN}+\text{FN})]}{(\text{TP}+\text{FN}+\text{FP}+\text{TN})^2}$ $\text{Cohen's } \kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$ |
| 8 | Mathews correlation coefficient (MCC) | $\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP}+\text{FP}) \times (\text{TP}+\text{FN}) \times (\text{TN}+\text{FP}) \times (\text{TN}+\text{FN})}}$ |

$P_r(a)$: relative observed agreement between the predicted classification of the model and the known classification; $P_r(e)$: hypothetical probability of chance agreement

statistical metrics explaining the quality of a classification model are given below and in Table 4.

In classification QSAR modeling, the compounds are classified into four main categories: a) true positives (TP), b) true negative (TN), c) false positive (FP), and d) false negative (FN) (Table 3). Researchers have used a variety of statistical tests to assess the classifier model performance and classification capability. Sensitivity (Sn) is the percentage of active compounds correctly predicted and is expressed as the ratio of true-positive results to the total number of positive data. Specificity (Sp) is the ratio of true-negative results to the total number of negative data. Accuracy (Acc) implies the fraction of correctly predicted compounds. The precision indicates the accuracy of a predicted class (ratio between the true positives and total positives) and F -measure refers to the harmonic mean of Recall (or Sensitivity) and Precision. Higher values for recall and precision give higher values for F -measure, thereby implying better classification.

G-means is a combination term that includes Sn and Sp into a single parameter merged via the geometric mean. This allows an easy assessment of the model's ability to distinguish between active or inactive samples.

Cohen's kappa (κ) can be utilized to determine the concordance between classification (predicted) models and known classifications (Cohen 1960). It is a measure of the degree of agreement. It returns value from -1 (total disagreement) to 0 (random classification) to 1 (total agreement).

Mathews correlation coefficient (MCC) measures the quality of binary classifications and compares different classifiers. In any case, where the number of positive and negative compounds is not equal, the terms sensitivity, specificity, and accuracy are not reliable. MCC uses all four values (TP, TN, FP, and FN) and is directly calculated from the confusion matrix to provide a more-balanced prediction evaluation. Like Cohen's kappa, the value for MCC also ranges from -1 to 1 .

Prediction reliability detection tools

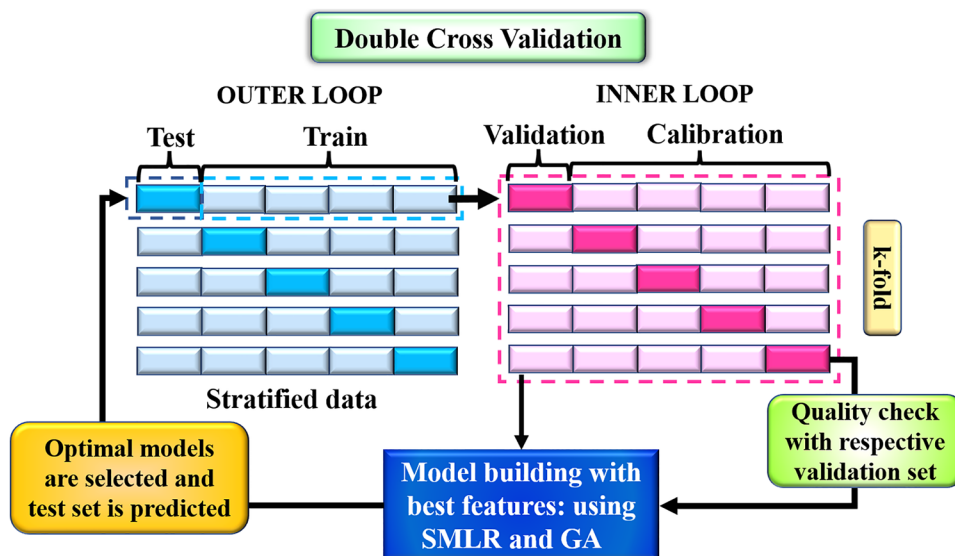
As discussed earlier, the process of QSAR modeling consists of three important steps: model development, model selection, and model interpretation. The model development process involves various feature selection practices including stepwise-multiple linear regression (S-MLR), genetic algorithm, genetic function approximation, etc. Model selection is based on the identification of the finest model (based on validation metric values) from a set of alternative models. When it comes to the reliability of QSAR/QSPR models, validation is essential. After a model has been selected, several internal and external validation metrics are assessed which help in demonstrating the actual

predictive performance of the chosen model. Several groups of researchers in QSAR suggested external validation to be the gold standard in demonstrating the predictive ability of a model (Golbraikh and Tropsha 2002; Gramatica and Sangion 2016; Gramatica 2020). Multiple modeling in consensus form has been introduced to achieve a lower degree of predicted residuals for query compounds (Roy et al. 2015b; Khan et al. 2019a; Roy et al. 2019). In the following sections, we will discuss various tools from the DTC Laboratory (<https://sites.google.com/site/kunalroyindia/home/qsar-model-development-tools>) that help understand the prediction ability of one or more QSAR models.

(i) Double cross-validation (version 2.0) tool

The most common scheme of external validation is by introducing the hold-out method. Here, the original dataset is divided into training and test sets, where the training set is used for model-building purposes followed by model selection based on internal validation metrics, and the test set is used for model validation through external validation metrics. This approach ensures that the test set is never applied during the model-building procedure and it remains unseen by the developed model. However, a single training set does not confirm feature optimization, since a fixed training set composition leads to a bias in feature selection. This issue is more apparent in the case of MLR models than partial least-squares (PLS) or principal component regression (PCR) models which are more robust and generalized methods. Baumann and Baumann (Baumann and Baumann 2014) discussed the concept of double cross-validation (DCV) which Roy and Ambure implemented in a tool (Roy and Ambure 2016) where the training set is further divided into 'n' number of calibration and validation sets. The tool is freely available from <http://dtclab.webs.com/software-tools> and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. The algorithm comprises two nested cross-validation loops (Bates et al. 2021), namely, the outer loop and the inner loop (Fig. 2). The outer loop consists of data points that are split arbitrarily into disjoint subsets known as training set compounds and test set compounds. The training set is utilized in the inner loop for model development and model selection, and the test set is used exclusively for checking model predictivity. The training set in the inner loop is further split into k number of calibration and validation sets in the inner loop by applying the k -fold cross-validation technique (Wainer and Cawley 2021). In the k -fold cross-validation method, the training data are initially segregated into k subsets followed by preparing k -iterations of calibration and validation sets. At each reiteration, different subset of data is excluded and used as validation set, while the remaining $k-1$ subsets are used as calibration sets. The data are passed through a stratification process, i.e., data rearrangement which helps

Fig. 2 Schematic diagram of double cross-validation algorithm (colour figure online)



maintain data uniformity (each fold is representative of the whole dataset). Each k -fold calibration set is then used to develop multiple linear regression (MLR) models, while the respective validation sets are applied to find the prediction errors. The tool provides two methods of feature selection: stepwise-multiple linear regression (S-MLR) (Maleki et al. 2014; Ojha and Roy 2018) and genetic algorithm-MLR (GA-MLR) (Leari 2001). The prediction error is checked using mean absolute error ($MAE_{95\%}$) (Roy et al. 2016). There is also a provision for the generation of PLS models in the tool. Furthermore, the models in the inner loop are selected based on three major criteria as follows:

- i) The models with the lowest MAE value (on the validation set) are selected.
- ii) Consensus predictions of the top model are selected based on the MAE value of the validation set.
- iii) Searching out the best descriptor combination from the top models.

Researchers found the DCV approach to be reliable and useful and thus successfully employed in various applications, for example, quantitative structure–property relationship (QSPR) modeling for sweetness potency of organic chemicals (Ojha and Roy 2018), developing nano-QSAR models for TiO_2 -based photocatalysts (Mikolajczyk et al. 2018), inhalational toxicity modeling (Nath et al. 2022), modeling of diagnostic agents (De et al. 2019; De et al. 2020, 2022; De and Roy 2020, 2021), etc.

(ii) Intelligent consensus predictor tool

A well-validated QSAR model engages different classes of descriptors, which accentuate many features of molecular structures. Individual QSAR models may exaggerate a few

important features, undervalue other features, and overlook some significant characteristics features. Roy et al. (2018b) proposed an “intelligent” selection of multiple models that would enhance the quality of predictions of query compounds (Roy et al. 2018b). This software helps judge the performance of consensus predictions compared to their quality obtained from the individual MLR models based on the MAE-based criteria (95%). The tool “Intelligent Consensus Prediction” is available from <http://dtclab.webs.com/software-tools> and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. The tool takes multiple individual models (M1, M2, M3, etc.) as input derived using a different combination of descriptors from the training set. There are four ways of consensus prediction explained in the work:

- (i) Consensus model 0 (CM0): it provides a simple average of predictions from all input individual models.
- (ii) Consensus model 1 (CM1): it is the average of predictions from all individual qualified models. It is calculated from the arithmetic average of predicted response values attained from the ‘ n ’ qualified models for test compounds rather than from all existing individual models.
- (iii) Consensus model 2 (CM2): it is the weighted average prediction (WAP) from all qualified individual models. In CM2, the average is evaluated by giving a proper weightage to the qualified models for a particular test set compound.
- (iv) Consensus model 3 (CM3): compound-wise best selection of predictions from qualified individual models. The best model for a particular test compound is selected based on its cross-validated mean absolute error (MAE_{CV}). A model with the lowest value MAE_{CV} is the best for a particular test set compound.

The tool further provides additional selection criteria which include:

- Euclidean distance cut-off: this is used to find a fitting model to predict the test set compound, where 10 most similar compounds are selected based on Euclidean Distance score. The user can set a Euclidean cut-off ranging from 0 to 1 to restrict the selection of only those training set compounds with a Euclidean distance score less than or equal to the set cut-off value.
- Applicability domain: AD helps to check whether the test/query compound is in the chemical space of the model or not. A simple standardization approach is used for AD determination.
- Dixon Q test: this test can be employed to spot and remove a response outlier out of selected similar training set compound.

The complete calculation method is demonstrated in the published article by Roy et al. and the methodology is given in Fig. 3. The ICP method has found good application in the prediction of pharmaceuticals (Khan et al. 2019a), organic chemicals and dyes (Roy et al. 2019; Khan and Roy 2019; Ghosh and Roy 2019; Ojha et al. 2020), determining aquatic toxicity (Hossain and Roy 2018), inhalational toxicity (Nath et al. 2022), polymer properties (Khan et al. 2018), etc.

(iii) Prediction Reliability Indicator tool

A QSAR model is developed based on the physicochemical features of an appropriately designed training set having experimentally derived response data. In contrast, the model is validated using one or more test set(s) for which the experimental response data are available. The suitability of this model for a completely new data set (true external set) for providing a reliable prediction is quite an interesting study.

Fig. 3 “Intelligent Consensus Prediction” algorithm

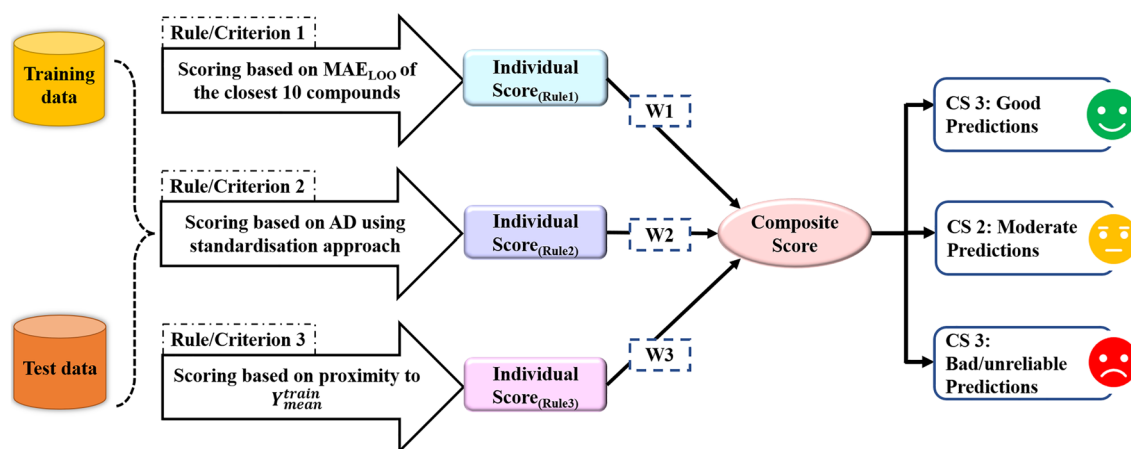
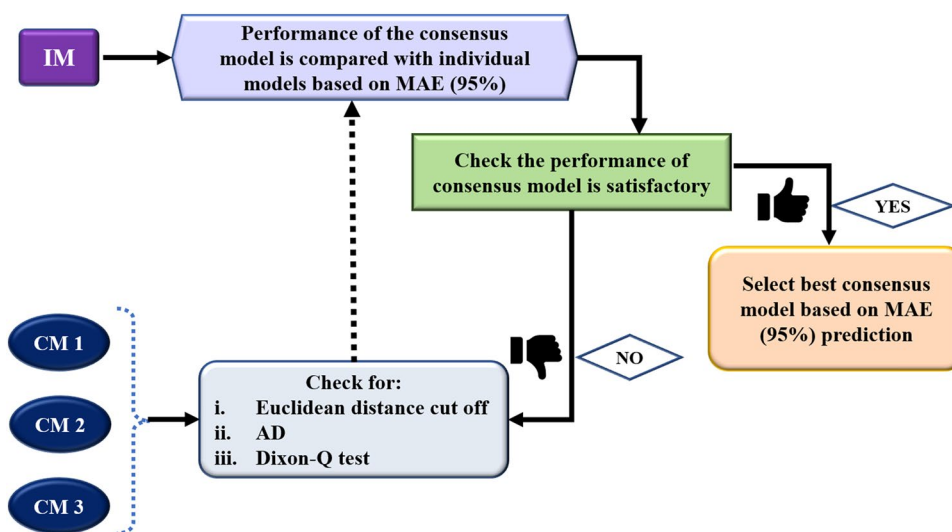


Fig. 4 Methodology applied for scoring test/query compounds in “Prediction Reliability Indicator” tool

Roy et al. (2018a, b) have developed a new scheme (Fig. 4) to define the reliability of predictions from QSAR models for new query compounds and implemented the method in a new tool called “Prediction Reliability Indicator” freely available from <http://dtclab.webs.com/software-tools> and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/. This tool is applicable for predictions from MLR and PLS models. The work aimed at formulating a set of rules/criteria that will ultimately empower the user to estimate the quality of predictions for individual test (external) compounds. Prediction of test/external sets can have varying quality. It might not be good predictions in all cases, while the model can show moderate to bad/unreliable predictions for some of the external set compounds. By keeping the variation of prediction quality, the authors have hypothesized three rules/criteria which might assist in classifying the quality of predictions for individual test/external set compounds into good, moderate, and poor/unreliable ones. We have now discussed the three rules briefly in the following segment:

- (a) Rule/criterion 1: the scoring is based on the quality of leave-one-out predictions of the closest 10 training compounds to a test/external compound. Here, 10 most similar compounds are identified for each test/query compound (based on Euclidean distance similarity), followed by which mean of absolute LOO prediction error (MAE_{LOO}) is calculated for the selected closest 10 compounds. For a test/query compound whose MAE_{LOO} is lowest corresponding to its closest training compounds is predicted well and gets the highest prediction score (Prediction Score = 3). Test/query compounds that have medium MAE_{LOO} values with corresponding close training compounds should get a moderate score (Prediction Score = 2), and those test compounds with corresponding close training compounds having high MAE_{LOO} values should get the least score (Prediction Score = 1). The MAE-based criteria (Roy et al. 2016) are applied here for scoring the compounds which involve MAE_{LOO} and standard deviation (σ_{LOO}) of the absolute prediction error values.
- (b) Rule/criterion 2: scoring based on the similarity-based AD using standardization method. The applicability domain (AD) of a model plays an important role in identifying uncertainty in the prediction of a specific chemical (test/query) by that model. This is based on how similar is the test/query compound with those in the training set. When a test/query compound is similar to a small fraction or none of the training compounds, the prediction is considered unreliable or fails to fall under the training set AD. Here, a simple AD based on the standardization approach (Roy et al. 2015a, b) is employed.

- (c) Rule/Criterion 3: scoring based on the proximity of predictions to the training set observed/experimental response mean. Earlier, the quality of fit or prediction of compounds is better when compounds possess experimental response values (training and test compounds) close to the training set observed response mean. Thus, in rule/criterion 3, the authors have proposed to assess the prediction quality of a test compound based on the closeness of predicted response value to the training set observed/experimental response mean. The predicted response value (Y_{pred}^{test}) of each test compound is first measured using the training set model, and then, this Y_{pred}^{test} value is compared with the training set experimental response mean (Y_{mean}^{train}) and the corresponding standard deviation (σ^{train}). The scoring is based on the following manner:

(i) A test compound with Y_{pred}^{test} value falling within the range inside $Y_{mean}^{train} \pm 2\sigma^{train}$, that is, $(Y_{mean}^{train} + 2\sigma^{train}) \geq Y_{pred}^{test} \geq (Y_{mean}^{train} - 2\sigma^{train})$, can be assumed to be well (good) predicted by the model and thus have a score 3.

(ii) A test compound with Y_{pred}^{test} value falling within the range $(Y_{mean}^{train} + 3\sigma^{train}) \geq Y_{pred}^{test} \geq (Y_{mean}^{train} - 3\sigma^{train})$ and $(Y_{mean}^{train} + 2\sigma^{train}) < Y_{pred}^{test} < (Y_{mean}^{train} - 2\sigma^{train})$ can be presumed to be predicted moderately by the model and thus gets a score 2.

(iii) A test compound with Y_{pred}^{test} value falling within the range $(Y_{mean}^{train} + 3\sigma^{train}) < Y_{pred}^{test} < (Y_{mean}^{train} - 3\sigma^{train})$ can be assumed to be predicted poorly by the model and thus gets a score 1.

Furthermore, after these three criteria are checked, a weighting scheme is employed to compute a composite score for judging the prediction quality of each test compound using all three individual scores. The composite score is defined as follows:

$$\begin{aligned} \text{Composite score} &= W_1 \times \text{score}_{\text{rule1}} \\ &+ W_2 \times \text{score}_{\text{rule2}} \\ &+ W_3 \times \text{score}_{\text{rule3}}. \end{aligned}$$

Here, $\text{score}_{\text{rule1}}$, $\text{score}_{\text{rule2}}$, and $\text{score}_{\text{rule3}}$ represent the scores obtained after applying respective rules, whereas W_1 , W_2 , W_3 indicate the weightage (automatic or user-provided) given to each of the three individual scores. The PRI tool offers a unique composite score which can act as a marker of prediction quality of true external test compound. This tool has found application for the prediction of external set/query compounds in many areas, viz., endocrine disruptor chemicals (Khan et al. 2019b), metal oxide nanoparticles (De et al. 2018), organic chemicals (Khan and Roy 2019; Khan et al. 2019c; De et al. 2020; 2022; Nath et al. 2022), etc.

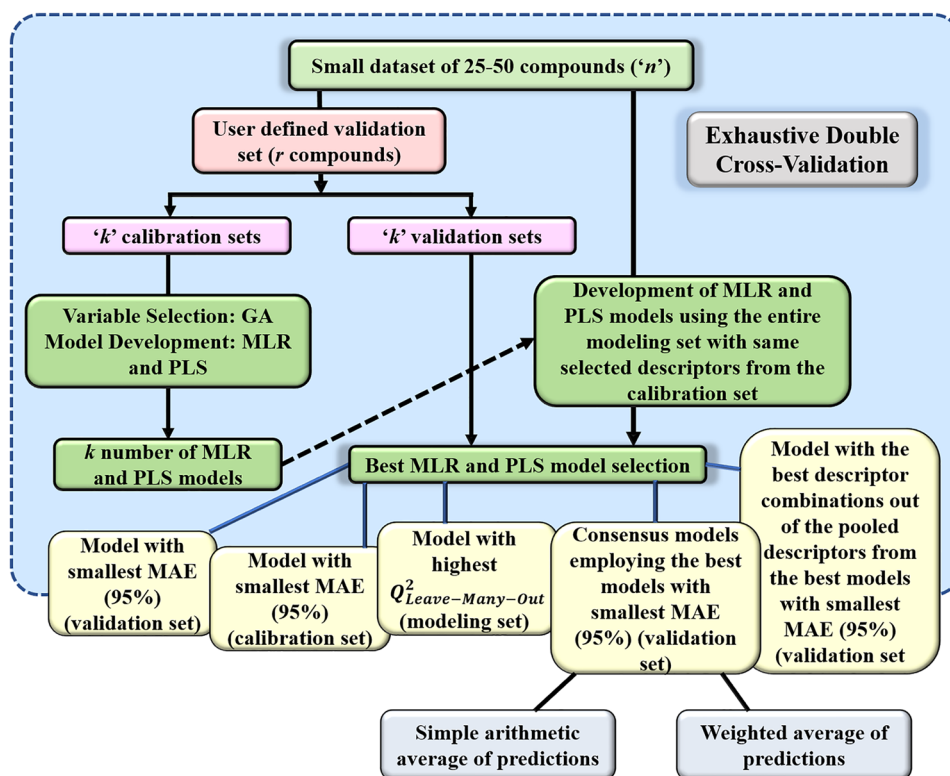
(iv) **Small dataset modeler (version 1.0.0) tool**

Various specialized datasets involving nanomaterials, properties of catalysts, radiosensitizer molecules, etc. have smaller number of data points where the division of data into training and test sets may not produce robust and predictive models. A small dataset with 25–50 compounds cannot be used for conventional double cross-validation as dividing the data set into training and test sets and further into calibration and validation sets is not possible. Ambure et al. have developed a new tool called the Small Dataset Modeler, version 1.0.0 (<http://dtclab.webs.com/software-tools> and http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/) solely for small datasets which includes a double cross-validation approach to develop a model for a small number of data points without training and test sets division of the dataset (Ambure et al. 2019) (Fig. 5). Here, the whole input set (containing n number of compounds) goes into a loop where it is repeatedly split up into calibration and validation sets (same as in the inner loop of DCV). The best possible combinations (k) are tried to obtain using validation sets of r compounds and calibration sets of $n-r$ compounds. The tool asks for the number of compounds (i.e., r) in the validation set from the user based on which all probable combinations of calibration and validation sets are produced. The Multiple Linear Regression (MLR) models are generated using the calibration set compounds employing the Genetic

Algorithm-Multiple Linear Regression (GA-MLR) method (Devilleers 1996; Venkatasubramanian and Sundaram 2002) of variable selection, while the validation sets are employed to judge the predictive ability of the models. Numerous important internal (R^2 , R^2_{adj} , Q^2_{LMO} , MAE_{LOO} , $r^2_m(\text{LOO})$ metrics) and external (Q^2_{F1} , Q^2_{F2} , $r^2_m(\text{test})$, CCC, MAE_{test}) validation metrics are measured in the exhaustive DCV method for all the chosen models. The tool is designed in such a way that it also develops Partial Least Squares Regression (PLS-R) models based on the descriptors selected in MLR models. The final top model selection can be done in any five of the following recommended ways:

- (i) Any model (MLR/PLS) with the smallest MAE (95%) in the validation set is chosen.
- (ii) Any model (MLR/PLS) with the smallest MAE (95%) in the modeling set is chosen.
- (iii) Any model (MLR/PLS) with the lowest $Q^2_{\text{Leave-Many-Out}}$ (modeling set) is chosen.
- (iv) Implementing consensus predictions using the best models that are chosen depending on the MAE (95%) in the validation sets. Consensus predictions can be of two types: (a) Using a simple arithmetic average of predictions of the best models. (b) Using a weighted average of predictions (WAP) by assigning proper weights to the top chosen models depending on the mean abso-

Fig. 5 Methodology behind the “Small Dataset Modeler” (version 1.0.0) tool to perform QSAR modeling for a small set of data points



lute error obtained from leave-one-out cross-validation, $MAE_{cv}(95\%)$.

(v) A pool of exclusive descriptors from the best models with the smallest $MAE(95\%)$ obtained from the validation set is again employed to build models. In the case of MLR, the best descriptor combinations are put through the *Best Subset Selection* method. However, in the case of a PLS model, descriptors nominated in the top models are pooled together for a PLS run.

The method proposed in the “Small Dataset Modeler” tool confirms internal divisions of small datasets within the DCV technique without taking any test set into account. The approach of “Small Dataset Modeler” tool integrates data curation, exhaustive DCV technique, and ideal modeling techniques entailing consensus predictions to develop models, principally for a small set of data points. The methodology behind the “Small Dataset Modeler” tool is schematically presented in Fig. 5. Small dataset modeling has found use in environmental toxicity modeling including acute toxicity of antifungal agents toward fish species (Nath et al. 2021) and soil ecotoxicity (Lavado et al. 2022), radiosensitization modeling (De and Roy 2020), modeling of Hepatitis C virus inhibitor protein (Ejeh et al. 2021), and modeling anesthetics causing GABA inhibition (Stošić et al. 2020).

(v) Read-Across-v3.1 tool

The read-across methodology has gained immense attention in recent years, because it is a non-testing approach

that can be utilized for data-gap filling. The basic aim of the read-across technique is to predict endpoint information for one or more chemicals (i.e., the target chemicals) using data from the same endpoint from another substance (the source chemicals) using the similarity principle. The method is widely used as an alternative tool for hazard assessment to fill data gaps (ECHA 2011). Read-across based predictions seem to be more fitting for small data sets (limited source compounds). Hence, it has provided promising results in nanosafety assessment possessing limited data. Chatterjee and co-workers (2022) developed a new prediction-oriented quantitative read-across approach based on certain similarity principles. The reported work verifies the efficiency of the newly developed read-across algorithm in filling nanosafety data gaps. A tool has been developed to facilitate the implementation of the approach (Fig. 6) for quantitative read-across which is available from <https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home>. The tool allows the users to optimize different hyperparameters including similarity kernel functions and distance and similarity thresholds to get the best quality of quantitative predictions. Mainly, three types of similarity estimation techniques were introduced involving Euclidean distance, Gaussian kernel function, and Laplacian kernel function. The algorithm developed in this study was optimized using three small nanotoxicity datasets ($n \leq 20$). The algorithm is based on two basic steps: (a) finding the 10 most similar training compounds for each query or test compound; (b) calculating the weighted average prediction of test set compounds from the most similar training set compounds. Different

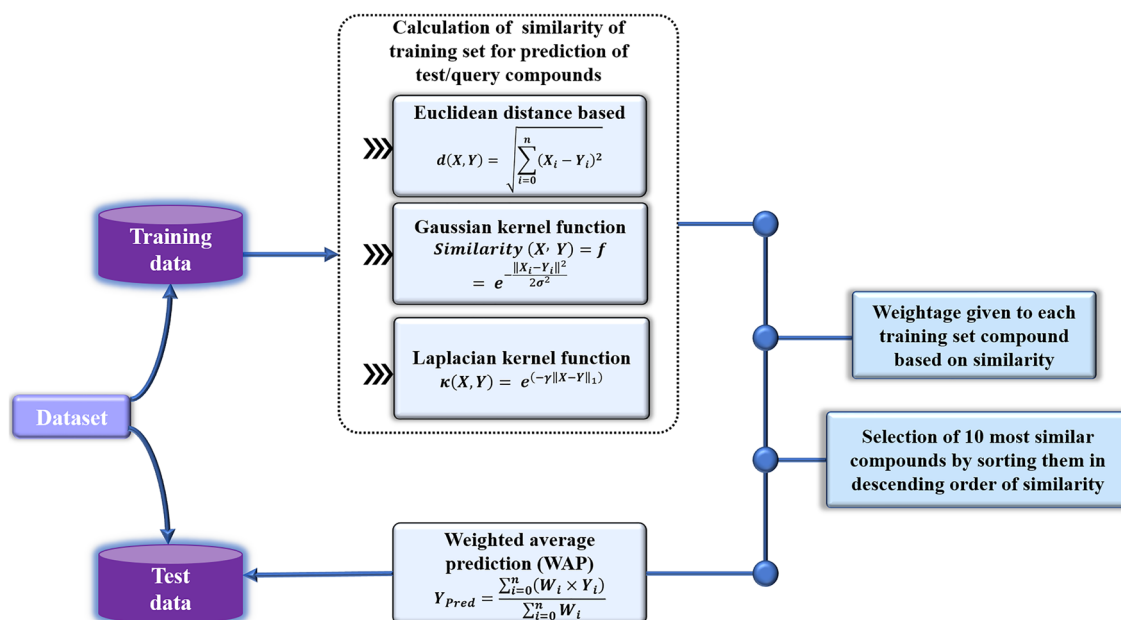


Fig. 6 Quantitative read-across algorithm

hyperparameters like sigma and gamma values in Gaussian and Laplacian kernel functions have been optimized. The effect of the number of close training compounds on the prediction quality has been evaluated; 2–5 close training compounds can efficiently predict the toxicity of query compounds. Another feature incorporation in the tool involves a distance threshold for the Euclidean distance similarity estimation and a similarity threshold for the Gaussian and Laplacian kernel function similarity estimations. This generated better prediction at the distance threshold of 0.4–0.5 and a similarity threshold of 0.00–0.05. This algorithm is easy to use, proficient, and an expert independent alternative method for the nanoparticle toxicity prediction which can further assist in data-gap filling and prioritization. Version 3.1 of this tool also computes classification-based validation metrics and generates receiver operating curve (ROC) for predictions which can be used to estimate the uncertainty of predictions. The tool is also applicable for several endpoints other than nanotoxicity, for example activity/toxicity/property of organic compounds in general.

Future perspectives

Over the past few decades, the QSAR methodology has received both praise and criticism in connection to its reliability, limitations, successes, and failures. The above discussion of the aforementioned tools from the DTC Laboratory provides methods and information relating to QSAR model development and validation, pointing out current trends, unresolved problems, and persistent challenges associated with evolution of QSAR. Furthermore, there are few scopes of further refining the present tools like inclusion of computation of Golbraikh and Tropsha's (Golbraikh and Tropsha 2002) criteria in the Double Cross Validation tool and computation of leave-many-out cross-validation (Q^2_{LMO}) criteria for both the Double Cross Validation tool and Small Dataset Modeler tool (PLS version), etc. Additionally, there is an opportunity to incorporate an uncertainty measure of predictions in the read-across tool which will improve the reliability for quantitative predictions of untested molecules.

Conclusion

The QSAR domain has been expanded substantially in the past few years as databases and their applications have grown. As the field of QSAR evolves through decades, it is necessary to evaluate the effectiveness of the QSAR models in predicting the behavior of new molecules. A QSAR model stands on the pillars of various validation metrics used to assess the quality of a predictive model that portrays the true

picture of the prediction errors. The present review explains various internal and external validation metrics necessary for model predictivity assessment. Furthermore, a brief explanation of various innovative QSAR modeling tools developed by Drug Theoretics and Cheminformatics (DTC) laboratory (<https://sites.google.com/site/kunalroyindia/home/qsar-model-development-tools>) is given for better selection and development of models. These tools are aimed at addressing various features like selection of training set, model development methodology, model selection techniques, the use of multiple models, scoring of query compounds, etc. These improvisations helped in enhancing the quality of predictions of QSAR models. The tools highly assist in the reliability estimation of untested chemicals when experimental data are unavailable. However, most of these tools cannot be used for classification-based/graded data, but are well suited for quantitative models like MLR and PLS regression. Furthermore, the tools have a major role in different fields for predicting chemicals associated with the pharmaceutical industry, cosmeceuticals, polymer chemistry, diagnostic agents, dyes, nano-chemistry, food chemistry, etc.

Acknowledgements PD thanks Indian Council for Medical Research, New Delhi for Senior Research Fellowship.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

- Abdi H, Williams LJ (2010) Principal component analysis. *Wiley Interdiscip Rev* 2(4):433–459
- Ambure P, Cordeiro MNDS (2020) Importance of data curation in QSAR studies especially while modeling large-size datasets. In: Roy K (ed) *Ecotoxicol QSARs*. Springer, New York, pp 97–109
- Ambure P, Gajewicz-Skretna A, Cordeiro MND, Roy K (2019) New workflow for QSAR model development from small data sets: small dataset curator and small dataset modeler integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques. *J Chem Inform Model* 59(10):4070–4076
- Bates S, Hastie T, Tibshirani R (2021) Cross-validation: what does it estimate and how well does it do it? *arXiv:210400673*
- Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform* 6(1):1–19
- Chatterjee M, Banerjee A, De P, Gajewicz-Skretna A, Roy K (2022) A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environ Sci Nano* 9(1):189–203
- Chirico N, Gramatica P (2011) Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J Chem Inf Model* 51(9):2320–2335

- Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46
- Consonni V, Todeschini R (2010) Molecular descriptors Recent advances in QSAR studies. Springer, New York, pp 29–102
- Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q₂ parameter for QSAR validation. *J Chem Inf Model* 49(7):1669–2167
- De P, Roy K (2018) Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. *SAR QSAR Environ Res* 29(4):319–337
- De P, Roy K (2020) QSAR modeling of PET imaging agents for the diagnosis of Parkinson's disease targeting dopamine receptor. *Theor Chem Acc* 139:176
- De P, Roy K (2021) QSAR and QSAAR modeling of nitroimidazole sulfonamide radiosensitizers: application of small dataset modeling. *Struct Chem* 32(2):631–642
- De P, Kar S, Roy K, Leszczynski J (2018) Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environ Sci Nano* 5(11):2742–2760
- De P, Bhattacharyya D, Roy K (2019) Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease. *Struct Chem* 30(6):2429–2445
- De P, Bhattacharyya D, Roy K (2020) Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling. *Struct Chem* 31(3):1043–1055
- De P, Bhayye S, Kumar V, Roy K (2022) In silico modeling for quick prediction of inhibitory activity against 3CLpro enzyme in SARS CoV diseases. *J Biomol Struct* 40(3):1010–1036
- Dearden JC (2016) The history and development of quantitative structure-activity relationships (QSARs). *Int J Quant Struct-Property Relat* 1(1):1–44
- Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, NY
- Directorate E (2007) Environment health and safety publications series on testing and assessment No. 69, Guidance document on the validation of (quantitative) structure-activity relationships [(Q) SAR] models. OECD, Paris, France
- ECHA (2011) The Use of Alternatives to Testing on Animals for the REACH Regulation. European Chemicals Agency Helsinki, Finland
- Ejeh S, Uzairu A, Shallangwa GA, Abechi SE (2021) Computational insight to design new potential hepatitis C virus NS5B polymerase inhibitors with drug-likeness and pharmacokinetic ADMET parameters predictions. *Future J Pharm Sci* 7(1):1–13
- Ghosh S, Ojha PK, Roy K (2019) Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs. *Chemosphere* 228:545–555
- Ghosh K, Bhardwaj B, Amin S, Jha T, Gayen S (2020) Identification of structural fingerprints for ABCG2 inhibition by using Monte Carlo optimization, Bayesian classification, and structural and physicochemical interpretation (SPCI) analysis. *SAR QSAR Environ Res* 31(6):439–455
- Golbraikh A, Tropsha A (2002) Beware of q₂! *J Mol Graph Model* 20(4):269–276
- Golmohammadi H, Dashtbozorgi Z, Acree WE Jr (2012) Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur J Pharm Sci* 47(2):421–429
- Goodarzi M, Dejaegher B, Heyden YV (2012) Feature selection methods in QSAR studies. *J AOAC Int* 95(3):636–651
- Gramatica P (2020) Principles of QSAR modeling: comments and suggestions from personal experience. *IJQSPR* 5(3):61–97
- Gramatica P, Sangion A (2016) A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. *J Chem Inf Model* 56(6):1127–1131
- Héberger K, Rác A, Bajusz D (2017) Which performance parameters are best suited to assess the predictive ability of models? *Advances in QSAR Modeling*. Springer, New York, pp 89–104
- Höltje H-D, Sippl W (2001) Rational approaches to drug desing: proceedings of the 13th European symposium on quantitative structure-activity relationships, August 27-September, 1, 2000. JR Prous Science
- Hossain KA, Roy K (2018) Chemometric modeling of aquatic toxicity of contaminants of emerging concern (CECs) in *Dugesia japonica* and its interspecies correlation with daphnia and fish: QSTR and QSTTR approaches. *Ecotoxicol Environ Saf* 166:92–101
- Hsu H-H, Hsu Y-C, Chang L-J, Yang J-M (2017) An integrated approach with new strategies for QSAR models and lead optimization. *BMC Genom* 18(2):1–9
- Kennard RW, Stone LA (1969) Computer aided design of experiments. *Technometrics* 11(1):137–148
- Khan K, Roy K (2019) Ecotoxicological QSAR modelling of organic chemicals against *Pseudokirchneriella subcapitata* using consensus predictions approach. *SAR QSAR Environ Res* 30(9):665–681
- Khan PM, Rasulev B, Roy K (2018) QSPR modeling of the refractive index for diverse polymers using 2D descriptors. *ACS Omega* 3(10):13374–13386
- Khan K, Benfenati E, Roy K (2019a) Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds. *Ecotoxicol Environ Saf* 168:287–297
- Khan K, Roy K, Benfenati E (2019b) Ecotoxicological QSAR modeling of endocrine disruptor chemicals. *J Hazard Mater* 369:707–718
- Khan PM, Roy K, Benfenati E (2019c) Chemometric modeling of *Daphnia magna* toxicity of agrochemicals. *Chemosphere* 224:470–479
- Kononov DA, Coomans D, Deconinck E, Vander Heyden Y (2007) Benchmarking of QSAR models for blood-brain barrier permeation. *J Chem Inf Model* 47(4):1648–1656
- Kononov DA, Llewellyn LE, Vander Heyden Y, Coomans D (2008) Robust cross-validation of linear regression QSAR models. *J Chem Inf Model* 48(10):2081–2094
- Kubinyi H, Hamprecht FA, Mietzner T (1998) Three-dimensional quantitative similarity—activity relationships (3d qsar) from seal similarity matrices. *J Med Chem* 41(14):2553–2564
- Lavado GJ, Baderna D, Carnesecci E, Toropova AP, Toropov AA, Dorne JLC, Benfenati E (2022) QSAR models for soil ecotoxicity: development and validation of models to predict reproductive toxicity of organic chemicals in the collembola *Folsomia candida*. *J Hazard Mater* 423:127236
- Leardi R (2001) Genetic algorithms in chemometrics and chemistry: a review. *J Chemom* 15(7):559–569
- Likas A, Vlassis N, Verbeek JJ (2003) The global k-means clustering algorithm. *Pattern Recognit* 36(2):451–461
- Maleki A, Daraei H, Alaei L, Faraji A (2014) Comparison of QSAR models based on combinations of genetic algorithm, stepwise multiple linear regression, and artificial neural network methods to predict K_d of some derivatives of aromatic sulfonamides as carbonic anhydrase II inhibitors. *Russ J Bioorganic Chem* 40(1):61–75
- Mikolajczyk A, Gajewicz A, Mulkiwicz E, Rasulev B, Marchelek M, Diak M, Hirano S, Zaleska-Medynska A, Puzyn T (2018) Nano-QSAR modeling for ecosafe design of heterogeneous TiO₂-based nano-photocatalysts. *Environ Sci Nano* 5(5):1150–1160
- Nath A, De P, Roy K (2021) In silico modelling of acute toxicity of 1, 2, 4-triazole antifungal agents towards zebrafish (*Danio rerio*)

- embryos: application of the small dataset modeller tool. *Toxicol in Vitro* 75:105205
- Nath A, De P, Roy K (2022) QSAR modelling of inhalation toxicity of diverse volatile organic molecules using no observed adverse effect concentration (NOAEC) as the endpoint. *Chemosphere* 287:131954
- Norinder U (1996) Single and domain mode variable selection in 3D QSAR applications. *J Chemom* 10(2):95–105
- Novellino E, Fattorusso C, Greco G (1995) Use of comparative molecular field analysis and cluster analysis in series design. *Pharm Acta Helv* 70(2):149–154
- Ojha PK, Roy K (2018) Development of a robust and validated 2D-QSPR model for sweetness potency of diverse functional organic molecules. *Food Chem Toxicol* 112:551–562
- Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm2 metrics for validation of QSPR models. *Chemometr Intell Lab Syst* 107(1):194–205
- Ojha PK, Kar S, Roy K, Leszczynski J (2020) Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy. *Nano Energy* 70:104537
- Organisation for Economic Co-operation and Development (OECD) (2004) The Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q) SARs] on the Principles for the Validation of (Q) SARs. Series on Testing and Assessment, p 206
- Pope P, Webster J (1972) The use of an F-statistic in stepwise regression procedures. *Technometrics* 14(2):327–340
- Roy K (2007) On some aspects of validation of predictive quantitative structure–activity relationship models. *Expert Opin Drug Discov* 2(12):1567–1577
- Roy K (2018) Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. *J Indian Chem Soc* 95(12):1497–1502
- Roy K, Ambure P (2016) The “double cross-validation” software tool for MLR QSAR model development. *Chemom Intell Lab Syst* 159:108–126
- Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 52(2):396–408
- Roy K, Kar S, Ambure P (2015a) On a simple approach for determining applicability domain of QSAR models. *Chemom Intell Lab Syst* 145:22–29
- Roy K, Kar S, Das RN (2015b) Statistical methods in QSAR/QSPR A primer on QSAR/QSPR modeling. Springer, New York, pp 37–59
- Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometr Intell Lab Syst* 152:18–33
- Roy K, Ambure P, Kar S (2018a) How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? *ACS Omega* 3(9):11392–11406
- Roy K, Ambure P, Kar S, Ojha PK (2018b) Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J Chemom* 32(4):e2992
- Roy J, Ghosh S, Ojha PK, Roy K (2019) Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). *Environ Sci Nano* 6(1):224–247
- Rücker C, Rücker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47(6):2345–2357
- Schüürmann G, Ebert R-U, Chen J, Wang B, Kühne R (2008) External validation and prediction employing the predictive squared correlation coefficient—test set activity mean vs training set activity mean. *J Chem Inf Model* 48(11):2140–2145
- Stošić B, Janković R, Stošić M, Marković D, Stanković D, Sokolović D, Veselinović AM (2020) In silico development of anesthetics based on barbiturate and thiobarbiturate inhibition of GABAA. *Comput Biol Chem* 88:107318
- Todeschini R (2010) Milano Chemometrics. University of Milano Bicocca, Milano, Italy (personal communication)
- Todeschini R, Ballabio D, Grisoni F (2016) Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models. *J Chem Inf Model* 56(10):1905–1913
- Veeramany R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK (2011) Validation of QSAR models-strategies and importance. *Int J Drug Des Discov* 3:511–519
- Venkatasubramanian V, Sundaram A (2002) Genetic algorithms: introduction and applications. In: Encyclopedia of computational chemistry 2. Wiley, New Jersey
- Wainer J, Cawley G (2021) Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Syst Appl* 182:115222
- White Paper on a Strategy for a Future Chemicals Policy. Commission of the European Communities. (2001) Brussels, Belgium
- Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58(2):109–130

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.