Application of quantitative Read-Across Structure-Property Relationship (q-RASPR) approach in materials property predictions

Thesis submitted in partial fulfillment for the requirements of the Degree of

MASTER OF PHARMACY

Faculty of Engineering and Technology

Thesis submitted by

SHUBHAM KUMAR PANDEY

B. PHARM.

Registration No. **163659 of 2022-23**

Examination Roll. No**. M4PHC24006**

Under the Guidance of

DR. KUNAL ROY

Professor

Drug Theoretics & Cheminformatics Laboratory

Department of Pharmaceutical Technology

Jadavpur University

Kolkata – 700 032

India

2024

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research as part of my work on "Application of quantitative Read-Across Structure-Property Relationship (q-RASPR) approach in materials property predictions".

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

NAME: **SHUBHAM KUMAR PANDEY**

EXAMINATION ROLL NUMBER: **M4PHC24006**

REGISTRATION NUMBER: **163659 of 2022-23**

THESIS TITLE: "**Application of quantitative Read-Across Structure-Property Relationship (q-RASPR) approach in materials property predictions"**

Shubham Kumar Pandey 27/8/2024

SIGNATURE WITH DATE

CERTIFICATE

Department of Pharmaceutical Technology
Jadavpur University
Kolkata - 700 032

This is to certify that **Mr. Shubham Kumar Pandey**, B. Pharm. (CSJM University, Kanpur), has carried out the research work on the subject entitled **"Application of quantitative Read-Across Structure-Property Relationship (q-RASPR) approach in materials property predictions"** under my supervision in Drug Theoretics & Cheminformatics Laboratory (DTC LAB) in the Department of Pharmaceutical Technology of this university. He has incorporated his findings into this thesis of the same title, being submitted by him, in partial fulfillment of the requirements for the degree of Master of Pharmacy of Jadavpur University. He has carried out this research work independently and with proper care and attention to my entire satisfaction.

27.08.2024.

Dr. Kunal Roy
Professor,
Drug Theoretics and Cheminformatics Laboratory,
Department of Pharmaceutical Technology,
Jadavpur University,
Kolkata-700 032

Kunal Roy, PhD, FRSC
Professor & Ex-Head
Department of Pharmaceutical Technology,
JADAVPUR UNIVERSITY,
Kolkata 700 032 (INDIA)
EiC: *Molecular Diversity* (Springer Nature)

27/8/24.

Head, Dept. of Pharmaceutical Technology,
Jadavpur University, Kolkata

Prof. Amalesh Samanta, Ph.D.
Head
Dept. of Pharmaceutical Technology
Jadavpur University, Kolkata, India

Dipak Laha 27.8.24

Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata

DEAN
Faculty of Engineering & Technology
JADAVPUR UNIVERSITY
KOLKATA-700 032

Acknowledgements

I deem it a pleasure and privilege to work under the guidance of ***Dr. Kunal Roy****, Professor, Drug Theoretics & Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata-32. I express my deep gratitude and regards to my revered mentor 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 for his valuable guidance throughout the work that enabled me to complete the work. With a deep sense of thankfulness and sincerity, I acknowledge the continuous encouragement, perpetual assistance, and co-operation from my seniors Dr. Vinay Kumar, Dr. Mainak Chatterjee, Joyita Roy, Arkaprava Banerjee, Sapna Kumari Pandey, and* Souvik Pore*. 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 junior Mobarak Hossain who extended his 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 Late Ram Gopal Pandey and Mrs. Shabitri Pandey for all the love and inspiration without which my dissertation work would remain incomplete. I would also like to thank my brothers Sagar Pandey and Sachin Pandey for their constant support and belief.

Shubham Kumar Pandey 27/8/2024

Shubham Kumar Pandey

Examination Roll No.: ***M4PHC24006***

Department of Pharmaceutical Technology,

Jadavpur University, Kolkata-700032

Preface

This dissertation is presented to partially fulfill the Master of Pharmacy degree in Pharmaceutical Technology. The work spanned over one year and six months. The present study was executed by developing in silico predictive models to derive better RASPR descriptors in the quantitative structure-property relationship (QSPR) paradigm. We considered various material properties for the development of the predictive models.

The general comprehension of the various material types mostly depends on understanding microstructure and atomic and molecular structure. Earlier, experimentation was the only route for detecting material properties that utilized a lot of time, capital, and resources. Despite all this, these experimental procedures are prone to error and sometimes fail to explore the desired outcomes, leading to a loss of capital and resources. Prior knowledge of materials' intrinsic and extrinsic properties would benefit their application in the required field of interest. Knowledge about the structural chemistry/features of the compounds that correspond to the materials property can provide a brief account of how to improve the required property and/or reduce the redundant property. Development of new materials with desired properties is the need of hour in various fields due to increased applications of materials in electrical, energy, health care, and manufacturing industries. So, to reduce the cost of failure and resources, different computational methods are used to develop new efficient materials before their synthesis.

The application of various computational approaches for the prediction of the properties of chemical substances has been an effective alternative to experimental methods. Quantitative structure-property relationship (QSPR) is a statistical method widely used to predict different property-based endpoints. Read-across (RA) is a similarity-based approach for predictions and data gap-filling. It does not involve the development of a mathematical model and, thus, is not a statistical technique. It simply generates the consensus-based predictions of the query compounds. Recently, the concepts of quantitative structure-property relationship (QSPR) and read-across (RA) methods were merged to develop a new emerging cheminformatic tool: read-across structure-property relationship (RASPR).

In the present study, we have modeled different properties of materials (especially, energetic compounds and p-type semiconductors) by using the q-RASPR method. The models developed

have shown acceptable statistical significance. The models developed were also validated rigorously based on internal and external validation strategies. The following analyses have been performed in this dissertation:

Study 1: Machine learning-based q-RASPR predictions of detonation heat for nitrogen-containing compounds.

Study 2: Predicting the performance and stability parameters of energetic materials (EMs) using a machine learning-based q-RASPR approach.

Study 3: Predictive cheminformatics modeling of reorganization energy (RE) for p-type organic semiconductors: Integration of quantitative read-across structure-property relationship (q-RASPR) and stacking regression analysis.

The accomplished 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: Result and Discussion

Chapter 5: Conclusion

Chapter 6: References

Appendix: Reprints

Abbreviations

HEDMs	High Energy Density Materials	CoMFA	Comparative Molecular Field Analysis
MI	Materials Informatics	CoMSIA	Comparative Molecular Similarity Indices Analysis
DFT	Density Functional Theory	OECD	Organisation for Economic Co-operation and Development
OQMD	Open Quantum Materials Database	WAP	Weighted Average Prediction
NoMaD	Novel Material Discovery	AI	Artificial Intelligence
MD	Molecular Dynamics	RF	Random Forest
ML	Machine Learning	GB	Gradient Boosting
QSPR	Quantitative Structure-Property Relationship	XGB	Extreme Gradient Boosting
QSAR	Quantitative Structure-Activity Relationship	SVM	Support Vector Machine
QSTR	Quantitative Structure-Toxicity Relationship	LSVM	Linear Support Vector Machine
RA	Read-Across	RR	Ridge Regression
RASPR	Read-Across Structure-Property Relationship	PLS	Partial Least Squares
RASAR	Read-Across Structure-Activity Relationship	AB	Adaptive Boosting
LFER	Linear Free Energy Relationship	ED	Euclidean Distance
MSA	Molecular Shape Analysis	GK	Gaussian Kernel
BSS	Best Subset Selection	LK	Laplacian Kernel
MLR	Multiple Linear Relationship	CTC	Close Training Compound
LV	Latent Variables	AD	Applicability Domain
CV	Cross-Validation	GA	Genetic Algorithm
MAE	Mean Absolute Error	RMSEC	Root Mean Square Error of Calibration
LOO	Leave-One-Out	RMSEP	Root Mean Square Error of Prediction
SHAP	Shapley Additive explanation	OSCs	Organic Semiconductors
EMs	Energetic Materials	RE	Reorganization Energy

Contents

Chapter 1

Introduction

1. INTRODUCTION

1.1 Materials science

Materials have always played a crucial role in the development of human civilization. The development of new objects can't be processed without any prior knowledge of the properties of the materials to be used. Materials possess unique physical, chemical, mechanical, thermodynamic, and electronic properties that can be harnessed to create new materials or improve existing ones, resulting in innovative and practical applications (Yu et. al., 2021). The nature of chemical bonds, atom ordering, and microstructure of the materials are the key components for determining materials' properties. So, one can consider that the behavior of materials limits the development and performance of the machine and/or equipment. Understanding materials' properties is required to develop new technologies and improve the quality of life worldwide. The chemical space of materials is so vast due to their broad composition and configurational degree of freedom (Pilania et. al., 2013).

Materials science is a multidisciplinary field that combines chemistry, physics, engineering, and many other sciences to study the properties of solid materials and how the material's composition and structure are linked to those properties. Materials properties can be specified using either the microscopic or the macroscopic attributes. Features like electron affinity, band energy, molecular atomization energy, lattice constant, etc. are used to define the microscopic attributes. The link between the physical and mechanical properties of the materials characterizes the macroscopic view of the materials. The microscopic features influence the macroscopic performance of the materials (Stergiou et. al., 2023.). The primary objective of materials science is to determine the relationships between a material's composition, atomic or molecular structure, microstructure, and macroscopic characteristics. Knowledge about materials and their respective intrinsic or extrinsic properties is necessary for their application in a particular field. Materials science is enhanced by materials engineering, as it deals with processes that involve manufacturing, transformation, and shaping of materials. As shown in **Figure 1.1**, the four major aspects related to materials science and technology are (Mercier et. al., 2002):-

i. Composition, structure
ii. Synthesis, manufacturing, processing
iii. Properties

iv. Performances

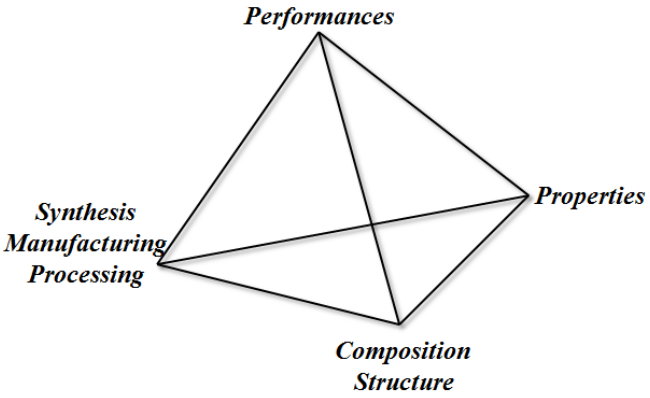

Figure 1.1: Major aspects of materials science

With the fast development and technological advancement in recent years, the rate of material discovery has increased to multiples. Materials property prediction is a complex task as materials' properties depend on various factors like geometry, material constitution, electronic characteristics, etc. Prediction of property and its optimization is essential for the development of new advanced/innovative entities. A long time ago, experiments were the only way to analyze different properties associated with compounds. There is no doubt that experiments provide us with the exact knowledge of the property, but at the same time, it requires a lot more time, capital, manpower, resources, and proper experimental set-up to carry out the experiments. Experimental procedures may give rise to errors (mechanical, human, or instrumental) in predictions arising due to instrument faults, inappropriate testing procedures, changes in environmental factors, etc. (Yu et. al., 2021).

Table 1.1 lists some of the commonly used material types in materials science and their relative properties of concern**.**

Type of material	Related properties	References
Ceramics	Hardness, thermal conductivity, thermal expansion, porosity, creep, chemical stability, optical properties, brittleness, entropy	Wachtman et. al., 2009
Polymers	Glass transition temperature, refractive index, Young's modulus, transparency, Melt Flow Index (MFI), absorption and swelling, decomposition rate	Askadskiĭ, 2003

Light absorbing materials	Band gap, absorption spectra, power conversion efficiency, optical density	Yu et. al., 2012
Energetic materials/ High energy density materials (HEDMs)	Detonation velocity, detonation pressure, density, thermal decomposition, impact sensitivity, melting point, detonation heat, heat of formation	Agrawal, 2010
Nanomaterials	Shape and size, surface area to volume ratio, magnetic properties, electronic properties, catalytic properties, toxicity and biocompatibility, surface charge	Asha and Narain, 2020
Semiconductors	Band gap, carrier mobility, thermal conductivity, photoelectric effect characteristics	Peter and Cardona, 2010
Composites	High strength-to-weight ratio, corrosion resistance, electrical insulation, ductility, hardness, temperature resistance (high/low), damping capacity, biocompatibility, porosity	Clyne and Hull, 2019; Chawla, 2012

1.2 Materials Informatics (MI)

As the experimental procedures are prone to errors and are long time consuming, researchers nowadays have shifted towards data-driven approaches to predict the materials' properties. *In-silico* methods have revolutionized the field of material science. Advancements in computational power and the development of new software tools enable researchers to access the materials on a large scale. Scientists have relied on different computational approaches because of the high cost of material synthesis and poor success rate (Ramakrishna et al., 2019). Materials science, in collaboration with information science, has led to the development of a new branch of materials science called "Materials informatics" (MI) (Takahashi and Tanaka, 2016). MI aims to develop a relationship between the molecular structures and properties associated with materials (Agarwal and Choudhary, 2019). Accessibility to the publicly available large databases generated through experimental results and/or computational simulations is advantageous for the development of MI. These large databases containing information on the properties of materials help the researchers to identify and correlate the patterns of the compounds. These correlations are further used for the development of predictive models to determine the behaviors of the materials (Lopez-Bezanilla and Littlewood, 2020). MI leverages advanced computational techniques and data-driven methods to accelerate the discovery, development, and optimization of materials. This emerging field addresses the

challenges faced in traditional materials science approaches, such as the time-consuming and costly nature of experimental trial and error. MI is applied across various industries, including electronics, energy, healthcare, and manufacturing, where the demand for innovative materials with specific functionalities continues to grow. Collaborations between materials scientists, chemists, physicists, computer scientists, and engineers are crucial for advancing the field and realizing its full potential.

The density functional theory (DFT) is one of the oldest computational methods used to predict the physical and chemical characteristics of crystalline materials (Kohn, 1999; Hafner et. al., 2006). Using DFT, approximately 10^4-10^6 materials properties have been calculated that are stored in large databases like Open Quantum Materials Database (OQMD) (Saal et. al., 2013; Kirlin et. al., 2015), the Automatic Flow of Materials Discovery Library (AFLOWLIB) (Curtarolo et. al., 2012), the Materials Project (Jain et. al., 2013), Joint Automated Repository for Various Integrated Simulations (JARVIS) (Choudhary et al., 2017; Choudhary et al., 2018), and the Novel Materials Discovery (NoMaD) (http://nomad-repository.eu/cms/.). With the availability of such large data sets, *in-silico* approaches can be used to design, optimize, and/or discover properties of de novo designed or untested compounds. Applying the cheminformatics approach in materials science helps to analyze and model the structural and electronic characteristics of materials for a particular physical, chemical, or mechanical property. Computational methodologies such as DFT, MD (molecular dynamics), Monte Carlo techniques, phase-field method, etc., are some of the existing theories that can be used to predict the property. Cheminformatics has been used in different fields for materials property prediction of nanomaterials (Malkiel et. al., 2018), microplastics (Li et. al., 2022), polymers (Doan Tran et. al., 2020), composites (Liu et. al., 2022), ceramics (Han et. al., 2022), photovoltaic cells (Gregg and Hanna, 2003), energetic materials, light-emitting diodes, etc. Due to more experimental and simulation data availability, ML (machine learning) provides an interesting platform for determining material behavior under different conditions and property predictions (Tercan et. al., 2018). Investigation of physical, chemical, and mechanical properties like Young's modulus of elasticity, yield strength, thermal conductivity, high thermal stability, and impact sensitivity are being calculated using computer simulations (Xie et. al., 2021). Some electrical, optical, phase-transitions, and crystal structure characteristics of materials can also be identified using simulation techniques (Stein et. al., 2019).

1.2.1 Quantitative structure-property relationship (QSPR)

Most molecular discoveries today are the results of an iterative, three-phase cycle of design, synthesis, and testing. Analysis of the results from one phase provides knowledge that enables the next cycle of discovery to be initiated and further improvement to be achieved. A common feature of this analysis stage is the construction of some form of model that enables the observed activity or properties to be related to the molecular structure. Such models are often referred to as Quantitative Structure-Activity Relationships.

The Quantitative Structure-Activity Relationship (QSAR) paradigm is based on the hypothesis that a fundamental relationship exists between the molecular structure and biological activity. Based on this assumption, QSAR attempts to establish a correlation between the various molecular properties of a set of molecules and their experimentally known biological activity. According to the type of response, or "endpoint," there are three main classes of studies: quantitative structure-property/activity/toxicity relationship (QSPR/QSAR/QSTR) studies that take into account the modeling of physicochemical property, biological activity, and toxicological data, respectively (Roy et. al., 2015). However, the term QSAR can be used in general to refer to all three studies. The QSPR (Ferreira, 2001) study deals with the molecular features governing their physicochemical properties. The descriptors measure the properties of the molecules and their hydrophobic, steric, and electronic features in addition to the various structural patterns. The QSTR (Carlsen et. al., 2009) technique determines the structural attributes of the molecules responsible for their toxicity profile. The pharmacophoric features and descriptors obtained from the developed QSAR models may also be utilized for the virtual screening of large numbers of diverse compounds for a definite response parameter. Besides this, identifying the prime features providing improved activity to the molecules under a particular study facilitates the *in-silico* design of new molecules with enhanced potency. Thus, a focused library (Tikhonova et. al., 2004) may be developed by compiling the newly designed molecules with a specific response.

This kind of relationship between molecular structures and changes in their property developed on a quantitative basis is the focus for quantitative relationship-based studies. Such correlation represents predictive models derived from applying statistical tools correlating response data of molecules (including therapeutic activity, property, and toxicity) of chemicals with descriptors representative of molecular structure and/or property (Selassie and Verma, 2003). These correlations may be qualitative (simple SAR) or quantitative (QSAR). This quantitative

technique of analyzing the structure-based analysis of molecules enables us to identify the structure-property relationships of molecules in a precise way. QSAR analysis is based on the notion that activity (A) depends on structure (C) and physicochemical properties (P) of the molecules:

Chemical Response (Chemical attributes) =f (Chemical attributes) = f (Structure, Property) (1.1)

The fact that a molecular structure determines its physicochemical properties is well imitated from Mendeleev's periodic table. The advent of QSAR can be dated back to the era of Hansch when Hansch and co-workers correlated the plant growth regulatory activity of phenoxyacetic acids to Hammett constants and partition coefficient (Hansch et. al., 1962). They showed that biological activity could be correlated linearly with free-energy-related terms, a model referred to as the Linear Free Energy Relationship (LFER) model. The introduction of Hansch's linear and parabolic models considerably impacted the understanding of how chemical structures influenced biological activity. The Free-Wilson approach determined the contributions made by various structural fragments to the molecules' overall biological activity (Heritage and Lowis, 1999). Hansch and Free-Wilson analyses thus proposed the concepts of classical QSAR involving structure-activity relationships in terms of physicochemical parameters, steric properties, and certain structural features. Later, Fujita-Ban (Leonard and Roy, 2004) modified the approach of the Free-Wilson model and proposed a substituent-based structure-activity relationship that determines the type and position of the substituents exerting the prime influence on the activity profile of these molecules. QSAR models are pattern recognition models that identify trends in structural features correlating with the experimental property. QSAR models are useful in several cases, such as suggesting structural modifications to enhance molecular property. Such quantitative approaches are being applied in many disciplines like risk assessment, toxicity prediction, and regulatory decisions (Tong et. al., 2005) apart from drug discovery and lead optimization. In '80s, several 3D (three dimensional) quantitative relationship approaches like molecular shape analysis (MSA), distance geometry, comparative molecular field analysis (CoMFA) comparative molecular similarity indices analysis (CoMSIA), hypothetical active site lattice (HASL), receptor surface analysis (RSA), molecular similarity matrices, comparative binding energy (COMBINE) have emerged (Geronikaki et. al., 2004).

Quantitative structure-property relationships (QSPRs) studies undeniably are of great importance in the field of materials science. Quantitative structure–property relationship (QSPR) models are quantitative regression methods that endeavor to relate chemical structure to property. Quantitative structure-property relationship and related methods have been applied extensively in a wide range of scientific disciplines, including material informatics, drug discovery, chemical property prediction, etc. (Wu et. al., 2013). QSPR models are now regarded as scientifically credible tools for predicting and classifying the properties of untested chemicals. QSPR method has become an essential tool in different industries, from discovering new material with desired properties to developing that material (Sukumar et al., 2012; Du et al., 2021; Le and Winkler, 2018). For example, a growing trend is to use QSPR early in the material development process as a screening and enrichment tool to eliminate from further development those chemicals lacking desired properties or predicted to have poor outcomes.

1.2.1.1 Objectives of QSPR

The principal objectives of QSPR analysis are:

1. Prediction of new analogues of compounds with better property

2. Better understanding and exploration of the effect of molecular structure on material property

3. Optimization of the chemical structure to get the desired properties.

4. Reduction of cost, time, and manpower requirements by developing more effective compounds using a scientifically less exhaustive approach.

To achieve the objectives as mentioned earlier, it is necessary to have a detailed knowledge of the following aspects:

(i) Various factors controlling the experimental condition of the molecules.

(ii) A thorough examination of molecular structures and their properties. Quantitative structure-property relationship is an interdisciplinary study of chemistry, statistics, and computer science. By the prediction of the essential structural requirements needed for obtaining a molecule with optimized properties, QSPR analysis provides a good platform for the synthesis of a relatively lower number of chemicals with an improved property of interest.

1.2.1.2 Descriptors

Molecular descriptors are terms that characterize specific information about a studied molecule. They are the "numerical values associated with the chemical constitution for correlation of chemical structure with various physical properties, chemical reactivity, or biological activity" (Van de Waterbeemd et. al., 1997; Randic, 1997). In other words, the modeled property is represented as a function of quantitative values of structural features or properties that are termed descriptors for a QSPR model. Cheminformatics methods depend on generating chemical reference spaces into which new chemical entities are predictable by the developed QSPR model. The definition of chemical spaces significantly depends on the use of computational descriptors of studied molecular structure, physical or chemical properties, or specific features.

Response (property) = f (information in the form of chemical structure or property) = f (descriptors) (1.2)

The type of descriptors used and the extent to which they can encode the structural features of the molecules correlated to the property are critical determinants of the quality of any QSPR model. The descriptors may be physicochemical (hydrophobic, steric, or electronic), structural (based on the frequency of occurrence of a substructure), topological, electronic (based on molecular orbital calculations), geometric (based on a molecular surface area calculation), or simple indicator parameters (dummy variables).

It is interesting to point out that the efficacy of a descriptor can rely heavily on the problem being considered. More precisely, specific endpoints may need to consider exact molecular features. The best possible features that make a descriptor ideal for the construction of a QSPR model are summarized here:

1. A descriptor must be correlated with the structural features for a specific endpoint and show negligible correlation with other descriptors.

2. A descriptor should apply to a broad class of compounds.

3. A descriptor that can be calculated rapidly and does not depend on experimental properties can be considered more suitable than one that is computationally exhaustive and relies heavily on experimental results.

4. A descriptor should generate dissimilar values for structurally different molecules, even if the structural differences are small. This means that the descriptor should show minimal degeneracy. In addition to degeneracy, a descriptor should be continuous. It signifies that small structural changes should lead to small changes in the value of the descriptor.

5. It is always important that the descriptor has some form of physical interpretability to encode the query features of the studied molecules.

6. Another significant aspect is the ability to map descriptor values back to the structure for visualization purposes (Segall et. al., 2009). These visualizations are only sensible when descriptor values are associated with structural features.

1.2.2 Read-Across (RA)

Among the various in-silico approaches, the QSPR method is one of the most popular methods for developing predictive models. QSPR is a statistical model-building approach that requires significant data points to build a meaningful model. In addition, the whole dataset needs to be divided into training and test sets for validation purposes to fulfill the requirements as recommended by OECD guidelines (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm) (Cherkasov et. al., 2014). Thus, a part of the dataset is kept aside for model validation that cannot be used for model building. In the case of small datasets, this type of data loss may lead to the development of a statistically unreliable model due to lower degrees of freedom. In such cases, different similarity-based approaches are used for the prediction that involves simple algebraic operations, and no data points are wasted (Chatterjee et al., 2022).

Read-across is a similarity-based grouping technique that involves simple algebraic operations and uses the similarity between two chemical compounds to make predictions (Berggren et al., 2015). It is a non-experimental data gap-filling method that provides information for the property of a target compound derived from known property data of source compounds with a similar chemical profile. It is one of the most essential *in-silico* methods used for data generation, data gap-filling, and regulatory decision-making (Kovarich et. al., 2019). Read-across method can be classified into two groups, one is qualitative read-across and the other is quantitative read-across (Patlewicz et. al., 2018). The target compounds are generally known as query chemicals and structural analogues which have known property data are known as source compounds. The predictions from this method are generally obtained by either analogue

or category approach. The analogue approach considers only one source compound but in the category approach multiple close source compounds are considered depending on the availability of data which makes it more reliable and robust (Patlewicz et. al., 2017).

Although this method involves only simple algebraic calculations, the algorithm becomes computationally inexpensive and can be used for small datasets. The read-across prediction of a compound can be calculated in different ways one of the methods is by taking the similarity weightage of the response value of the close source compound (Chatterjee et. al., 2022), which is calculated by using the following equation:

$$Weighted\ Average\ Prediction\ (\overline{x_{wt}}) = \frac{\sum W_i \times Y_i}{\sum W_i} \quad (1.3)$$

W_i = weightage of ith source compounds which is calculated based on the similarity with the target compound, Y_i = property value of the ith source compound

1.2.3 Read-across structure-property relationship (RASPR)

Although the read-across method is useful for the dataset with a limited number of data points with experimental data, the main disadvantage of this method is that it does not provide any information on the quantitative contribution of each descriptor. Another similarity-based approach like the read-across structure-property relationship (RASPR) – similar to the read-across structure-activity relationship (RASAR), generates a mathematical model using the similarity and error-based measures as descriptors and has been used for predictive modeling. The RASAR method was first introduced by Luechtefeld et al. (Luechtefeld et al., 2018) who developed the classification-based RASAR models. In contrast, Banerjee and Roy were the first to develop the regression-based quantitative RASAR (q-RASAR) models (Banerjee and Roy, 2022). The RASPR method is a combined method of read-across and QSPR that encapsulates the advantages of both of these methods and generates enhanced predictivity. This method uses selected structural and physicochemical descriptors to generate different similarity and error-based measures (known as RASPR descriptors) from the similarity-based read-across approach (Banerjee et. al., 2022). These measures are merged with the initial structural and physicochemical descriptors, and further feature selection algorithms are employed to develop RASPR models. The description of the RASPR descriptors is shown in **Table 1.2**.

Table 1.2: Definition of RASPR descriptors

RASPR descriptors	Description	Mathematical Equation
SD_similarity	It represents the standard deviation of the similarity levels of the selected close training compounds (CTCs).	$SD_{similarity} = \sqrt{\frac{\sum_{i=1}^{n}(f_i - \bar{f})^2}{n-1}}$ n = number of CTC f_i = similarity level of selected CTC $\bar{f}$ = mean similarity levels of CTC
CV similarity (CVsim)	It represents the coefficient of variation of the similarity levels of the selected CTCs	$CVsim = \frac{SD\_similarity}{\bar{f}}$
Avg.Sim	It is the mean of the similarity levels to the selected CTCs	$Avg.Sim.(\bar{f}) = \frac{\sum_{i=1}^{n} f_i}{n}$
Pos.Avg.Sim	It is the mean of the similarity levels to the positive CTCs	
Neg.Avg.Sim	It is the mean of the similarity levels to the negative CTCs	
MaxPos	It is the maximum similarity level to the CTC with response value of more than the average response value of the training set.	

MaxNeg	It is the minimum similarity level to the CTC with response value of less than the average response value of the training set.	
Abs MaxPos-MaxNeg	It is the absolute difference of the MaxPos and the MaxNeg values.	$Abs\ Diff = \lvert MaxPos - MaxNeg \rvert$
s_m^1	Banerjee-Roy similarity coefficient 1 (can be used to analyze modelability of a set)	$s_m^1 = \frac{MaxPos - MaxNeg}{argmax\ (MaxPos,\ MaxNeg)}$
s_m^2	Banerjee-Roy similarity coefficient 2 (can be used to analyze modelability of a set)	$s_m^2 = \frac{Pos.Avg.Sim - Neg.Avg.Sim}{Avg.Sim}$
RA function	It is a composite function of all the selected molecular features that is derived from read-across.	$RA\ function = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ $w_i = \frac{S_i}{\sum_{i=1}^{n} S_i}$ Where, w_i= weightage of each CTC, S_i = similarity between each CTC and query compound, andx_i = observed response values of CTC
SD_activity	It represents the weighted standard deviation of response values of the selected CTCs.	$SD_{activity} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_{wt}})^2}{\sum_{i=1}^{n} w_i}} \times \frac{n}{n-1}$

CV_activity	It represents the coefficient of variance of response values of the selected CTCs.	$CV_{activity} = \frac{SD\_activity}{\overline{x_{wt}}}$
g_m	Banerjee-Roy concordance coefficient	$g_m = (-1)^n \lvert Posfrac - 0.5 \rvert$ Where, n = 1 if MaxPos<MaxNeg, n = 2 if MaxPos>MaxNeg, and Posfrac is the fraction of CTC with response value more than the mean response of the training set.
g_m**SD_similarity*	It is the product of g_m and *SD_Similarity*	
g_m**Avg.Sim*	It is the product of g_m and *Avg. Sim*	
g_m*_class*	A modified form of g_m describing the propensity of a query compound to be positive or negative	
SE	It represents the weighted standard error of the response values of the selected CTCs.	$SE = \frac{SD\_activity}{\sqrt{n}}$

1.2.4 Machine learning (ML)

ML is a part of artificial intelligence (AI) that enables machines to learn from previous data, improve performance based on previous experiences, and predict new data points. At its core, ML involves the development of algorithms and models that enable computers to recognize patterns, make predictions, and make decisions based on data. The process begins with the collection and preparation of relevant data serving as the foundation for training these algorithms. Through exposure to this data, ML models can identify underlying patterns and relationships, allowing them to generalize their understanding and make accurate predictions or classifications when tested with new and unseen data (Jordan and Mitchell, 2015). For different types of data problems, ML relies on different types of algorithms that are classified into three main groups – supervised, unsupervised, and reinforcement ML algorithms. In the supervised ML algorithm, the labelled data is used to train the algorithm, whereas, in unsupervised ML, the data is unlabelled. The reinforcement algorithm is a feedback-based learning method where the learning agent is rewarded for every right action and gets a penalty for the wrong action (Geron, 2022). Currently, ML algorithms have moved beyond purely theoretical applications to practical applications like the creation of new molecules (Lo et. al., 2018). ML models and methods have proven to be effective for solving complicated problems by learning from the data; however, there are also some disadvantages associated with different ML models including the need for large amounts of high-quality data, complex algorithms, and difficulty in interpretation of results (Geron, 2022). Despite these challenges, ML methods have grown rapidly with more powerful algorithms and techniques. Currently, the field of "explainable AI" has attracted lots of attention, which helps ML models to provide interpretable explanations for particular predictions (Linardatos et. al., 2020). SHAP or Shapley additive explanation analysis is one of the important methods used for the interpretation of the ML models (Yosipof et. al., 2016).

Chapter 2

Present Work

2. PRESENT WORK

The limited resources and cost involved in experimentation have slowed down the process of development of new materials. The development of new materials is successful through a lot of trial and error during experimentation. Due to the loss of resources, time, money, manpower, etc., different molecular modeling techniques are being increasingly used as an alternative to experimentation. In the last few decades, the advent of computational tools such as quantitative structure-property relationship (QSPR) has provided significant insight into materials science (Yosipof et. al., 2016). QSPR can be simply defined as mathematical relationships linking a compound's property with its chemical structure in a qualitative/quantitative manner. The guidelines specified by the Organization for Economic Cooperation and Development (OECD) (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm) are followed to develop a QSPR model. Performing QSPR analysis for small datasets is always possible, so one can use different similarity-based prediction approaches to develop predictive models. Read-across (RA) is one of the most popular similarity-based methods that can be used for data generation and data gap-filling (Chatterjee et. al., 2022). The read-across structure-property relationship (RASPR) is another approach that incorporates both structural and physiochemical features of QSPR and similarity and error measures of RA for the development of the model. The RA-based similarity and error measures are also known as RASPR descriptors.

The development of predictive models in the form of RASPR analysis provides a well-validated rational platform for the determination of the properties of all the new chemicals and to fill data gaps.

In this present study, we have utilized the quantitative read-across structure-property relationship (q-RASPR) algorithm to determine or predict the properties of materials like energetic compounds and p-type semiconductors. For the energetic materials, various properties related to their performance and stability were calculated using the q-RASPR models developed in the studies. For the p-type semiconductors, the mobility of charge carriers is determined through the prediction of reorganization energy (RE) using the q-RASPR model developed via stacking regression. During the analysis of the models, we found that the incorporation of the similarity and error measures derived from RA had led to the enhancement in the external predictivity of the models. We have used the Euclidean distance-based, Gaussian

kernel-based, and Laplacian kernel-based similarity algorithms for the calculation of RASPR descriptors.

2.1 Study 1

High energy density materials (HEDMs) are a class of compounds or combinations of compounds with explosive groups or oxidants and incendiary materials (He et. al., 2021). The performance and stability of these energetic materials (EMs) depend on several parameters such as detonation heat, detonation pressure, detonation velocity, density, the heat of formation, impact sensitivity, chemical degradation/decomposition, electrostatic discharge, etc. (Huang et al., 2021). The wide applications of these EMs are in civil, military, and industrial fields.

In this study, we have opted for the quantitative read-across structure-property relationship (q-RASPR) approach [an analog to quantitative read-across structure-activity relationship (q-RASAR)] to develop a predictive model for the prediction of detonation heat of different N-containing compounds. The heat of detonation (Q) refers to the quantity of heat energy liberated by an energetic compound per unit when detonated (Infante-Catillo and Hernandez-Rivera, 2012). The incorporation of nitrogen into the parent structure or the addition of a nitrogen-containing substituent enhances the heat of detonation of the EMs as their energy content is predominantly derived from the heat of formation due to a large number of dynamic N–N and C–N bonds instead of coming thoroughly from the heat of combustion. Also, the final detonation product of these nitrogenous compounds is dinitrogen (N2), which is less toxic to the environment (Jaidann et. al., 2010; Yin et. al., 2016). A set of 162 nitrogenous compounds was used in this study, collected from the work of He et. al. (He et. al., 2021). The data set contains information on detonation heat (expressed in KJ/kg) for 162 compounds, both aromatic and non-aromatic.

2.2 Study 2

In this study, we had developed several predictive models for the prediction of different properties of EMs corresponding to their performance and stability. The predictive models were developed using the RASPR approach. Here, performance parameters such as density and heat of formation were used while for the thermal stability decomposition temperature and melting point were used for the modeling purpose. The datasets were collected from 2 different literature sources containing the experimental values for each dataset (Wespiser and Mathieu, 2023). The in-house data derived by Wespiser et. al. was used for the decomposition temperature, the Bradley melting point dataset was used in previous work by Wespiser et. al.

for melting point, the Crystallography Open Database was used to collect the data for the density dataset, and experimental data for the heat of formation was collected from the work of D. Mathieu (Mathieu, 2018).

2.3 Study 3

Organic semiconductors (OSCs), being light in weight, decomposable, cheap, and flexible, can be an excellent replacement for inorganic semiconductors. The p-type semiconductors are a crucial component in semiconductor physics and device engineering. They represent a class of semiconductors where most charge carriers responsible for electrical conduction are positively charged "holes" rather than negatively charged electrons (n-type SCs). The p-type semiconductors play a fundamental role in semiconductor technology, offering versatility and enabling the design and fabrication of diverse electronic devices essential to modern life. Organic semiconductors' reorganization energy (RE) is a critical parameter that influences their charge transport properties. RE (λ) can be defined as the energy required for the geometric relaxation during charge transfer. Since OSCs are used as an active layer for many OLEDs, OFETs, etc., they can contribute to developing efficient renewable energy sources with better energy efficiency, and reduced toxicity (as it does not contain any heavy material).

In this study, we have used a set of 173 molecular p-type OSCs which contains a diverse set of organic compounds having moieties of acenes, thiophenes, thienoacenes, and some anti-aromatic pantalenes. The experimental RE values for the compounds were collected from previously published literature (Atahan-Evrenk, 2018).

Table 2.1 provides a brief overview of the type of materials, their related properties, and the number of data points used in the above-mentioned studies**.**

Table 2.1: Description of the datasets.

Study	Material type	Property	Unit	No. of compounds	Reference
1	N-containing EMs	Heat of detonation	kJ/kg	162	Pandey et. al., 2023
2	Energetic materials	Decomposition temperature	^{o}C	565	Pandey and Roy, 2024
		Melting point	^{o}C	19667	
		Density	g/cm^3	12805	
		Enthalpy of formation	kJ/mol	2565	
3	p-type OSCs	Reorganization energy	LogmeV	173	-

Chapter 3

Materials & Methods

3. MATERIALS & METHODS

The main aim of the present study is the implementation of a transparent methodological framework for the development of predictive models using RASPR descriptors. We have endeavored 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. The section has been divided into the following parts:

- Details of datasets.
- Study-wise specific description of methodologies utilized in each study.

3.1 Details of datasets

3.1.1 Dataset for the nitrogen-containing energetic compounds (Study 1)

This dataset includes 162 nitrogen-containing energetic compounds. 122 compounds were present in the training set, and 40 compounds were in the test set. The detailed dataset used in the study is given in **Table 3.1.**

Table 3.1: Details of N-containing energetic compounds.

S. No.	Observed value of detonation heat (kJ/kg)	SMILES strings
1	3446.08	NNC1=NN=C(NN)N=N1
2	5042.26	[O-][N+]1=C(N)C(C[O-])=NC([N+]([O-])=O)=C1N
3	5380.67	O=[N+](C1=C(N)C([N+]([O-])=O)=NN1)[O-]
4	432.83	ClC1=NC(Cl)=NC(Cl)=N1
5	4316.69	O=C1C=NN([N+]([O-])=O)N1
6	9040.43	O=[N+](N1N=C(/N=N/C2=NN([N+]([O-])=O)C=N2)N=C1)[O-]
7	5079.97	O=[N+](C1=C(N3N=C([N+]([O-])=O)N=C3N)N=CN=C1N2N=C([N+]([O-])=O)N=C2N)[O-]

8	5142.22	O=[N+](C1=NNC=N1)[O-]
9	4332.7	O=[N+](C1=NC(N)=NN1)[O-]
10	4141.69	[O-][N+]1=C(N)[N+]([O-])=C(N)C([N+]([O-])=O)=C1N
11*	5769.03	O=[N+](C1=NC(/N=N/C2=NNC([N+]([O-])=O)=N2)=NN1)[O-]
12	3926.25	NN1NC(N3N=NN=C3)=NN=C1N2N=NN=C2
13	4587.84	NC1=NN=C(/N=N/C2=NN=C(N)N=N2)N=N1
14*	4583.73	[O-][N+]1=NN(N)[NH+]([O-])C=C1N
15	5978.72	O=[N+](C1=NNC2=C1NN=C2[N+]([O-])=O)[O-]
16	4679.37	OCC(C(OC(C)=O)N=[N+]=[N-])(CO)C(N=[N+]=[N-])(N=[N+]=[N-])ON=[N+]=[N-]
17	5893.03	O=[N+](C1=CC([N+]([O-])=O)=CC=C1)[O-]
18	6479.23	O=[N+](C1=CC([N+]([O-])=O)=CC([N+]([O-])=O)=C1)[O-]
19	5685.92	O=[N+](C1=CC=C(C)C([N+]([O-])=O)=C1)[O-]
20	6342.3	CC1=C([N+]([O-])=O)C=C([N+]([O-])=O)C=C1[N+]([O-])=O
21	6241.88	O=[N+](C1=C(C)C([N+]([O-])=O)=C(C)C([N+]([O-])=O)=C1)[O-]
22*	4734.16	O=[N+](C1=CC=C(Cl)C([N+]([O-])=O)=C1)[O-]
23	5602.02	O=[N+](C1=CC([N+]([O-])=O)=C(Cl)C([N+]([O-])=O)=C1)[O-]
24	6269.8	O=[N+](C1=C(O)C([N+]([O-])=O)=CC([N+]([O-])=O)=C1)[O-]
25	5291.44	O=[N+](C1=C(O)C(C)=CC([N+]([O-])=O)=C1)[O-]

26	6218.08	O=[N+](C1=C(O)C([N+]([O-])=O)=C(C)C([N+]([O-])=O)=C1)[O-]
27	5774.82	COC1=CC=C([N+]([O-])=O)C=C1[N+]([O-])=O
28	6384.33	O=[N+](C1=CC([N+]([O-])=O)=C(OC)C([N+]([O-])=O)=C1)[O-]
29*	6620.19	NC1=C([N+]([O-])=O)C=C([N+]([O-])=O)C([N+]([O-])=O)=C1[N+]([O-])=O
30	5277.62	O=[N+](C1=C(N)C([N+]([O-])=O)=C(N)C([N+]([O-])=O)=C1)[O-]
31*	4849.82	O=[N+](C1=C(N)C([N+]([O-])=O)=C(N)C([N+]([O-])=O)=C1N)[O-]
32	5400.78	NNC1=CC=C([N+]([O-])=O)C=C1[N+]([O-])=O
33	6013.23	O=C(O)C1=C([N+]([O-])=O)C=C([N+]([O-])=O)C=C1[N+]([O-])=O
34	5627.11	O=[N+](C1=CC=CC2=C([N+]([O-])=O)C=CC=C12)[O-]
35	5727.06	O=[N+](C1=CC=CC2=CC=CC([N+]([O-])=O)=C12)[O-]
36	6523.65	O=[N+](C1=CC([N+]([O-])=O)=CC2=CC([N+]([O-])=O)=CC([N+]([O-])=O)=C12)[O-]
37*	6897.36	O=[N+](C1=C([N+]([O-])=O)C([N+]([O-])=O)=C([N+]([O-])=O)C([N+]([O-])=O)=C1OC2=CC=CC=C2[N+]([O-])=O)[O-]
38	6212.12	O=[N+](C1=C([N+]([O-])=O)C([N+]([O-])=O)=C([N+]([O-])=O)C([N+]([O-])=O)=C1SC2=CC=CC=C2[N+]([O-])=O)[O-]
39	6784.2	NC1([N+]([O-])=O)C=C([N+]([O-])=O)C(C2=CC=C(N)C=C2)=C([N+]([O-])=O)C1

40	7105.37	O=[N+](C1=C([N+]([O-])=O)C([N+]([O-])=O)=C([N+]([O-])=O)C([N+]([O-])=O)=C1/N=N/C2=CC=CC=C2[N+]([O-])=O)[O-]
41	6426.36	O=[N+](C1=CC([N+]([O-])=O)=CC([N+]([O-])=O)=C1NC2=NON=C2NC3=C([N+]([O-])=O)C=C([N+]([O-])=O)C=C3[N+]([O-])=O)[O-]
42	5271.44	O=[N+](N2CN([N+]([O-])=O)C1=NON=C1N([N+]([O-])=O)C2)[O-]
43	7153.57	O=[N+](N(CCN3[N+]([O-])=O)C2C3N([N+]([O-])=O)C1=NON=C1N2[N+]([O-])=O)[O-]
44	7445.56	O=[N+](N1C(N([N+]([O-])=O)C2=NON=C2N3[N+]([O-])=O)C3N([N+]([O-])=O)C1)[O-]
45	7390.33	O=C3N([N+]([O-])=O)C2NC1=NON=C1N([N+]([O-])=O)C2N3[N+]([O-])=O
46	7603.01	O=[N+](N2C1=NON=C1N([N+]([O-])=O)C4C2N([N+]([O-])=O)C3=NON=C3N4[N+]([O-])=O)[O-]
47	7830.64	O=[N+](N(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)C1=NON=C1C2=NON=C2N([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]
48	6795.58	O=[N+](N(C1=NON=C1N([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]
49	7706.95	[O][N]1=C([N+]([O-])=O)C([N+]([O-])=O)=NO1
50	6326.21	[O][N]1=C2C(C(N)=C([N+]([O-])=O)C=C2[N+]([O-])=O)=NO1
51	7588.94	[O][N]4=C3C1=NO[N]([O])=C1C2=NO[N]([O])=C2C3=NO4

52	5787.93	NC2=C([N+]([O-])=O)C(N)=C([N+]([O-])=O)C1=NO[N]([O])=C12
53	7710.74	NC1=NON=C1/[N]([O])=N\C2=NON=C2N
54	6645.14	N#CC1=NO[N]([O])=C1C#N
55	8250.98	[O][N]1=C([N+]([O-])=O)C(/N=N\C2=NO[N]([O])=C2[N+]([O-])=O)=NO1
56*	7331.25	O=C(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)OC1=NO[N]([O])=C1OC(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O
57*	5502.45	O=[N+](N(CC([N+]([O-])=O)(F)[N+]([O-])=O)CC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
58	5667.22	FC([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)CCN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)F
59	5964	O=C1N([N+]([O-])=O)C(N(CC([N+]([O-])=O)(F)[N+]([O-])=O)[N+]([O-])=O)C(N(CC([N+]([O-])=O)(F)[N+]([O-])=O)[N+]([O-])=O)N1[N+]([O-])=O
60	6226.58	FC([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)F
61	4142.13	O=C(N(CN(F)F)C1[N][N+]([O-])=O)N([N+]([O-])=O)C1(CN(F)F)NC=O
62*	4958.57	O=[N+](C([N+]([O-])=O)(F)COCOCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
63*	1181.38	FC(C1(C(F)(F)F)OC[C]([N+]([O-])=O)([N+]([O-])=O)(F)CO1)(F)F

64	3242.55	O=[C](F)(CC([N+]([O-])=O)(F)[N+]([O-])=O)(CC([N+]([O-])=O)(F)[N+]([O-])=O)F
65	127.94	O=S(C(F)(F)F)(OCC(F)(F)[N+]([O-])=O)=O
66	4840.04	FC([N+]([O-])=O)([N+]([O-])=O)OC(C)OC(F)([N+]([O-])=O)[N+]([O-])=O
67	3162.74	O=[N+](C([N+]([O-])=O)(F)OCC(F)(F)[N+]([O-])=O)[O-]
68	5069.14	O=[N+](C([N+]([O-])=O)(F)COCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
69*	5802.63	O=[N+](C([N+]([O-])=O)(F)COCC([N+]([O-])=O)([N+]([O-])=O)COCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
70	3913.33	O=[N+]([O-])OCC([N+]([O-])=O)(F)[N+]([O-])=O
71	4329.86	O=S(OCC([N+]([O-])=O)(F)[N+]([O-])=O)(OCC([N+]([O-])=O)(F)[N+]([O-])=O)=O
72	665.02	O=[N+](C([N+]([O-])=O)(F)OCC(F)(F)F)[O-]
73	5225.75	O=[N+](C([N+]([O-])=O)(F)COCOCCOCOCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
74*	3726.09	O=[N+](C([N+]([O-])=O)(F)COC(C)OCC(F)(F)[N+]([O-])=O)[O-]
75	3242.38	O=[N+](C([N+]([O-])=O)(F)COCOCC(F)(F)[N+]([O-])=O)[O-]
76	4097.77	O=C(OC)CCC([N+]([O-])=O)(F)[N+]([O-])=O
77	4822.29	O=[N+](C([N+]([O-])=O)(F)COC(OC)OCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
78*	1610.1	O=[N+](C(F)(F)COC(F)(F)OCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]

79	1270.04	O=[N+](C(COC(F)(F)F)([N+]([O-])=O)COC(F)(F)OCC([N+]([O-])=O)([N+]([O-])=O)COC(F)(F)F)[O-]
80	1172.33	O=C(OCC(F)(F)F)CCC([N+]([O-])=O)(F)[N+]([O-])=O
81	3068.09	O=[N+](C([N+]([O-])=O)([N+]([O-])=O)COC(F)(F)F)[O-]
82	641.43	O=[N+](C([N+]([O-])=O)(F)COC(F)(F)F)[O-]
83	1503.22	CC([N+]([O-])=O)([N+]([O-])=O)COC(F)(F)F
84*	624.78	O=[N+](C(COC(F)(F)F)([N+]([O-])=O)COC(F)(F)F)[O-]
85*	5956.91	COCC([N+]([O-])=O)([N+]([O-])=O)CC([N+]([O-])=O)(F)[N+]([O-])=O
86	4997.84	O=[N+](C([N+]([O-])=O)(F)COC(N(F)F)(N(F)F)OCC([N+]([O-])=O)(F)[N+]([O-])=O)[O-]
87	2244.93	O=[N+](C(COC(F)(F)F)([N+]([O-])=O)COCOCC([N+]([O-])=O)([N+]([O-])=O)COC(F)(F)F)[O-]
88*	2656.09	FC(C(OCC([N+]([O-])=O)(F)[N+]([O-])=O)OCC([N+]([O-])=O)(F)[N+]([O-])=O)(F)F
89*	1276.16	FC(C1OC[C]([N+]([O-])=O)([N+]([O-])=O)(F)CO1)(F)F
90	1295	CCOC(C(F)(F)F)(C(F)(F)[N+]([O-])=O)OCC([N+]([O-])=O)(F)[N+]([O-])=O
91	651.1	O=[N+](C([N+]([O-])=O)(F)COC(F)(F)OCC(F)(F)F)[O-]
92*	3272.66	O=[N+](C([N+]([O-])=O)([N+]([O-])=O)COCOCC(F)(F)F)[O-]
93	4724.64	O=[N+](N1CC([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)CC(N(F)F)(N(F)F)C1)[O-]

94	7147.42	O=[N+]([O-])N(C1C2N(C4C3N2[N+]([O-])=O)[N+]([O-])=O)C(N3[N+]([O-])=O)C(N4[N+]([O-])=O)N1[N+]([O-])=O
95	6422.31	O=[N+](N1N3C([N+]([O-])=O)([N+]([O-])=O)C2([N+]([O-])=O)N([N+]([O-])=O)N(C3)N([N+]([O-])=O)N1C2)[O-]
96*	7112.75	O=[N+](N([N+]([O-])=O)[NH]1([N+]([O-])=O)[NH]([N+]([O-])=O)([N+]([O-])=O)N=NN=C1[N+]([O-])=O)[O-]
97	9275.51	O=[N+](C12C3([N+]([O-])=O)C5([N+]([O-])=O)C([N+]([O-])=O)1C4([N+]([O-])=O)C([N+]([O-])=O)2C([N+]([O-])=O)3C45[N+]([O-])=O)[O-]
98*	8691.12	O=[N+](C1(C2)CC(C3)([N+]([O-])=O)CC2([N+]([O-])=O)CC3([N+]([O-])=O)C1)[O-]
99	5740.49	O=[N+](N1C([N+]([O-])=O)C1)[O-]
100*	4162.22	NC(N[N+]([O-])=O)=N
101	6555.39	O=[N+](N1CN([N+]([O-])=O)CN([N+]([O-])=O)C1)[O-]
102	5901.58	O=NN1C(N=O)(N=O)CC1
103	6626.72	O=[N+](N1CN([N+]([O-])=O)CN([N+]([O-])=O)CN([N+]([O-])=O)C1)[O-]
104	5947.78	NC1([N+]([O-])=O)C=C([N+]([O-])=O)C(C2=CC=C(N)C=C2)=C([N+]([O-])=O)C1
105*	6808.48	N[N+]([O-])=O.O=[N+]([O-])OCC.O[N+]([O-])=O
106*	6510.31	O=[N+](N1CN([N+]([O-])=O)CN([N+]([O-])=O)C1=C=O)[O-]
107	5412.47	O=C(NC1N2[N+]([O-])=O)N([N+]([O-])=O)C1NC2=O
108	6558.11	O=C(N([N+]([O-])=O)C([N+]([O-])=O)1N2[N+]([O-])=O)N([N+]([O-])=O)C1NC2=O

109	6923.13	O=C(N([N+]([O-])=O)C2N1[N+]([O-])=O)N([N+]([O-])=O)C(C2)N([N+]([O-])=O)C1=O
110	7202.02	O=[N+](C1([N+]([O-])=O)CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)C1)[O-]
111	6577.12	O=[N+]([O-])NC1=CC=C([N+]([O-])=O)C([N+]([O-])=O)=C1[N+]([O-])=O
112	7001.75	O=[N+]([O-])N([N+]([O-])=O)C1=CC=CC([N+]([O-])=O)=C1[N+]([O-])=O.O=[N+]([O-])OCC
113	6049.45	O=[N+](C([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]
114	7362.91	O=[N+](C1([N+]([O-])=O)CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)CN(N=O)C1)[O-]
115*	7613.27	O=[N+](C1([N+]([O-])=O)CN(N=O)CC([N+]([O-])=O)([N+]([O-])=O)CN(N=O)C1)[O-]
116	6996.43	O=[N+](C1([N+]([O-])=O)CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)CN(C(O[N+]([O-])=O)C)C1)[O-]
117*	6934.62	O=[N+](C1([N+]([O-])=O)CN([N+]([O-])=O)CN([N+]([O-])=O)C=C1)[O-]
118	6745.61	O=[N+](N1CC([N+]([O-])=O)([N+]([O-])=O)CN([N+]([O-])=O)CC1)[O-]
119*	5516.18	N/C(N[N+]([O-])=O)=N\[N+]([O-])=O
120	6043.37	[N-]=[N+]=NC1CN([N+]([O-])=O)C1
121	7350.6	O=[N+](C1([N+]([O-])=O)CN([N+]([O-])=O)C1)[O-]
122	3579.88	O=[N+](C([N+]([O-])=O)([N+]([O-])=O)C([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]

123	7413.06	CN([N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O
124	7724.79	NCC(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)N[N+]([O-])=O
125	7296.96	O=C(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O
126*	7131	O=[N+](C([N+]([O-])=O)([N+]([O-])=O)COCOCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]
127	7476.78	O=[N+](CCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]
128*	6051.35	O=[N+](N(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[O-]
129	7456.29	O=[N+](C1=C(C([N+]([O-])=O)=CC([N+]([O-])=O)=C1)N(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[N+]([O-])=O)[O-]
130	7000.29	O=C(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CN(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)[N+]([O-])=O
131	6644.73	O=C(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CCCCC(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O
132	6991.77	O=C(N)N(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O
133*	6361.9	O=C(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)C(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O

134	6953.54	O=C(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)CCC(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O
135	7356.67	O=[N+]([O-])OCC(CO[N+]([O-])=O)(CO[N+]([O-])=O)CO[N+]([O-])=O
136	7007.4	O=[N+]([O-])OCC(O[N+]([O-])=O)CO[N+]([O-])=O
137	7400.5	O=[N+]([O-])OC(C([N+]([O-])=O)C(C)C)C(O[N+]([O-])=O)CO[N+]([O-])=O
138*	6510.14	O=[N+](C([N+]([O-])=O)(OS(O)=O)C[N+]([O-])=O)[O-]
139*	4779.26	O=S(C1=CC=CC=C1)(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O
140	6742.93	O=S(C1=CC=CC([N+]([O-])=O)=C1)(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O
141	6342.09	CC1=CC=C(S(=O)(OCC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=O)C=C1
142	7159.54	CC1=CC(CC([N+]([O-])=O)([N+]([O-])=O)[N+]([O-])=O)=C(S(=O)(O[N+]([O-])=O)=O)C([N+]([O-])=O)=C1
143	5052.5	O=[N+](C1C([N+]([O-])=O)N([N+]([O-])=O)C([N+]([O-])=O)C([N+]([O-])=O)N1[N+]([O-])=O)[O-]
144*	7207	O=[N+](/N=C1/N([N+]([O-])=O)C(CCO[N+]([O-])=O)([N+]([O-])=O)C(CCO[N+]([O-])=O)([N+]([O-])=O)N1[N+]([O-])=O)[O-]
145*	6598.08	OCC(O[N+]([O-])=O)CO[N+]([O-])=O
146	5374.26	OCC(O[N+]([O-])=O)C(O[N+]([O-])=O)Cl
147	5984.08	CC(OCC(O[N+]([O-])=O)CO[N+]([O-])=O)=O

148*	7100.3	O=[N+]([O-])O[C@H](CO[N+]([O-])=O)COC[C@@H](O[N+]([O-])=O)CO[N+]([O-])=O
149*	7064.18	O=[N+]([O-])OCCO[N+]([O-])=O
150	6465.27	O=[N+]([O-])OCCOCCO[N+]([O-])=O
151*	6151.87	O=[N+]([O-])OCCOCCOCCO[N+]([O-])=O
152	6676.02	O=[N+]([O-])OCCCO[N+]([O-])=O
153	6680.69	CC(O[N+]([O-])=O)CO[N+]([O-])=O
154*	6390.64	CC(O[N+]([O-])=O)CCO[N+]([O-])=O
155	7121.1	O=[N+]([O-])OCC(O[N+]([O-])=O)CCO[N+]([O-])=O
156	6869.41	CC(C)CC(O[N+]([O-])=O)(O[N+]([O-])=O)O[N+]([O-])=O
157*	6790.3	CCC(CO[N+]([O-])=O)(CO[N+]([O-])=O)CO[N+]([O-])=O
158	6452.55	O=[N+]([O-])OC
159	5947.78	O=[N+]([O-])OCC
160	5303.14	O=[N+]([O-])OCCC
161*	5124.94	O=[N+]([O-])OC(C)C
162*	6323.65	O=[N+]([O-])OC([N+]([O-])=O)([N+]([O-])=O)COC1=CC=CC=C1

'*' represent the test set compound

3.1.2 Datasets used in Study 2

We have used 4 datasets for four different properties (i.e. decomposition temperature, melting point, density, and heat of formation) that were studied in this work. The overview on the number of compounds in each dataset is already given in **Table 2.1**. The dataset used in this work can be retrieved from the supplementary material section of our published literature entitled "Predicting the performance and stability parameters of energetic materials (EMs) using a machine learning-based q-RASPR approach" (Pandey and Roy, 2024).

3.1.3 p-type organic semiconductors (OSCs) dataset (Study 3)

This dataset consists of 171 compounds, among which 129 compounds were present in the training set and 42 compounds were in the test set. Detailed information on compounds in the study is given in **Table 3.2.**

Table 3.2: Details of p-type OSCs dataset.

S. no.	**Smiles**	**RE (meV)**
1	S1C=CC=C1	403
2*	C1=CC2=CC=CC=C2C=C1	185
3*	S1C=CC2=C1C=CS2	409
4	C1=CC=C(C=C1)C1=CC=CC=C1	358
5	S1C=CC=C1C1=CC=CS1	420
6	C1=CC=C2C(C=CC3=CC=CC=C23)=C1	218
7	C1=CC2=CC3=CC=CC=C3C=C2C=C1	138
8	S1C=CC2=C3SC=CC3=CC=C12	230
9*	S1C=CC2=C3C=CSC3=CC=C12	288
10	S1C=CC2=CC3=C(SC=C3)C=C12	108
11*	S1C=CC2=CC3=C(C=CS3)C=C12	165
12	S1C=C2SC3=C(C=CS3)C2=C1	193
13	S1C=CC2=C1C1=C(S2)C=CS1	352
14	S1C=C2SC3=C(SC=C3)C2=C1	209
15	S1C=C2SC3=CSC=C3C2=C1	187
16	C1=C2C(=CC3=CC=CC=C23)C2=CC=CC=C12	279
17	C1=CC=C2C(C=CC3=C2C=CC2=CC=CC=C32)=C1	165
18	C1=CC2=CC3=CC4=CC=CC=C4C=C3C=C2C=C1	111
19	S1C=CC2=CC3=CC4=CC=CC=C4C=C3C=C12	110
20	S1C=CC2=C3C=CC4=C(C=CS4)C3=CC=C12	243
21	S1C=CC2=C1C1=C(C=CS1)C1=CC=CC=C21	238
22	S1C=CC2=CC3=CC4=C(SC=C4)C=C3C=C12	100
23	S1C=CC2=CC3=CC4=C(C=CS4)C=C3C=C12	105
24*	S1C=CC2=C1C1=CC=C3C=CSC3=C1C=C2	280

25*	S1C2=CC=CC=C2C2=C1C1=C(S2)C=CC=C1	225
26	S1C=C2SC3=CSC4=C3C2=C1\C=C/C=C\4	378
27	S1C=CC=C1C1=CC=C(S1)C1=CC=CS1	373
28	C1=CC2=C3C(C=CC=C3C3=C4C(C=CC=C24)=CC=C3)=C1	145
29	S1C=CC2=C1SC1=C2C2=C(SC=C2)S1	301
30*	S1C=CC2=C1C1=C(S2)C2=C(S1)C=CS2	326
31*	S1C2=C(SC(=C2)C2=CC=CC=C2)C2=CC=CC=C12	299
32	S1C=C(C2=CSC3=CC=CC=C23)C2=CC=CC=C12	302
33	S1C=C2C3=CSC4=C3C(=CS4)C3=CSC1=C23	183
34	C\C=C\C1=CC2=C(S1)SC1=C2C=C(S1)\C=C\C	215
35	C1=CC=C2C(C=CC3=C4C=CC5=CC=CC=C5C4=CC=C23)=C1	185
36	C1=CC=C2C(C=CC3=CC4=C(C=CC5=CC=CC=C45)C=C23)=C1	168
37*	C1=CC=C2C=C3C(C=CC4=CC5=CC=CC=C5C=C34)=CC2=C1	178
38	C1=CC2=CC3=CC4=CC5=CC=CC=C5C=C4C=C3C=C2C=C1	93
39	S1C=CC2=C1C=C(S2)C1=CC2=C(S1)C=CS2	365
40	S1C=CC2=C1C(=CS2)C1=CSC2=C1SC=C2	256
41*	S1C=CC=C1C1=CC2=C(S1)C1=C(S2)C=CS1	359
42	S1C=CC2=CC3=CC4=CC5=CC=CC=C5C=C4C=C3C=C12	96
43*	S1C2=C(C=C3C=CC=CC3=C2)C2=C1C=C1C=CC=CC1=C2	118
44*	S1C=CC2=C3C=C4C=CC5=C(C=CS5)C4=CC3=CC=C12	155
45	S1C=CC2=CC3=C4C=C5C=CSC5=CC4=CC=C3C=C12	200
46	S1C=CC2=CC3=CC4=CC5=C(SC=C5)C=C4C=C3C=C12	94
47*	S1C=CC2=CC3=CC4=CC5=C(C=CS5)C=C4C=C3C=C12	95
48	S1C=CC2=CC3=CC=C4C=C5C=CSC5=CC4=C3C=C12	182
49*	S1C=CC2=C1C1=CC3=CC=C4C=CSC4=C3C=C1C=C2	134
50	S1C2=CC3=CC=CC=C3C=C2C2=C1C1=C(S2)C=CC=C1	153
51	S1C2=CC=CC=C2C2=CC3=C(C=C12)C1=C(S3)C=CC=C1	117
52*	S1C2=CC=CC=C2C2=CC3=C(SC4=C3C=CC=C4)C=C12	87
53	S1C=C(C2=C1SC=C2C1=CC=CC=C1)C1=CC=CC=C1	266
54	S1C=CC2=CC3=C(SC4=C3C=C3C=CSC3=C4)C=C12	149
55	S1C=CC2=CC3=C(C=C12)C1=C(S3)C=C2C=CSC2=C1	118
56	C1=C2C(C=CC3=CC=CC=C23)=C2C=C3C(C=CC4=CC=CC=C34)=C12	208

57	C1=C2C(=CC3=CC4=CC=CC=C4C=C23)C2=CC3=CC=CC=C3C=C12	115
58	S1C=CC2=C1C=C(S2)C#CC1=CC2=C(S1)C=CS2	293
59*	S1C=CC2=C1C1=CC3=C(C=C1S2)C1=C(S3)C=CS1	231
60	C1C2=CC=CC=C2C2=CC(=CC=C12)C1=CC2=C(SC=C2)S1	320
61	S1C=CC=C1C1=CC=C(S1)C1=CC2=C(S1)C=CS2	362
62	C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	309
63*	S1C=CC2=C1C1=C(S2)C2=C(S1)C1=C(S2)C=CS1	307
64*	S1C(=CC=C1C1=CC=CC=C1)C1=CC=C(S1)C1=CC=CC=C1	318
65*	S1C(=CC2=CC=CC=C12)C1=CC2=C(S1)C1=C(S2)C=CC=C1	266
66	S1C=CC=C1C1=CC=C(S1)C1=CC=C(S1)C1=CC=CC=C1	339
67	C1=CC=C2C(=C1)C1=CC=CC=C1C1=C2C2=CC=CC=C2C2=CC=CC=C12	193
68*	C1=CC=C2C(C=CC3=C4C=CC5=C(C=CC6=CC=CC=C56)C4=CC=C23)=C1	148
69	C1=CC2=CC3=CC4=CC5=CC6=CC=CC=C6C=C5C=C4C=C3C=C2C=C1	79
70	C1=CC=C(C=C1)C1=C2C=CC=CC2=C(C2=CC=CC=C2)C2=CC=CC=C12	255
71*	S1C=CC=C1C1=CC=C(S1)C1=CC=C(S1)C1=CC=CS1	348
72	S1C=CC2=CC3=CC4=CC5=CC6=CC=CC=C6C=C5C=C4C=C3C=C12	85
73	S1C=CC2=CC3=CC4=CC5=CC6=C(C=CS6)C=C5C=C4C=C3C=C12	87
74*	S1C2=CC3=CC4=C(C=CC=C4)C=C3C=C2C2=C1C1=C(S2)C=CC=C1	114
75*	S1C2=C(C3=C1C1=C(S3)C=CC3=CC=CC=C13)C1=CC=CC=C1C=C2	196
76*	S1C2=C(C3=C1C=CC1=CC=CC=C31)C1=C(S2)C=CC2=CC=CC=C12	187
77	S1C2=C(SC3=C2C=CC2=CC=CC=C32)C2=C1C1=CC=CC=C1C=C2	189
78	S1C2=CC3=CC=CC=C3C=C2C2=C1C1=C(S2)C=C2C=CC=CC2=C1	130
79	S1C2=CC3=C(SC(=C3)C3=CC=CC=C3)C=C2C=C1C1=CC=CC=C1	267
80	S1C(\C=C\C2=CC=CC=C2)=CC2=C1C=C(S2)\C=C\C1=CC=CC=C1	252
81	S1C2=C(SC(=C2)C2=CC=CC=C2)C2=C1C=C(S2)C1=CC=CC=C1	312
82	S1C2=C(C=C(S2)C2=CC=CC=C2)C2=C1SC(=C2)C1=CC=CC=C1	225
83	S1C=C(C2=C1SC1=C2C(=CS1)C1=CC=CC=C1)C1=CC=CC=C1	212
84*	S1C=CC2=C3C(SC4=C3C3=C5C=CSC5=CC=C3S4)=CC=C12	211

85	C1=CC=C(C=C1)C1=C2C3=CC=CC=C3C(=C2C2=CC=CC=C12)C1=CC=CC=C1	320
86	C1=CC2=CC3=CC=C(C=C3C=C2C=C1)C1=CC2=CC3=CC=CC=C3C=C2C=C1	103
87	S1C=CC=C1C1=CC2=C(S1)C1=C(S2)C=C(S1)C1=CC=CS1	328
88*	S1C=CC2=C1C1=C(S2)C2=C(S1)C1=C(S2)C2=C(S1)C=CS2	291
89	S1C=CC2=C(C3=C4SC=CC4=C(C3=C12)C1=CC=CC=C1)C1=CC=CC=C1	414
90	S1C2=C3C(C=CC4=C3C(C=C2)=C(S4)C2=CC=CC=C2)=C1C1=CC=CC=C1	160
91	C1C2=C(C3=C(S2)C2=C(S3)C3=C(CC4=CC=CC=C34)S2)C2=CC=CC=C12	237
92*	S1C2=C(SC(=C2)C2=CC=C3C=CC=CC3=C2)C2=C1C1=CC=CC=C1S2	262
93	C1=CC=C2C(C=CC3=C4C=CC5=C6C=CC7=CC=CC=C7C6=CC=C5C4=CC=C23)=C1	152
94	S1C2=C(SC(=C2)C2=CC3=C(S2)C2=C(S3)C=CC=C2)C2=CC=CC=C12	281
95	S1C(=CC2=CC=CC=C12)C1=CC2=C(S1)C1=C(S2)C2=CC=CC=C2S1	264
96	S1C=CC2=CC3=C(C=C(S3)C3=CC4=CC5=C(C=CS5)C=C4S3)C=C12	230
97	C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	288
98*	S1C=CC2=CC3=CC4=CC5=CC6=CC7=C(C=CS7)C=C6C=C5C=C4C=C3C=C12	79
99*	S1C2=CC3=CC4=CC=CC=C4C=C3C=C2C2=C1C1=C(S2)C=C2C=CC=CC2=C1	103
100*	S1C=CC2=C1SC1=C2C=C(S1)C1=CC2=C(SC3=C2C=CS3)S1	372
101	S1C=CC2=C1C1=C(S2)C=C(S1)C1=CC2=C(S1)C1=C(S2)C=CS1	337
102*	S1C(=CC2=C1C1=CC=C3C=C(SC3=C1C=C2)C1=CC=CC=C1)C1=CC=CC=C1	253
103	S1C2=CC3=CC4=C(SC(=C4)C4=CC=CC=C4)C=C3C=C2C=C1C1=CC=CC=C1	155
104	S1C2=CC3=CC4=C(C=C(S4)C4=CC=CC=C4)C=C3C=C2C=C1C1=CC=CC=C1	116

105	S1C2=CC=C3C4=C(SC(=C4)C4=CC=CC=C4)C=CC3=C2C=C1C1=CC=CC=C1	232
106	S1C2=CC3=C(SC4=C3C=CC=C4)C=C2C2=CC3=C(C=C12)C1=C(S3)C=CC=C1	106
107	S1C2=CC3=C(C=C2C2=CC4=C(C=C12)C1=C(S4)C=CC=C1)C1=C(S3)C=CC=C1	110
108	S1C2=C(SC(=C2)C2=CC=C(C=C2)C2=CC=CC=C2)C2=C1C1=CC=CC=C1S2	292
109	S1C=CC2=C1C1=C(C=CS1)C1=CC3=C4C=CSC4=C4SC=CC4=C3C=C21	132
110	S1C=CC2=C1C1=CC3=C4SC=CC4=C4C=CSC4=C3C=C1C1=C2C=CS1	124
111	S1C2=CC=CC=C2C2=C1C1=C(S2)C=C2C(SC3=C2SC2=C3C=CC=C2)=C1	179
112	S1C=CC2=C1C=C(S2)C1=CC2=CC=C(C=C2C=C1)C1=CC2=C(S1)C=CS2	305
113*	S1C2=CC=CC=C2C2=C1C1=C(S2)C2=C(S1)C1=C(S2)C2=C(S1)C=CC=C2	241
114*	S1C=CC=C1C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=CS1	309
115*	S1C=CC2=C1C1=C(S2)C2=CC3=C(C=C2S1)C1=C(S3)C2=C(S1)C=CS2	236
116	S1C=CC2=C1SC1=C2C=C(S1)\C=C\C1=CC2=C(SC3=C2C=CS3)S1	308
117	S1C=CC2=C1C(=CS2)C1=CSC2=C1SC=C2C1=CSC2=C1SC=C2	160
118*	S1C=CC2=C1SC1=C2SC2=C1C1=C(S2)SC2=C1SC1=C2C=CS1	207
119	S1C=CC2=C1SC1=C2C2=C(S1)SC1=C2C2=C(SC3=C2C=CS3)S1	210
120	S1C=CC2=C1C1=C(S2)C2=C(S1)C1=C(S2)C2=C(S1)C1=C(S2)C=CS1	280
121	C1C2=C(C3=C(S2)C2=C(S3)C3=C(S2)C2=C(CC4=C2C=CC=C4)S3)C2=C1C=CC=C2	134
122	S1C2=C(SC3=C2C2=CC=CC=C2C2=CC=CC=C32)C2=C1C1=CC=CC=C1C1=CC=CC=C21	186
123	S1C2=CC=C3C(C=CC4=CC=CC=C34)=C2C2=C1C1=C(S2)C=CC2=C1C=CC1=CC=CC=C21	181
124	S1C2=C(SC3=C4C=CC5=CC=CC=C5C4=CC=C23)C2=C1C1=C(C=C2)C2=CC=CC=C2C=C1	181

125	S1C2=CC3=CC4=CC=CC=C4C=C3C=C2C2=C1C1=C(S2)C=C2C=C3C=CC=CC3=CC2=C1	85
126	S1C2=C(SC(=C2)C2=CC=C(C=C2)C2=CC=CC=C2)C=C1C1=CC=C(C=C1)C1=CC=CC=C1	311
127	S1C2=C(SC(=C2C2=CC=CC=C2)C2=CC=CC=C2)C(=C1C1=CC=CC=C1)C1=CC=CC=C1	290
128	S1C2=C(SC(=C2)C2=CC=C3C=CC=CC3=C2)C2=C1C=C(S2)C1=CC=C2C=CC=CC2=C1	261
129*	S1C2=CC=CC=C2C2=C1C1=C(S2)C=C2C=C3C(SC4=C3SC3=C4C=CC=C3)=CC2=C1	124
130	C1=CC=C(C=C1)C1=C2C(C=CC3=CC=CC=C23)=C2C1=C1C=CC3=CC=CC=C3C1=C2C1=CC=CC=C1	242
131*	C1=CC=C(C=C1)C1=C2C3=CC4=CC=CC=C4C=C3C(=C2C2=CC3=CC=CC=C3C=C12)C1=CC=CC=C1	141
132	S1C=CC2=C1C=C(S2)C1=CC2=CC3=CC=C(C=C3C=C2C=C1)C1=CC2=C(S1)C=CS2	240
133	S1C(\C=C\C2=CC=CC=C2)=CC2=C1C1=C(S2)C2=C(S1)C=C(S2)\C=C\C1=CC=CC=C1	232
134	S1C(=CC2=C1C=C(S2)C1=CC=C(S1)C1=CC=CC=C1)C1=CC=C(S1)C1=CC=CC=C1	301
135	C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	254
136*	S1C2=C(SC(=C2)C2=CC3=CC=CC=C3S2)C2=C1C=C(S2)C1=CC2=CC=CC=C2S1	257
137	S1C2=CC=CC=C2C2=C1C1=C(C3=C(SC4=C3C=CC=C4)C1=C2C1=CC=CC=C1)C1=CC=CC=C1	348
138	S1C2=C3C(C=CC4=C3C(C=C2)=C(S4)C2=CC3=C(C=CC=C3)C=C2)=C1C1=CC2=CC=CC=C2C=C1	165
139	C1C2=CC=CC=C2C2=CC=C(C=C12)C1=CC2=C(S1)C=C(S2)C1=CC=C2C(CC3=CC=CC=C23)=C1	300
140	S1C=CC=C1C1=C2C(=S=C(C3=CC=CS3)C2=C(S1)C1=CC=CS1)C1=CC=CS1	182

141	S1C=CC=C1C1=C(C2=CC=CS2)C2=C(S1)C(C1=CC=CS1)=C(S2)C1=CC=CS1	452
142	S1C(=CC=C1C1=CC=C(C=C1)C1=CC=CC=C1)C1=CC=C(S1)C1=CC=C(C=C1)C1=CC=CC=C1	309
143	S1C=CC2=C1C1=C(S2)C2=C(S1)C1=C(S2)C2=C(S1)C1=C(S2)C2=C(S1)C=CS2	268
144	S1C(\C=C\C2=CC=CC=C2)=CC2=C1C=C(S2)C1=CC2=C(S1)C=C(S2)\C=C\C1=CC=CC=C1	232
145	S1C=CC=C1C1=CC=C(C=C1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(C=C1)C1=CC=CS1	323
146	c1csc(c1)-c1ccc(s1)-c1ccc(cc1)-c1ccc(cc1)-c1ccc(s1)-c1cccs1	287
147	S1C2=C(SC3=C4C=CC5=CC=CC6=CC=C(C=C23)C4=C56)C2=C1C1=CC=C3C=CC=C4C=CC(=C2)C1=C34	76
148	S1C2=CC3=CC=C4C=CC=C5C=CC(=C2C2=C1C1=C(S2)C=C2C=CC6=CC=CC7=CC=C1C2=C67)C3=C45	140
149	S1C2=C(SC3=C2C2=CC=CC4=CC=C5C=CC=C3C5=C24)C2=C1C1=CC=CC3=CC=C4C=CC=C2C4=C13	123
150	S1C(=CC2=C1C1=C(S2)C2=CC=CC=C2S1)C1=CC2=C(S1)C1=C(S2)C2=C(S1)C=CC=C2	270
151	S1C2=CC3=CC(=CC=C3C=C2C2=C1C1=C(S2)C=C2C=C(C=CC2=C1)C1=CC=CC=C1)C1=CC=CC=C1	145
152*	S1C2=CC3=CC=C(C=C3C=C2C2=C1C1=C(S2)C=C2C=CC(=CC2=C1)C1=CC=CC=C1)C1=CC=CC=C1	108
153	S1C=CC=C1C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=CS1	258
154	S1C2=C(SC(=C2)C2=CC=C(C=C2)C2=CC=CC=C2)C2=C1C=C(S2)C1=CC=C(C=C1)C1=CC=CC=C1	293
155	S1C2=C(SC(=C2)C2=CC=CC=C2C2=CC=CC=C2)C2=C1C=C(S2)C1=CC=CC=C1C1=CC=CC=C1	305
156	S1C=CC=C1C1=C(C2=C(S1)C1=C(S2)C(=C(S1)C1=CC=CS1)C1=CC=CC=C1)C1=CC=CC=C1	481

157	S1C2=CC=CC=C2C2=C1C1=C(S2)C2=C(S1)C=C1C(SC3=C1SC1=C3SC3=C1C=CC=C3)=C2	205
158*	S1C2=C3C(C=CC4=C3C(C=C2)=C(S4)C2=CC=C(C=C2)C2=CC=CC=C2)=C1C1=CC=C(C=C1)C1=CC=CC=C1	181
159	S1C=CC2=C1SC1=C2C=C(S1)\C=C\C1=CC=C(\C=C\C2=CC3=C(S2)SC2=C3C=CS2)C=C1	257
160	S1C=CC2=C3C=CSC3=C3C(SC4=C3SC3=C4SC4=C3C3=C(C=CS3)C3=C4SC=C3)=C12	199
161	C1C2=CC=CC=C2C2=CC=C(C=C12)C1=CC2=C(S1)C1=C(S2)C=C(S1)C1=CC=C2C(CC3=CC=CC=C23)=C1	275
162	S1C=CC=C1C1=CC=C(S1)C1=CC2=C(S1)C1=C(S2)C=C(S1)C1=CC=C(S1)C1=CC=CS1	290
163	C1=CC2=CC3=CC=C(C=C3C=C2C=C1)C1=CC2=CC3=CC=C(C=C3C=C2C=C1)C1=CC2=CC3=CC=CC=C3C=C2C=C1	83
164	C1=CC=C(C=C1)C1=C2C(C3=CC=CC=C3)=C3C=CC=CC3=C(C3=CC=CC=C3)C2=C(C2=CC=CC=C2)C2=CC=CC=C12	147
165*	S1C=CC2=C1C=C(S2)C1=CC2=C(S1)C=C(S2)C1=CC2=C(S1)C=C(S2)C1=CC2=C(S1)C=CS2	301
166	S1C=CC=C1C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=CS1	265
167	S1C2=C(C=CC=C2)C2=C1C1=C(C3=C4SC5=C(SC6=C5C=CC=C6)C4=C(C3=C1S2)C1=CC=CC=C1)C1=CC=CC=C1	308
168	S1C2=CC=C(C=C2C2=C1C1=C(S2)C=C2C=C3C(SC4=C3SC3=C4C=CC(=C3)C3=CC=CC=C3)=CC2=C1)C1=CC=CC=C1	133
169	S1C2=C(SC(=C2)C2=C(C=CC=C2C2=CC=CC=C2)C2=CC=CC=C2)C2=C1C=C(S2)C1=C(C=CC=C1C1=CC=CC=C1)C1=CC=CC=C1	363
170*	S1C=CC=C1C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=C(S1)C1=CC=CS1	268
171	S1C2=CC=CC=C2C2=C1C1=C(S2)C2=C(C3=C4SC5=C(SC6=C5SC5=CC=CC=C65)C4=C(C3=C2S1)C1=CC=CC=C1)C1=CC=CC=C1	266

3.2 Study wise specific description of methodologies utilized in each study

3.2.1 Study -1

3.2.1.1 Data collection

The values of detonation heat (expressed in KJ/kg) of 162 N-containing compounds were collected from previously published literature (He et. al., 2021) and are listed in **Table 3.1.** The structures were prepared in MarvinSketch (version- 5.5.0.1) https://www.chemaxon.com, added the explicit hydrogen, cleaned the structure, and aromatized the aromatic rings as applicable. A **chemical diversity plot (Figure 3.1)** was prepared using the molecular weight and $logP_{cons}$ which shows the diversity in the chemical nature of the compounds.

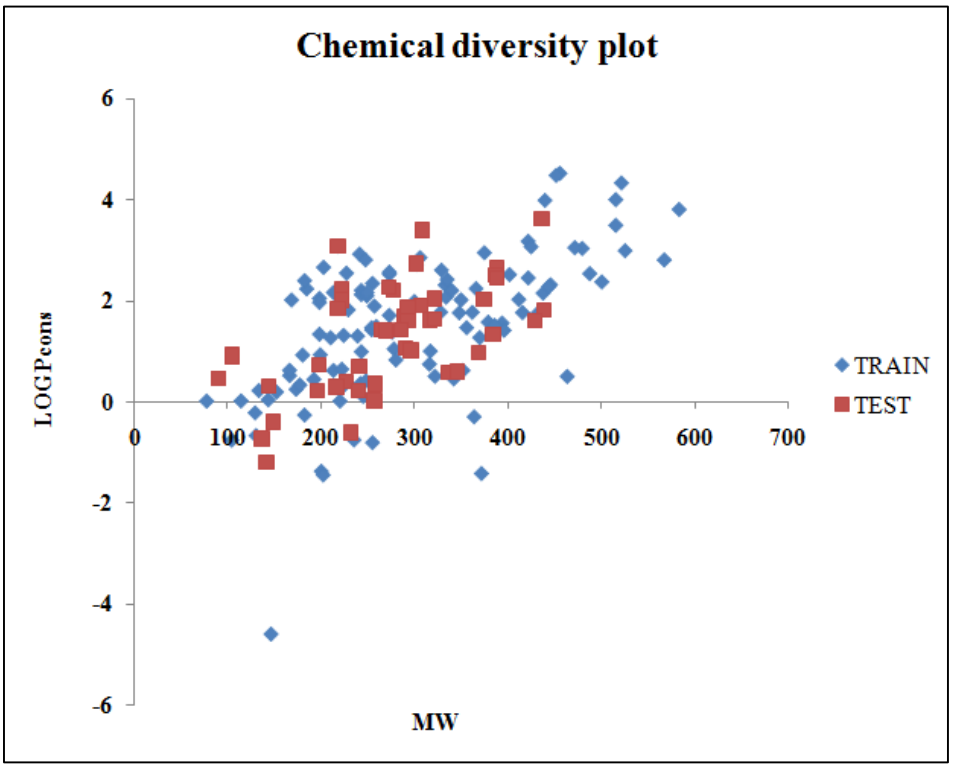

Figure 3.1: Chemical diversity plot

3.2.1.2 Descriptor calculation and data pre-treatment

Molecular descriptors are the quantitative values derived from the structural information of the molecules. Different classes of 2D descriptors like molecular properties, 2D atom pairs, atom type E-state indices, atom-centered fragments, functional group counts, connectivity indices, ring descriptors, constitutional indices, and extended topochemical atom (ETA) indices were calculated using alvaDesc v2.0.6 (Mauri, 2020). These different classes of descriptors are so chosen as they are highly interpretable and also are efficient in the development of models as

evident from our previous experiences. A total of 689 molecular descriptors were calculated initially.

The obtained descriptors were then subjected to a pretreatment process using a java-based tool DataPreTreatmentGUI 1.2 available from http://teqip.jdvu.ac.in/QSAR_Tools/ to remove the intercorrelated descriptors with a variance cut-off of 0.0001 and a correlation coefficient cut-off value of 0.95. In this process, descriptors that are highly inter-correlated to each other and descriptors with null or constant values for each data point are obviated. After the pre-treatment process, a total of 473 descriptors were left which were used for further study.

3.2.1.3 Data division

The division of the dataset is a necessary step prior to the model development. To establish a powerful QSPR model with good predictive ability the data set is divided into a training set and a test set. In this work, the dataset was divided in a ratio of 75:25, constituting 122 compounds in the training set and 40 compounds in the test set using the Euclidean Distance-based division algorithm (Danielsson, 1980) with the help of a java-based tool datasetDivisionGUI1.2 available from http://teqip.jdvu.ac.in/QSAR_Tools/. After division, the training and the test sets were subjected to pretreatment with the help of dataPreTreatmentTrainTest1.0 tool from http://teqip.jdvu.ac.in/QSAR_Tools/ to remove intercorrelated descriptors. The development of the model is done using the training set whereas the test set is used to check the predictive ability and external validation of the developed model.

3.2.1.4 Feature selection and QSPR model development

The selection of important features contributing to the property of compounds is a crucial step during the development of a QSPR model (Bursac et. al., 2008). We have prepared several Genetic Algorithm (GA) (Katoch et. al., 2021) models using a java-based tool GeneticAlgorithm_v4.1 from http://teqip.jdvu.ac.in/QSAR_Tools/ and selected the descriptors that appeared frequently in a maximum number of models. The generation of GA models and feature selection is done using the training set only without the involvement of the test set. The training set and test set matrices with the selected features were prepared. Further, we have used the Best Subset Selection v2.1 tool available from http://teqip.jdvu.ac.in/QSAR_Tools/ to generate different MLR models with all possible combinations of a given number of descriptors. A good robust model was selected based on the cross-validation result which is used for further q-RASPR analysis.

3.2.1.5 Optimization of the Read-Across hyperparameters

Identifying the optimized setting of hyperparameters (σ, γ, number of close source/training compounds, and best similarity-based algorithm) is an essential step for Read-Across based prediction. Per the QSPR prediction principles, hyperparameter optimization should be done based on training/source set only without any involvement of the test/query set. The training set containing the descriptors involved in the QSPR model was further divided into corresponding sub-train and sub-test sets. With the help of a java-based tool Auto_RA_Optimizer-v1.0, available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, we have selected the values for σ and γ to be 0.5, number of close training compounds be 8, and Gaussian kernel-based similarity as our best similarity-based algorithm. Here, the selection of hyperparameters was based on the maximum occurrence frequency of individual hyperparameters obtained during optimization using different sub-training and sub-test sets prepared through the division of the training set via different algorithms.

3.2.1.6 Calculation of the RASPR descriptors

Before proceeding with the q-RASPR study, the prominent step is calculating the similarity and error-based RASPR descriptors (Banerjee and Roy, 2023) for the individual training and test sets. Unlike structural and physiological descriptors, the RASPR descriptors are calculated after the division process. This is so because the RASPR descriptors are calculated based on the similarity of test set compounds to the training set compounds. The Gaussian kernel-based similarity descriptors with σ value 0.5 were calculated using a java-based tool RASAR-Desc-Calc-v2.0, available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. For the calculation of RASPR descriptors for the test set, we have used the training set and the test set containing the selected physiochemical descriptors as input, whereas for the computation of training set RASPR descriptors, only the training set is used as input.

3.2.1.7 Feature selection and development of the q-RASPR model

Since, the q-RASPR study is the combination of both QSPR and RA-based predictions, it is necessary to combine the structural and physiological descriptors with the similarity and error-based RASPR descriptors. The 15 similarity and error-based descriptors are fused with the previously selected structural and physiological descriptors for respective training and test sets. A grid search was performed to generate a MLR q-RASPR model with all the possible combinations of a given number of descriptors using the Best Subset Selection v2.1 tool available from http://teqip.jdvu.ac.in/QSAR_Tools/. Descriptor optimization was based on the

Q^2_{LOO} (cross-validation) metric. The final PLS q-RASPR model was developed with the selected features.

3.2.1.8 Application of other machine learning (ML) algorithms

The predictive performance of the developed q-RASPR model was further evaluated by applying various supervised Machine Learning (ML) algorithms. We have used 7 different ML algorithms to develop various regression models such as Random Forest (RF) (Breiman, 2001), Adaptive Boosting (AdaBoost/AB) (Wu wt. al., 2010), Gradient boosting (GB) (Friedman et. al., 2002), Extreme Gradient Boosting (XGB) (Chen and Guestrin, 2016), Support Vector Machine (SVM) (Noble, 2006), Linear Support Vector Machine (LSVM), and Ridge Regression (RR) (Hoerl and Kennard, 1970). Scaling of the training and test sets data values was achieved using a Java-based tool Scale1.0 fromhttps://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. With the help of a Python-based tool Hyperparameter Optimizer v1.2 and the scaled data of the training set, we have calculated the optimized hyperparameters for each ML algorithm. The selection of the hyper-parameters was based on the MAE results. Using the optimized settings of the hyperparameters and the scaled training and test sets, we have developed several ML models using a Python-based tool Machine Learning Regressor v 2.0 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The final selection of the best predictive model was done based on MAE_{Test} results.

3.2.1.9 Statistical validation metrics

The developed models were evaluated for their predictability and reliability in terms of various internal and external validation parameters. Internally the model was evaluated on the basis of determination coefficient (R^2), adjusted R^2 (R^2_{adj}), Leave-One-Out cross-validated Q^2 (Q^2_{LOO}), and root mean squared error of calibration ($RMSE_C$) while the external statistical parameters involve the calculation of R^2_{pred} or $Q^2_{F1,}$ $Q^2_{F2,}$ and root mean squared error of prediction ($RMSE_P$) (Roy, 2007). Both internal and external validation tests were done using the mean absolute error (MAE) based criteria (Roy et. al., 2016) as Q^2_{ext} does not always provide exact prediction quality because of its dependence on the response range and response value distribution in the training and test set compounds.

3.2.1.10 Applicability domain (AD)

The validity of the q-RASPR model is denoted by a defined domain of applicability (OECD principle 3) (Roy et. al., 2015a). AD (Roy et. al., 2015b) represents the response and chemical

structure space which is defined by the chemicals used in the development of the model (in the training set). The distance to model X (DModX) approach (Roy et. al., 2015c) was used with a 99% confidence level with the help of SIMCA software (https://landing.umetrics.com/downloads-simca) to check whether the compounds in the sets are within the AD. In the DModX technique, the residuals of X and Y act as diagnostic values for the quality of the model. The standard deviation (SD) of X-residuals corresponds to the respective row of residual matrix E. As SD is directly proportional to the distance between the data points and the model plane in X-space, it is commonly called DModX (distance to the model in X-space). Those compounds which are present in the chemical space can be predicted precisely and those lying outside the AD are termed as outliers.

The detailed workflow is represented in **Figure 3.2.**

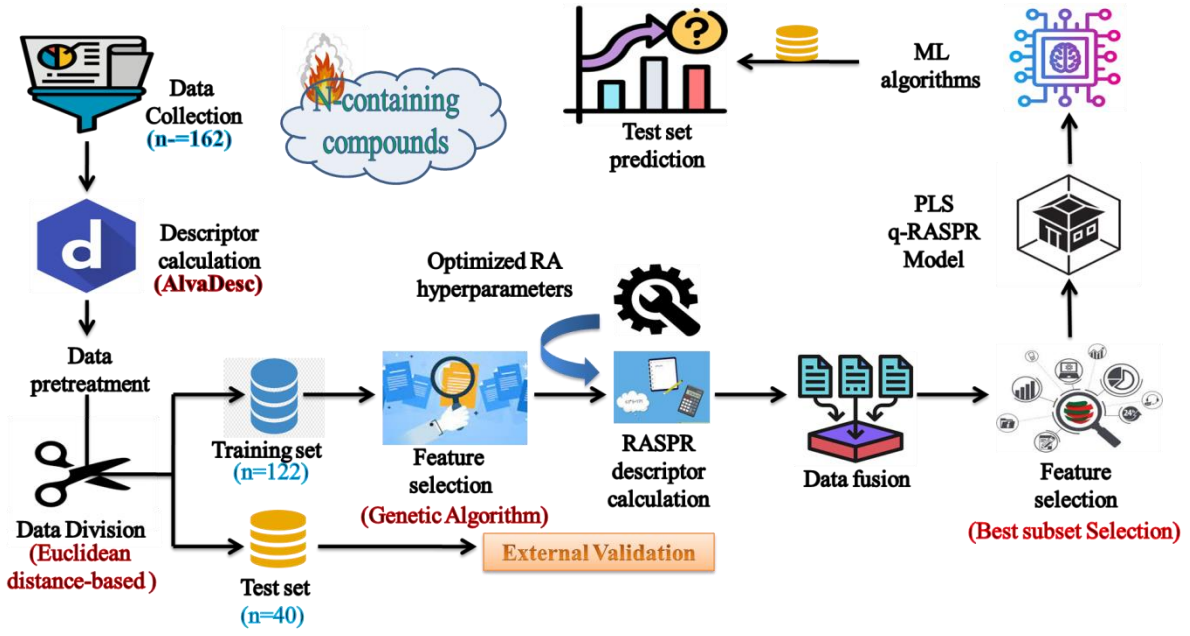

Figure 3.2: Workflow of the q-RASPR model development to estimate the detonation heat of N-containing compounds

3.2.2 Study 2

3.2.2.1 Data set preparation, curation, and structural representation

It is crucial to have high-quality data while building computational models. Therefore, we collected four data sets with their experimental data, each containing information about the one of the properties like decomposition temperature, melting point, density, and heat of formation, from previously published literature sources (Mathieu, 2018; Wespiser and Mathieu, 2013).

The data taken from the 2 literature sources are all experimental data. The T_{dec} data was derived in-house by Wespiser et. al., the Bradley melting point data set was used by Wespiser et. al. for the melting point data set, the density data set was collected from Crystallography open database by Wespiser et. al., and the heat of formation data contains different types of compounds with their experimental data which is also clearly mentioned in the literature (Mathieu, 2018). The data set used by Wespiser et al. contains some other organic compounds also with their experimental data for the heat of formation and densities. This was done so to extract the features which correspond to high positive heat of formation and higher densities of the compounds. These features can help to get insights into how the densities and heat of formation are affected by the presence of certain features in the compounds. The determination of these features will help to design new better performing EMs with less sensitivity.

To ensure accuracy, we curated the collected data to remove any duplicates, inorganic compounds, or mixtures, if present. After the curation process, we were left with 656, 19667, 12805, and 2565 data points for the decomposition temperature (˚C), melting point (˚C), density (g/cm^3), and gas phase enthalpy of formation data (kJ/mol) sets, respectively. We made all the curated data sets available in the Excel sheets of Supplementary Materials **(SI-1**) (Pandey and Roy, 2024). The SMILES (Simplified Molecular Identity Line Entry System) notation was used for the representation of all data points, and MarvinSketch v-5.11.5 https://www.chemaxon.com was used to prepare the structures, which were then subjected to aromatization, the addition of explicit hydrogens and 2D cleaning as necessary.

3.2.2.2 Descriptor calculation and data pre-treatment

The molecular structures so prepared were used to calculate the descriptors (quantitative values derived from the molecular structural information) for the respective data sets using the AlvaDesc software v2.0.6. (Mauri, 2020). Nine different classes of highly interpretable 2D descriptors like molecular properties, functional group counts, atom type E-state indices, atom-centered fragments, 2D atom pairs, connectivity indices, constitutional indices, ring descriptors, and Extended Topochemical Atom (ETA) indices were calculated for all data sets.

The calculated descriptors set was then subjected to the pre-treatment process where the descriptors having high inter-correlation (>0.8) or having constant/null values were removed from the descriptors set. The final pre-treated files were used for further division of the data set into training and test sets.

3.2.2.3 Dataset division

To check the predictive power of the model, there is a requirement to check the predictions for external compounds in addition to those included in the development of the model. To do so, the data set was divided into training and test sets. The training set was used for the development of the model while the test set validates the predictivity of the developed model. We have divided all the data sets into respective training and test sets in a 3:1 ratio. Based on different algorithms, the data sets were divided with the help of the Dataset-DivisionGUI1.2tool freely available from http://teqip.jdvu.ac.in/QSAR_Tools/. The information on the number of compounds in the individual training and test sets after the division, along with the division algorithm applied, is enlisted in **Table 3.3.** The details of the data sets are provided in **Supplementary Information SI-1** (Pandey and Roy, 2024)**.**

Additionally, for the density data set, we have also prepared a true external set of 37 energetic compounds with their experimental density (g/cm^3) collected from Rice and Brydr. (Rice et. al., 2007).

Table 3.3: List of training and test compounds in data sets and the applied division algorithm

Data Set	No. of compounds		Division algorithm
	Training	Test	
Decomposition temperature (T_{dec})	424	141	Property-sorted
Melting point (T_m)	14750	4917	Property-sorted
Density	9604	3201	Property-sorted
Heat of formation (ΔH°_f) (gas phase)	1923	642	Kennard-Stone

After the division of the dataset into respective training and test sets, we further pre-treated the training and test set descriptor matrix to remove the null/constant descriptors, and the final training and test set so obtained were used for the feature selection process.

Figure 3.3 presents the chemical diversity plot (MW vs LOGPcons) prepared using the molecular weight and LOGPcons for all the data sets to see the diversity in the chemical nature of the compounds present in the respective training and test sets of the individual data set.

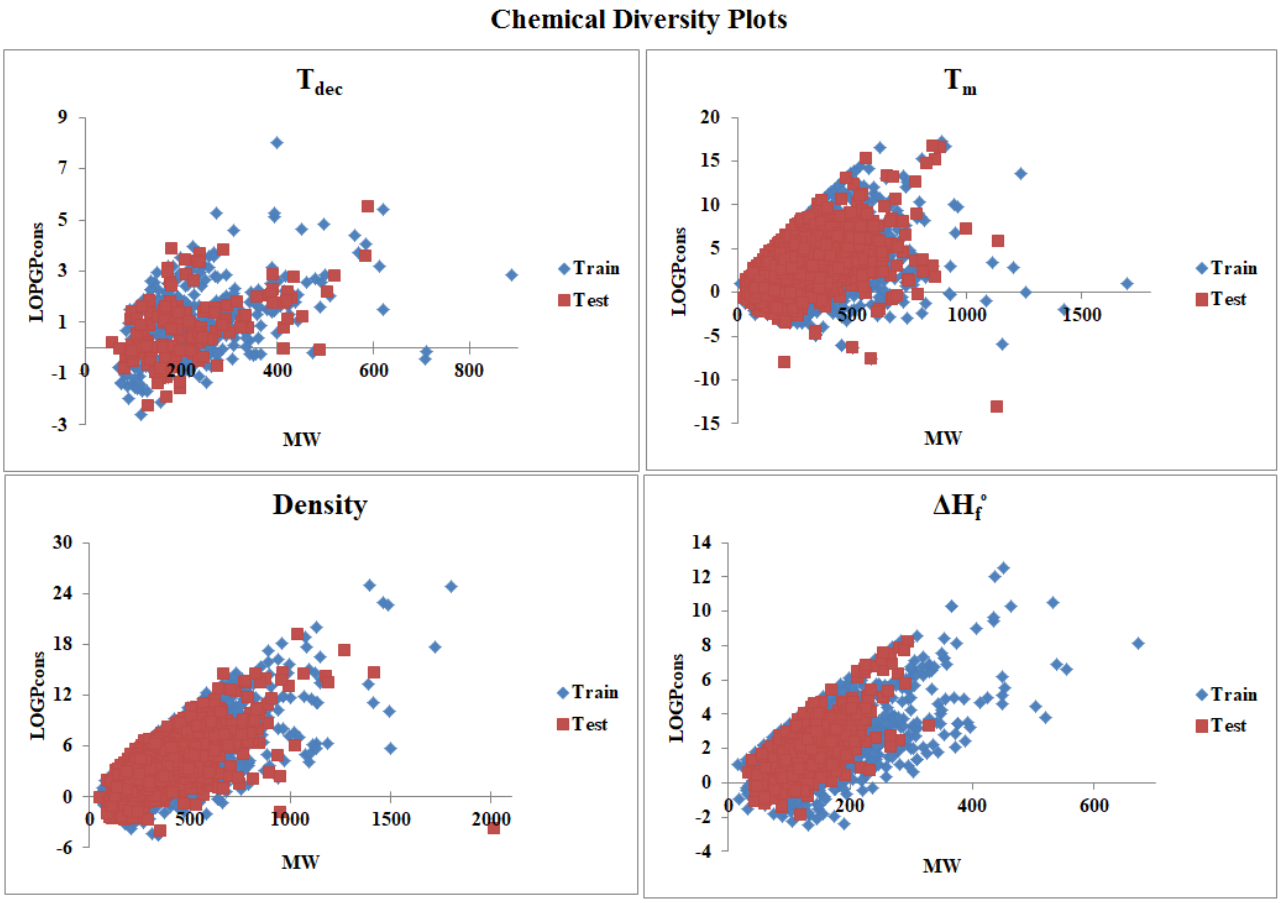

Figure 3.3. Chemical diversity plots

3.2.2.4 Feature selection and QSPR model development

The selection of the potential features from the descriptor pool that are closely related to the activity/property/toxicity of the compound is a key step during the development of a QSAR model (Bursac et. al., 2008). There are several variable selection methods like step-wise selection, all possible subset selection, genetic algorithm, factor analysis, etc. (Roy et. al., 2015c). In this work, we used step-wise and genetic algorithms to prepare a pool of important descriptors and then used the all-subset selection method to finalize the set of descriptors for the final models. The features are selected based on the MAE-based criteria (training set only without any involvement of the test set. A pool of features was prepared through various feature selection processes. A grid search was performed using the pool of selected features for the generation of several MLR models using the Best Subset Selection tool v2.1 available

from http://teqip.jdvu.ac.in/QSAR_Tools/. The final robust PLS QSPR model was selected based on the cross-validation (Q^2_{LOO}) result with a lower number of latent variables (LVs). The final model so obtained was then used for Read-across-based similarity prediction.

3.2.2.5 RA predictions

For the calculation of RA-based similarity predictions, we have used the default values of the hyperparameters, i.e. σ=1, γ=1, and the number of closed training/source compounds (CTC) to be 10. Using the default hyperparameters and a Java-based tool Read-Across-v4.2 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, we have calculated the similarity predictions of the test set compounds for different similarity algorithms such as Gaussian kernel-based, Laplacian kernel-based, and Euclidean distance-based similarity. Further, based on the MAE_{test} results, we have selected the best similarity measure for the individual data set.

3.2.2.6 RASPR descriptor calculation

The calculation of the similarity and error-based RASPR descriptors is the first and foremost step needed to build a q-RASPR model (Banerjee and Roy, 2023). The calculation of the RASPR descriptors (for the best similarity measure obtained from RA prediction) is done after the division process which is different from the calculation of structural and physiochemical descriptors that are calculated before the data set division. This is because here the test/query set RASPR descriptors are calculated based on their similarity to the training/query set compounds. For the calculation of the test set RASPR descriptors, both the training as well as test sets (containing the structural and physiochemical descriptors) were used while the training set RASPR descriptors were calculated from itself only.

3.2.2.7 Feature selection and q-RASPR model development

The descriptor matrix of the QSPR model was fused with the 18 calculated similarity and error-based RASPR descriptors. The prepared descriptor pool was then used for the feature selection using a step-wise process or performing a grid search through the Best Subset Selection tool v2.1 available from http://teqip.jdvu.ac.in/QSAR_Tools/. The optimal number of descriptors selected in the model was based on the leave-one-out cross-validated (Q^2_{LOO}) results, and the same features were used to develop the final PLS model. The PLS model was developed for all sets except the melting point data set where a univariate model was developed.

3.2.2.8 Statistical quality and validation metrics

After the development of a model, the model needs to be validated internally as well as externally. The OECD principle 4 describes the different validation metrics needed to judge the predictive potential of a model (Gramatica, 2007). To check the statistical quality and validate the model internally, we have used the determination coefficient (R^2), leave-one-out cross-validated Q^2 (Q^2_{LOO}), mean absolute error (MAE_{train}), and root mean squared error of calibration set ($RMSE_C$) (Roy, 2007). The external validation was done based on Q^2_{F1}, Q^2_{F2}, mean absolute error (MAE_{test}), and root mean squared error of prediction set ($RMSE_P$). Both the internal and external validation tests were done based on the MAE-based criteria as Q^2 metrics do not always provide a good reflection of the prediction quality (Roy et. al., 2016).

3.2.2.9 Application of ML algorithms

We have also applied different machine learning algorithms to check the predictivity of our developed PLS q-RASPR model. Here, we have used 7 different supervised ML algorithms such as Random Forest (RF), Support Vector Machine (SVM), Linear Support Vector Machine (LSVM), Adaptive Boosting (AB), Gradient Boosting (GB), Extreme Gradient Boosting (XGB), and Ridge Regression (RR) to build various regression models. These machine learning modeling methods are described in **Supplementary Materials SI-2**(Pandey and Roy, 2024)**.** The training and test set descriptors and response values of the developed PLS model were scaled before the application of ML algorithms using a Java-based tool Scale1.0 freely available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. Different ML models were developed for each property data set (except Tm) with the help of a Python-based tool RSLv2.2 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. We have used the default setting of the hyperparameters for the development of the ML models.

3.2.2.10 Applicability Domain (AD)

As per the OECD principle 3, the defined applicability domain (AD) represents the validity of the developed q-RASPR model. The chemicals employed in the model development define the chemical structure space, which is represented by AD (Roy et. al., 2015b). To check whether the compounds in the test set are within the chemical space of the training set used for the modeling, we have used the DModX (distance to model X) approach with 99% confidence level (only for the PLS models) using the SIMCA software https://landing.umetrics.com/downloads-simca. (Wold et. al., 2001). The compounds within the

AD can be predicted precisely whereas the compounds outside the AD are termed outliers. The DModX approach was used for defining the AD of T_{dec}, density, and ΔH_f° data sets, while for the T_m data set, we used the leverage approach (Roy et. al., 2015c) for determining the AD.

The detailed workflow we have used during the model development is represented in **Figure 3.4.**

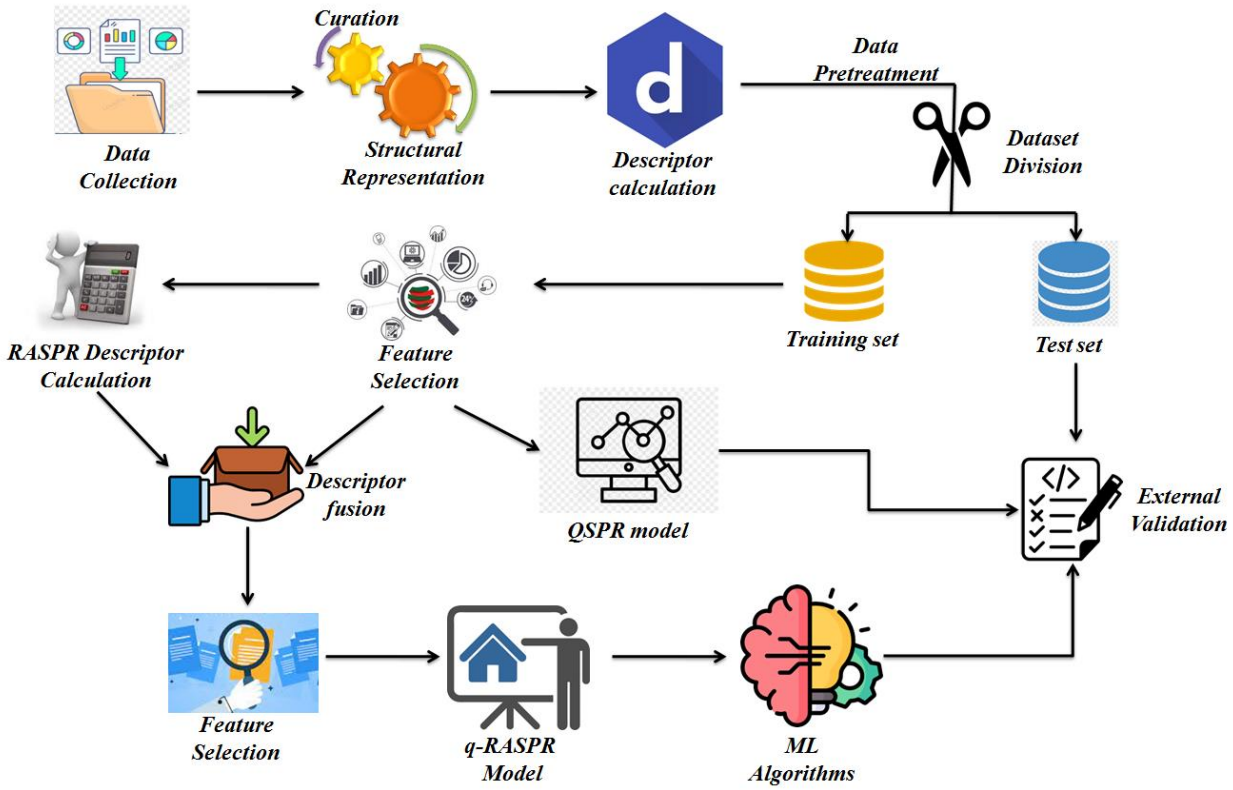

Figure 3.4: Schematic workflow for the model development

3.2.3 Study 3

3.2.3.1 Dataset preparation and molecular representations

The authors have collected experimental data on the RE for 173 molecular p-type OSCs from the previously published literature (Atahan-Evrenk, 2018). The data set contains a diverse set of organic compounds with acenes, thiophenes, thienoacenes, and anti-aromatic pentalenes with their experimental RE measured in mili electron-volt (meV). The logarithmic transformation of the RE was performed to reduce the range of the response value. Simplified molecular identity line-entry system (SMILES) notations were used for the molecular representation of the entities in the data set, which were then used to prepare the molecular structures of the compounds by using MarvinSketch (https://www.chemaxon.com) v-5.11.5.40.

The prepared structures were aromatized in appropriate cases, added with explicit hydrogen, and cleaned in the 2D-space.

3.2.3.2 Calculation of structural and physiochemical descriptors

Molecular descriptors are required to generate mathematical correlations between the molecular structure information and the response values. Molecular descriptors are the quantitative values used to define/quantify/represent various structural features and are derived from the structural representation of the molecules. In this study, we have calculated a total of nine classes of highly interpretable 2D structural and physiochemical descriptors using the AlvaDesc software v2.0.641, (Mauri, 2020) namely 2D atom pairs, molecular properties, functional group counts, constitutional indices, atom-centered fragments, connectivity indices, ring descriptors, Extended Topochemical Atom (ETA) indices, and atom type E-state indices.

After calculating the above-mentioned descriptors, they were subjected to a data pre-treatment process to remove the descriptors with null and/or constant values and features having high inter-correlation between them. Here, we have used the inter-correlation cut-off of 0.95. The descriptor file after the pre-treatment process was further used for the dataset division purpose.

3.2.3.3 Division of the dataset

Splitting the dataset into a training set and a test set is a very important step required for the development of a well-validated model. The division of the dataset should ensure that the compounds in the training and the test sets are distributed within the entire descriptor space of the compounds in the whole dataset. In this work, we have applied the property-sorted response-based division to divide the data set into a 3:1 ratio using a java based tool Dataset-DivisionGUI1.2 freely available from http://teqip.jdvu.ac.in/QSAR_Tools/. The compounds in the training set were used for the development of the model whereas the test set compounds were used to validate the model externally.

3.2.3.4 Variable selection and QSPR model development

The process of variable selection refers to the extraction of important features from the whole descriptor pool that are highly correlated to the response (here, RE). The feature selection is performed using only the training set and does not involve the test set (Bursac et. al., 2008). Among various feature selection techniques, we have used the step-wise feature selection and genetic algorithm (GA) method to pool out significant descriptors (Roy et. al., 2015c; Rogers and Hopfinger, 1994). Through GA feature selection, the descriptors that frequently appeared

were selected via the generation of several GA models. The descriptor pool formed after the feature selection was then used for performing a grid search using the Best Subset Selection tool v2.1 available from http://teqip.jdvu.ac.in/QSAR_Tools/ to generate different MLR models. Soon after the selection of the best MLR model, the same descriptor combination was used to develop the final PLS QSPR model, the latter being more robust and generalized version of the former. The PLS QSPR model was developed with a lower number of latent variables (LVs) on the basis of the cross-validation (Q^2_{LOO}) result. The descriptors appearing in the developed QSPR model were then used for further read-across (RA) based similarity predictions.

3.2.3.5 Read-across similarity predictions

Before proceeding with the similarity calculations, tuning the hyperparameters associated with different similarity measures is necessary. Per the QSPR prediction principles, the hyperparameters are optimized using only the source/training set. The training set of the final QSPR model was further divided into several sub-training and validation sets. Using the sub-training and validation sets as input files in a Java-based tool Auto_RA_Optimizer-v1.0, available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, we have optimized the hyperparameters for Gaussian kernel-based similarity, Laplacian kernel-based similarity, and Euclidean distance-based similarity measures. The hyperparameters such as the number of close training compounds (CTC), sigma (σ) value [for Gaussian kernel], and gamma (γ) value [for Laplacian kernel] were selected based on the frequency of the value occurring the maximum number of times when ran with different sub-train and validation sets (Chatterjee et. al., 2022).

After the selection of the hyperparameters for the individual similarity measure, these tuned hyperparameters were then used to calculate the prediction of the query/test set previously obtained from the division of the whole dataset. The prediction of the individual query set compound is done based on its similarity with the close source compound in the training set. The RA predictions were performed using a Java-based tool Read-Across-v4.1 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The prediction of the query/test set was obtained individually for the above-mentioned three similarity measures.

3.2.3.6 Computation of RASPR descriptors

To develop a q-RASPR model, similarity and error-based features, also known as RASPR descriptors, are calculated for each similarity measure (with their optimized hyperparameters)

for the individual training/source set and for the test/query set (Banerjee and Roy, 2023). The calculation of the RASPR descriptors is done after the division process which differs from the calculation of the structural and physiochemical descriptors in a conventional QSPR analysis before the division of training and test sets. With the help of a Java-based tool RASAR-Desc-Calc-v3.0.2 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home, we have calculated the RASPR descriptors for each similarity measure. To calculate the RASPR descriptor for the training set, the training set with structural and physiochemical features of the final QSPR model itself was used as an input, whereas the RASPR descriptors for the test set were calculated using both the training set and test set files of the QSPR model.

3.2.3.7 Development of the q-RASPR models

A q-RASPR model contains information on both the structural and physicochemical features and similarity information. Therefore, the amalgamation of the structural and physiochemical descriptors of the QSPR model with the similarity and error-based RASPR descriptors becomes a necessary step. The newly prepared descriptor matrix of the training set and the test set were then used for performing a grid search for descriptors with the help of the Best Subset Selection v2.1 tool available from http://teqip.jdvu.ac.in/QSAR_Tools/ where different MLR models were developed with a certain number of features. The best model was selected based on the cross-validation metric; Q^2_{LOO}. The same descriptors were then used to develop the PLS q-RASPR model with a lower number of LVs.

It should be noted here that we have developed three different q-RASPR models for the three different similarity measures. To do this, we combined the structural and physiochemical features of the QSPR model's training and test sets with the RASPR descriptors for each similarity measure individually to obtain three different sets of training and test sets.

The predictions of the compounds present in both the training and the test sets were calculated using the above 3 models separately. Furthermore, we have used the predictions obtained from the individual PLS models to perform stacking. The final stacking q-RASPR model was developed using the PLS regression algorithm as the stacking regressor. The developed stacking PLS q-RASPR model contains information on the structural and physiochemical features along with the different similarities (Euclidean, Gaussian, and Laplacian) between the source and the query compounds.

3.2.3.8 ML predictions

We have also applied several ML algorithms to perform the stacking regression to enhance the quality of the developed model. Tree-based methods (RF, AB, GB, XGB), and kernel-based methods (SVM, LSVM, RR) were used to evaluate our developed models. Before applying the supervised ML algorithms, the training and the test sets were scaled with the help of a Java-based tool Scale1.0 freely available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The ML models were developed using the above-mentioned ML algorithms with the help of RSLv2.2 (a Python-based tool) available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The ML models use the default hyperparameters during the learning process.

3.2.3.9 Validation of the developed models

As per the OECD principle 4, the acceptance of a developed model relies on validating the model both internally (based on the training set) as well as externally (based on the test set). The evaluation of the predictivity, goodness of fit, and robustness of the developed model was done through the internal and external validation of the models. Statistical quality and validation metrics like the determination coefficient (R^2), adjusted R^2 (R^2_{adj}), and leave-one-out squared correlation coefficient (Q^2_{LOO}) were used to judge the goodness of fit and robustness of the developed model. For the external validation Q^2_{F1} (or R^2_{pred}), Q^2_{F2}, Q^2_{F3}, and concordance correlation coefficient (CCC) were calculated to determine the predictivity of the model (Roy, 2007). Error metrics such as mean absolute error (MAE) and root mean squared error (RMSE) were also used for the validation of the models both internally and externally (Roy et. al., 2016).

3.2.3.10 Applicability domain

The applicability domain (AD) (Roy et. al., 2015b) is defined as a chemical structure space represented by the chemicals that are present in the training set. According to OECD principle 3, one should perform the AD study to validate their developed model (Roy et. al., 2015a). In this study, we have used the distance to model in X space (DModX) approach (Roy et. al., 2015c) with a 99% confidence level to evaluate whether the compounds in the training and test sets are within the domain of applicability. SIMCA software (https://landing.umetrics.com/downloads-simca) was used for performing the DModX-AD analysis. For the precise prediction of a compound, the compound must lie within the AD of

the model, and if they do not, their predictions are not reliable and hence, termed as outliers or outside the applicability domain.

The detailed workflow of the current study is shown in **Figure 3.5.**

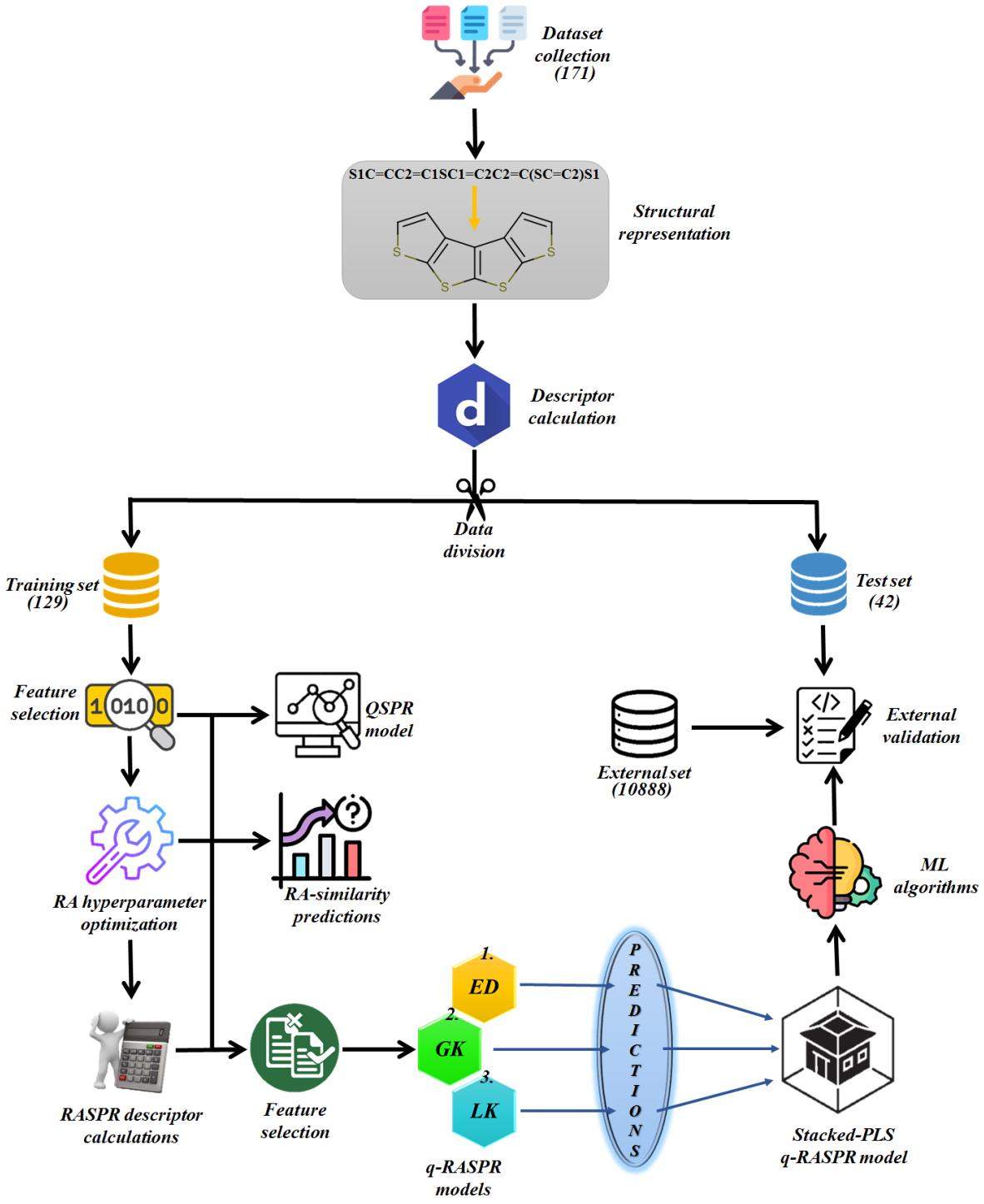

Figure 3.5: Sequential steps for model development and its validation

Chapter 4

Result & Discussion

4. RESULT & DISCUSSION

4.1 Study 1: Machine learning-based q-RASPR predictions of detonation heat for nitrogen-containing compounds

4.1.1 QSPR model development

The data set comprising 162 compounds with the detonation heat energy and computed descriptors is provided in the Supplementary Materials section. The training set consists of 122 compounds, while the predictions and external validation were done using a test set having 40 compounds. After the feature selection process, a total of 6 descriptors were used to develop the final PLS QSAR model with 5 latent variables as shown in **Equation (4.1)**

$$Q = 2504.432 + 264.478 \times F01[N-O] - 151.749 \times X\% + 156.626 \times SddsN + 297.997 \times nCt + 2393.524 \times Eta_{epsi_D} - 284.446 \times F01[C-F] \quad (4.1)$$

$$n_{(Training)} = 122, n_{(Test)} = 40$$

$$R^2_{(Train)} = 0.851, Q^2_{(LOO)} = 0.832, R^2_{(adj)} = 0.843, MAE_{(Train)} = 482.451$$

$$Q^2_{F1} = 0.921, Q^2_{F2} = 0.920, Q^2_{F3} = 0.916, CCC = 0.960, MAE_{(Test)} = 430.542$$

The developed model was statistically reliable as the internal as well as external validation metrics were far above the required threshold values.

4.1.2 Chemical Read-Across (RA) prediction

To perform the similarity-based Read-Across predictions, the structural and physiochemical parameters of the developed QSPR model were used. Hyper-parameters (similarity approach, the number of close source compounds, σ, and γ) optimization was done using the training set containing the selected variables. The training and test sets with the selected features were used as the inputs for the RA predictions based on the different similarity approaches like Euclidean distance-based similarity, Gaussian kernel-based similarity, and Laplacian kernel-based similarity. The results obtained show that the Gaussian kernel-based similarity has the best predictive quality for the test set (or query set) using the default hyper-parameters (close source compounds=8, σ=0.5, and γ=0.5) with Q^2_{F1}=0.906, Q^2_{F2}=0.905, MAE_{Test}=418.004, and $RMSE_P$=580.938. The same information of the hyper-parameters and Laplacian kernel-based

similarity were used to calculate the similarity and error-based RASPR descriptors for individual training and test sets respectively.

4.1.3 q-RASPR model development

Clubbing of the structural and physiochemical features with the similarity and error-based measures was done before further model development. The new descriptor matrix contains information on both QSPR and RA-based predictions. The training set formed after clubbing the features is used for the selection of the important contributing descriptors for the development of the models. A 5 descriptors combination MLR model was prepared based on internal validation metrics. Finally, a PLS model was developed using the selected 5 descriptors with 4 latent variables and was evaluated for its robustness, reliability, and predictive ability using various internal and external validation parameters. **Equation (4.2)** (*vide infra*) shows the corresponding q-RASPR model and the descriptors involved. The detailed information on the descriptors is listed in **Table 4.1**. The **Scatter plot (Figure 4.1)** represents the observed and predicted detonation heat energy values of individual training and test set compounds. The graph infers that there is a low difference between observed and corresponding predicted values of compounds present in both the training set and the test set.

$$Q = 1930.622 + 217.106 \times F01[N-O] - 78.832 \times X\% + 130.881 \times SddsN + 237.814 \times nCt + 0.536 \times RAfunction(GK) \quad (4.2)$$

$$n_{(Training)} = 122, n_{(Test)} = 40$$

$$R^2_{(Train)} = 0.846, Q^2_{(LOO)} = 0.828, R^2_{(adj)} = 0.839$$

$$Q^2_{F1} = 0.927, Q^2_{F2} = 0.927, Q^2_{F3} = 0.923, CCC = 0.963$$

$$MAE_{(Train)} = 489.865, MAE_{(Test)} = 395.705, RMSE_c = 723.177, RMSE_P = 510.755$$

Table 4.1: List of descriptors and their contribution in the final PLS q-RASPR model

S. No.	Descriptor	Type	Description	Contribution
1.	X%	Constitutional indices	Percentage of halogen atoms	**Negative (-ve)**
2.	F01[N-O]	2D Atom Pairs	Frequency of N - O at topological distance 1	**Positive (+ve)**
3.	nCt	Functional group counts	Total number of tertiary carbon	**Positive (+ve)**
4.	SddsN	Atom-type E-state indices	Sum of ddsN E-states (-N==)	**Positive (+ve)**
5.	RA function (GK)	RASPR descriptor	All structural information	**Positive (+ve)**

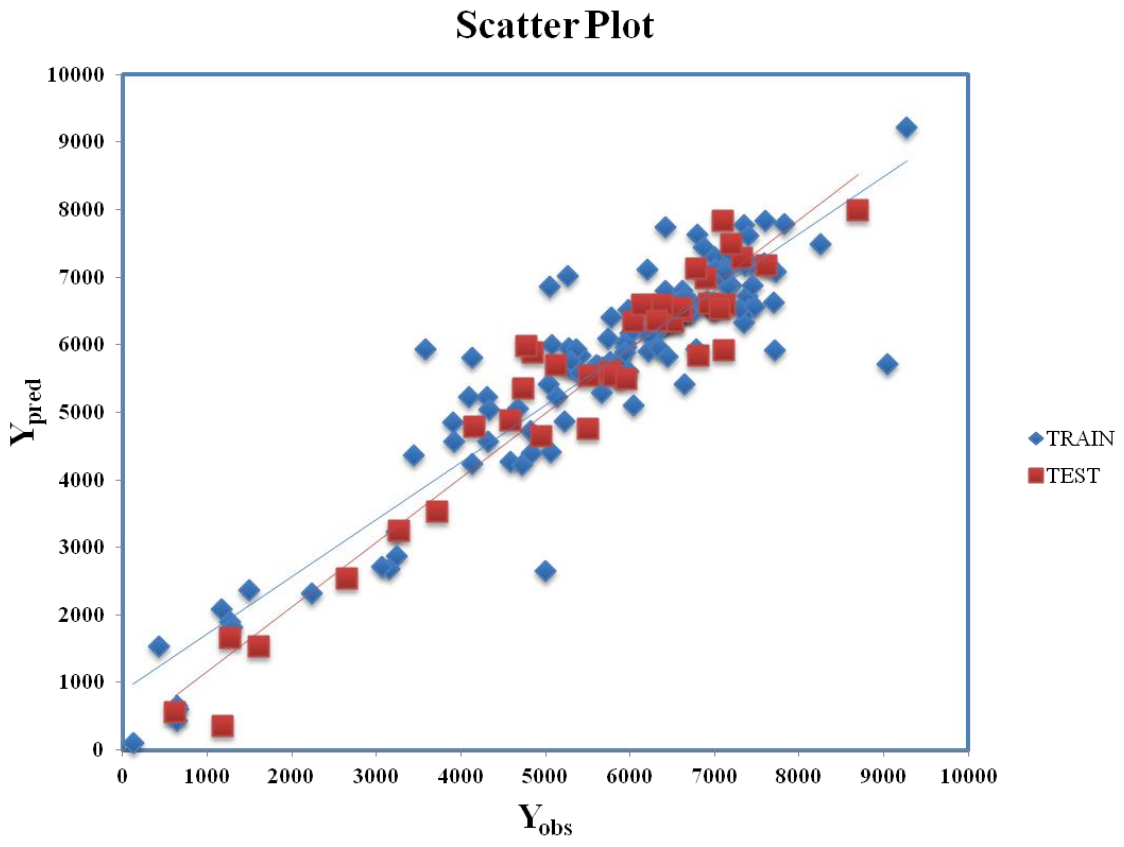

Figure 4.1: Scatter Plot (Y_{obs} vs Y_{pred}) for Eq. (4.2)

Additionally, we have also checked for the structural outliers in the training and test sets using the **Williams Plot (Figure 4.2).** The plot infers that two of the compounds from the training set and one compound from the test set are structural outliers.

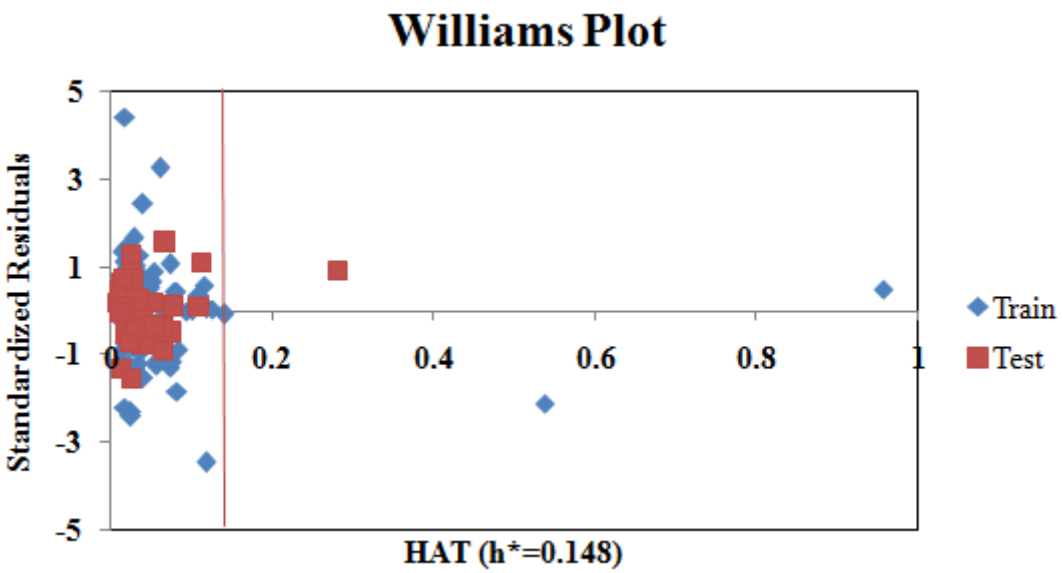

Figure 4.2: Williams plot (standardized cross-validated residuals vs. leverage values)

4.1.4 Descriptors interpretation of the PLS q-RASPR model

The descriptor ***RA function (GK)*** is a composite RASPR descriptor that contains all the selected atomic as well as structural information of the compounds. The *RA function (GK)* descriptor contributes positively to the prediction of detonation heat energy of N-containing compounds which is easily visualized in *3,6-Bis(1H-1,2,3,4-tetrazolyl-5-amino)-1,2,4,5-tetrazine* (**12**) where the value of *RA function (GK)* is more resulting in high detonation heat energy while in *3,3'-Azobis(6-amino-1,2,4,5-tetrazine)* (**13**), *RA function (GK)*is low resulting in a low detonation heat energy.

The descriptor **nCt** defines the number of tertiary carbons in the compound and it contributes positively to the prediction of detonation heat energy. *Octanitrocubane* (**97**) due to its cage-like structure represents a total of 8 such tertiary carbons in its structure present at the vertices. Compounds having a ring/cage structures can liberate more energy at the time of detonation because of the excess strain energy associated with the ring (Li, 2009). In *Isopentanetrioltrinitrate* **(156)**, the value of detonation heat energy is less as it contains only a single tertiary-Carbon.

The descriptor **F01[N-O]** defines the frequency of N-O bonds at the topological distance 1. This descriptor contributes positively to the value of detonation heat energy which can be seen in *4,4'-heavy (N-trinitroethyl-N-nitro)-3,3'-difurazan* (**47**) and *Heavy (N-trinitroethyl-N-nitro)furazan* (**48**) having 20 and 18 N-O bonds respectively and high detonation heat values, while *3-nitro-1,2,4-triazole* (**8**) and *1-methyl-2,4-dinitrobenzene* (**19**) have 2 and 4 N-O in their structures respectively; hence, they have low values of detonation heat. In the compounds, F01[N-O] corresponds to the presence of explosophores in the form of nitro (NO_2), nitrito (ONO_2), furazan ring, furaxan ring, etc. leading to the production of more detonation heat energy (Wang et. al., 2022).

The descriptor **X%** depicts the percentage of halogen present in the compound. This descriptor contributes negatively to the value of detonation heat energy. This can be seen in *2,2-Difluoro-2-nitroethyl trifluoromethane-sulfonate* (**65**) having a high halogen percentage and showing the least value of detonation heat among all the 162 compounds whereas *Methyl 4-fluoro-4,4-dinitrobutyrate* (**76**) has the lowest halogen percentage and have more value of detonation heat energy. In *trifluoromethane-sulfonate* (**65**), the electronegative fluorine atom is situated close to the positively charged nitrogen (more energy, less stable), therefore stabilizing its energy due to ion-dipole interaction resulting in a decrease in detonation energy.

The descriptor **SddsN** describes the atom-type E-state index for -N== groups (nitro) and contributes positively to the detonation energy. The nitrogen present in the form of the nitro group is ina high energy state (higher oxidation state in nitro) which after explosion forms inert N_2 gas (lowest oxidation state) and hence releases more energy (Kumar and Elias, 2019). *Pentaerythritoltetranitrate* (**135**) and*1-Nitropiperazine-2,3-co(1',3'-dinitroimidazolidinone-2')-5,6-nafurazan* (**45**) have higher SddsN values compared to *Hexanitrodiphenylsulfide* (**38**) and *Tetranitroglycoluril* (**108**) respectively, having lower E-state index for the –N== group showing lower detonation energy.

The descriptors with their respective VIP levels and compounds with higher and lower detonation heat energy values associated with individual descriptors are represented in **Figure 4.3.**

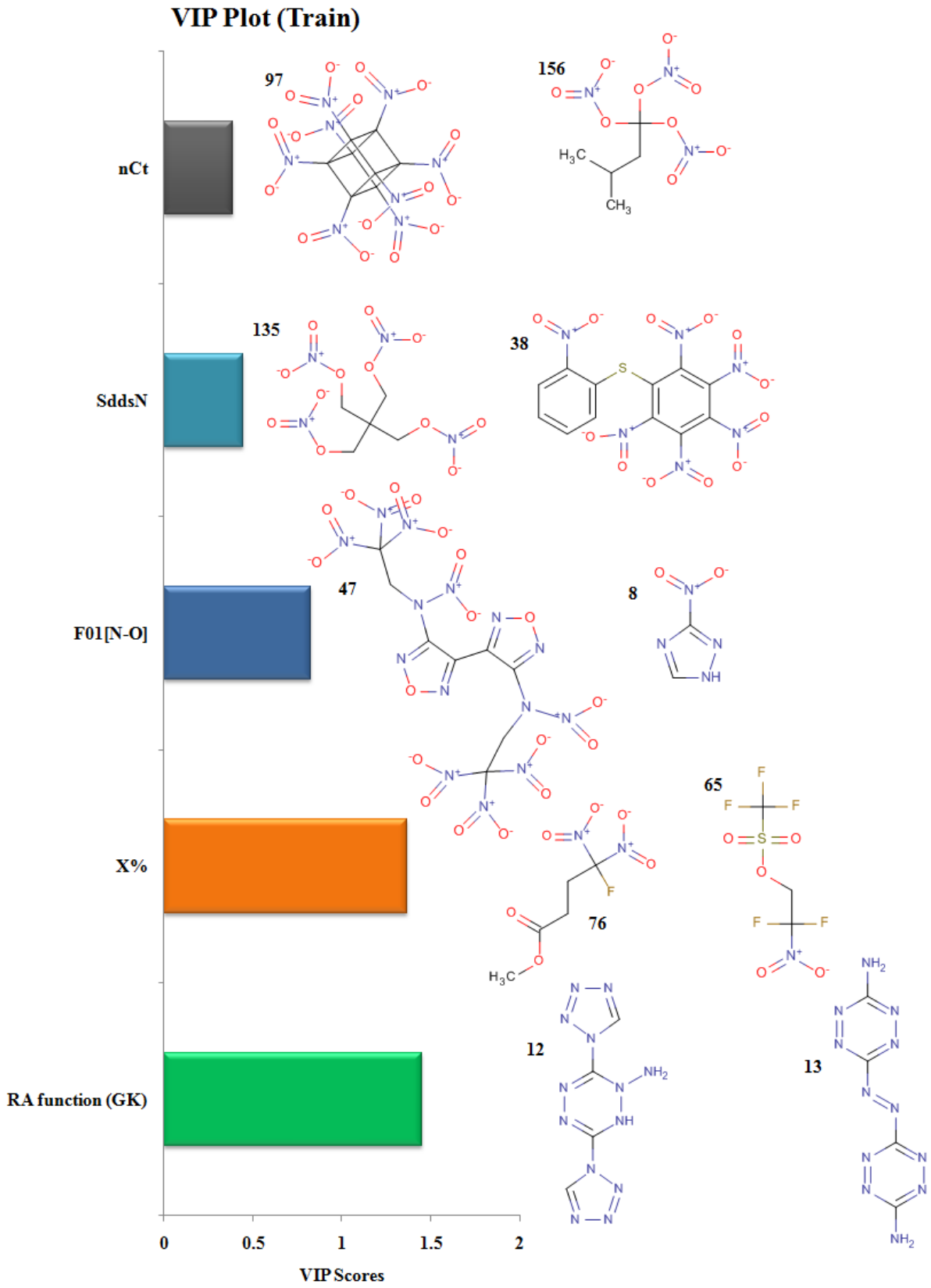

Figure 4.3: Variable importance plot with structural representations of molecules with higher and lower Q values

4.1.5 Predictions through various ML models

We have also employed different machine-learning algorithms for the prediction of the detonation heat energy of N-containing compounds. Here, in this work, we have applied 7 different ML algorithms to develop our models and check their predictive performance. Before applying different ML methods, we have scaled both the descriptor matrix and the response values of individual training and test sets using a java-based tool Scale1.0 available fromhttps://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. For the optimization process, we have used a python-based tool Hyperparameter Optimizer v1.2 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home and performed a grid search for optimizing the hyper-parameters of each method using the scaled training set as input. The results of RF and Adaboost/AB show that these models are not robust as the difference between the values of R^2 and Q^2_{LOO} is high and hence are not reliable. The predictive performance of Gradient boost, XGBoost, and ridge regression are almost similar to our developed PLS model. Based on the MAE_{Test} results, the Gradient boost model shows the best predictive performance with the lowest error. To check the quality of the models we have performed the MAE cross-validation (CV), i.e. leave-one-out CV, 20 times 5-fold CV, and shuffle-split CV with n_splits =1000. The MAE CV results of RF, AB, GB, and SVM models have increased significantly which shows the models are of inferior quality in comparison to other models. On comparison, it was found that the PLS and RR models have efficient predictive performance in terms of Q^2_{F1} and $MAE_{P.}$ So, on the basis of $RMSE_P$ criteria, we have selected the PLS q-RASPR model as the best model for the prediction of both the training and test sets. The validation metrics of all the models are represented in **Table 4.2**.

Table 4.2: Comparison between performances of different q-RASPR models

q-RASPR MODELS	Training Set Statistics					Test Set Statistics				Optimized Hyperparameters
	R^2	Q^2_{LOO}	MAE_C	MAE_{LOO}	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$	
PLS	0.846	0.828	0.265	0.28	0.391	0.927	**0.927**	0.214	**0.276**	(LV=4)
RF	0.957	0.722	0.142	0.36	0.206	0.885	0.884	0.242	0.347	(n=120, leaf=1, split=3, depth=none)
AB	0.864	0.677	0.301	0.41	0.367	0.859	0.858	0.284	0.385	(n=60, loss=linear)
GB	0.878	0.750	0.226	0.33	0.349	0.925	0.925	**0.199**	0.280	(n=150, leaf=1, split=2, depth=1)
XGB	0.840	0.825	0.267	0.28	0.399	0.926	0.925	0.213	0.279	(n=60, depth=5, booster=gblinear, learning rate=0.1)
SVM	0.885	0.747	0.212	0.31	0.337	0.854	0.853	0.224	0.391	(C=5.0, Degree=2, Gamma=auto)
LSVM	0.831	0.824	0.270	0.28	0.409	0.916	0.915	0.223	0.297	(C=25.0)
RR	0.847	0.829	0.264	0.28	0.390	0.927	0.926	0.214	0.277	(α=1.0)

4.1.6 Interpretation of the PLS plots

To identify the outliers in the respective training set and test set, the **DModX** (distance to model X) AD plots (**Figure 4.4**) were prepared for each training set and each test set, and it shows that there are 2 outlier compounds in the training set while no compounds from the test set were outside the applicability domain (AD). To find the relation between the X-variables (descriptors) and the Y-variable (property) and also get an idea about the variable importance, we have prepared the **loading plot (Figure 4.5)** developed using the first and second PLS components. The interpretation of the plot depicts that the descriptors situated at a greater distance from the origin have more impact on the Y-variable (here property). In the plot, *RA function (GK)* and X% descriptors were the farthest from the origin showing their larger impact on the prediction of detonation heat which can also be verified from the VIP plot **(Figure 4.3)** showing their VIP score >1. The **coefficient plot** (**Figure 4.6)** shows the standardized regression coefficient values of each descriptor of the model. The **bubble plot (Figure 4.7)** shows the standardized regression coefficient of the descriptors on the Y-axis and the size of the bubble corresponds to their importance (VIP levels). The **score plot (Figure 4.8)** was prepared using the first two PLS components for the training set. The score plot for the training set contains a total of 4 outliers. We have also performed the Shapley Additive exPlanations **(SHAP)** analysis (Rodriguez-Perez and Bajorath, 2020) **(Figures 4.9)** to see the contribution of each feature to the outcome of the model (i.e. detonation heat). The SHAP analysis for the training set shows that the F01[N-O] is the most important descriptor for the prediction of detonation heat while in the case of the test set, the *RA function (GK)* has the highest impact on the detonation heat prediction. The nCt descriptor is of the least importance for both the training and test set.

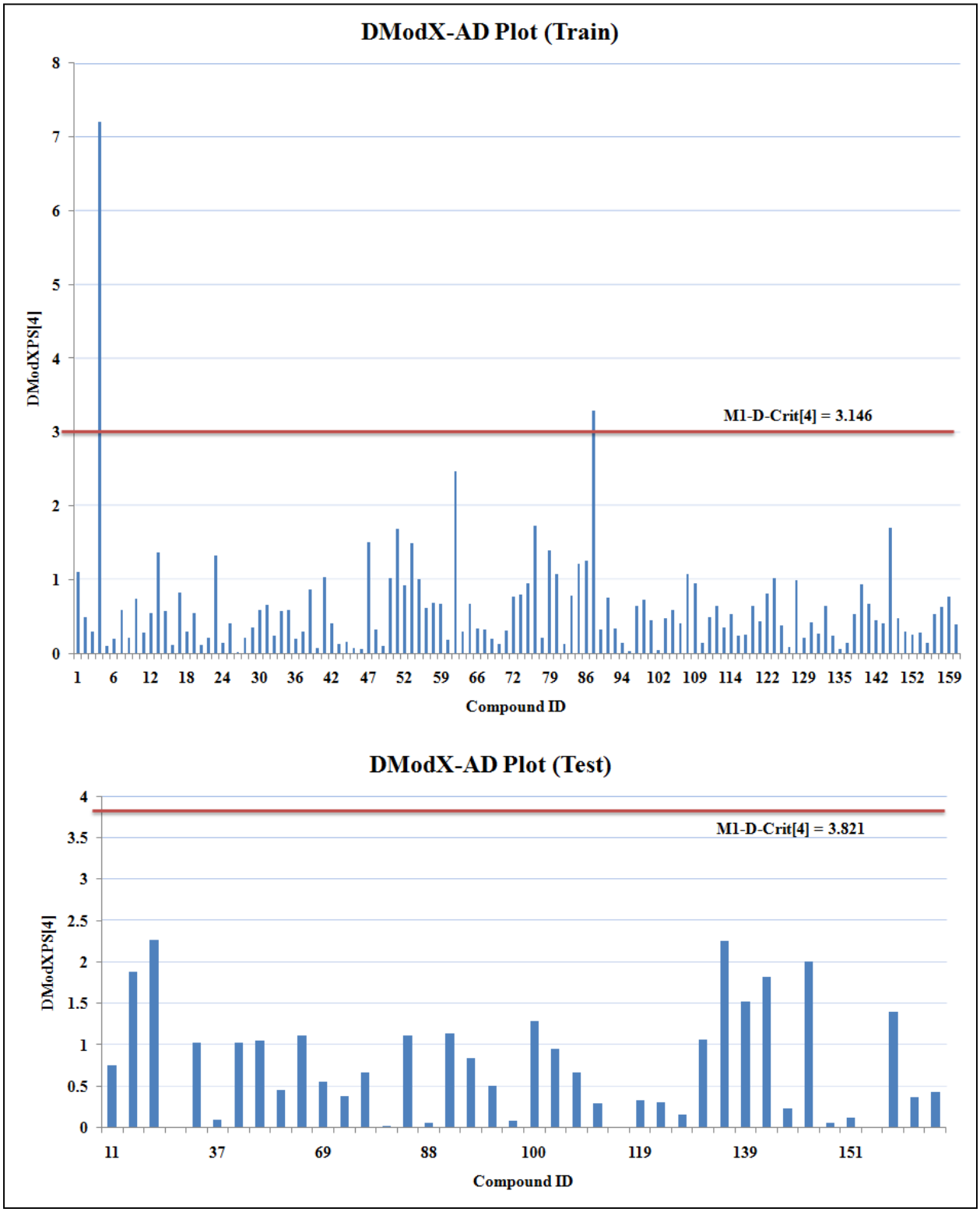

Figure 4.4: DModX AD plot

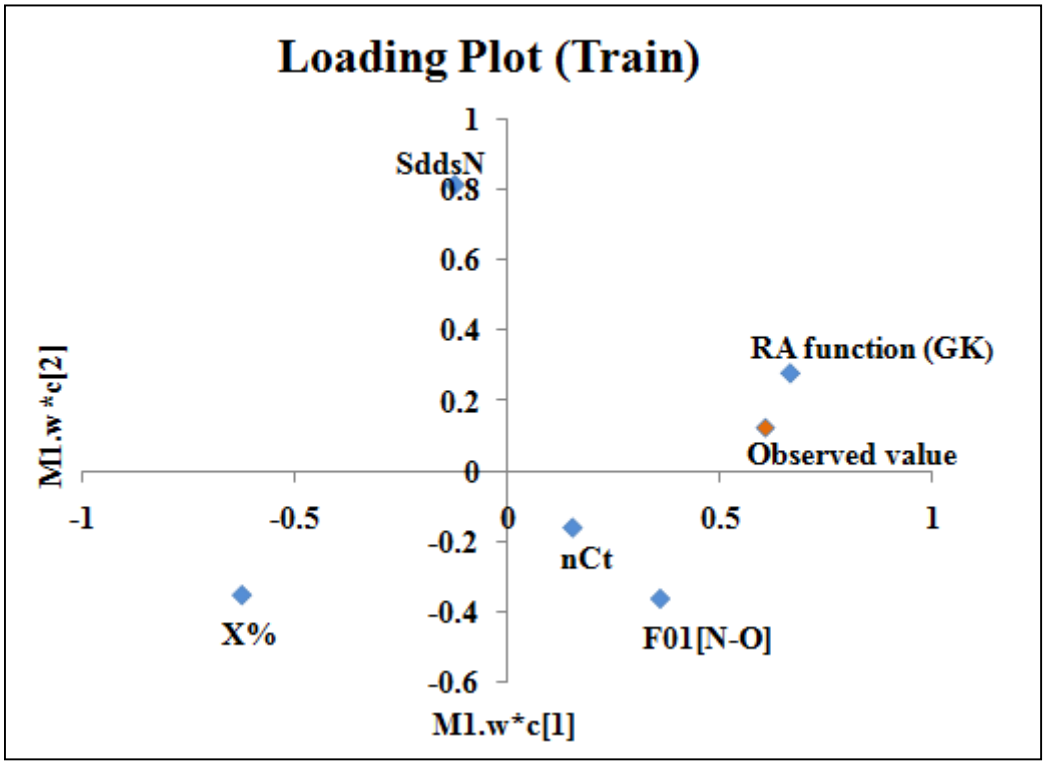

Figure 4.5: Loading plot

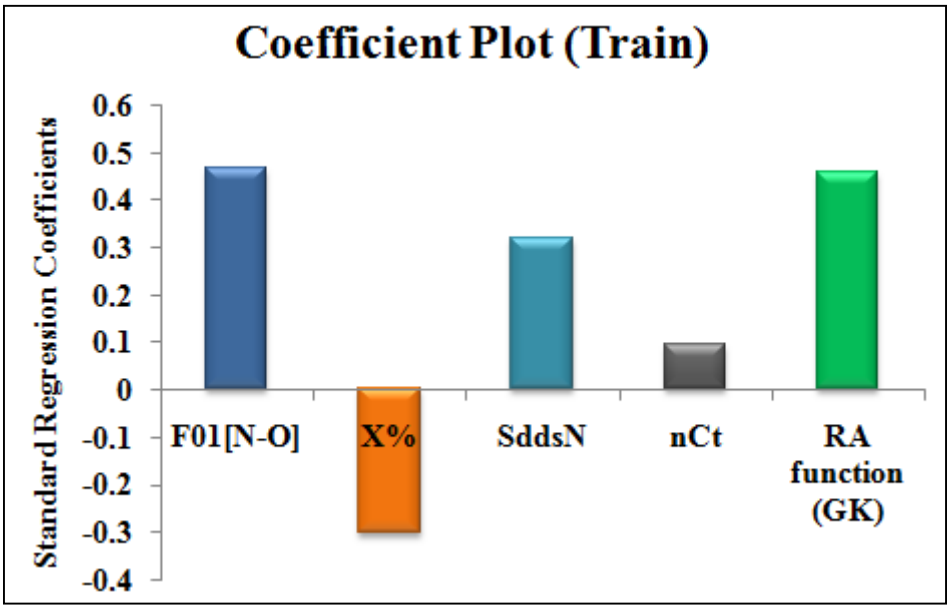

Figure 4.6: Coefficient plot

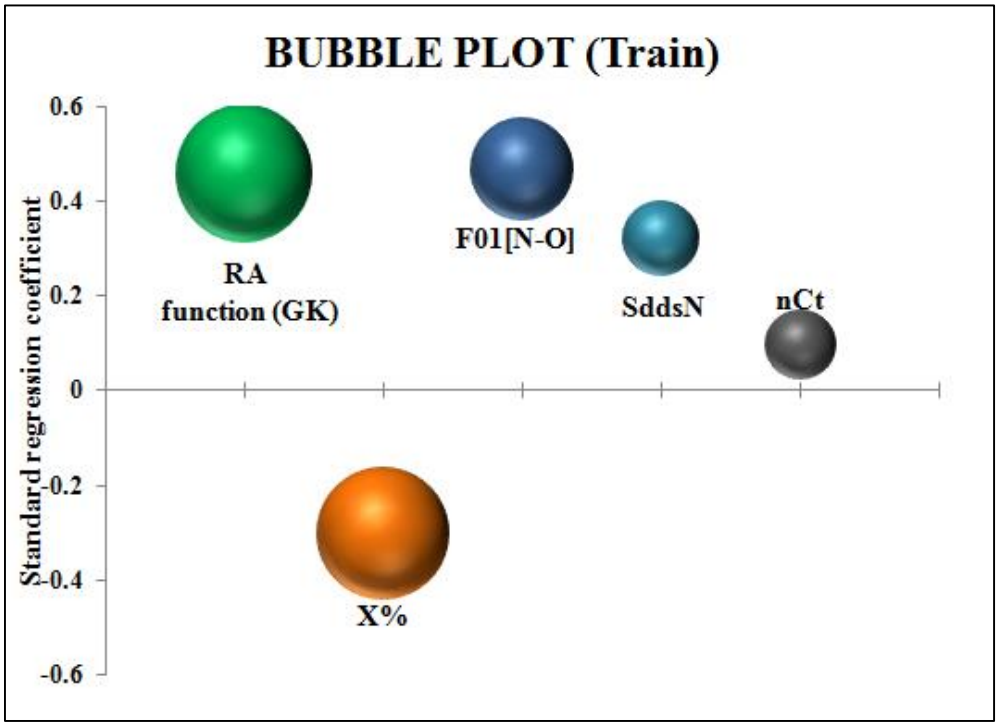

Figure 4.7: Bubble plot of the q-RASPR model depicting the contribution of the descriptors

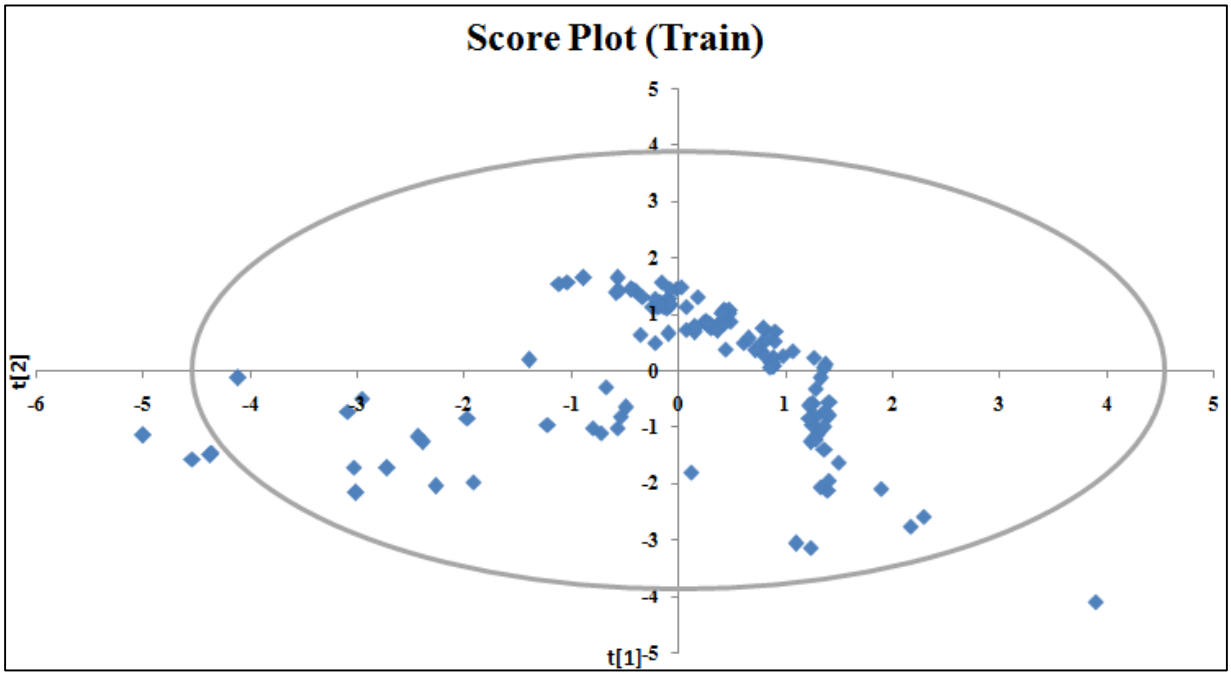

Figure 4.8: Score plot of q-RASPR model for training set

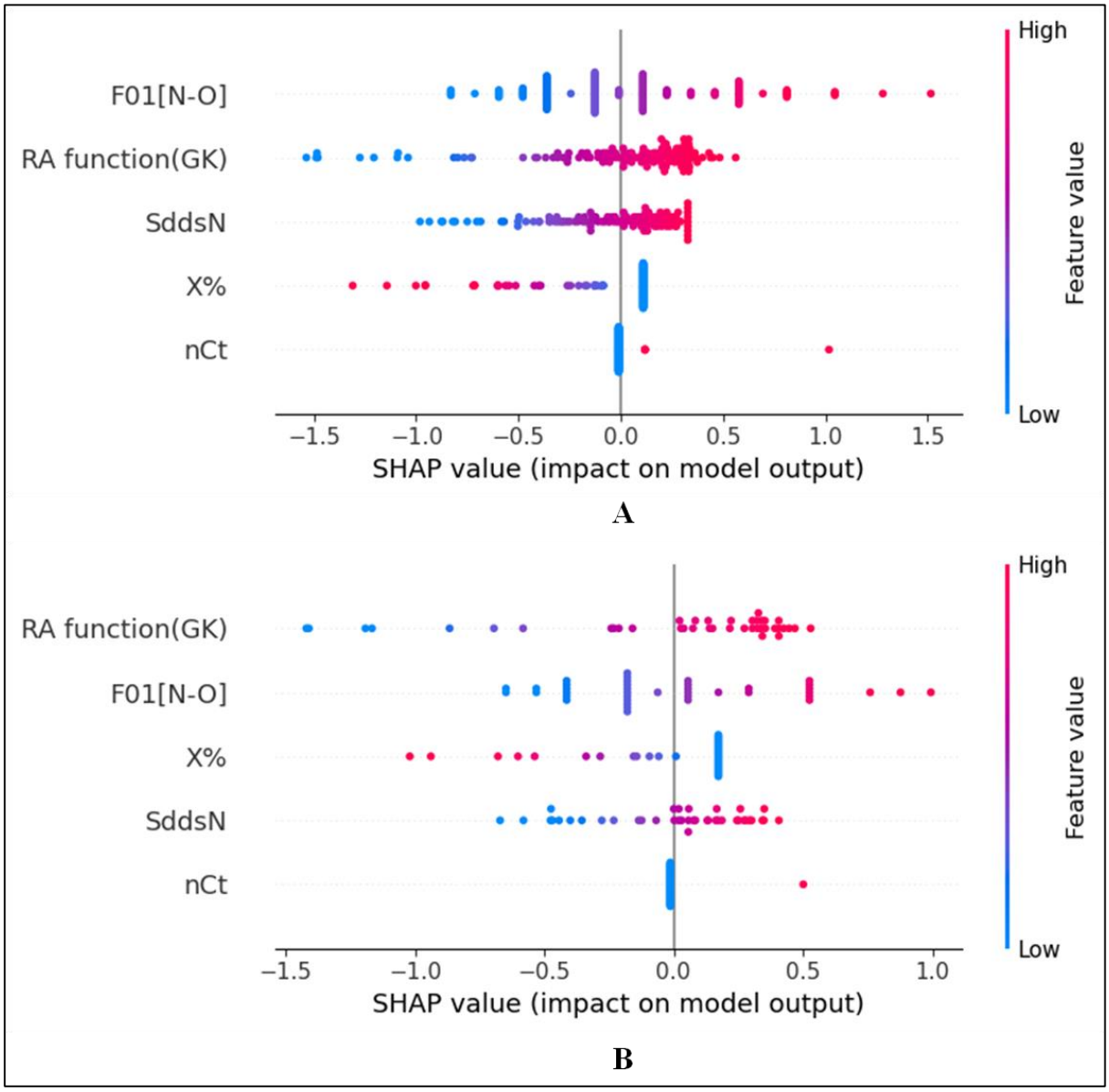

Figure 4.9: SHAP analysis for training set (A) and test set (B) for the developed PLS model

4.1.7 Comparison of the q-RASPR model with other models

4.1.7.1 Comparison with the present QSPR model

We have compared the results of the developed q-RASPR model with our own QSPR model (section 3.1). The chemical information associated with both the models is same as the features appearing in the QSPR model were used for the RASPR descriptor calculation and further model development. Although the internal validation metrics were comparable for both QSPR ($R^2_{(Train)}$ =

0.851, $Q^2_{(LOO)}$ =0.832, $MAE_{(Train)}$ = 482.451) and q-RASPR ($R^2_{(Train)}$ = 0.846, $Q^2_{(LOO)}$ = 0.828, $MAE_{(Train)}$ = 489.865) models, the results of the test set prediction of the q-RASPR model (Q^2_{F1} =0.927, Q^2_{F2} = 0.927, $MAE_{(Test)}$ = 395.705) were better than the QSPR model (Q^2_{F1} =0.921, Q^2_{F2} =0.920, $MAE_{(Test)}$ = 430.542) in terms of $MAE_{(Test)}$. The external validation results show that there is an enhancement in the prediction quality of the q-RASPR model. It should also be noted that the q-RASPR model is developed using 5 descriptors while the QSPR model has 6 descriptors. This depicts that the q-RASPR model with a lower number of descriptors is more efficient in the prediction of detonation heat with same type of chemical information.

4.1.7.2 Comparison with the previous model

The previous QSPR study was performed using the random forest (RF) algorithm using a set of 3D-descriptors. Our q-RASPR model shows better predictive results in terms of Q^2_{F1} and $RMSE_P$ with a lower number of descriptors. It should also be noted here that we have only used the 2D-descriptors, which do not need prior structure optimization, unlike computing3D-descriptors. A comparison of our model's different validation metrics with those of the previously developed model is given in **Table 4.3.**

Table 4.3: Comparative results of previous model with our q-RASPR model

Models	No. of descriptors	R^2	$RMSE_C$	Q^2F_1	$RMSE_P$
He et al., 2021	7	0.965	377.8	0.880	641.8
Our q-RASPR model	5	0.846	723.177	**0.927**	**510.755**

4.2 Study 2: Predicting performance and stability parameters of energetic materials (EMs) using the machine learning-based q-RASPR approach

4.2.1 QSPR model development

We have developed 4 different QSPR models for the prediction of 4 different properties of energetic compounds. Three models ($T_{dec,}$ density, and ΔH_f°) were developed using the PLS regression algorithm while one of the models [for the melting point (T_m)] was developed using Multiple Linear Regression (MLR).

A 10-descriptor MLR model for decomposition temperature (T_{dec}) was selected after the feature selection process by performing a grid-search using the Best Subset Selection tool v2.1 available from http://teqip.jdvu.ac.in/QSAR_Tools/. The same descriptor set was used to develop the final PLS QSAR model with 5 latent variables (LVs) which are optimized by LOO Q^2. The equation for the model is given in **Table 4.4.** The training set of the melting point (T_m) temperature data set was subjected to a forward step-wise feature selection process to enlist the prominent features closely related to the melting point. A 29-descriptor MLR QSPR model was developed to predict the melting point temperature of the compounds. The MLR equation for the model is shown in **Table 4.4.** The feature selection of the density data set was performed through step-wise selection using the training set. After the feature selection process, a 6-descriptor MLR model was prepared and further, PLS regression was used to develop the QSPR model with 5 LVs. The PLS equation of the model is given in **Table 4.4.** For the enthalpy of formation ($\Delta H_f°$), a step-wise feature selection process was performed after the division of the data set. The pool of descriptors so obtained from the step-wise selection was then used to develop several MLR models through a grid-search approach using a java based tool Best Subset Selection tool v2.1 available from http://teqip.jdvu.ac.in/QSAR_Tools/. An 11-descriptor MLR model was selected based on the cross-validation result (Q^2_{LOO}), and further with the same set of descriptors, a PLS QSPR model was developed with 3 LVs. The PLS equation is given in **Table 4.4.**

Table 4.4: Model equations and validation metrics of the developed QSPR models

Property	Model equation	Training set metrics	Test set metrics
T_{dec} (PLS model)	$T_{dec} = 436.990 + 3.952 \times C\% - 142.266 \times B01[O-O] - 28.762 \times B03[N-O] + 9.558 \times Hy - 14.993 \times LOGP99 + 34.492 \times nArNO2 + 24.399 \times C-005 - 25.504 \times nN \pm 39.061 \times B01[N-N] - 34.360 \times B01[N-O]$ $Descriptors = 10, LVs = 5$	$n_{training} = 424$ $R^2 = 0.578$ $Q^2_{LOO} = 0.557$ $MAE_{tr} = 45.257$ $RMSE_C = 57.971$	$n_{test} = 141$ $Q^2_{F1} = 0.621$ $Q^2_{F2} = 0.621$ $MAE_{te} = 44.919$ $RMSE_P = 54.814$
T_m (MLR model)	$T_m = 291.1 + 13.46 \times Ui + 22.98 \times nHDon + 15.08 \times Rbrid + 26.5 \times B03[C-O] + 19.12 \times nN + 50 \times nArCOOH + 2.1 \times AMW - 0.212 \times T(N..O) + 5.27 \times Rprim + 23 \times nRCOOH - 0.28 \times F10[C-O] + 6.95 \times NdssC - 31.3 \times nR = Cp - 4.25 \times F07[C-N] - 38.1 \times minsssB + 1.539 \times MLOGP2 - 350 \times Mi - 3.69 \times nCbH - 16.7 \times MaxssCH2 + 11.57 \times N-072 + 1.79 \times O\% - 3.76 \times F05[O-O] - 1.3 \times F10[C-C] + 32.7 \times B02[C-C] - 14.8 \times F02[O-Cl] + 86 \times NssssN^+ + 1.79 \times StN - 4.64 \times F10[O-O] - 9.4 \times nOHs$ $Descriptors = 29$	$n_{training} = 14750$ $R^2 = 0.679$ $Q^2_{LOO} = 0.676$ $MAE_{tr} = 39.633$ $RMSE_C = 51.686$	$n_{test} = 4917$ $Q^2_{F1} = 0.670$ $Q^2_{F2} = 0.670$ $MAE_{te} = 39.626$ $RMSE_P = 52.501$

Density (PLS model)	$Density = 1.235 + 0.120 \times AMW - 1.409 \times Mp + 0.015 \times nX - 0.008 \times X\% + 0.196 \times MCD - 0.015 \times NRS$ $Descriptors = 6, LVs = 5$	$n_{training} = 9604$ $R^2 = 0.924$ $Q^2_{LOO} = 0.922$ $MAE_{tr} = 0.037$ $RMSE_C = 0.053$	$n_{test} = 3201$ $Q^2_{F1} = 0.928$ $Q^2_{F2} = 0.928$ $MAE_{te} = 0.037$ $RMSE_P = 0.051$
$\Delta H_f°$ (PLS model)	$\Delta H_f° = -25.420 - 196.661 \times nF - 71.385 \times F01[C - O] - 23.045 \times nCsp3 + 91.062 \times nCIC + 187.180 \times F01[N - F] - 115.277 \times O - 058 + 57.671 \times F01[N - N] - 83.572 \times NsOH + 32.203 \times NdsCH + 128.918 \times nCsp + 32.832 \times nN$ $Descriptors = 11, LVs = 3$	$n_{training} = 1924$ $R^2 = 0.967$ $Q^2_{LOO} = 0.966$ $MAE_{tr} = 53.553$ $RMSE_C = 78.571$	$n_{test} = 643$ $Q^2_{F1} = 0.932$ $Q^2_{F2} = 0.931$ $MAE_{te} = 47.903$ $RMSE_P = 67.412$

4.2.2 Chemical Read-Across (RA) predictions

The structural and physiochemical features of the developed QSPR model were used to evaluate the similarity-based RA predictions. The default setting of the hyperparameters (σ=1, γ=1, no. of close source/training compounds=10) was used to perform the Read-across predictions for the 3 different similarity approaches like Laplacian kernel-based (LK), Gaussian kernel-based (GK), and Euclidean distance-based (ED) similarity. The prediction results show that the Laplacian kernel-based similarity has the best predictivity for T_{dec}, T_m, and ΔH_f° whereas the Gaussian kernel-based similarity has the best performance for the density data set. The results of RA predictions are shown in **Table 4.5**. The default hyperparameters of each similarity measure were used to calculate the RASPR descriptors for each of the data sets.

Table 4.5: Read-across predictions for different data sets

Metrics / Property	Q^2_{F1}	Q^2_{F2}	MAE_P^*	$RMSE_P^*$	Similarity measure
T_{dec}	0.645	0.645	41.756	53.037	LK
T_m	0.736	0.736	34.075	46.520	LK
Density	0.925	0.925	0.039	0.052	GK
ΔH_f°	0.924	0.924	49.100	70.787	LK

*Non-standardized values

4.2.3 q-RASPR model development

The motive behind the development of the q-RASPR model is to increase the external predictivity of the model over the traditional QSPR model. The calculated RASPR descriptors are composed of different similarity, error, concordance as well as predictive functions from the structural and physiochemical descriptors. These calculated RASPR descriptors were clubbed with the previously selected structural and physiochemical descriptors to form the new descriptor matrix for the individual training and test set. The prepared training set was further used for the selection of the prominent features for the development of the model. To develop the q-RASPR model for

T_{dec} and ΔH_f°, a grid search was performed on the fused descriptor matrix (obtained from the fusion of QSPR and RASPR descriptors)to develop several MLR models using the Best Subset Selection tool v2.1 freely available from http://teqip.jdvu.ac.in/QSAR_Tools/. The best MLR model was selected based on the leave-one-out (LOO) cross-validation result, and the same was used further to develop the final PLS q-RASPR model with a lower number of LVs which are optimized using LOO Q^2. For the density dataset, a forward step-wise feature selection method was used to develop the MLR model, and further, the PLS algorithm was applied to obtain the final PLS q-RASPR model. Both grid-search and step-wise selection were performed for the T_m dataset, and in both cases a univariate q-RASPR model with *RA function (LK)* as the only descriptor was obtained. The final model equations for individual models with their internal and external validation metrics are tabulated in **Table 4.6.**

Additionally, to evaluate the predictivity of the developed PLS q-RASPR model for the density dataset, we have collected a true external set of 37 energetic compounds from Rice and Byrd[43]and calculated the validation metrics for the same. The result shows that our model can predict new compounds accurately.

$$Q^2_{F1} = 0.883, MAE = 0.073, RMSE = 0.088$$

The scatter plots shown in **Figure 4.10** represent that there is a high correlation between the observed and predicted values. As in the individual plots, the scattering is not much which represents that the quality of the developed models was good. The distribution of the heat of formation data set in **Figure 4.10** shows that only a few (approx. 14) compounds are present far from the clusters of training (1924) and test (643) sets which are very small in number w.r.t. the whole training set compounds. Also, the division algorithm used here was based on the Kennard-Stone method which divides the data set based on the descriptor matrix, and not based on property/response.

Table 4.6: Model equations and validation metrics for the developed q-RASPR models

Property	Model equation	Training set metrics*	Test set metrics*
T_{dec} (PLS model)	$T_{dec} = 144.449 + 2.684 \times C\% - 43.374 \times B01[O-O] - 15.109 \times B03[N-O] + 8.425 \times Hy - 8.311 \times LOGP99 + 19.520 \times nArNO2 + 16.965 \times C-005 - 8.233 \times B01[N-N] + 0.596 \times RA\ function\ (LK) - 0.870 \times SE\ (LK)$ $Descriptors = 10, LVs = 5$	$n_{training} = 424$ $R^2 = 0.620$ $Q^2_{LOO} = 0.600$ $MAE_{tr} = 42.313$ $RMSE_C = 55.013$	$n_{test} = 141$ $Q^2_{F1} = 0.676$ $Q^2_{F2} = 0.676$ $MAE_{te} = 41.383$ $RMSE_P = 50.683$
T_m (Univariate model)	$T_m = 9.081 + 0.952 \times RA\ function(LK)$ $Descriptor = 1$	$n_{training} = 14750$ $R^2 = 0.746$ $Q^2_{LOO} = 0.746$ $MAE_{tr} = 33.959$ $RMSE_C = 46.005$	$n_{test} = 4917$ $Q^2_{F1} = 0.741$ $Q^2_{F2} = 0.741$ $MAE_{te} = 34.297$ $RMSE_P = 46.520$

Density (PLS model)	$Density = 0.425 + 0.042 \times AMW - 0.690 \times Mp + 0.082 \times MCD + 0.741 \times RA\ function(GK) - 0.049 \times CVsim(GK)$ $Descriptors = 5, LVs = 4$	$n_{training} = 9604$ $R^2 = 0.940$ $Q^2_{LOO} = 0.940$ $MAE_{tr} = 0.035$ $RMSE_C = 0.047$	$n_{test} = 3201$ $Q^2_{F1} = 0.939$ $Q^2_{F2} = 0.939$ $MAE_{te} = 0.035$ $RMSE_P = 0.047$
$\Delta H_f°$ (PLS model)	$\Delta H_f° = 28.972 + 1.020 \times RA\ function(LK) - 0.298 \times SD\ Activity(LK) - 1.884 \times nCsp3$ $Descriptors = 3, LVs = 2$	$n_{training} = 1924$ $R^2 = 0.943$ $Q^2_{LOO} = 0.942$ $MAE_{tr} = 61.718$ $RMSE_C = 103.603$	$n_{test} = 643$ $Q^2_{F1} = 0.931$ $Q^2_{F2} = 0.931$ $MAE_{te} = 47.158$ $RMSE_P = 67.630$

*Non-standardized MAE and RMSEP values are shown

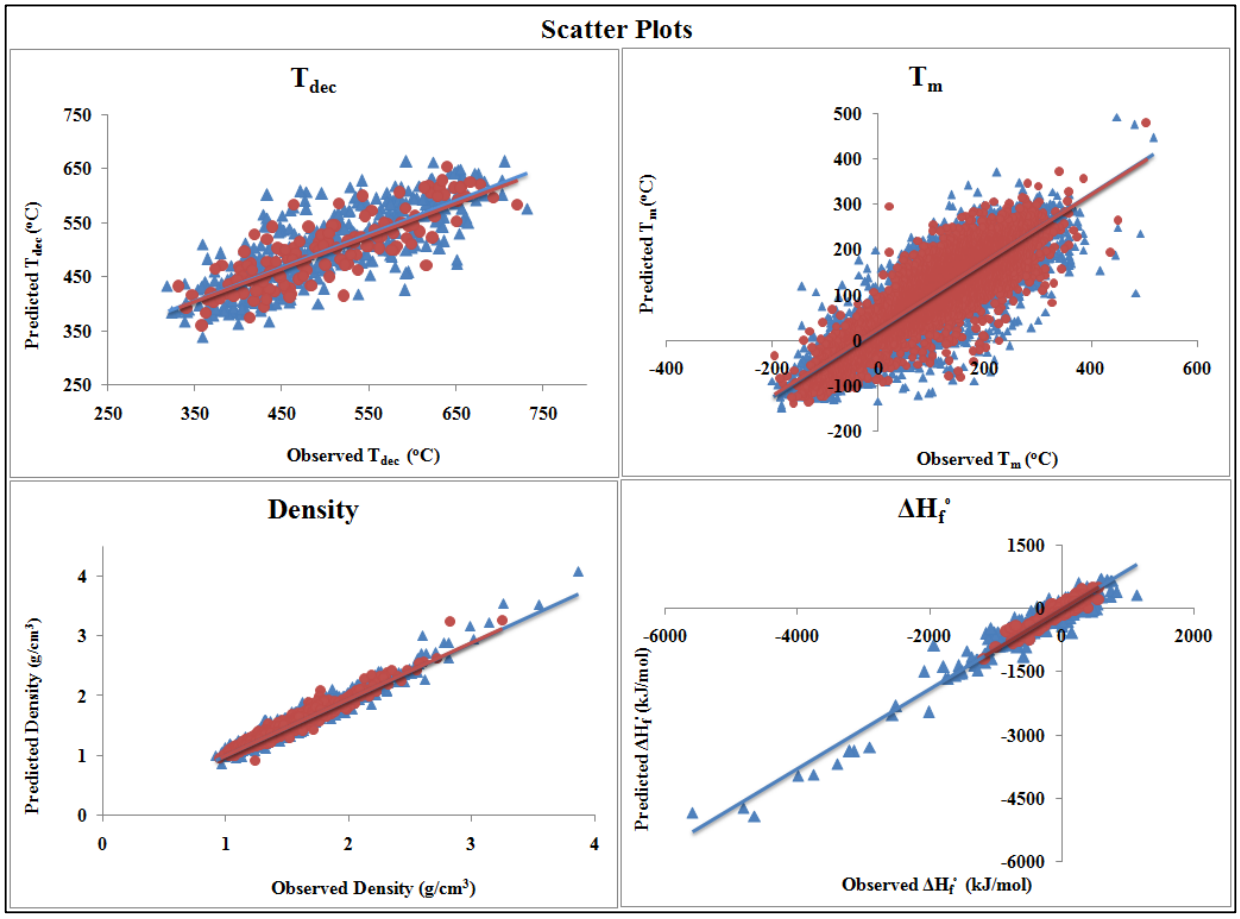

Figure 4.10: Scatter plots for the individual PLS models

The violin plots shown in **Figure 4.11** represent the frequency of compounds with the residual values (i.e. observed – predicted) in the training and test sets of respective models for each property. The graph seems to be more flattened in the middle portion representing that there are more compounds in the training and test sets with lower residual values, and the tapered end at both the ends of the violin represents the lower number of compounds with high residuals.

4.2.4 PLS plot interpretation

Models were developed from all the datasets, except for the melting point (Tm) dataset, using PLS regression, as the final model of the Tm data set contains only a single descriptor. Hence, a univariate model has been reported got Tm instead of reporting it in the form of a PLS model, which represents several original descriptors into a lower number of latent variables (LVs).

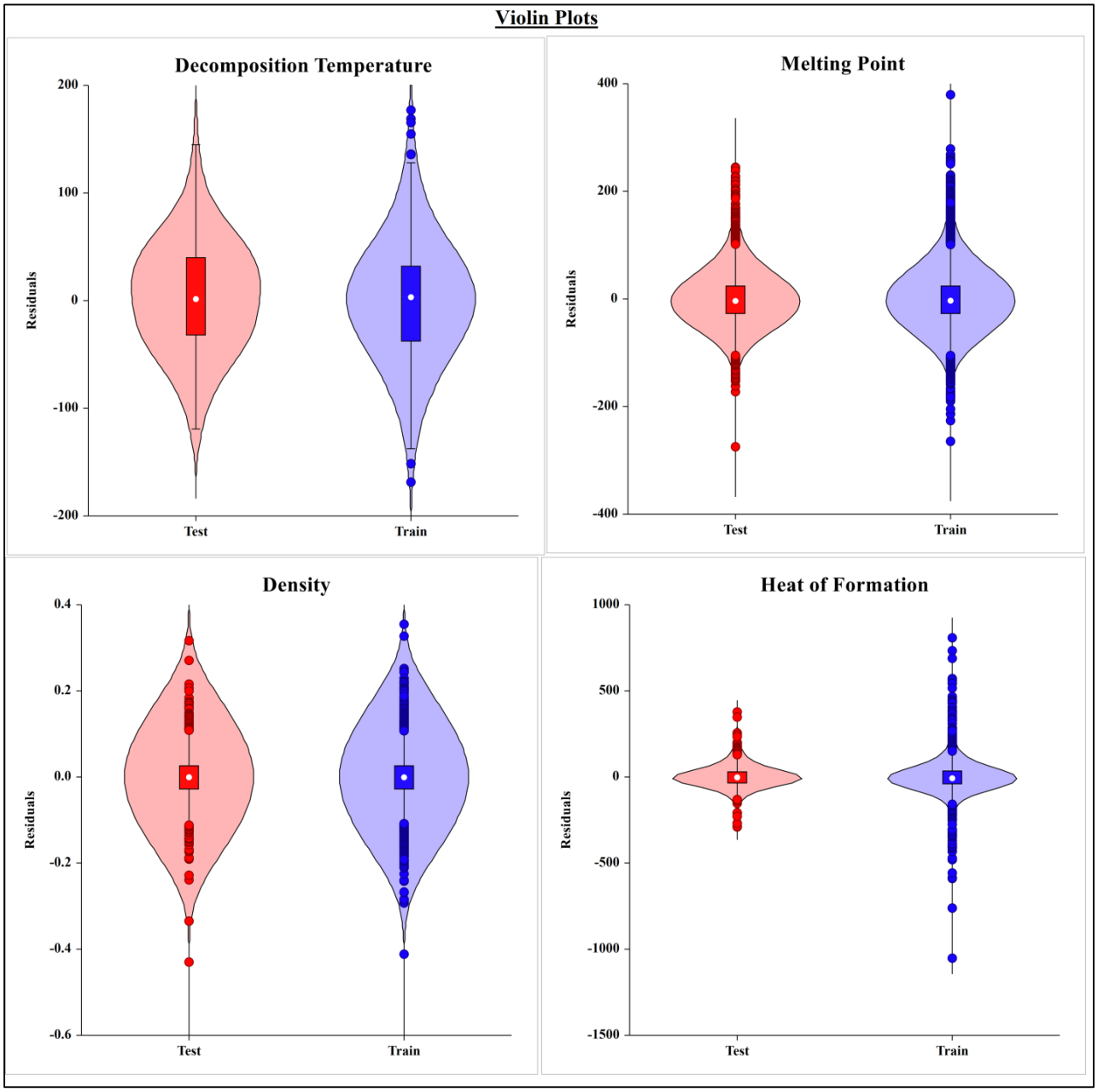

Figure 4.11: The violin plot of each model represents the variation in the residual values for compounds in the respective training and test sets. The width of the plot represents the frequency/number of data points for the given residuals.

We have used the DModX (Distance to Model X) approach to check the numbers of outliers present in the training and test sets, respectively (except for the melting point data set). The DModX-AD plots of the developed PLS models are given in supplementary materials (**Figures**

4.12, 4.13, and 4.14). The applicability domain of the univariate model for melting point was calculated using the leverage approach. The leverage values for the individual data points of training and test set were calculated using the Java-based tool Hi_Calculator-v2.0 (accessible from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). William's plot (**Figures 4.15)** represents the outliers from the training and test sets of the melting point data set with leverage values higher than the critical h* value (0.0004). The percentage (%) of compounds as outliers in the training and test sets of the respective models is shown in the bar graph in **Figure 4.16**.

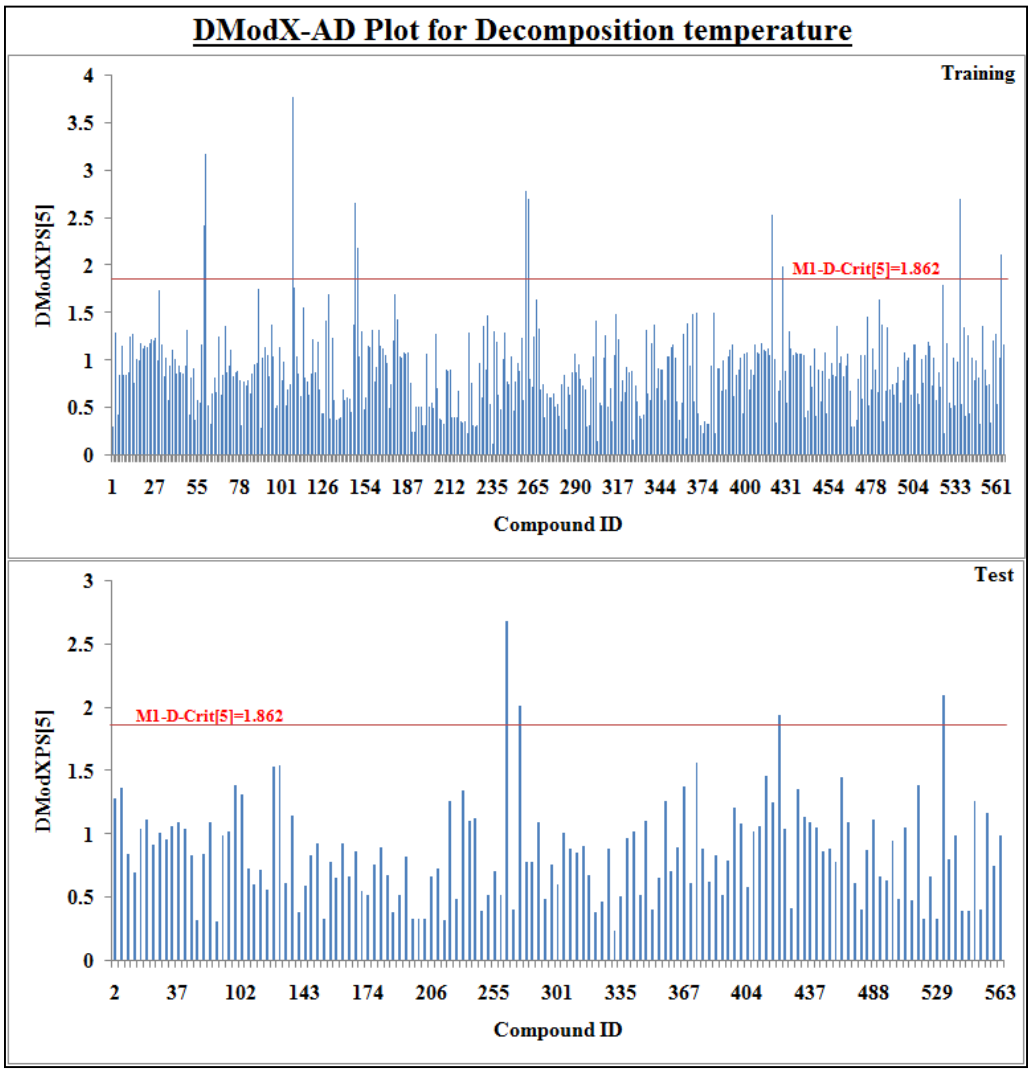

Figure 4.12: AD plot for T_{dec}

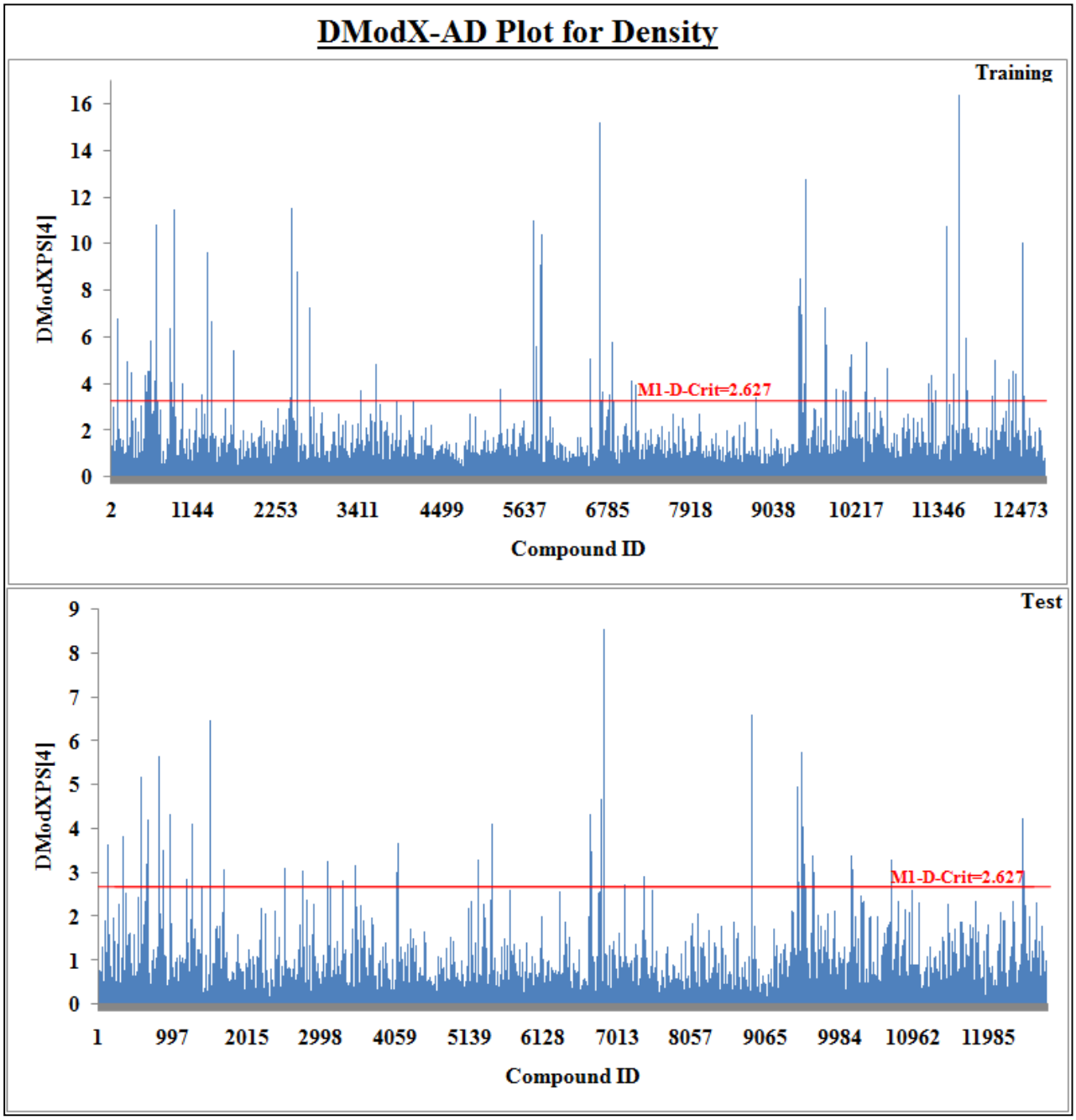

Figure 4.13: AD plot for Density

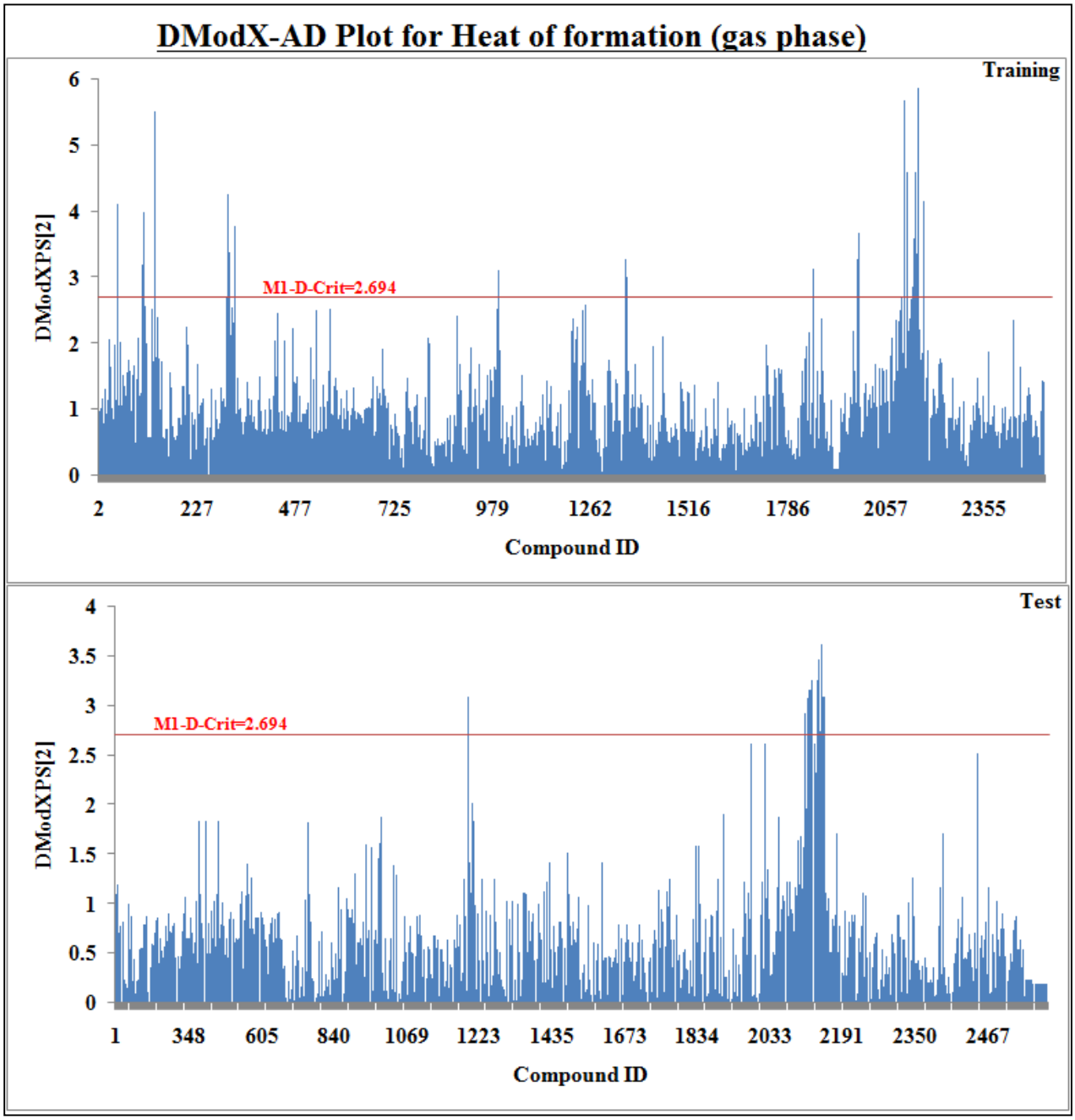

Figure 4.14: AD plot for ΔH_f°

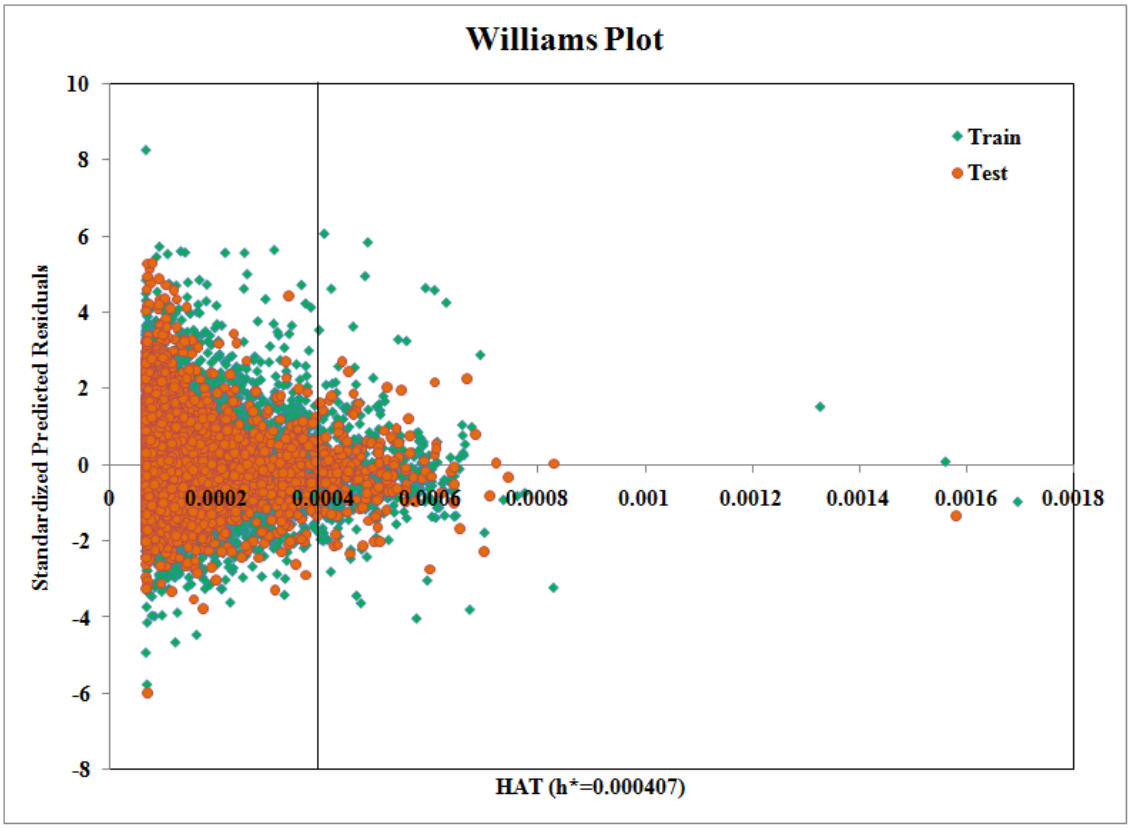

Figure 4.15: Williams plot for T_m

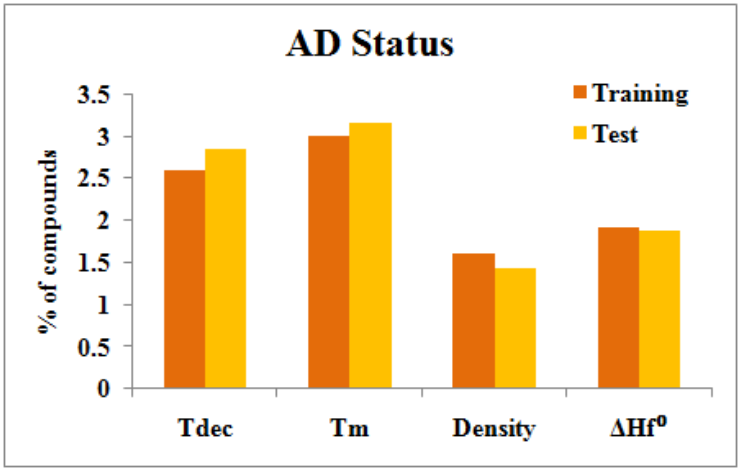

Figure 4.16: AD status for individual models. It represents the percentage (%) of compounds as outliers in training and test sets of the respective model.

To check the impact of the descriptors (i.e. X-variables) on the property (Y-variable), we have developed the loading plot (**Figure 4.17)** using the first 2 PLS components. The variables that are more dispersed from the origin have a high impact on the model. We have also used the VIP plot (**Figure 4.18)** to interpret the importance of respective descriptors according to their VIP values in the model. The coefficient plot representing the standardized regression coefficient values for each descriptor of the individual model and the score plots for each model are given in the supplementary materials (**Figures 4.19** and **4.20,** respectively**).** As the score plot for each model (**Figure 4.20**) has been developed using the first 2 components (t1 and t2) of the model, the compounds outside the ellipse can be considered outliers for the model with 2 latent variables. The ellipse indicates the model's applicability domain, as defined by Hotelling's t^2 (a multivariate generalization of Student's- *t*-tests). The AD study shows that the compounds present far away from the ellipse are just not outliers based on the two components of the model. Still, they are also outliers for the whole descriptor space shown in the DModX applicability domain (AD) plots (**Figures 4.12, 4.13 and 4.14**).

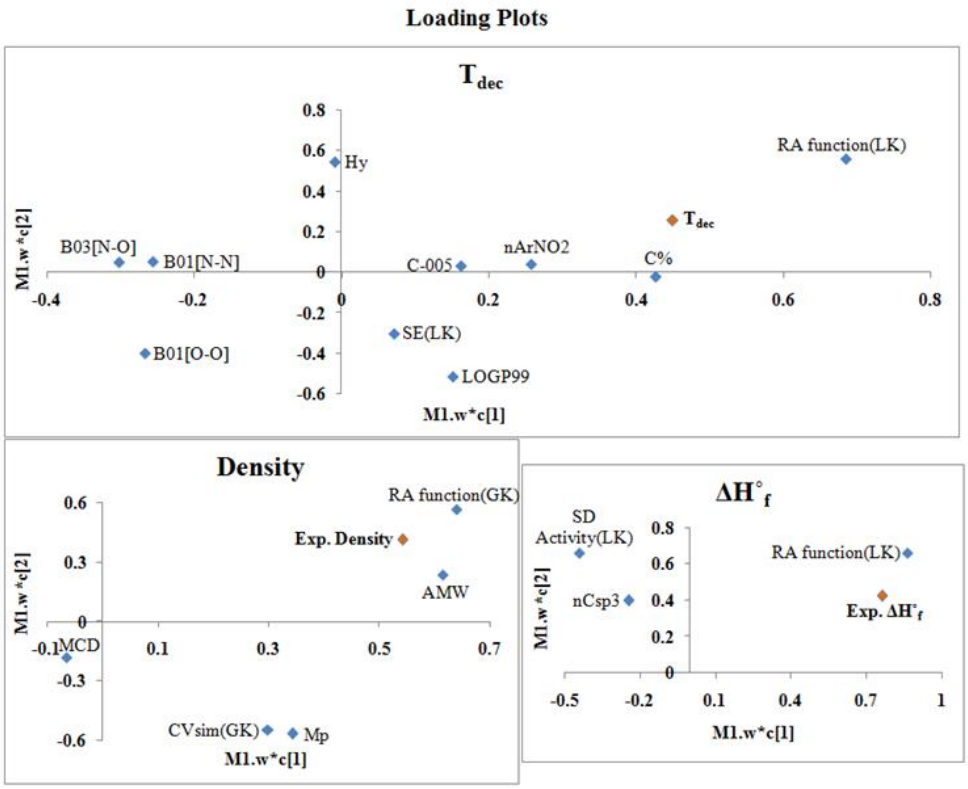

Figure 4.17: Loading Plots for different PLS q-RASPR models

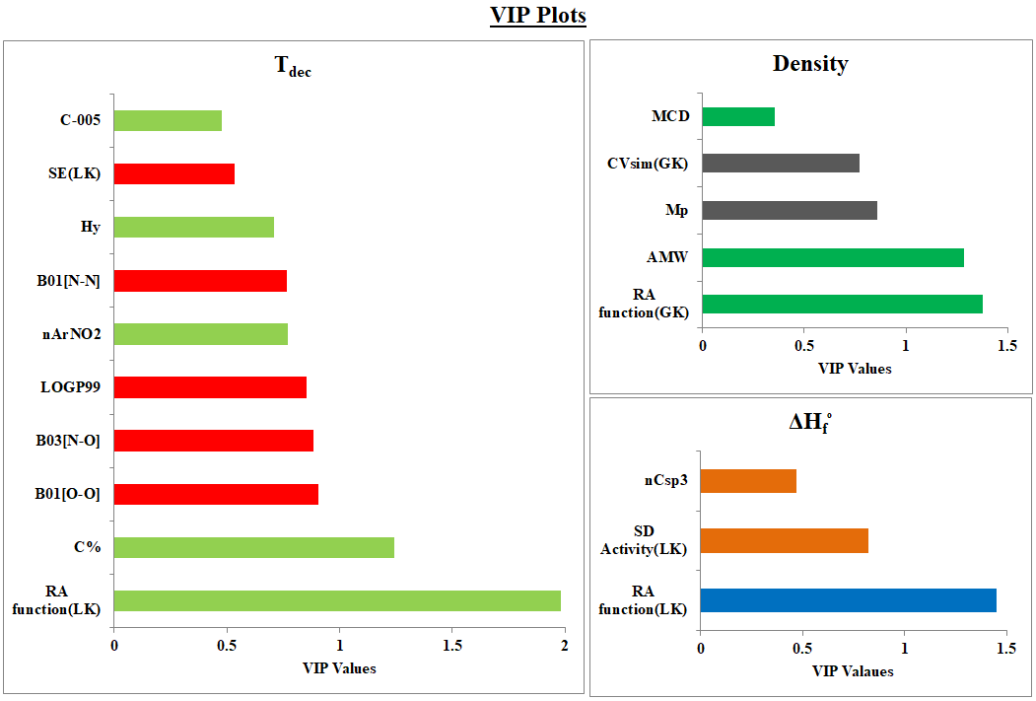

Figure 4.18: VIP plots for different PLS models

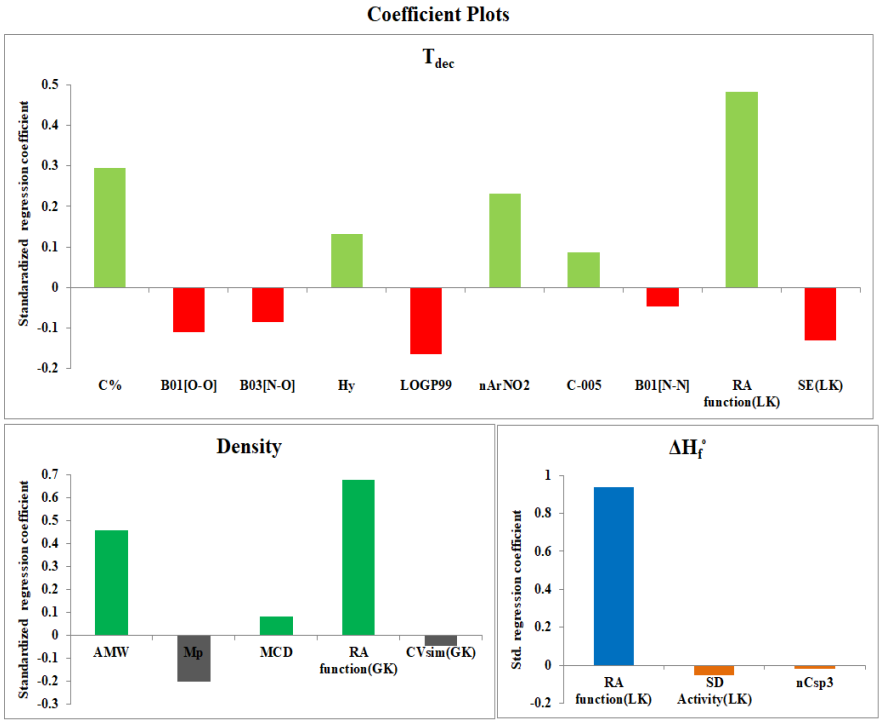

Figure 4.19: Coefficient Plots for each PLS model

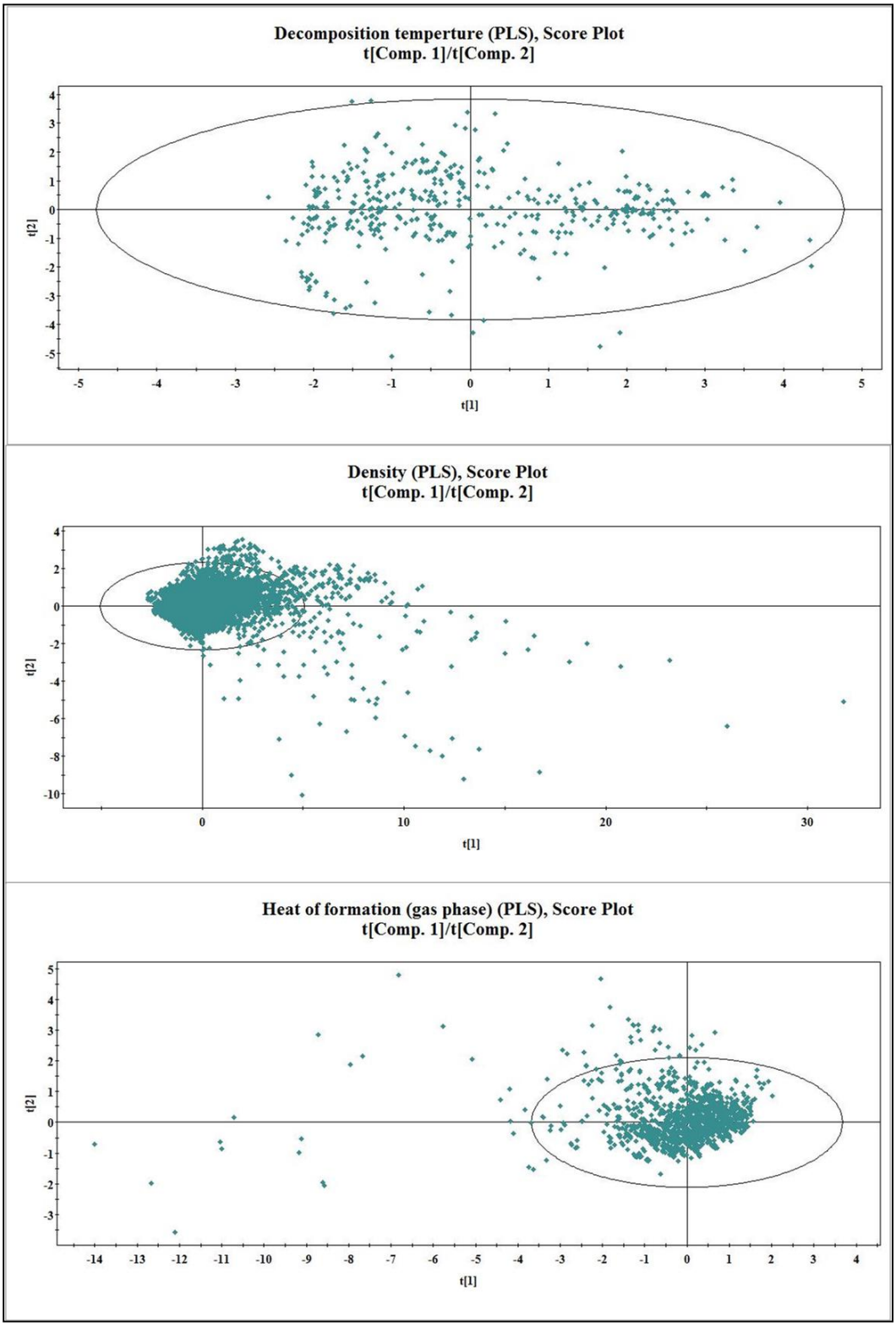

Figure 4.20: PLS Score Plots for respective models

The bubble plot (**Figure 4.21**) collectively represents the VIP values (size of bubble) of the descriptors with their standardized regression coefficient values (Y-axis) of the PLS models.

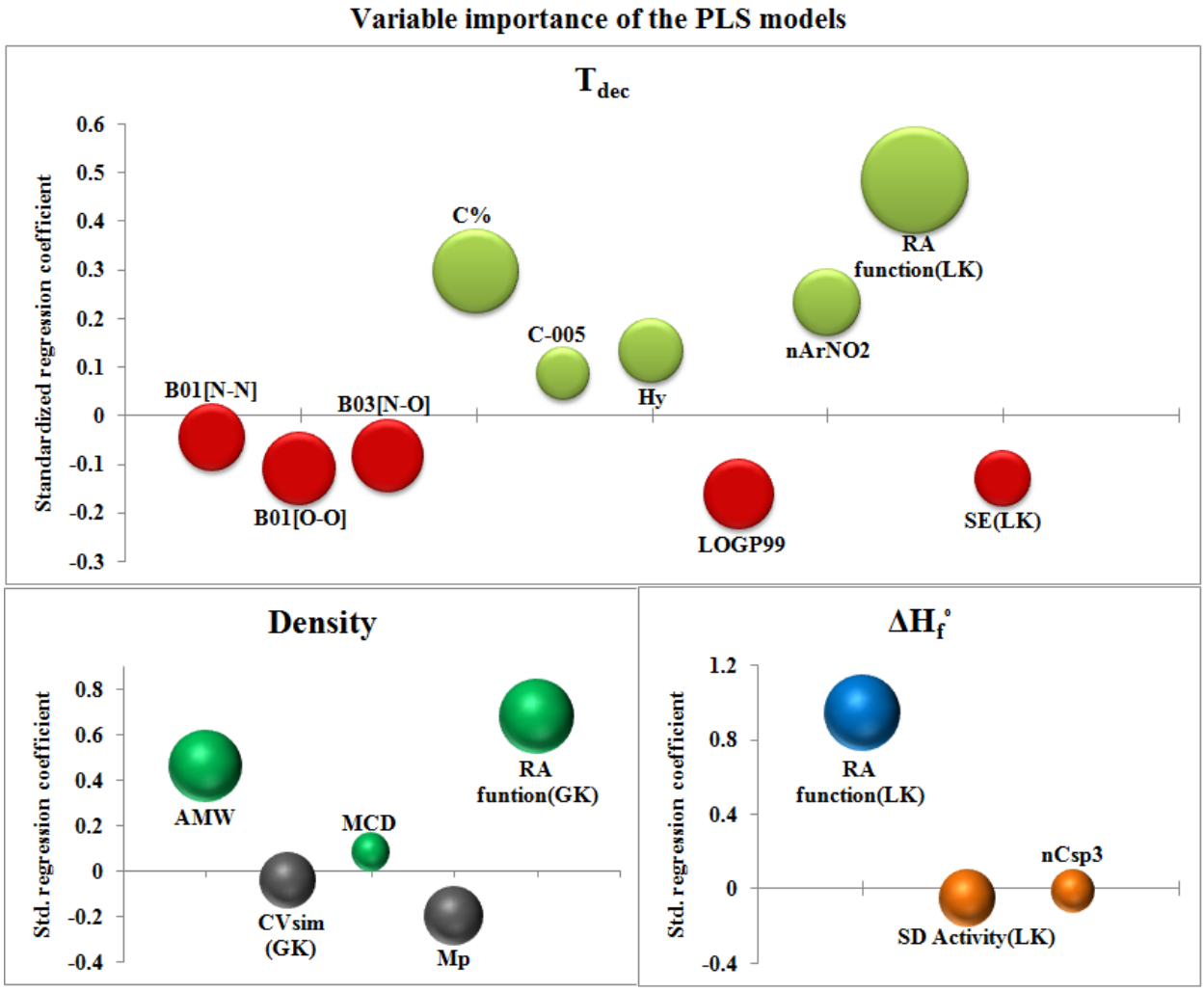

Figure 4.21: Bubble Plots for the respective PLS models representing variable importance and standardized regression coefficients

4.2.5 Prediction through ML models

We have also developed various ML models for the individual data sets (except T_m) to predict the respective properties. Here, 7 different ML algorithms were used to develop the models. Scale1.0 (a Java-based tool) was used to scale the descriptors and response values of both the training and test sets. The default values of the hyperparameters for each algorithm were used during the model development process. The statistics for the model quality and predictivity are reported in **Tables 4.7, 4.8, and 4.9** given below. We have also performed 5-fold and 10-fold cross-validation and noted MAE_C (CV) to check the quality of our developed models. For the density and ΔH_f° data sets, 5-fold and 10-fold cross-validated R^2 values were determined to check the robustness of the developed models, as LOO-CV is not appropriate for such large data sets. The graphical representation of various quality and error metrics for different ML-based q-RASPR models is shown in **Figure 4.22.**

In the case of T_{dec}, the external validation metrics of the **PLS** model infer that it has better predictivity in comparison to the other developed ML models in terms of Q^2_{F1}, Q^2_{F2}, and $RMSE_P$.

For the density data set, the external predictions of the LSVM, RR, and PLS models were similar in terms of Q^2_{F1} and Q^2_{F2} but the error for the LSVM model in terms of MAE_P was the least among all the models. Therefore, the **LSVM** model can be considered to be the best-performing model for the prediction of density.

For the prediction of gas-phase heat of formation, the **RR** model shows its better predictivity with the least error in terms of MAE_P and cross-validated MAE_C.

We have also performed the Shapley Additive exPlanations (SHAP) analysis (Rodroguez-Perez and Bajorath, 2020) (**Figure 4.23)** for the final ML models to see the impact/importance of the descriptors on the model predictions. It was found in all the 3 models that the descriptors having high feature values and positive SHAP values contribute positively to the predictions and vice-versa. The features which are more dispersed along the X-axis have a high impact on the model.

Table 4.7: Comparison between the performances of different q-RASPR models for decomposition temperature (T_{dec})

T_{dec}	Training set statistics						Test set statistics			
Models	R^2	Q^2_{LOO}	MAE_C	MAE_C± SEM (5-foldCV)	MAE_C± SEM (10-foldCV)	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$
RF	0.935	0.527	0.187	0.54 ± 0.036	0.53 ± 0.035	0.254	0.633	0.633	0.477	0.604
AB	0.632	0.496	0.505	0.58 ± 0.036	0.56 ± 0.028	0.606	0.564	0.564	0.557	0.658
GB	0.853	0.559	0.295	0.54 ± 0.036	0.53 ± 0.038	0.383	0.594	0.594	0.507	0.635
XGB	0.937	0.501	0.189	0.56 ± 0.040	0.55 ± 0.035	0.250	0.591	0.591	0.523	0.637
SVM	0.687	0.544	0.409	0.54 ± 0.031	0.54 ± 0.032	0.559	0.674	0.674	**0.456**	0.569
LSVM	0.613	0.605	0.469	0.49 ± 0.031	0.48 ± 0.028	0.621	0.662	0.662	0.468	0.574
RR	0.621	0.600	0.474	0.50 ± 0.027	0.49 ± 0.028	0.615	0.674	0.674	0.468	0.569
PLS	0.620	0.600	0.474	0.49 ± 0.027	0.49 ± 0.028	0.616	**0.676**	**0.676**	0.463	**0.567**

Table 4.8: Comparison between the performances of different q-RASPR models for density (Den)

Density	Training set statistics						Test set statistics				
Models	R^2	R^2 ± SEM (5-fold CV)	R^2± SEM (10-fold CV)	MAE_C	MAE_C ± SEM (5-fold CV)	MAE_C ± SEM (10-fold CV)	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$
RF	0.991	0.92 ± 0.004	0.92 ± 0.006	0.066	0.19±0.009	0.19±0.006	0.931	0.936	0.931	0.182	0.250
AB	0.913	0.89 ± 0.013	0.88 ± 0.009	0.224	0.23±0.004	0.23±0.006	0.295	0.905	0.905	0.227	0.305
GB	0.947	0.92 ± 0.004	0.92 ± 0.006	0.172	0.19±0.004	0.19±0.006	0.230	0.932	0.932	0.184	0.257
XGB	0.911	0.87 ± 0.004	0.88 ± 0.009	0.205	0.23±0.009	0.22±0.009	0.298	0.905	0.905	0.215	0.303
SVM	0.915	0.87 ± 0.022	0.88 ± 0.016	0.172	0.19±0.009	0.19±0.009	0.292	0.916	0.916	0.178	0.286
LSVM	0.940	0.93± 0.004	0.92 ± 0.003	0.178	0.18±0.004	0.18±0.006	0.247	**0.939**	**0.939**	**0.177**	0.245
RR	0.940	0.93 ± 0.004	0.93 ± 0.006	0.179	0.18±0.004	0.18±0.006	0.244	**0.939**	**0.939**	0.178	**0.243**
PLS	0.940	0.93 ± 0.004	0.92 ± 0.006	0.180	0.18±0.004	0.18±0.006	0.246	**0.939**	**0.939**	0.180	0.244

Table 4.9: Comparison between the performance of different q-RASPR models for the heat of formation (ΔH_f°)

ΔH_f°	Training set statistics								Test set statistics			
Models	R^2	Q^2_{LOO}	R^2 ± SEM (5-fold CV)	R^2 ± SEM (10-fold CV)	MAE_C	MAE_C ± SEM (5-foldCV)	MAE_C ± SEM (10-foldCV)	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_P	$RMSE_P$
RF	0.991	0.934	0.86 ± 0.004	0.87 ± 0.013	0.054	0.18± 0.0031	0.17± 0.028	0.096	0.913	0.913	0.123	0.1758
AB	0.926	0.905	0.82 ± 0.022	0.83 ± 0.016	0.190	0.22± 0.027	0.21± 0.022	0.271	0.879	0.879	0.156	0.207
GB	0.968	0.933	0.88 ± 0.009	0.88 ± 0.016	0.118	0.17± 0.027	0.16± 0.025	0.180	0.925	0.925	0.114	0.163
XGB	0.935	0.897	0.82 ± 0.027	0.79 ± 0.028	0.146	0.20± 0.036	0.20± 0.028	0.255	0.899	0.899	0.137	0.189
SVM	0.827	0.761	0.74 ± 0.094	0.79 ± 0.054	0.154	0.21± 0.058	0.15± 0.044	0.416	0.928	0.928	0.110	0.159
LSVM	0.942	0.942	0.91 ± 0.013	0.90 ± 0.013	0.141	0.14± 0.018	0.19±0.019	0.240	0.930	0.930	**0.108**	0.157
RR	0.943	0.942	0.91 ± 0.013	0.90 ± 0.013	0.142	0.14± 0.018	0.14± 0.016	0.239	**0.931**	**0.931**	**0.108**	**0.156**
PLS	0.943	0.942	0.91 ± 0.013	0.90 ± 0.013	0.143	0.15± 0.018	0.14± 0.016	0.239	**0.931**	**0.931**	0.109	**0.156**

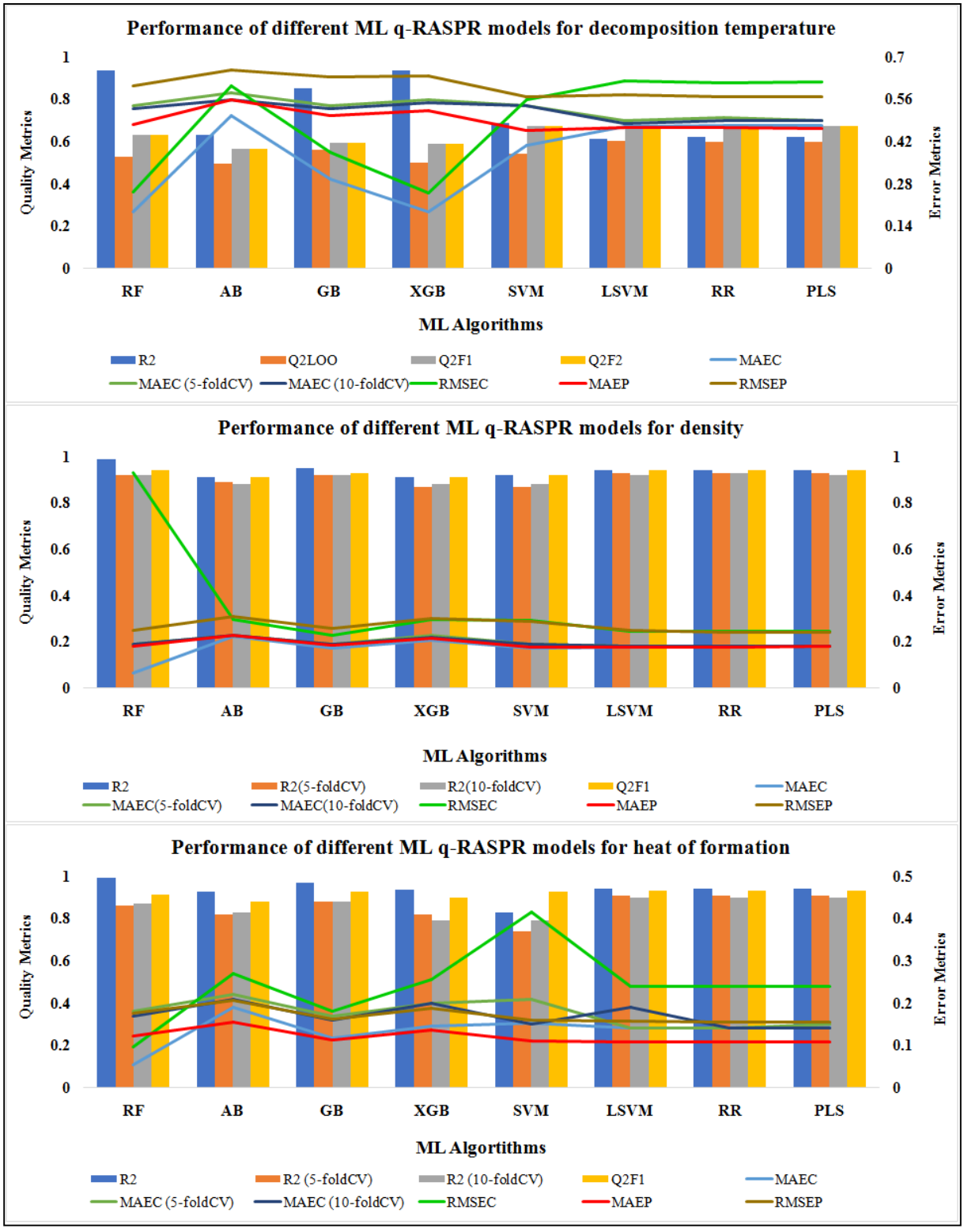

Figure 4.22: Comparison of quality and error metrics of different q-RASPR models

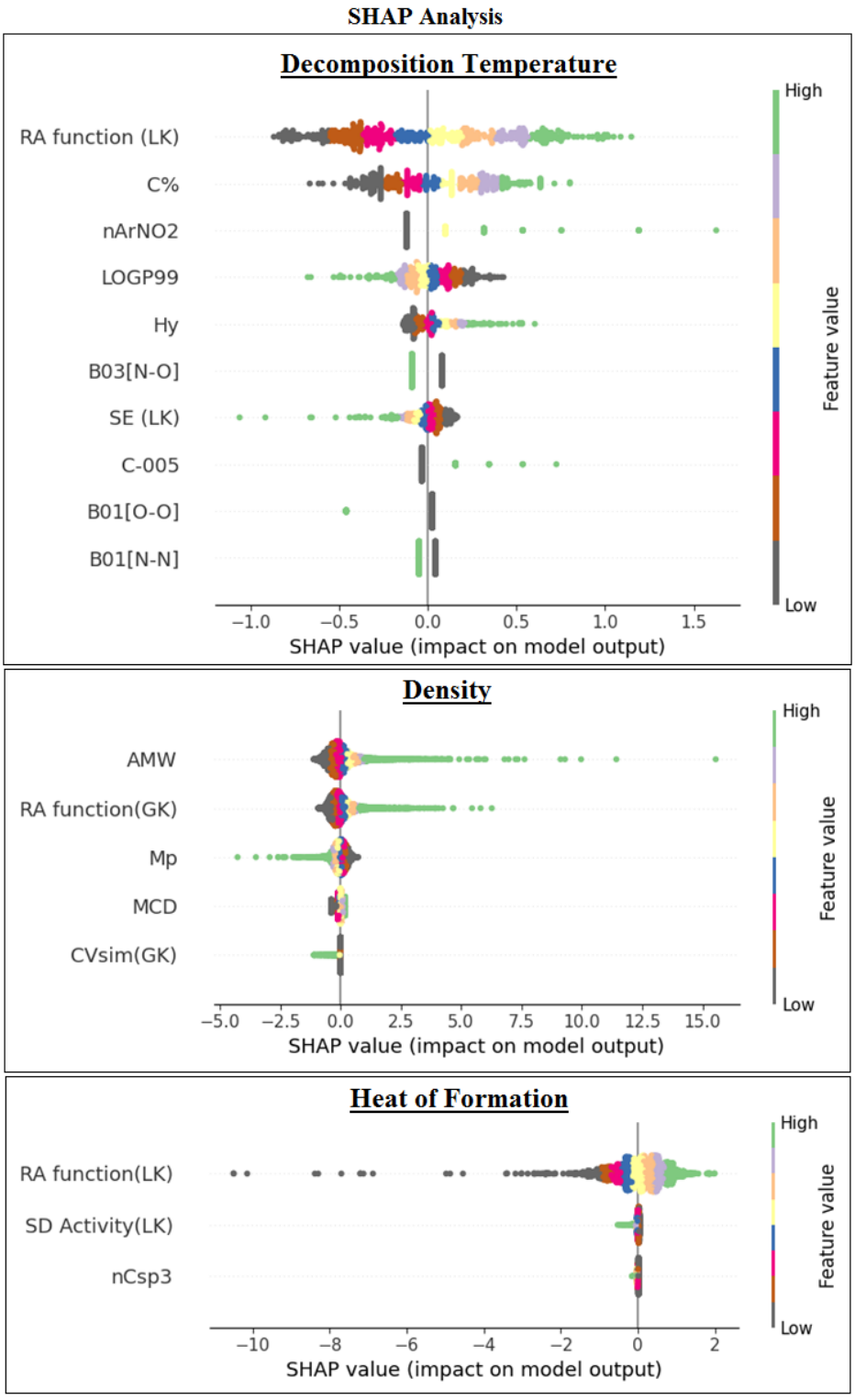

Figure 4.23: Determination of feature importance through the SHAP summary plots

Table 4.10: List of descriptors with their definition and contribution to the PLS q-RASPR models

Descriptor	Definition	Type	Model	Contribution
C%	Percentage of Carbon atom	Constitutional indices	T_{dec}	Positive (+ve)
B01[O-O]	Presence/absence of O-O at topological distance 1	2D atom pairs	T_{dec}	Negative (-ve)
B01[N-O]	Presence/absence of N-O at topological distance 3	2D atom pairs	T_{dec}	Negative (-ve)
Hy	Hydrophilic factor	Molecular property	T_{dec}	Positive (+ve)
LOGP99	Wildmann-Crippen octanol-water coefficient (LogP)	Molecular property	T_{dec}	Negative (-ve)
$nArNO_2$	Number of nitro (-NO2) groups (Aromatic)	Functional group count	T_{dec}	Positive (+ve)
C-005	CH3X	Atom centered fragment	T_{dec}	Positive (+ve)
B01[N-N]	Presence/absence of N-N at topological distance 1	2D atom pairs	T_{dec}	Negative (-ve)
AMW	Average molecular weight	Constitutional indices	Density	Positive (+ve)
Mp	Mean atomic polarizability (scaled on C-atom)	Constitutional indices	Density	Negative (-ve)
MCD	Molecular cyclized degree	Ring Descriptor	Density	Positive (+ve)

nCsp3	Number of sp3 hybridized C-atom	Constitutional indices	ΔH_f°	Negative (-ve)
RA function	A composite function derived from Read-Across	RASPR descriptor	T_{dec}, T_m, Density, ΔH_f°	Positive (+ve)
SE (LK)	Weighted standard error of the close source compounds' response values	RASPR descriptor	T_{dec}	Negative (-ve)
CVsim(GK)	Coefficient of variance of similarity values of close source compounds'	RASPR descriptor	Density	Negative (-ve)
SD_Activity (LK)	Weighted standard deviation of the close source compounds' observed response values	RASPR descriptor	ΔH_f°	Negative (-ve)

4.2.6 Descriptor Interpretation of the PLS q-RASPR models

The final PLS q-RASPR models for different properties of EMs have been presented in the form of mathematical equations in **Table 4.6**. In contrast, the descriptions of the descriptors with their contribution to the models are listed in **Table 4.10**. The descriptor influences on the properties with suitable examples are discussed below:

4.2.6.1 Interpretation of descriptors for the T_{dec} model

In the decomposition temperature (T_{dec}) model, the descriptors *RA function (LK)*, C%, $nArNO_2$, Hy, and C-005 are contributing positively to the decomposition temperature which means that any increase or decrease in the values of the descriptors mentioned above will result in the simultaneous increase or decrease, respectively, in the T_{dec} of the compounds. On the other hand, the descriptors B01[N-N], B01[O-O], B03[N-O], LOGP99, and *SE(LK)* have negative contributions to the T_{dec}. The positive contribution of the ***RA function (LK)*** can be represented by compound **452**(*RA function (LK*) = 673.168, T_{dec} = 608.15˚C), **151**(*RA function (LK*) = 587.517, T_{dec} = 573.15˚C), and **19**(*RA function (LK*) = 378.162, T_{dec} = 397.15˚C). The presence of 55.56% and 6.67% of carbon in compounds**187** (T_{dec} = 536.55˚C) and **78** (T_{dec} = 383.15˚C) confirms the positive contribution of the descriptor **C%**. The presence of 8 nitro groups in **300** (T_{dec} = 658.15˚C), 3 in **262** (T_{dec} = 587.15˚C), and none in **343** (T_{dec} = 526.65˚C) shows the positive contribution of the descriptor **$nArNO_2$** in the model. The hydrophilic factor **Hy**, contributes positively to the model which can be represented by the compound **113** (Hy = 6.992, T_{dec} = 511.15˚C) and **11** (Hy = -0.200, T_{dec} = 468.15˚C). The atom-centered fragment **C-005** represents the fragment CH_3X (where X is an electronegative atom, here oxygen). The positive contribution of CH_3X can be represented by the compound **536** (CH_3X = 3, T_{dec} = 655.15˚C) and **223** (CH_3X = 0, T_{dec} = 623.15˚C). The T_{dec} value of **180** is 620.95˚C, and it does not contain any **N-N, O-O,** and **N-O** bonds at the topological distances of 1, 1, and 3, respectively. But in compounds **51** (T_{dec} = 461.15˚C), **184**(T_{dec} = 471.15˚C), and **103** (T_{dec} = 381.15˚C) the presence of these bonds corresponds to a decrease in their T_{dec}. The negative contribution of LOGP99 can be presented by the compound **177** (LOGP99 = 7.830, T_{dec} = 359.15˚C) and **443** (LOGP99 = -0.882, T_{dec} = 503.65˚C). Also, the negative contribution of the RASPR descriptor *SE (LK)* can be described by the compound 364 (*SE (LK)* = 88.991, T_{dec} = 364.65˚C) and 277 (*SE (LK)* = 22.036, T_{dec} = 448.15˚C)

4.2.6.2 Interpretation of the RA function descriptor in the T_m model

The RASPR descriptor, *RA function(LK)*is the only descriptor in the univariate model for melting point. This RA-derived composite function contributes positively towards the property prediction. The positive contribution of *RA function(LK)* can be represented by compounds **19458** (T_m= 481˚C), **12637** (T_m = 360˚C), **17948** (T_m= 117.5˚C), and **16** (T_m = -100.67˚C) with their respective feature values 491.162, 328.358, 114.91, and -108.684.

4.2.6.3 Interpretation of descriptors for the density model

The density of a compound can be calculated as the ratio of molecular mass to its volume. The descriptor **AMW** in the developed model stands for the Average Molecular Weight of the compound and contributes positively to the prediction of the density. As we know density is directly correlated with the mass of the compound, as the AMW increases the density of the molecule also increases simultaneously. The compound **223** and **551**with molecular densities of 3.866 and 3.546, have an average molecular weight of 53.57 and 41.53, respectively. Again, compounds **12764** and **12765**, with densities of 1.027 and 1.03, have AMW of 4.88 and 4.89, respectively. The constitutional descriptor **Mp** represents the mean atomic polarizability (Scaled on C-atom) and contributes negatively to the model prediction. The polarizability is directly proportional to the compound's volume, which is indirectly related to the density. So, the increase in the polarizability indicates a decrease in the density of the compound. It can be easily illustrated by**337**with a mean polarizability value of 0.532, having a molecular density of 1.859 g/cm3, while 12351 has a molecular density of 1.696 g/cm^3 with only 0.852 Mp value. The descriptor **MCD** (Molecular Cyclized Degree) positively impacts the model predictivity. MCD represents the ratio of number of atoms present in the ring to the total number of atoms in the molecule. The cyclic molecules have a higher density due to the stronger London forces because the ring system allows for a larger area of contact. The density of **11446** is 1.254 g/cm^3 with a degree of cyclization of 0.857 whereas with 0.75 degree of cyclization, **8403** has a density of 1.171 g/cm^3. The RASPR descriptor, ***RA function (GK)*** is a composite descriptor derived from the Read-Across and is contributing positively to the prediction of density. It can be seen in **223, 7347,8127,** and **12773** having descriptor values of 3.546, 1.715, 1.268, and 1.024 corresponding to their densities in the order of 3.866, 1.764, 1.325, and 1.041, respectively. ***CVsim (GK)*** indicates the coefficient of variance of the similarity values of the close source compounds and shows a negative contribution

in the model. When the variation between the similarity values increases among the close training compounds, it indicates that the prediction is not so reliable for the test set compound. The compounds **9129** (*CVsim (LK)* = 0.005, d = 1.323 g/cm^3) and **1335** (*CVsim (LK)* = 3.162, d = 1.184 g/cm^3) verify the negative contribution of *CVsim (LK).*

4.2.6.4 Interpretation of descriptors for the ΔH_f° model

In the ΔH_f° model, the descriptor *RA function LK)* contributes positively to the model. The compounds**849, 569,** and **102** with the descriptor value of 693.341, 407.732, and -4455.65 have their enthalpy of formation 681.4 kJ/mol, 364 kJ/mol, and -4806.4 kJ/mol respectively. Another RASPR descriptor *SD_Activity (LK)* has a negative contribution to the model. The compounds **120** (*SD_Activity (LK)* = 876.004, ΔH_f° = -1551 kJ/mol), **2353** (*SD_Activity (LK)* = 62.293, ΔH_f° = -272 kJ/mol), and **1825** (*SD_Activity (LK)* = 6.991, ΔH_f° = -227.4kJ/mol) confirms that the increase in the weighted standard deviation of close source compounds response values results in the decrease in the amount of ΔH_f°. The descriptor nCsp3 represents the number of sp3 hybridized C-atom in the molecule and represents a negative contribution to the model. The ΔH_f° of compound **279** (nCsp3 = 0, ΔH_f° = 147.45 kJ/mol) and **280**(nCsp3 = 6, ΔH_f° = -48.9 kJ/mol) shows that the hydrogenation in the later compound increases the number of sp3 hybridized carbon from 0 to 6 which leads to decrease in the value of ΔH_f° of the molecules.

4.2.7 Comparison of the quality of q-RASPR models with QSPR models

4.2.7.1 Comparison with our QSPR models

We have compared the q-RASPR models with our own developed QSPR models for all 4 properties. The validation metrics for all the developed models are shown in **Table 4.4** (QSPR model) and **Table 4.6** (q-RASPR model). The comparative results depict that the prediction quality has been enhanced for all the q-RASPR models when compared to their corresponding QSPR models. The number of descriptors in the q-RASPR models was also lower than the descriptors present in the QSPR models which shows that with a lower number of regressors (except in the case of decomposition temperature), our q-RASPR models can efficiently predict the compounds having identical chemical information.

4.2.7.2 Comparison with the previous models

The process of performing curation is most important to obtain a noise-free data set, to develop a relevant model with a high degree of acceptance. While performing curation on the obtained data set, we have found that the data set used by the authors (Wespiser and Mathieu, 2023) contains several duplicate compounds and mixtures as well. Previously, the authors (Wespiser and Mathieu, 2023) prepared two QSPR models for the T_{dec} and T_m data sets, and two semi-empirical additivity scheme models for the density and ΔH_f° data sets. Apart from this, they developed deep-learning models using the MPNN (Message Passing Neural Network) algorithm for all the data sets. The validation metrics of the training sets were not reported by the authors and at the same time, the feature selection process or the final features in the developed models were also not reported. Also, for the T_{dec} and ΔH_f° data sets, only the external test set results were reported.

For easy interpretability and reproducibility of our developed models, we have mentioned the descriptors (both the number and types) of our QSPR as well as of q-RASPR models (**Table 4.10)**. This information can be used for the prediction of properties of newly developed compounds or compounds whose properties are not known yet using our models. Wespiser et. al. did not mention the descriptor number and type for the models, which challenges the reproducibility of their developed models.

A comparison of the results for the test set prediction quality of our QSPR and q-RASPR models with the previously developed QSPR and MPNN models is presented in **Table 4.11**. We can state that our T_{dec} q-RASPR model reports a lower $RMSE_P$ error compared to the QSPR and MPNN models developed previously. The q-RASPR model for T_m shows a good predictive quality with only a single descriptor [i.e. *RA function (LK)*] for a very large data set. Although the prediction quality of our q-RASPR model does not exceed the previous QSPR and/or MPNN models, a model with a single descriptor with this much accuracy for a large data set is quite remarkable. Comparing the results for the density data set, we infer that with only 5 descriptors in the final model, the model shows a very minute difference in the error estimation both with respect to MAE and RMSE. Also, the quality and prediction of our PLS q-RASPR model for ΔH_f° was almost similar to the MPNN DL model.

Therefore, we can infer that, with much less model complexity, our q-RASPR models with few features can efficiently predict the enlisted properties, and the developed models are also easily reproducible.

Table 4.11: Comparison of our q-RASPR models with our own QSPR models and previously developed models

Property	Models	No. of descriptors	R^2	MAE_P	$RMSE_P$
T_{dec}	QSPR (Wespiser and Mathieu, 2023)	Not defined	0.82	39	53.6
	MPNN (Wespiser and Mathieu, 2023)	Not defined	0.83	40	53
	QSPR (our work)	10	0.621	44.919	54.814
	q-RASPR (our work)	10	0.676	41.383	50.683
T_m	QSPR (Wespiser and Mathieu, 2023)	Not defined	0.93	25.2	35.8
	MPNN (Wespiser and Mathieu, 2023)	Not defined	0.95	20.2	30.1
	QSPR (our work)	29	0.67	39.626	52.501
	q-RASPR (our work)	1	0.741	34.3	46.52
Density	QSPR (Wespiser and Mathieu, 2023)	Not defined	0.98	0.031	0.040
	MPNN (Wespiser and Mathieu, 2023)	Not defined	0.98	0.034	0.046
	QSPR (our work)	6	0.928	0.037	0.051
	q-RASPR (our work)	5	0.939	0.035	0.047
ΔH_f°	QSPR (Wespiser and Mathieu, 2023)	Not defined	0.972	23.4	30.8
	MPNN (Wespiser and Mathieu, 2023)	Not defined	0.94	47.9	67.4
	QSPR (our work)	11	0.932	47.903	67.412
	q-RASPR (our work)	3	0.931	47.158	67.63

4.3 Study 3: Predictive cheminformatics modeling of reorganization energy (RE) for p-type organic semiconductors: Integration of quantitative read-across structure-property relationship (q-RASPR) and stacking regression analysis

4.3.1 QSPR modeling

The feature selection process was applied to the training set with 129 compounds. A pool of 28 significant descriptors was prepared through step-wise and GA feature selection algorithms. The same pool of descriptors was then subjected to a grid search, and a 9 descriptor MLR model was selected based on the cross-validated (Q^2_{LOO}) result. Finally, the same descriptor combination was used to construct a PLS regression model with 7 LVs. The PLS equation (**Equation 4.3)** and the validation metrics are mentioned below:

$$LogRE = -18.051 - 0.453 \times RCI - 57.7 \times Eta_{B_A} + 48.316 \times Eta_{epsi_3} + 3.105 \times Eta_{D_{epsiB}} - 0.018 \times (nCb-) + 0.063 \times H - 046 + 0.0667 \times MaxaasC + 0.055 \times B03[S-S] - 0.029 \times F06[S-S] \quad (4.3)$$

$$N_{train} = 129, N_{test} = 42, Descriptors = 9, LVs = 7$$

$$R^2 = 0.731, Q^2_{LOO} = 0.688, MAE_C = 0.078, RMSE_C = 0.099$$

$$Q^2_{F1} = 0.741, Q^2_{F2} = 0.741, MAE_P = 0.075, RMSE_P = 0.095$$

4.3.2 Similarity predictions

The descriptors of the PLS QSPR model were used to perform RA-based similarity predictions of the query set compounds. The predictions for each compound were made using Euclidean distance, Gaussian kernel, and Laplacian kernel-based similarity of the query compound with its close source compounds. Following the optimization of RA hyperparameters for different similarity measures, we have obtained the values for σ be 2 for the Gaussian kernel, γ be 2 for the Laplacian kernel, and the number of CTC be 3. These hyperparameters were used to compute the similarity predictions of the query set compounds for each similarity measure, and the results are shown in **Table 4.12.** The results for the Laplacian kernel-based similarity were found to be superior to the other similarity parameters.

Table 4.12: Results for the RA-based similarity predictions

Validation Metrics↓	Similarity measures		
	Euclidean distance (ED)	**Gaussian kernel (GK)**	**Laplacian kernel (LK)**
Q^2_{F1}	0.640521	0.637524	**0.669772**
Q^2_{F2}	0.640302	0.637304	**0.66957**
$RMSE_P$	0.112152	0.112619	**0.107493**
MAE_P	0.088217	0.088818	**0.086095**

4.3.3 q-RASPR modeling

The q-RASPR model development aims to incorporate the advantages of both QSPR and RA-based similarity. The q-RASPR descriptor matrix was prepared by combining the structural and physiochemical features with the RASPR descriptors. The newly prepared descriptor matrix of the training set was further used for the variable selection process to select the significant features through a grid search. Based on the cross-validation (Q^2_{LOO}) results, we have selected three MLR models (one for each similarity measure) with 7-descriptors. The same descriptor combination was then used to generate the PLS regression models with the least number of LVs optimized using LOO Q^2. The PLS equations of the models for each similarity function are given in **Table 4.13** along with their validation metrics.

Table 4.13: Model equations and metrics for the PLS q-RASPR models

Model	PLS Equation	Validation metrics	
		Training set	Test set
q-RASPR (ED)	$LogRE = -20.855 - 0.442 \times RCI - 50.463 \times Eta\_B\_A + 53.217 \times Eta\_Epsi\_3 - 0.021 \times (nCb -) + 0.051 \times H - 046 - 0.022 \times F06[S - S] + 0.364 \times RA\ function(ED)$ Descriptors=7, LVs=6	$n_{training} = 129$ $R^2 = 0.707$ $Q^2_{LOO} = 0.668$ $MAE_C = 0.083$ $RMSE_C = 0.104$	$n_{test} = 42$ $Q^2_{F1} = 0.750$ $Q^2_{F2} = 0.750$ $MAE_P = 0.073$ $RMSE_P = 0.094$
q-RASPR (GK)	$LogRE = -20.492 - 0.428 \times RCI - 49.969 \times Eta\_B\_A + 52.316 \times Eta\_Epsi\_3 - 0.021 \times (nCb -) + 0.049 \times H - 046 - 0.022 \times F06[S - S] + 0.371 \times RA\ function(GK)$ Descriptors=7, LVs=6	$n_{training} = 129$ $R^2 = 0.707$ $Q^2_{LOO} = 0.667$ $MAE_C = 0.083$ $RMSE_C = 0.104$	$n_{test} = 42$ $Q^2_{F1} = 0.748$ $Q^2_{F2} = 0.748$ $MAE_P = 0.074$ $RMSE_P = 0.094$
q-RASPR (LK)	$LogRE = -23.544 - 0.587 \times RCI - 54.149 \times Eta\_B\_A + 60.285 \times Eta\_Epsi\_3 - 0.023 \times (nCb -) + 0.058 \times H - 046 - 0.027 \times F06[S - S] + 0.263 \times RA\ function(LK)$ Descriptors=7, LVs=6	$n_{training} = 129$ $R^2 = 0.706$ $Q^2_{LOO} = 0.666$ $MAE_C = 0.083$ $RMSE_C = 0.104$	$n_{test} = 42$ $Q^2_{F1} = 0.753$ $Q^2_{F2} = 0.753$ $MAE_P = 0.073$ $RMSE_P = 0.093$

The results of these models suggest that the prediction quality of the q-RASPR model is better than the previously developed QSPR model. Also, the q-RASPR models contains a lower number of variables compared to the number of descriptors in the QSPR model.

4.3.4 Predictions through stacking regressor

Soon after the development of q-RASPR models using different similarity approaches, the predictions from the individual models were used for stacking. Using the predictions as the

descriptors, we have calculated the RE (in logarithmic terms). The stacking regression was performed using the PLS algorithm and the number of LVs was optimized based on the leave-one-out squared correlation coefficient (Q^2_{LOO}). The validation metrics for the stacked-PLS model are given in **Table 4.14**:

Table 4.14: Statistical results of the stacked-PLS q-RASPR model

Training set	Test set
$n_{training} = 129$	$n_{test} = 42$
$R^2 = 0.708$	$Q^2_{F1} = 0.753$
$Q^2_{LOO} = 0.698$	$Q^2_{F2} = 0.753$
$MAE_C = 0.083$	$MAE_P = 0.073$
$RMSE_C = 0.104$	$RMSE_P = 0.093$

The scatter plot (**Figure 4.24**) represents the correlation between the observed and predicted RE of the molecules in the dataset for the stacking PLS q-RASPR model.

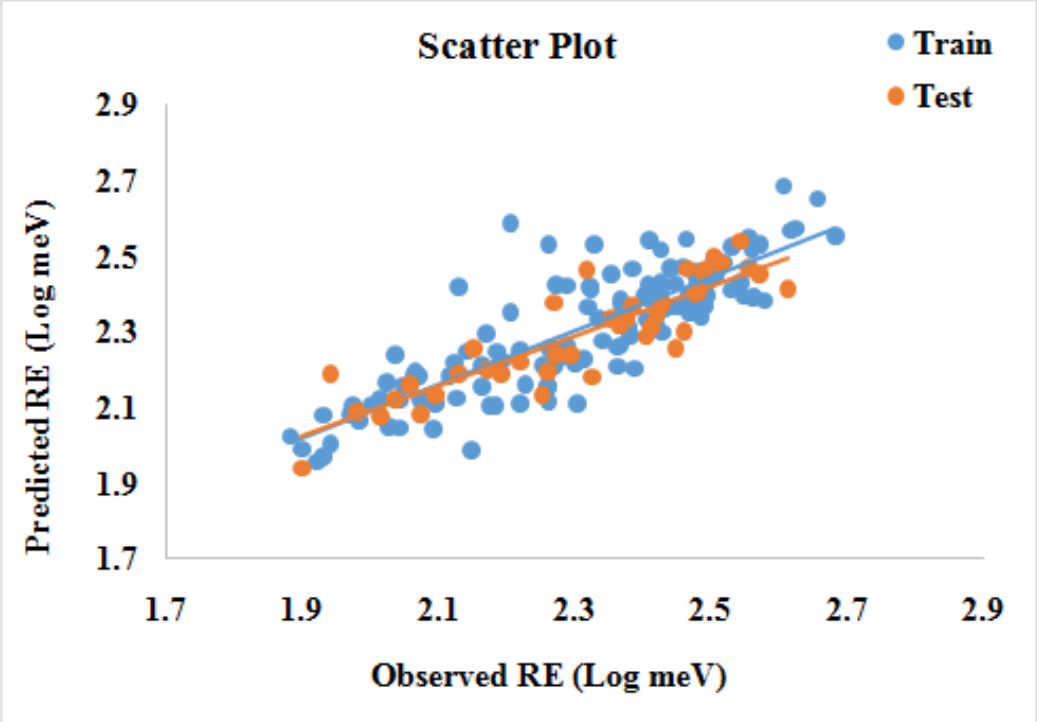

Figure 4.24: Scatter plot for the Stacking PLS q-RASPR model

4.3.5 Interpretation of the PLS plots

The final stacking PLS model was developed using the predictions of 3 PLS q-RASPR models with different similarity measures. We have analyzed the PLS plots for each PLS q-RASPR model, and the following conclusions were drawn:

i. The variable importance (VIP) scores of the descriptors in all the models (**Figure 4.25**) signifies that the RASPR descriptor *RA_function* is the most important descriptor followed by nCb-, RCI, eta_epsi_3, eta_B_A, and H-046 while the descriptor F06[S-S] was of the least importance.

ii. The loading plots (**Figure 4.26**) signify that the descriptors that are dispersed more away from the origin have more impact on the property. In all the 3 PLS q-RASPR models, the X-variables (descriptors) dispersion is almost similar w.r.t the Y-variable (property).

iii. The coefficient plots (**Figure 4.27**) represent the standardized regression coefficients of the descriptors and their respective contribution (+ve/-ve) to the models.

iv. In the score plots (**Figure 4.28**), compounds **1**, **119**, and **143** were found to be outliers for all the 3 models constructed using the first 2 PLS components. The AD study performed using the DModX approach (**Figure 4.29 and 4.30**) shows that only one compound (Compound **1**) is present out of the AD and is present in the training set. This can be because of the fact that compound **1** is the only monocyclic compound in the dataset whereas all other compounds consist of 2 or more rings. In the test set, all the compounds were present within the AD of the respective models.

v. We have also performed the Y-Randomization test (**Figure 4.31**) to check whether our model has any chance correlation or not.

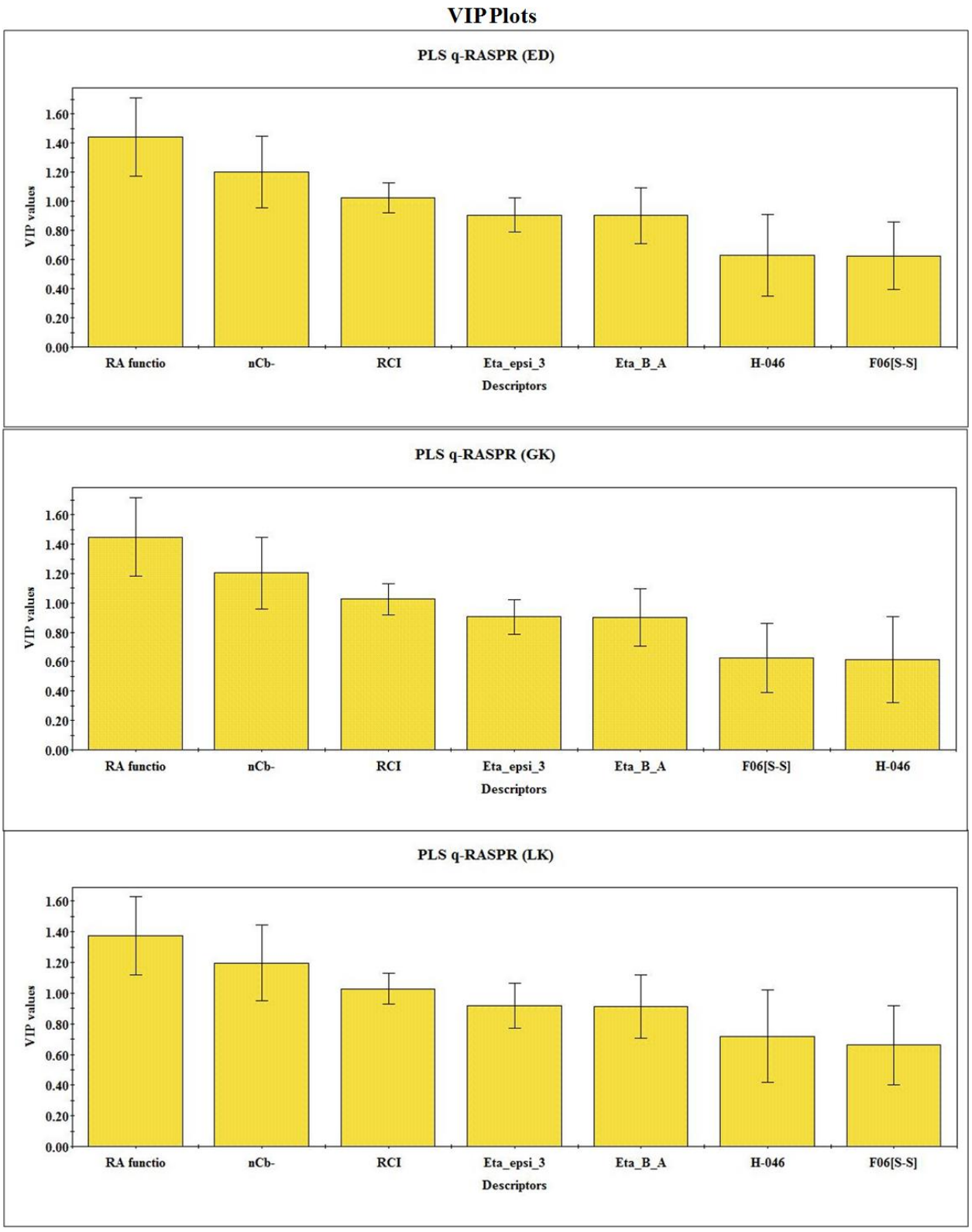

Figure 4.25: VIP plots for the individual PLS q-RASPR models

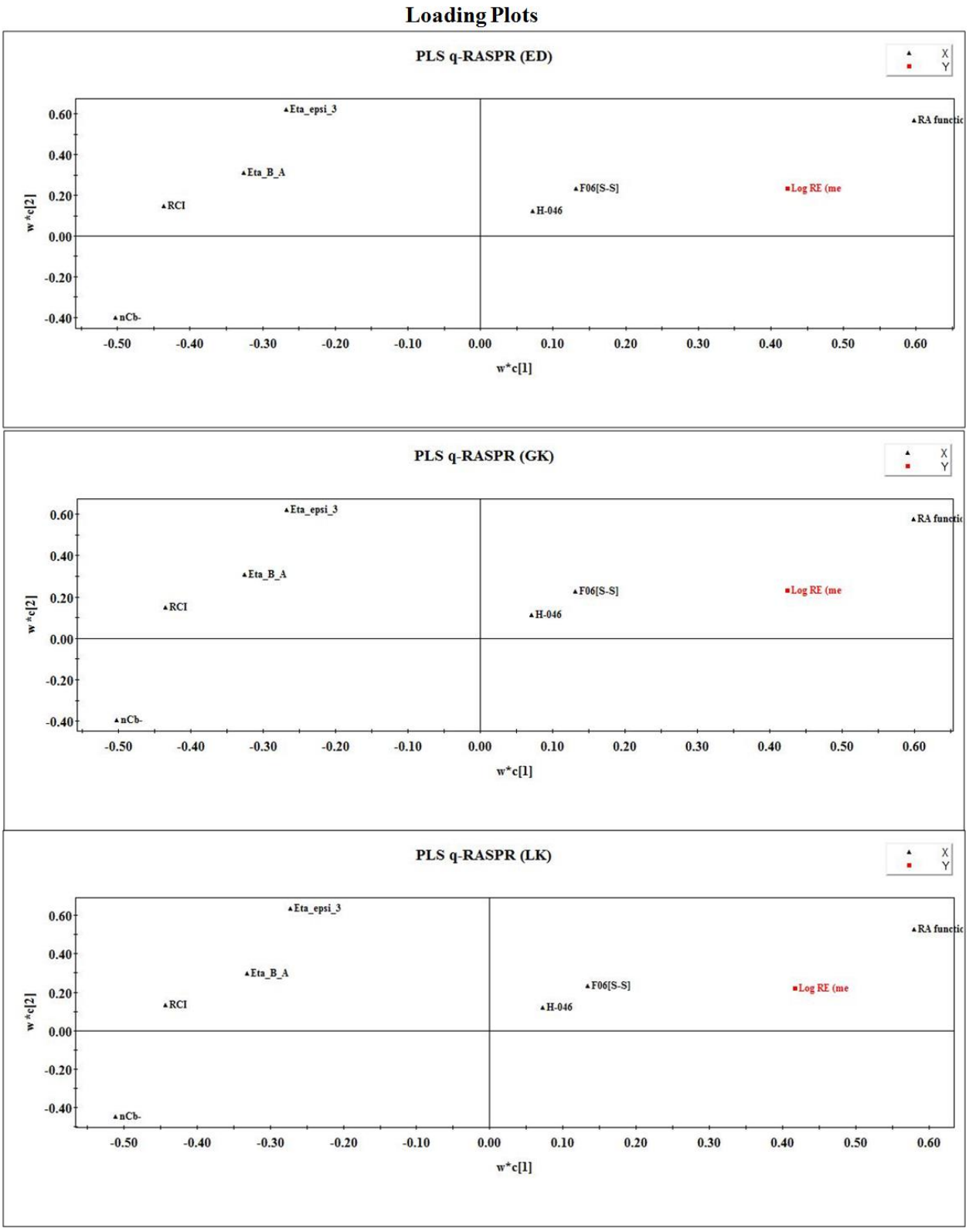

Figure 4.26: Loading plots for the individual PLS q-RASPR models

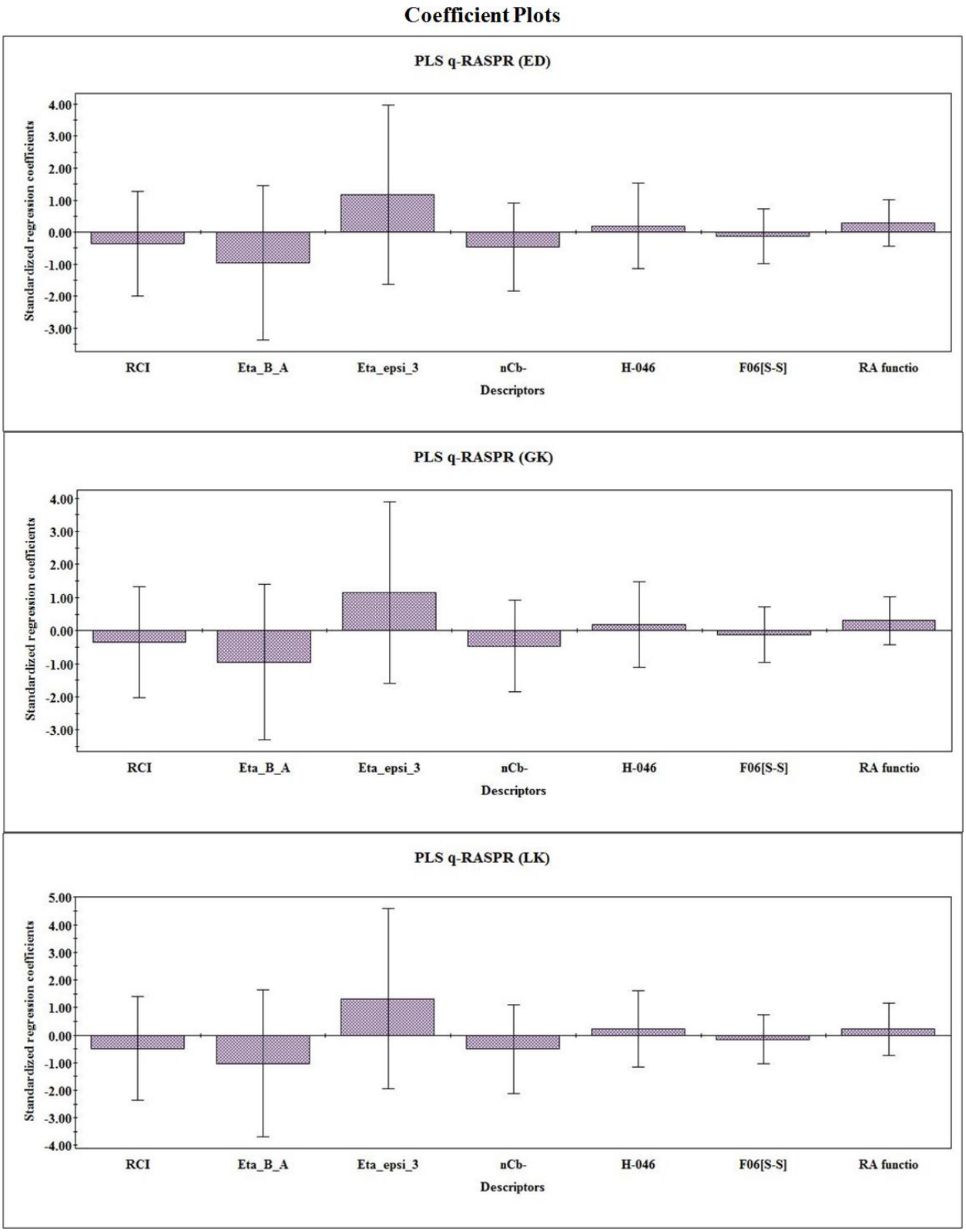

Figure 4.27: Coefficient plots for the individual PLS q-RASPR models

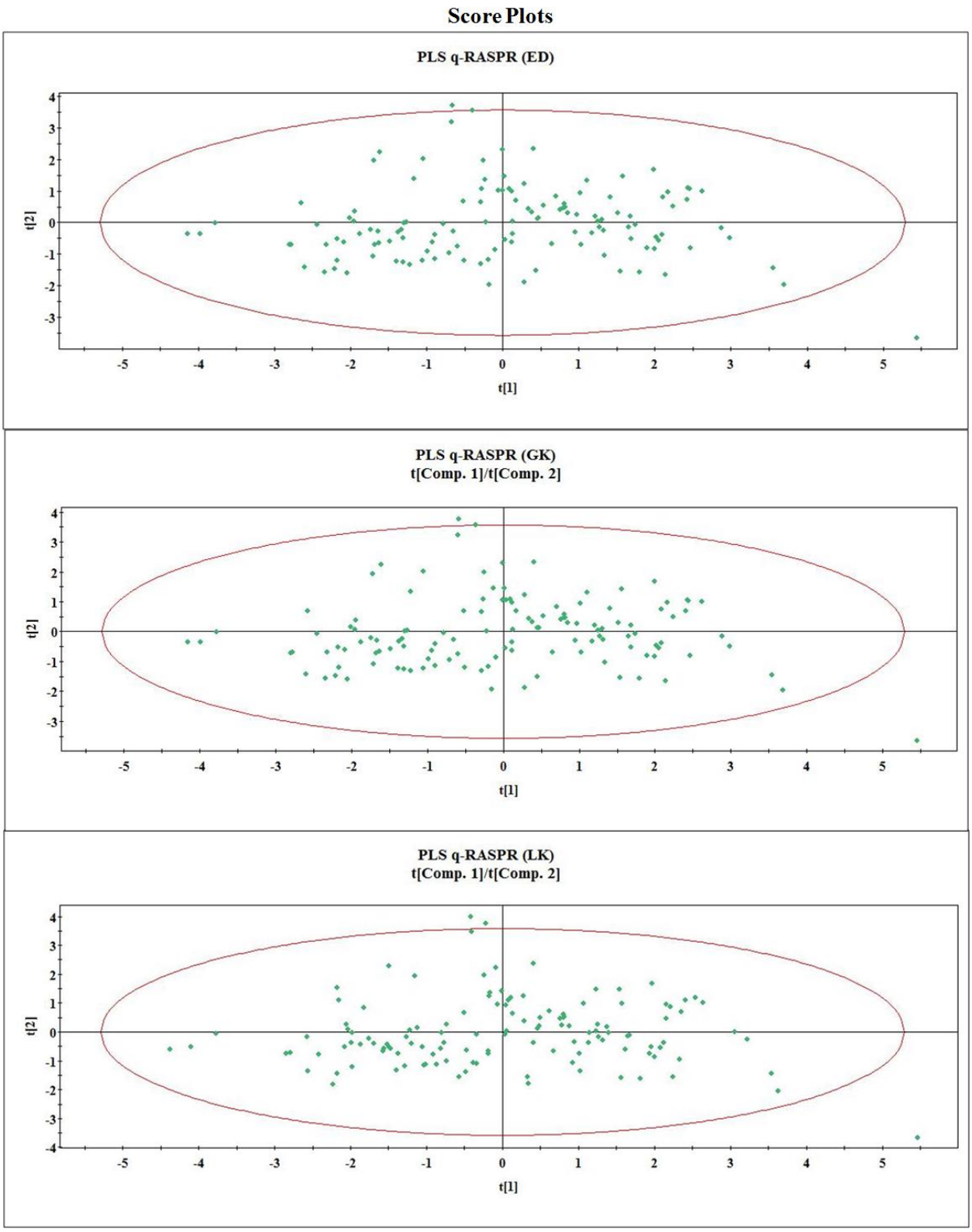

Figure 4.28: Score plots for the individual PLS q-RASPR models

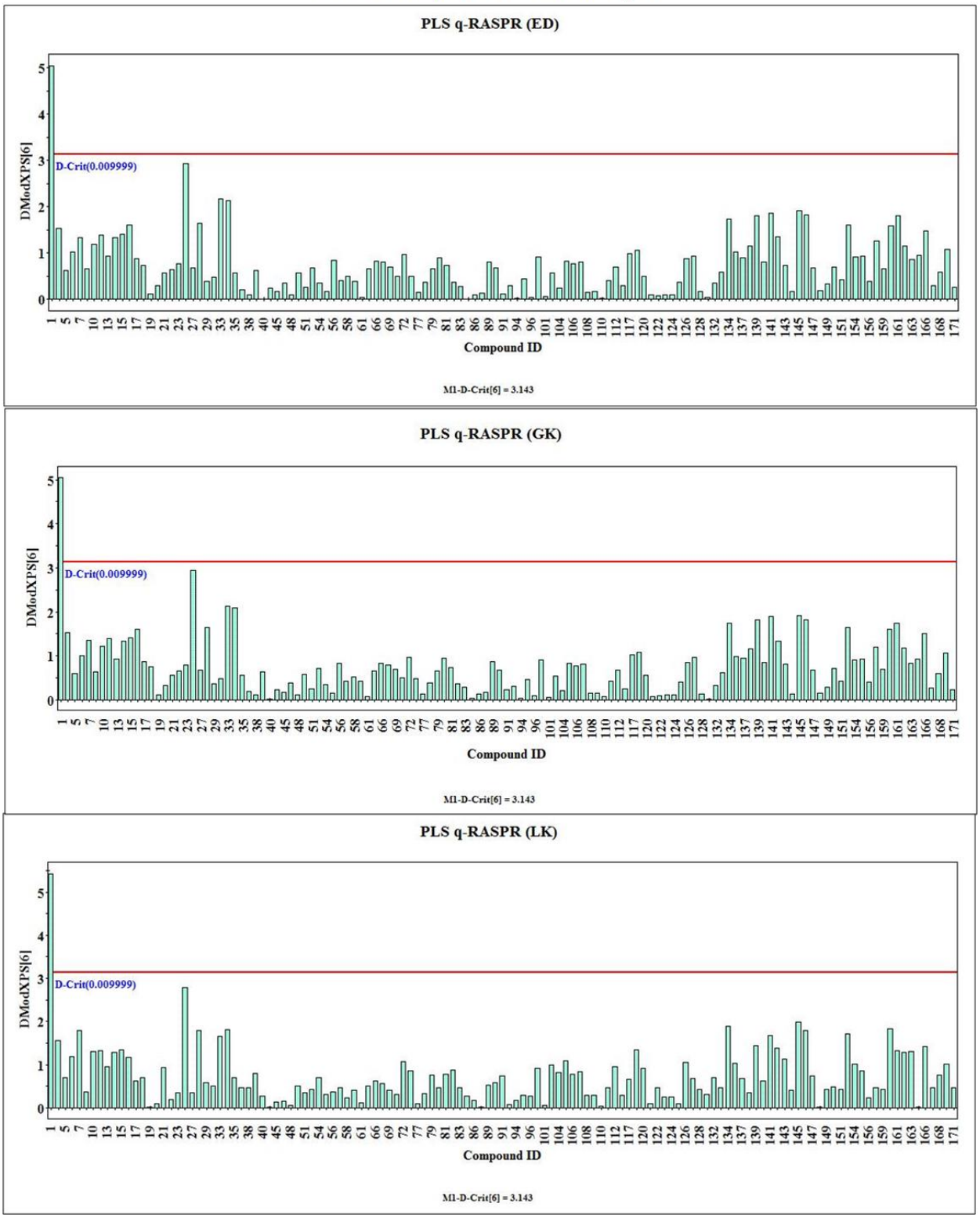

Figure 4.29: DModX-AD plots for the training set of individual PLS q-RASPR models

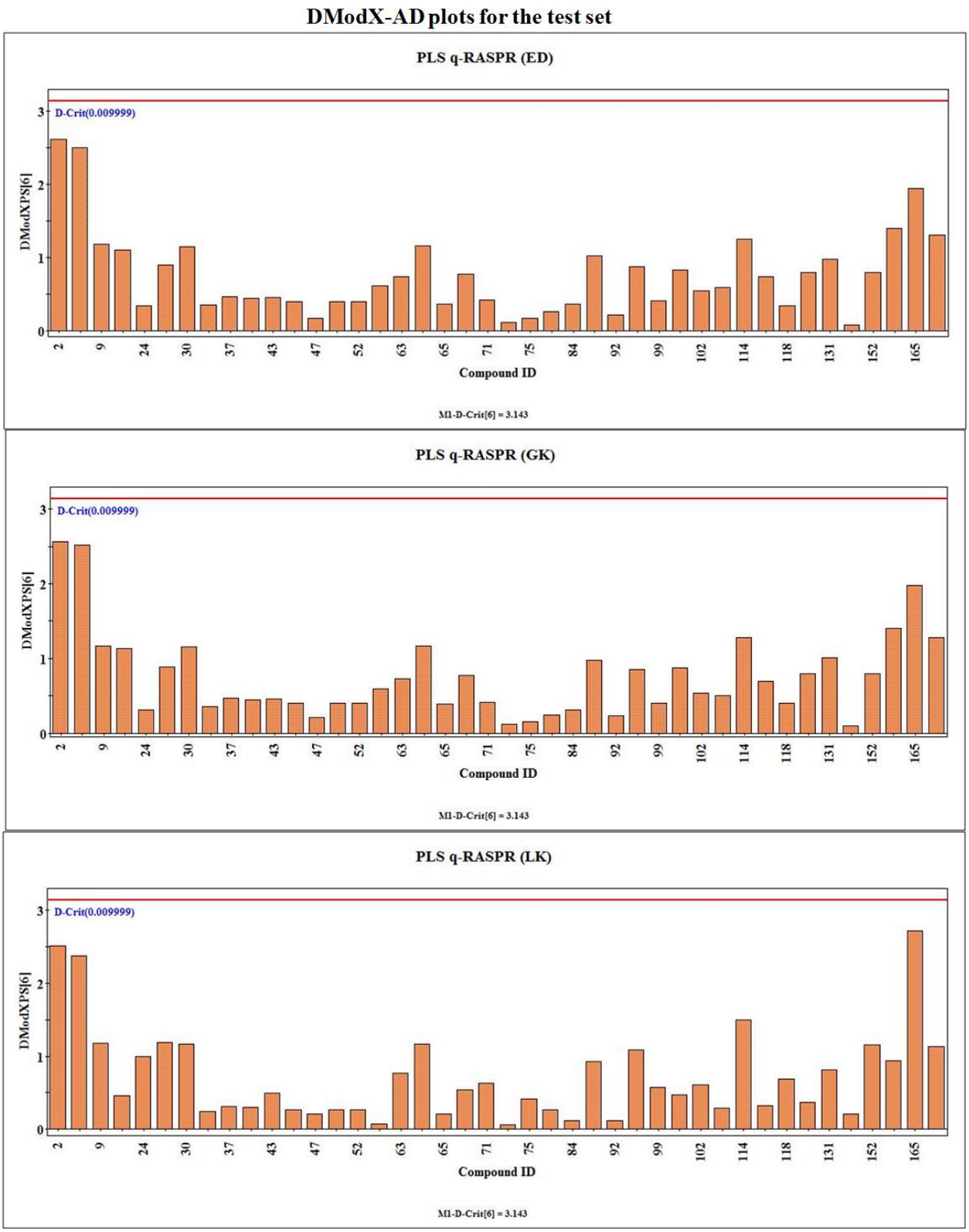

Figure 4.30: DModX-AD plots for the test set of individual PLS q-RASPR models

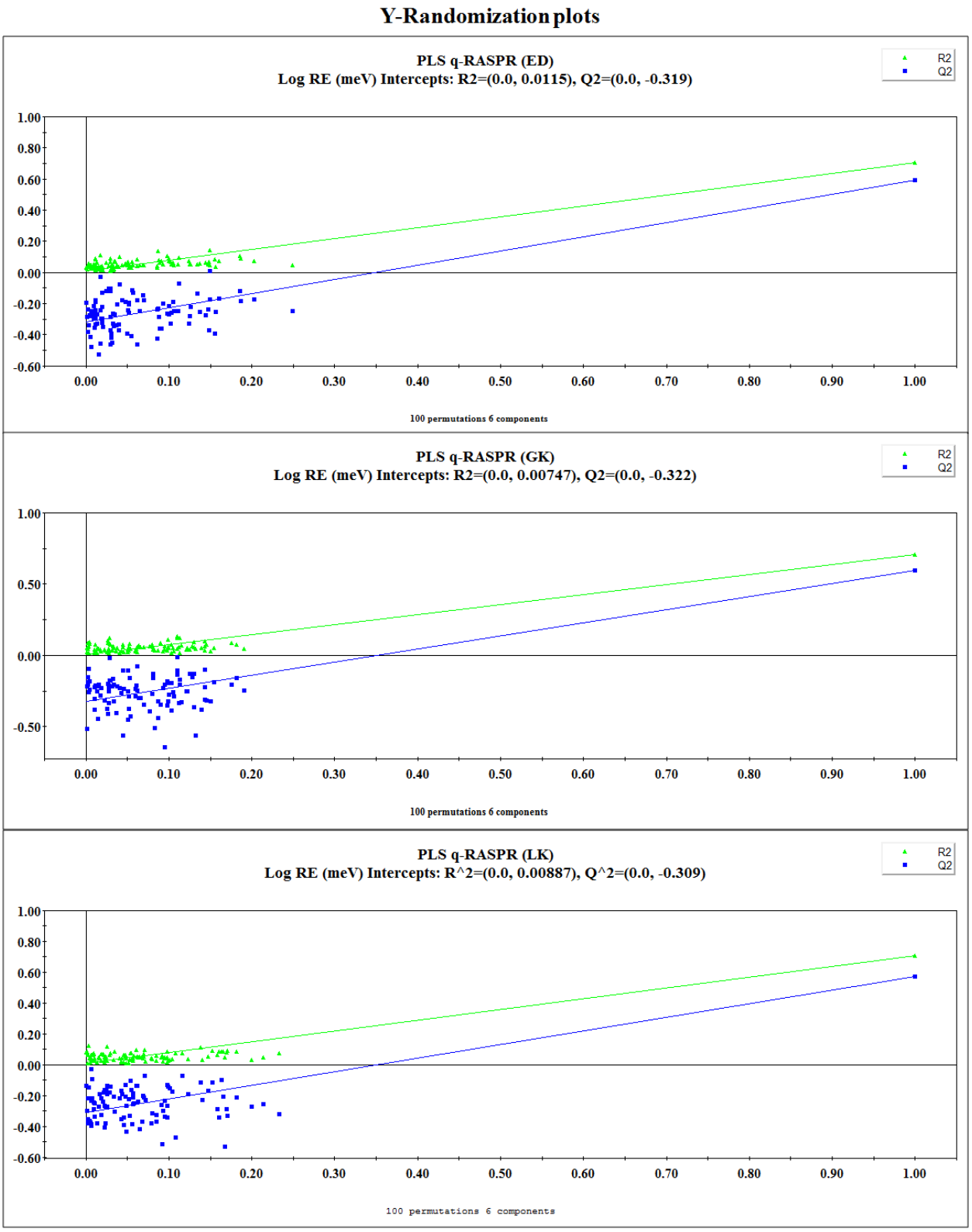

Figure 4.31: Y-randomization plots for the individual PLS q-RASPR models

4.3.6 Interpretation of the modeled features

Stacking was performed using the predictions of the different models that were developed using different structural and physiochemical features, and similarity measures. Here, in this section, we will discuss the contribution of different features influencing the RE of organic semiconductors. Among the descriptor set of the PLS q-RASPR models, the descriptors RCI, Eta_B_A, nCb-, and F06[S-S] contribute negatively to the prediction of RE whereas the descriptors Eta_epsi_3, H-046, and *RA_function* contribute positively. The detailed information on the modeled descriptors is given in **Table 4.15**.

Table 4.15: List of descriptors of the q-RASPR models

Descriptor	Description	Contribution
RA_function	RA-derived composite function	+ve
nCb-	Number of substituted benzene C (sp^2)	-ve
RCI	Ring complexity index	-ve
Eta_epsi_3	Eta electronegativity measure 3	+ve
Eta_B_A	Eta average branching index	-ve
H-046	H attached to C0 (sp^3), no X attached to next C	+ve
F06[S-S]	Frequency of S-S at topological distance 6	-ve

The RASPR descriptor ***RA_function*** is an RA-derived composite function that contains information of all the other structural and physiochemical features. This descriptor contributes positively to the prediction of the RE,as can be seen in molecule **67** (*RA_function* = 2.110, RE = 193 meV) and **69** (*RA_function* = 2.047, RE = 79 meV).

The descriptor **RCI** represents the Ring Complexity Index of the molecule. The OSCs constitute of conjugated π-systems and the electronic structures of these OSCs are affected due to the size and complexity of these conjugated π-systems. The longer conjugated systems provide a larger

surface area for the delocalization of e^-, thus reducing the energy required for electronic reorganization (Salaneck et. al., 2001). The negative impact of RCI can be seen in compounds**149** (RCI = 1.611, RE = 123 meV), **111** (RCI = 1.461, RE = 179 meV), and **97** (RCI = 1, RE = 288 meV) (see **Figure 4.32**).

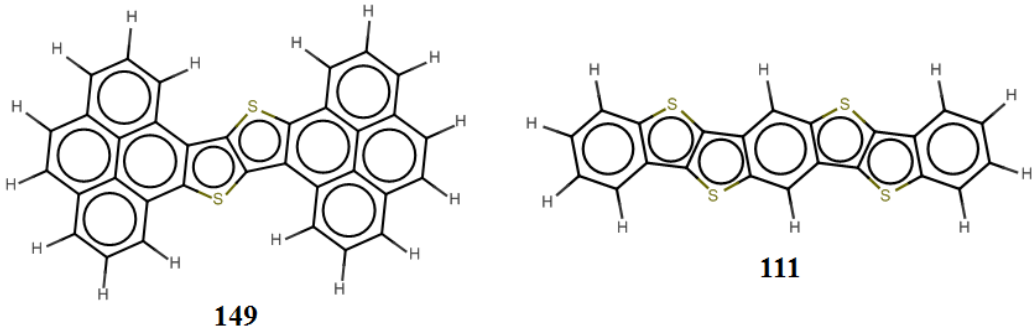

Figure 4.32: Compounds representing ring complexity

The substitution of benzene carbon represented by the descriptor **nCb**-has a negative contribution to the model predictivity. In the dataset, the molecules with benzene substitution show the fusion of the benzene ring with another conjugated ring system (i.e. thiophene) which further enhances the complexity of the molecule which results in lowering their RE,e.g., compound **50** (nCb- = 6, RE = 117 meV) and **51** (nCb- = 8, RE = 153 meV) (see **Figure 4.33**). The negative contribution of nCb- also supports/validates the contribution of the RCI descriptor as both these descriptors reflect the molecular complexity due to the increased π-conjugated system.

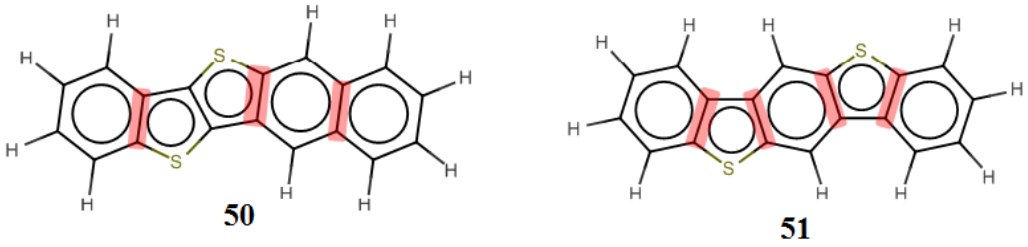

Figure 4.33: Compounds showing substitution of benzene carbon (highlighted)

The atom-centered fragment descriptor **H-046** shows the presence of attached H-atom to a sp^3 hybridized C-atom. The positive contribution of H-046 is represented in molecules**139** (H-046 = 4, RE = 300 meV), **161** (H-046 = 4, RE = 275 meV), and **60** (H-046 = 2, RE = 320 meV) (see

Figure 4.34). Due to the presence of sp^3 carbon between the 2 benzene rings, there is a discontinuity of conjugation between the rings which results in an increase in RE. Again, in molecule **86** (H-046 = 0, RE = 103 meV), no such hydrogen atom was there, so the π-conjugation is maintained throughout the molecule exhibiting a lower RE.

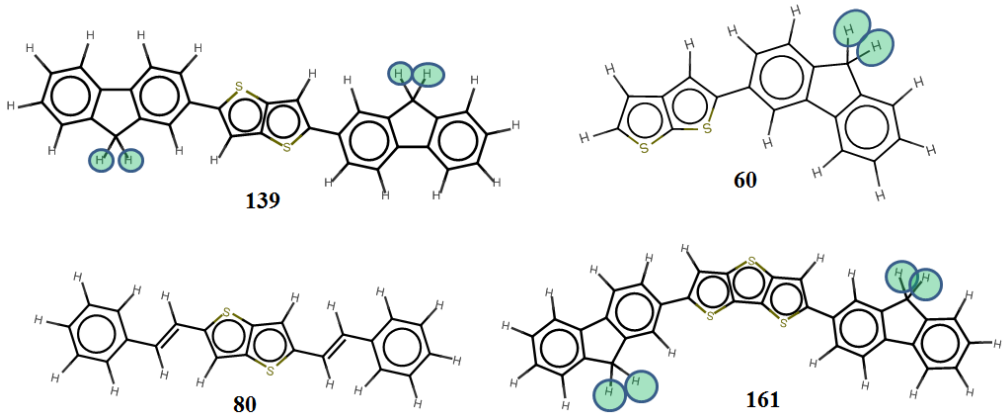

Figure 4.34: Compounds representing hydrogen substitution at C-atom (sp^3)

The 2D atom pair descriptor **F06[S-S]** which represents the frequency of S-S at the topological distance 6 contributes negatively to the model predictions. In both compounds**119** (RE = 210 meV) and **120** (RE = 280 meV), 7 thiophene rings are present but compound **119** has 4 F06[S-S] and compound **120**has 3 F06[S-S] atom pairs (see **Figure 4.35**). This shows that the arrangement of the thiophene ring within the molecule is an essential feature governing the RE. Since the sulfur atom in the thiophene rings lowers the HUMO-LOMO gap because of its electron-donating nature and presence of π-conjugation, it tends to lower the RE of the semiconductors (Mamada and Yamashita, 2015).

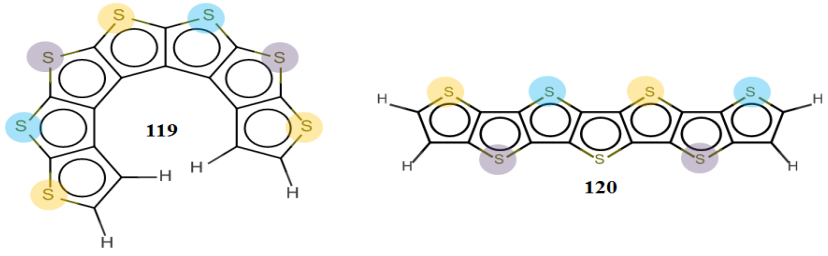

Figure 4.35: Compounds highlighting S-S pair at 6 topological distance. Sulfur highlighted with similar colours are paired with each other.

The descriptor **Eta_epsi_3** represents the ETA electronegativity measure of 3. This positive contribution of this descriptor can be visualized from compound **143** (RE = 268 meV, Eta_epsi_3 = 0.463) and **153** (RE = 258 meV, Eta_epsi_3 = 0.45) (see **Figure 4.36**). In compound **143**, the steric hindrance is more due to the presence of fused thiophene rings whereas in compound **153**, where each thiophene ring is separated by a single bond lowers the steric hindrance of the molecule thus reducing the RE.

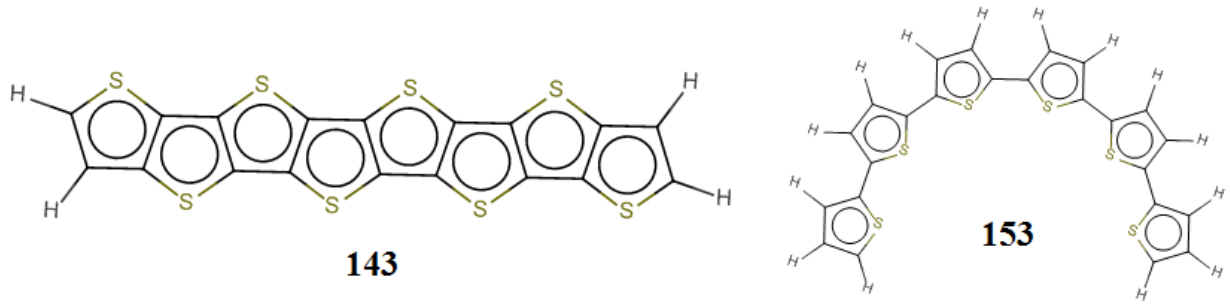

Figure 4.36: Compounds representing steric hindrance in the molecule

The descriptor **Eta_B_A** which shows the ETA average branching index (here fusion pattern) in the molecule has a negative impact on the model predictivity, and the same can be represented by the compound **125** (RE = 85 meV, Eta_B_A = 0.025), **122** (RE = 186 meV, Eta_B_A = 0.022), and **8** (RE = 230 meV, Eta_B_A = 0.019) (see **Figure 4.37**).

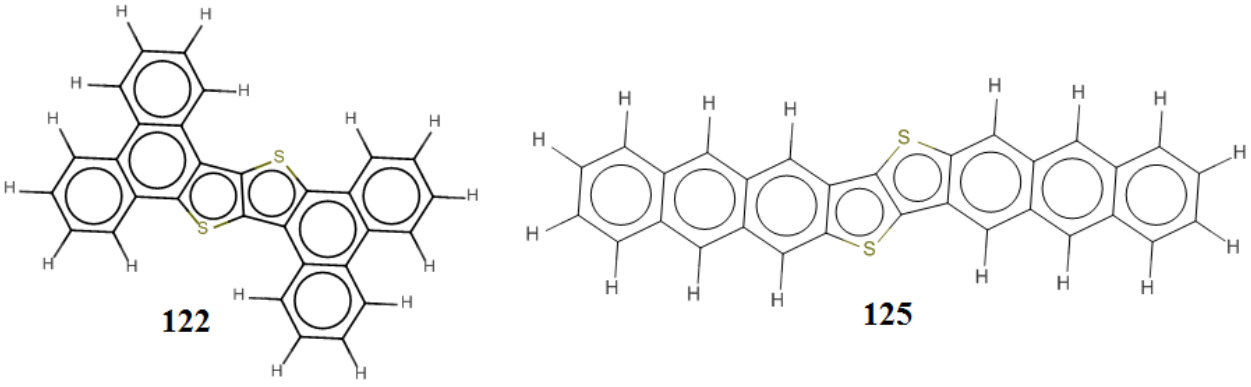

Figure 4.37: Compounds representing ETA_B_A indices

4.3.7 Predictions through different ML algorithms

We have also applied various ML algorithms to perform stacking regression. This was done so to improve the model's quality and predictivity. Before applying the ML algorithms, the data of the training set and the test set are needed to be scaled. Scaling of descriptors and response values was performed using the Java-based tool Scale1.0available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The machine learning models were developed using the default hyperparameters while performing the regression algorithms. The statistical values representing the model quality and predictivity of the individual ML models are given in **Table 4.16**, from which we can conclude that the tree-based methods such as RF, AB, GB, and XGB showed an excellent model quality in terms of R^2. Still, their respective LOO-CV results confirm that these models are not statistically robust. The results for the RR and PLS models are statistically similar but their predictivity is somewhat compromised when compared to the other models like SVM and LSVM. So, based on the cross-validated results and the predictive power of the models, the model developed using the stacking support vector regression (i,e. SVM) was selected as the best-performing model. This SVM model is of good statistical quality, robust, and highly predictive.

Table 4.16: ML prediction results of the models developed using different algorithms

Stacking regressor models	Validation Metrics							
	Training set				Test set			
	R^2	Q^2_{LOO}	MAE_C^*	$RMSE_C^*$	Q^2_{F1}	Q^2_{F2}	MAE_P^*	$RMSE_P^*$
RF	0.965	0.725	0.141	0.186	0.742	0.742	0.380	0.494
AB	0.842	0.695	0.343	0.396	0.737	0.736	0.395	0.499
GB	0.966	0.718	0.145	0.184	0.746	0.746	0.382	0.490
XGB	0.980	0.642	0.114	0.143	0.681	0.681	0.423	0.549

SVM	0.735	0.709	0.384	0.513	**0.801**	**0.801**	**0.312**	**0.433**
LSVM	0.699	0.686	0.419	0.547	0.774	0.774	0.347	0.462
RR	0.708	0.696	0.430	0.538	0.753	0.753	0.378	0.483
PLS	0.708	0.698	0.430	0.538	0.753	0.753	0.378	0.483

* Standardized values of MAE and RMSE are reported.

4.3.8 Validation of model using a true external set

We have also validated our model by using a set of 10888 compounds collected from the work of Chen et. al. (Chen et. al., 2022). This dataset comprises 10900 flexible π-conjugated organic molecules generated through molecular transformation operation on benzene. A total of 12 compounds were present in the dataset for which structural information in the form of SMILES strings was not available; therefore, such compounds were excluded from the dataset. For more information on the dataset, one can refer to (Chen et. al., 2022). The RE was calculated using DFT and generic force-field (GFN-FF) for the molecules that were collected. Using the DFT calculated RE as a reference we have calculated the MAE of the predictions of all compounds using our model and the prediction using the GFN model of Chen et. al.

For the total 10888 compounds, the MAE for our stacking-SVM q-RASPR model was found to be 87.946 meV whereas it was 150.083 meV for the GFN model (Chen et. al., 2022). The results thus obtained suggest that the prediction using our stacking-SVM q-RASPR model was better than those predicted using the GFN model.

4.3.9 Comparison of model quality with other developed models

We have compared the model quality of our q-RASPR models with the other models that were developed by Sule Atahan-Evrenk (Atahan-Evrenk, 2018). Previously, Sule Atahan-Evrenk developed several models by using signature descriptors and 3D molecular transforms calculated from molecular mechanics force-field (MMFF94) and DFT. The statistical parameters of the previous models have been compared with our own developed stacking q-RASPR models. The comparison of the model parameters is shown in **Table 4.17**, which shows that our stacking q-RASPR model developed using support vector regression has given the best results for both the

calibration set and the prediction set with low error measures (MAE and RMSE). Our model was developed from 2D descriptors only that do not involve the calculation of any signature descriptors for certain heights (σ_{03}, σ_{04}) or molecular optimization for calculating 3D molecular transforms that were done by the previous authors. Also, our PLS model was developed using a single LV while their best PLS model consists of 8 LVs.

Table 4.17: Statistical comparison between different models with the current model

Descriptor type	**Model (LVs)**	**R^2_{train}**	**R^2_{test}**	**RMSE#**	**MAE#**
Signatures (Atahan-Evrenk, 2018)	**σ_{03} PLS (5)**	0.96	0.69	55	41
	σ_{03} PCR (8)	0.62	0.57	57	43
	σ_{04} PLS (8)	0.99	0.70	54	39
	σ_{04} PCR (16)	0.67	0.58	56	42
3D-transforms (Atahan-Evrenk, 2018)	**DFT-PLS (7)**	0.85	0.66	60	43
	MM-PLS (5)	0.79	0.62	60	44

2D (our work)	**Stacking q-RASPR**			**$RMSE_C$#**	**$RMSE_P$#**	**MAE_C#**	**MAE_P#**
	PLS (1)	0.708	0.753	54.176	46.916	40.911	34.856
	SVM	**0.735**	**0.801**	**50.118**	**41.587**	**36.457**	**29.134**

MAE and RMSE are reported in meV units.

Chapter 5

Conclusion

5. CONCLUSION

In the present study, we have developed different predictive models using RASPR descriptors, derived from the similarity-based read-across (RA) method. The RASPR descriptors were calculated from different physicochemical and structural descriptors using various non-linear similarity functions. Here, we implemented a simple and straightforward yet robust formalism in computing descriptors, developing models, evaluating their prediction reliability in defined chemical space, and diagnosing chemical information in accordance with OECD guidelines. The models were developed using various chemometric tools and were subjected to internal and external validation to confirm their unbiased predictability. In some cases, developed models were also tested for validation using a Y-randomization test.

5.1 Machine learning-based q-RASPR predictions of detonation heat for nitrogen-containing compounds

The present work reports a q-RASPR model developed using a step-wise process of data point collection, computation of molecular structures, descriptor calculation, pre-treatment, data division, feature selection, QSPR model development, Read-Across predictions, calculation of RASPR descriptors, data fusion and finally feature selection to develop the final q-RASPR model. Initially, an MLR q-RASPR model was selected based on the cross-validation result and after that the corresponding PLS model was developed with fewer latent variables. The authors have also employed various ML algorithms for predicting the detonation heat through the generation of different ML-based models. Further, different cross-validation strategies such as leave-one-out (LOO), 20 times 5-fold CV, and shuffle-split CV (n-splits=1000) were performed for each model to detect any over-fitting in the models. A comparison between the predictive performances of all the developed models was done as shown in **Table 3**. The selection of the final model (here PLS) was done on the ground of an error-based measure, i.e. Root Mean Squared Error of Predictions (RMSEP) of the test set compounds, i.e. $RMSE_P$. The purpose of this study was to develop an efficient model to predict the detonation property of N-containing compounds in terms of detonation heat. The study represents the development of a novel q-RASPR model in accordance with the OECD guidelines and is highly robust, easily interpretable, and reproducible. The developed model can be used to prepare new and efficient nitrogenous compounds with better

detonation performance in measures of the detonation heat and to predict the detonation heat of a new compound.

5.2 Predicting performance and stability parameters of energetic materials (EMs) using the machine learning-based q-RASPR approach

In the present work, the authors report the development of q-RASPR models for predicting different properties of energetic compounds associated with their energetic performance and thermal stability. We have used properties like decomposition temperature and melting point for the prediction of the thermal stability of compounds. For the evaluation of performance, we have used density and gas phase heat of formation. Firstly, we developed QSPR models through a feature selection process for individual data sets and then used the developed models' structural and physiochemical features to calculate the RASPR descriptors. The calculated RASPR descriptors were then fused with those structural and physiochemical descriptors. Again for each modeled response, the feature selection process was employed to the fused descriptor matrix to develop an MLR q-RASPR model based on the cross-validated result. Finally, with a lower number of LVs, a PLS q-RASPR model was developed. Several ML-based models were also prepared to predict the properties associated with the energetic compounds. Furthermore, we have also checked the model quality by using 5-fold and 10-fold cross-validation tests (in terms of R^2 and MAE) which also reflect the absence of any over-fitting.

The models so developed in the study were found to be robust and predictive, and they can be used during the early developmental stages of energetic compounds for screening purposes. This will help to select the best compound with better performance and thermal stability. These models can also be used for the development of new efficient, energetic materials or the prediction of the property for newly developed molecules. Thus, the models can be useful for the designing and manufacturing of new energetic compounds at a low cost, and a fast rate with a decrease in the hazards associated with them during the experiments.

5.3 Predictive cheminformatics modeling of reorganization energy (RE) for p-type organic semiconductors: Integration of quantitative read-across structure-property relationship (q-RASPR) and stacking regression analysis

The current study describes the method for the development of a q-RASPR model via stacking regression for predicting the RE of OSCs. RE is an essential parameter to study the ease of charge transport in the semiconductors. The work presents the collection of the dataset, development of the QSPR model, RA predictions, RASPR descriptor calculation, q-RASPR predictions using different similarity measures, and stacking regression predictions through various regression algorithms. The authors used the features of the QSPR model to perform the RA-based similarity predictions, and further, the features were used to calculate the RASPR descriptors. The RASPR descriptors were calculated for three different similarity measures namely; Euclidean distance, Gaussian kernel, and Laplacian kernel-based similarity (Banerjee and Roy, 2024). After that, the RASPR descriptors for each similarity were fused with the descriptors of the QSPR model, and a grid search was performed using the fused descriptor matrix to get the q-RASPR model with good quality and predictivity. A total of 3 PLS q-RASPR models (one for each similarity measure) were selected, and the predictions from each model were used to perform final stacking. Initially, the PLS algorithm was used to develop the stacking model using 3 predictions (as variables) and only 1 LV. The PLS model developed using stacking shows an enhancement in the prediction compared to the individual q-RASPR models. To increase the quality of the predictions of the model, the authors have also applied several ML algorithms to train the model as a stacking regressor. It was found that when the stacking was performed using the SVM regression algorithm, there was an improvement in both model quality and predictivity showing a decrease in the model errors.

The study fulfills the aim of the authors, i.e., developing a high-quality, robust, interpretable, and reproducible statistical model that can efficiently predict the RE of the p-type OSCs with the least error. Thus, the study can be used further to evaluate the mobility of charge carriers by predicting the RE of the molecules (more precisely acenes, thiophenes, thienoacenes, and pantalenes). Screening of large databases or prediction of new compounds can be done using our model within a short time without any experimental procedure or high-end computations.

Chapter 6

References

6. References

Agrawal, A. and Choudhary, A., 2019. Deep materials informatics: Applications of deep learning in materials science. Mrs Communications, 9(3), pp.779-792.

Agrawal, J.P., 2010. High energy materials: propellants, explosives, and pyrotechnics. John Wiley & Sons.

Asha, A.B. and Narain, R., 2020. Nanomaterials properties. In Polymer science and nanotechnology (pp. 343-359). Elsevier.

Askadskiĭ, A.A., 2003. Computational materials science of polymers. Cambridge Int Science Publishing.

Atahan-Evrenk, S., 2018. A quantitative structure–property study of reorganization energy for known p-type organic semiconductors. RSC advances, 8(70), pp.40330-40337.

Banerjee, A. and Roy, K., 2022. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. Molecular Diversity, 26(5), pp.2847-2862.

Banerjee, A. and Roy, K., 2023. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity endpoints. Chemical Research in Toxicology, 36(3), pp.446-464.

Banerjee, A. and Roy, K., 2024. How to correctly develop q-RASAR models for predictive cheminformatics. Expert Opinion on Drug Discovery, pp.1-6.

Banerjee, A., Chatterjee, M., De, P. and Roy, K., 2022. Quantitative predictions from chemical read-across and their confidence measures. Chemometrics and Intelligent Laboratory Systems, 227, p.104613.

Berggren, E., Amcoff, P., Benigni, R., Blackburn, K., Carney, E., Cronin, M., Deluyker, H., Gautier, F., Judson, R.S., Kass, G.E. and Keller, D., 2015. Chemical safety assessment using read-across: assessing the use of novel testing methods to strengthen the evidence base for decision making. Environmental health perspectives, 123(12), pp.1232-1240.

Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

Bursac, Z., Gauss, C.H., Williams, D.K. and Hosmer, D.W., 2008. Purposeful selection of variables in logistic regression. Source code for biology and medicine, 3, pp.1-8.

Carlsen, L., Kenessov, B.N. and Batyrbekova, S.Y., 2009. A QSAR/QSTR study on the human health impact of the rocket fuel 1, 1-dimethyl hydrazine and its transformation products: Multicriteria hazard ranking based on partial order methodologies. Environmental toxicology and pharmacology, 27(3), pp.415-423.

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

Chawla, K.K., 2012. Composite materials: science and engineering. Springer Science & Business Media.

Chen, K., Kunkel, C., Reuter, K. and Margraf, J.T., 2022. Reorganization energies of flexible organic molecules as a challenging target for machine learning enhanced virtual screening. Digital Discovery, 1(2), pp.147-157.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R. and Consonni, V., 2014. QSAR modeling: where have you been? Where are you going to?. Journal of medicinal chemistry, 57(12), pp.4977-5010.

Choudhary, K., Cheon, G., Reed, E. and Tavazza, F., 2018. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. Physical Review B, 98(1), p.014107.

Choudhary, K., Kalish, I., Beams, R. and Tavazza, F., 2017. High-throughput identification and characterization of materials using density functional theory. Scientific reports, 7(1), p.5179.

Clyne, T.W. and Hull, D., 2019. An introduction to composite materials. Cambridge university press.

Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R.H., Nelson, L.J., Hart, G.L., Sanvito, S., Buongiorno-Nardelli, M. and Mingo, N., 2012. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations. Computational Materials Science, 58, pp.227-235.

Danielsson, P.E., 1980. Euclidean distance mapping. Computer Graphics and image processing, 14(3), pp.227-248.

Doan Tran, H., Kim, C., Chen, L., Chandrasekaran, A., Batra, R., Venkatram, S., Kamal, D., Lightstone, J.P., Gurnani, R., Shetty, P. and Ramprasad, M., 2020. Machine-learning predictions of polymer properties with Polymer Genome. Journal of Applied Physics, 128(17).

Du, J., Lu, X., Gin, S., Delaye, J.M., Deng, L., Taron, M., Bisbrouck, N., Bauchy, M. and Vienna, J.D., 2021. Predicting the dissolution rate of borosilicate glasses using QSPR analysis based on molecular dynamics simulations. Journal of the American Ceramic Society, 104(9), pp.4445-4458.

Ferreira, M.M., 2001. Polycyclic aromatic hydrocarbons: a QSPR study. Chemosphere, 44(2), pp.125-146.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis, 38(4), pp.367-378.

Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".

Geronikaki, A.A., Dearden, J.C., Filimonov, D., Galaeva, I., Garibova, T.L., Gloriozova, T., Krajneva, V., Lagunin, A., Macaev, F.Z., Molodavkin, G. and Poroikov, V.V., 2004. Design of new cognition enhancers: from computer prediction to synthesis and biological evaluation. Journal of medicinal chemistry, 47(11), pp.2870-2876.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR & combinatorial science, 26(5), pp.694-701.

Gregg, B.A. and Hanna, M.C., 2003. Comparing organic to inorganic photovoltaic cells: Theory, experiment, and simulation. Journal of Applied Physics, 93(6), pp.3605-3614.

Hafner, J., Wolverton, C. and Ceder, G., 2006. Toward computational materials design: the impact of density functional theory on materials research. MRS bulletin, 31(9), pp.659-668.

Han, T., Huang, J., Sant, G., Neithalath, N. and Kumar, A., 2022. Predicting mechanical properties of ultrahigh temperature ceramics using machine learning. Journal of the American Ceramic Society, 105(11), pp.6851-6863.

Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 194(4824), pp.178-180.

He, T., Lai, W., Li, M., Feng, Y., Liu, Y., Yu, T., Tang, H., Zhang, T. and Li, H., 2021. The detonation heat prediction of nitrogen-containing compounds based on quantitative structure-activity relationship (QSAR) combined with random forest (RF). Chemometrics and Intelligent Laboratory Systems, 213, p.104249.

Heritage, T.W. and Lowis, D.R., 1999. Molecular hologram QSAR.

Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: applications to nonorthogonal problems. Technometrics, 12(1), pp.69-82.

Huang, X., Li, C., Tan, K., Wen, Y., Guo, F., Li, M., Huang, Y., Sun, C.Q., Gozin, M. and Zhang, L., 2021. Applying machine learning to balance performance and stability of high energy density materials. Iscience, 24(3).

Infante-Castillo, R. and Hernández-Rivera, S.P., 2012. Predicting Heats of Explosion of Nitroaromatic Compounds through NBO Charges and 15N NMR Chemical Shifts of Nitro Groups. Advances in Physical Chemistry, 2012(1), p.304686.

Jaidann, M., Roy, S., Abou-Rachid, H. and Lussier, L.S., 2010. A DFT theoretical study of heats of formation and detonation properties of nitrogen-rich explosives. Journal of hazardous materials, 176(1-3), pp.165-173.

Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. and Persson, K.A., 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL materials, 1(1).

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.

Katoch, S., Chauhan, S.S. and Kumar, V., 2021. A review on genetic algorithm: past, present, and future. Multimedia tools and applications, 80, pp.8091-8126.

Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S. and Wolverton, C., 2015. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Computational Materials, 1(1), pp.1-15.

Kohn, W., 1999. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. Reviews of Modern Physics, 71(5), p.1253.

Kovarich, S., Ceriani, L., Fuart Gatnik, M., Bassan, A. and Pavan, M., 2019. Filling data gaps by read-across: A mini review on its application, developments and challenges. Molecular Informatics, 38(8-9), p.1800121.

Kumar, D. and Elias, A.J., 2019. The explosive chemistry of nitrogen: A fascinating journey from 9th century to the present. Resonance, 24(11), pp.1253-1271.

Le, T.C. and Winkler, D.A., 2018. Applications in Materials Science. Applied Chemoinformatics: Achievements and Future Opportunities, pp.547-569.

Leonard, J.T. and Roy, K., 2004. Classical QSAR modeling of CCR5 receptor binding affinity of substituted benzylpyrazoles. QSAR & Combinatorial Science, 23(6), pp.387-398.

Li, J., 2009. An evaluation of nitro derivatives of cubane using ab initio and density functional theories. Theoretical Chemistry Accounts, 122, pp.101-106.

Li, Y., She, Q., Wang, X., Ma, W., Yu, H., Yu, N. and Wei, S., 2022. Classification and identification of polar pollutants on microplastics from freshwater using nontarget screening strategy. Science of The Total Environment, 822, p.153468.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2020. Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), p.18.

Liu, J., Zhang, Y., Zhang, Y., Kitipornchai, S. and Yang, J., 2022. Machine learning assisted prediction of mechanical properties of graphene/aluminium nanocomposite based on molecular dynamics simulation. Materials & Design, 213, p.110334.

Lo, Y.C., Rensi, S.E., Torng, W. and Altman, R.B., 2018. Machine learning in chemoinformatics and drug discovery. Drug discovery today, 23(8), pp.1538-1546.

Lopez-Bezanilla, A. and Littlewood, P.B., 2020. Growing field of materials informatics: databases and artificial intelligence. MRS Communications, 10(1), pp.1-10.

Luechtefeld, T., Marsh, D., Rowlands, C. and Hartung, T., 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicological Sciences, 165(1), pp.198-212.

Malkiel, I., Mrejen, M., Nagler, A., Arieli, U., Wolf, L. and Suchowski, H., 2018. Plasmonic nanostructure design and characterization via deep learning. Light: Science & Applications, 7(1), p.60.

Mamada, M. and Yamashita, Y., 2015. S-Containing Polycyclic Heteroarenes: Thiophene-Fused and Thiadiazole-Fused Arenes as Organic Semiconductors. Polycyclic Arenes and Heteroarenes: Synthesis, Properties, and Applications, pp.277-308.

Mathieu, D., 2018. Atom pair contribution method: fast and general procedure to predict molecular formation enthalpies. Journal of chemical information and modeling, 58(1), pp.12-26.

Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs, pp.801-820.

Mercier, J.P., Zambelli, G. and Kurz, W., 2002. Introduction to materials science. Elsevier.

Noble, W.S., 2006. What is a support vector machine?. Nature biotechnology, 24(12), pp.1565-1567.

Pandey, S.K. and Roy, K., 2024. Predicting the performance and stability parameters of energetic materials (EMs) using a machine learning-based q-RASPR approach. Energy Advances, 3(6), pp.1293-1306.

Patlewicz, G., Cronin, M.T., Helman, G., Lambert, J.C., Lizarraga, L.E. and Shah, I., 2018. Navigating through the minefield of read-across frameworks: A commentary perspective. Computational Toxicology, 6, pp.39-54.

Patlewicz, G., Helman, G., Pradeep, P. and Shah, I., 2017. Navigating through the minefield of read-across tools: A review of in silico tools for grouping. Computational Toxicology, 3, pp.1-18.

Peter, Y.U. and Cardona, M., 2010. Fundamentals of semiconductors: physics and materials properties. Springer Science & Business Media.

Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. and Ramprasad, R., 2013. Accelerating materials property predictions using machine learning. Scientific reports, 3(1), p.2810.

Ramakrishna, S., Zhang, T.Y., Lu, W.C., Qian, Q., Low, J.S.C., Yune, J.H.R., Tan, D.Z.L., Bressan, S., Sanvito, S. and Kalidindi, S.R., 2019. Materials informatics. Journal of Intelligent Manufacturing, 30, pp.2307-2326.

Randić, M., 1997. On characterization of chemical structure. Journal of chemical information and computer sciences, 37(4), pp.672-687.

Rice, B.M., Hare, J.J. and Byrd, E.F., 2007. Accurate predictions of crystal densities using quantum mechanical molecular volumes. The Journal of Physical Chemistry A, 111(42), pp.10874-10879.

Rodríguez-Pérez, R. and Bajorath, J., 2020. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of computer-aided molecular design, 34(10), pp.1013-1026.

Rogers, D. and Hopfinger, A.J., 1994. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. Journal of Chemical Information and Computer Sciences, 34(4), pp.854-866.

Roy, K., 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert opinion on drug discovery, 2(12), pp.1567-1577.

Roy, K., Das, R.N., Ambure, P. and Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems, 152, pp.18-33.

Roy, K., Kar, S. and Ambure, P., 2015b. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems, 145, pp.22-29.

Roy, K., Kar, S. and Das, R.N., 2015a. A primer on QSAR/QSPR modeling: fundamental concepts. Springer.

Roy, K., Kar, S. and Das, R.N., 2015c. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press.

Saal, J.E., Kirklin, S., Aykol, M., Meredig, B. and Wolverton, C., 2013. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). Jom, 65, pp.1501-1509.

Salaneck, W.R., Seki, K., Kahn, A. and Pireaux, J.J., 2001. Conjugated Polymer and Molecular Interfaces: Science and Technology for Photonic and Optoelectronic Application. CRC press.

Segall, M., Champness, E., Obrezanova, O. and Leeding, C., 2009. Beyond profiling: using ADMET models to guide decisions. Chemistry & Biodiversity, 6(11), pp.2144-2151.

Selassie, C.D. and Verma, R.P., 2003. History of quantitative structure-activity relationships. Burger's medicinal chemistry and drug discovery, 1, pp.1-48.

Stein, H.S., Guevarra, D., Newhouse, P.F., Soedarmadji, E. and Gregoire, J.M., 2019. Machine learning of optical properties of materials–predicting spectra from images and images from spectra. Chemical science, 10(1), pp.47-55.

Stergiou, K., Ntakolia, C., Varytis, P., Koumoulos, E., Karlsson, P. and Moustakidis, S., 2023. Enhancing property prediction and process optimization in building materials through machine learning: A review. Computational Materials Science, 220, p.112031.

Sukumar, N., Krein, M., Luo, Q. and Breneman, C., 2012. MQSPR modeling in materials informatics: a way to shorten design cycles?. Journal of Materials Science, 47, pp.7703-7715.

Takahashi, K. and Tanaka, Y., 2016. Materials informatics: a journey towards material design and synthesis. Dalton Transactions, 45(26), pp.10497-10499.

Tercan, H., Guajardo, A., Heinisch, J., Thiele, T., Hopmann, C. and Meisen, T., 2018. Transfer-learning: Bridging the gap between real and simulation data for machine learning in injection molding. Procedia Cirp, 72, pp.185-190.

Tikhonova, I.G., Baskin, I.I., Palyulin, V.A. and Zefirov, N.S., 2004. Virtual screening of organic molecule databases. Design of focused libraries of potential ligands of NMDA and AMPA receptors. Russian chemical bulletin, 53, pp.1335-1344.

Tong, W., Hong, H., Xie, Q., Shi, L., Fang, H. and Perkins, R., 2005. Assessing QSAR limitations-A regulatory perspective. Current Computer-Aided Drug Design, 1(2), pp.195-205.

Van de Waterbeemd, H., Carter, R.E., Grassy, G., Kubinyi, H., Martin, Y.C., Tute, M.S. and Willett, P., 1997. Glossary of terms used in computational drug design (IUPAC Recommendations 1997). Pure and applied chemistry, 69(5), pp.1137-1152.

Wachtman, J.B., Cannon, W.R. and Matthewson, M.J., 2009. Mechanical properties of ceramics. John Wiley & Sons.

Wang, L., Zhai, L., She, W., Wang, M., Zhang, J. and Wang, B., 2022. Synthetic strategies toward nitrogen-rich energetic compounds via the reaction characteristics of cyanofurazan/furoxan. Frontiers in Chemistry, 10, p.871684.

Wespiser, C. and Mathieu, D., 2023. Application of machine learning to the design of energetic materials: preliminary experience and comparison with alternative techniques. Propellants, Explosives, Pyrotechnics, 48(4), p.e202200264.

Wold, S., Sjöström, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), pp.109-130.

Wu, K., Natarajan, B., Morkowchuk, L., Krein, M. and Breneman, C.M., 2013. From drug discovery QSAR to predictive materials QSPR: the evolution of descriptors, methods, and models. In Informatics for materials science and engineering (pp. 385-422). Butterworth-Heinemann.

Wu, Q., Burges, C.J., Svore, K.M. and Gao, J., 2010. Adapting boosting for information retrieval measures. Information Retrieval, 13, pp.254-270.

Xie, Q., Suvarna, M., Li, J., Zhu, X., Cai, J. and Wang, X., 2021. Online prediction of mechanical properties of hot rolled steel plate using machine learning. Materials & Design, 197, p.109201.

Yin, P., Zhang, Q. and Shreeve, J.N.M., 2016. Dancing with energetic nitrogen atoms: versatile N-functionalization strategies for N-heterocyclic frameworks in high energy density materials. Accounts of chemical research, 49(1), pp.4-16.

Yosipof, A., Shimanovich, K. and Senderowitz, H., 2016. Materials informatics: statistical modeling in material science. Molecular Informatics, 35(11-12), pp.568-579.

Yu, R., Lin, Q., Leung, S.F. and Fan, Z., 2012. Nanomaterials and nanostructures for efficient light absorption and photovoltaics. Nano energy, 1(1), pp.57-72.

Yu, Z., Ye, S., Sun, Y., Zhao, H. and Feng, X.Q., 2021. Deep learning method for predicting the mechanical properties of aluminum alloys with small data sets. Materials Today Communications, 28, p.102570.