q-RASPR modelling of the Power Conversion Efficiency (PCE) of the different classes of organic dyes of Dye Sensitized Solar Cells (DSSCs)

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

MASTER OF PHARMACY

Faculty of Engineering and Technology

Thesis submitted by

SOUVIK PORE

BPHARM

Registration No. **139976 of 2017-2018**

Examination Roll. No**. M4PHC23030**

Under the Guidance of

DR. KUNAL ROY

Professor

Drug Theoretics & Cheminformatics Laboratory

Department of Pharmaceutical Technology

Jadavpur University

Kolkata – 700 032

India

2023

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 "q-RASPR modelling of the Power Conversion Efficiency (PCE) of the different classes of organic dyes of Dye Sensitized Solar Cells (DSSCs)".

All information in this document have 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: **SOUVIK PORE**

EXAMINATION ROLL NUMBER: **M4PHC23030**

REGISTRATION NUMBER: **139976 of 2017-2018**

THESIS TITLE: "**q-RASPR modelling of the Power Conversion Efficiency (PCE) of the different classes of organic dyes of Dye Sensitized Solar Cells (DSSCs)**"

SIGNATURE WITH DATE:

CERTIFICATE

Department of Pharmaceutical Technology
Jadavpur University
Kolkata - 700 032

This is to certify that **Mr. Souvik Pore**, B. Pharm. (Jadavpur University), has carried out the research work on the subject entitled **"q-RASPR modelling of the Power Conversion Efficiency (PCE) of the different classes of organic dyes of Dye Sensitized Solar Cells (DSSCs)"** under my supervision in Drug Theoretics & Cheminformatics Laboratory 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.

_____

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

_____

(Prof. Sanmoy karmakar)
Head, Dept. of Pharmaceutical Technology,
Jadavpur University, Kolkata

_____

(Prof. Saswati Mazumdar)
Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata

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 Priyanka De, Joyita Roy, Mainak Chatterjee, Arkaprava Banerjee, Sapna Kumari Panday, Ankur Kumar and Vinay Kumar. 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 juniors Subham Panday and Doelima Bera who all have extended their helping hands and friendly cooperation all through my work.*

I am indeed glad to convey cordial thanks to all my friends specially Shilpayan Ghosh. I am thankful to the authority of Jadavpur University and Head of the Department Prof. Dr. Sanmoy Karmakar for providing all the facilities to carry out this work.

A word of thanks to all those people associated with this work directly or indirectly whose names I have been unable to mention here. Finally, I would like to thank my parents Mr. Chittaranjan Pore and Mrs. Kakoli Pore for all the love and inspirations without which my dissertation work would remain incomplete.

_____

SOUVIK PORE

Examination Roll No.: M4PHC23030

Department of Pharmaceutical Technology,

Jadavpur University,

Kolkata-700032

Preface

This dissertation is presented for the partial fulfillment of the degree of Master of Pharmacy in Pharmaceutical Technology. The presented work is spanned over one year. The present study has been executed by developing in silico predictive models in the quantitative structure-property relationship (QSPR) paradigm. We have considered power conversion efficiency (PCE) endpoints of organic dyes in dye-sensitized solar cells for the development of the predictive models.

In the recent past, the over-expansion of population led to an increase in the usage of fossil fuels to meet the increasing energy demand. This overuse of fossil fuels is the main reason behind global warming and climate change. Renewable energy source is the main important alternative to the fossil fuel which is a never-ending eco-friendly source of energy. There are various forms of renewable energy present in nature like solar energy, wind energy, tidal energy, geothermal energy, etc. One of the important sources of renewable energy is solar energy where energy coming from the sun in the form of heat and solar radiation is converted to electrical energy. The photovoltaic (PV) solar cell converts solar radiation into electrical energy when an incoming photon strikes the solar cell. The dye-sensitized solar cells (DSSCs) are the third-generation solar cell that contain dye molecules as photosensitizers for the generation of electron. The main advantages of the DSSCs compared to the previous solar cells are eco-friendly, cost effectiveness, easy to manufacture, flexible, etc. Extensive research is going on to find out new dye molecules with better power conversion efficiency. The development of a new dye molecule through experimental methods requires a lot of resources, money and time, but the success rate of getting a better dye molecule is uncertain. To reduce the cost of failure and resources, different computational methods are used for the development of new dyes even before the synthesis.

The application of various computational approaches for the prediction of properties of chemical substances has been an effective alternative to experimental methods. QSPR is a statistical method widely used for prediction purposes of different property-based endpoints. Read-Across (RA) is a similarity-based approach used 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 chemoinformatic tool: read-across structure-property relationship (RASPR).

In this present study, we have modeled the power conversion efficiency (PCE) of organic dyes used in dye-sensitized solar cells (DSSCs) by employing the q-RASPR method. The developed models have shown acceptable statistical significance. The models developed were validated rigorously based on various internal and external validation metrics. The following analyses have been performed in this dissertation:

Study 1: Machine learning - based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells.

Study 2: Application of machine learning-based read-across structure-property relationship (RASPR) as a new tool for predictive modelling: Prediction of power conversion efficiency (PCE) for selected classes of organic dyes in dye-sensitized solar cells (DSSCs).

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: Results and Discussion

Chapter 5: Conclusion

Chapter 6: References

Appendix: Reprints

Abbreviations

RES	Renewable Energy Source
PW	Petawatts
UV	Ultraviolet
PV	Photovoltaic
CSP	Concentrated Solar Power
QDSSC	Quantum-Dot-Sensitized Solar Cell
DSSC	Dye Sensitized Solar Cells
QD	Quantum Dot
OPV	Organic Photovoltaic cell
TCO	Transparent Conductive Oxide
SCO	Semi-Conductive Oxide
PA	Photoanode
LUMO	Lowest Unoccupied Molecular Orbital
HOMO	Highest Occupied Molecular Orbital
ICT	Intramolecular Charge Transfer
PCE	Power Conversion Efficiency
MI	Materials Informatics
QSPR	Quantitative Structure-Property Relationship
RA	Read-Across
ML	Machine Learning
QSAR	Quantitative Structure-Activity Relationship
QSTR	Quantitative Structure-Toxicity Relationship
LFER	Linear Free Energy Relationship
MSA	Molecular Shape Analysis
BSS	Best Subset Selection
MLR	Multiple Linear Regression
LV	Latent Variables
CV	Cross Validation
MAE	Mean Absolute Error
LOO	Leave One Out
SHAP	SHapley Additive exPlanation
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis
HASL	Hypothetical Active Site Lattice
RSA	Receptor Surface Analysis
COMBINE	Comparative Binding Energy
OECD	Organisation for Economic Co-operation and Development
WAP	Weighted Average Prediction
RASPR	Read-Across Structure-Property Relationship
RASAR	Read-Across Structure-Activity Relationship
AI	Artificial Intelligence
LSO	Leave-Same-Out
RF	Random Forest
GB	Gradient Boosting
XGB	Extreme Gradient Boosting
SVM	Support Vector Machine
LSVM	Linear Support Vector Machine
RR	Ridge Regression
PLS	Partial Least Squares
ADB	Adaptive Boosting
ED	Euclidean Distance
GK	Gaussian Kernel
LK	Laplacian Kernel
CTC	Close Training Compound
PCA	Principal Component Analysis
AD	Applicability Domain
GA	Genetic Algorithm
RMSEC	Root Mean Square Error of Calibration
RMSEP	Root Mean Square Error of Prediction
MSE	Mean Squared Error
LIME	Local Interpretable Model-Agonistic Explanations

Contents

Chapter – 1

INTRODUCTION

1. INTRODUCTION

1.1 Solar energy

The rapid expansion of population and technological advancements has led to an exponential increase in fossil fuel usage to meet the increasing energy demand. This rapid usage of fossil fuels has a significant impact on the environment, which is the main reason for global warming and climate change [1, 2]. Therefore, extensive research is going on to find out different methods to reduce or replace the usage of fossil fuels. There are various approaches used to reduce the usage of fossil fuels, like – (a) by increasing the efficiency of the conventional power conversion device through waste heat recovery;[3, 4] (b) by developing a power conversion device that is eco-friendly;[5, 6] (c) by using renewable energy source which has no or minimal effect on the environment [7-9]. A renewable energy source (RES) is most attractive due to their significantly low cost of generation, ease to access, inexhaustible and environment friendliness. The main sources of RES include solar energy, wind energy, tidal energy, etc., as shown in the **Figure 1.1**.

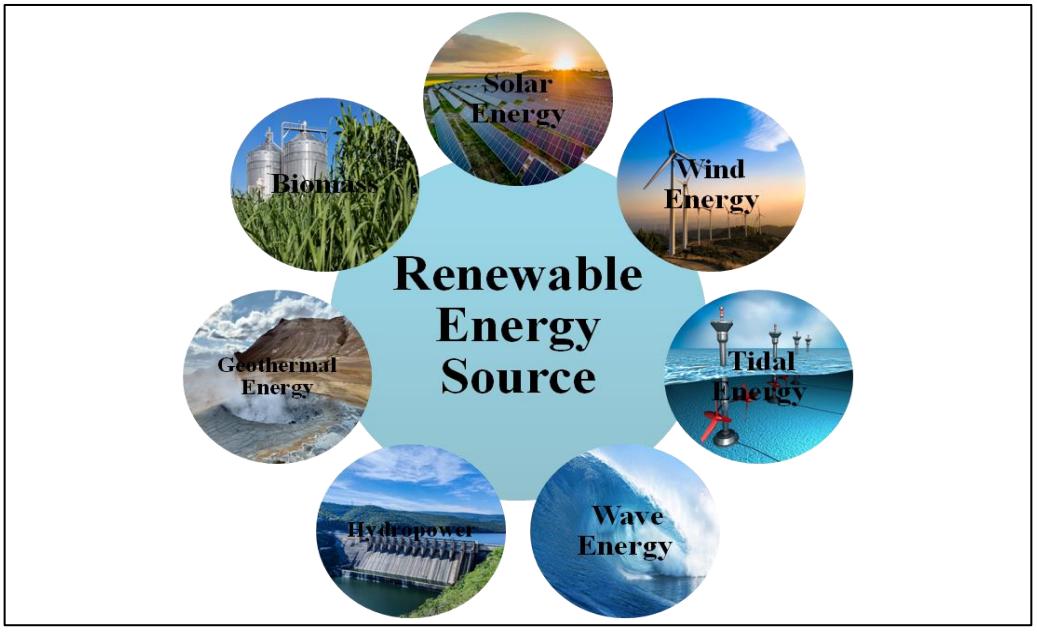


Figure 1.1: Different sources of renewable energy

One of the most important sources of renewable energy is solar energy in which energy coming from the sun is converted into electrical energy. From the sun, energy is coming in the form of heat and radiation which can initiate photochemical or photoelectrical reactions [10-12]. The

total solar energy incident on Earth is far greater than the global energy needs at the moment and in the future. This highly diffused source might be able to supply all future energy needs if properly utilized [10]. Due to its limitless supply and lack of pollution, solar energy is predicted to become a more appealing renewable energy source in the future than the finite fossil fuels coal, petroleum, and natural gas [13]. The utilization of solar energy has gained significant attention in recent years due to its potential to address the growing concerns of climate change and the need for sustainable energy options.

Even though the sun is a very potent energy source and sunlight is by far the most abundant energy source that Earth receives, the intensity of sunlight at the planet's surface is actually rather low. This happens because of the enormous radial spreading of radiation and additional loss is occur due to the scattering and absorption of radiation in the Earth's atmosphere and clouds [14]. The earth receives approximately 174 petawatts (PW) of incoming solar radiation at the upper atmosphere but 30 percent of this radiation got reflected to space and the rest 70 percent is absorbed by clouds, oceans, and lands. At Earth's surface, the sun's radiation is divided roughly equally between 3 to 5 percent UV (below 400 nm), 42 to 43 percent visible (400 to 700 nm), and 52 to 55 percent infrared (above 700 nm), as shown in **Figure 1.2**. At the top of the atmosphere, sunlight is about thirty percent more intense and contains 8% more ultraviolet (UV), the majority of which is short-wave ultraviolet, which is biologically harmful [15].

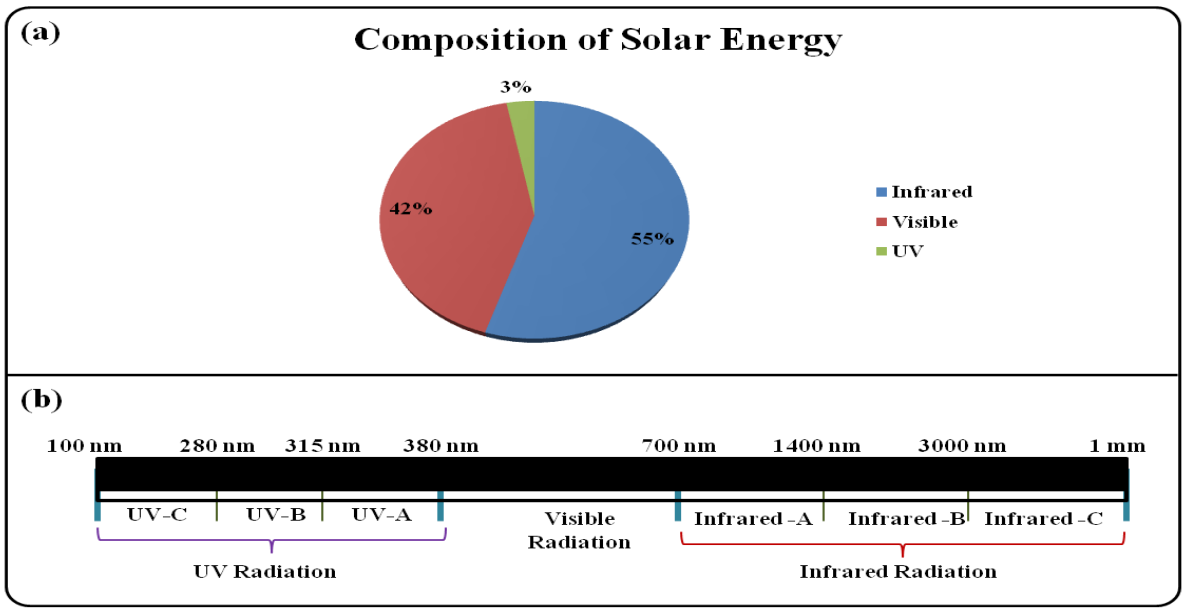


Figure 1.2: Composition of solar radiation coming to the earth.

Even though solar energy is free in itself, the high expense of collecting, converting, and storing it prevents widespread use of it. Solar radiation can be converted either into thermal energy (heat) or into electrical energy, though the former is easier to achieve [16]. One of the primary methods of harnessing solar energy is through the use of photovoltaic (PV) cells, also known as solar panels. These panels convert sunlight directly into electricity by utilizing the photovoltaic effect. When sunlight strikes the PV cells, electrons in the cells are excited, creating an electric current. This electricity can be used to power homes, businesses, and even entire communities [17]. Solar energy can also be harnessed using solar thermal systems. These systems use the sun's heat to generate electricity or provide hot water for domestic and industrial purposes. Concentrated Solar Power (CSP) plants use mirrors or lenses to focus sunlight onto a receiver, which in turn heats a fluid to produce steam. The steam is then used to drive a turbine and generate electricity [18].

The advantages of solar energy are numerous. First and foremost, solar energy is renewable and abundant. The sun provides an inexhaustible supply of energy, making solar power a sustainable alternative to fossil fuels. Additionally, solar energy production does not produce greenhouse gas emissions, helping to mitigate climate change [19]. Solar panels can be installed on rooftops, reducing the need for large land areas, and can even be integrated into building materials, such as solar windows or solar tiles. Another benefit of solar energy is its potential for decentralization. Unlike centralized power plants, solar energy systems can be installed at the point of consumption, allowing individuals and communities to generate their electricity. This can lead to energy independence, reduced reliance on the grid, and lower electricity bills [20]. Despite its advantages, solar energy does face some challenges. The intermittent nature of sunlight means that solar power generation is dependent on weather conditions and time of day. Energy storage technologies, such as batteries, are being developed to address this issue and ensure a constant supply of electricity. The initial cost of installing solar panels can also be a barrier for some, although the long-term savings on energy bills can offset this investment [19].

1.2 Solar cells

The sun sends huge amounts of energy every day, in the form of heat and radiation which is known as solar energy and it is available at no cost [21, 22]. The main advantage of solar energy over other conventional power generation method is that sunlight can be directly harvested into

solar energy by using small photovoltaic (PV) solar cells [23, 24]. The PV solar cells convert the solar radiations into electrical energy by the method known as photovoltaic effect, where electrons or other charge carriers are excited to higher energy state upon absorption of photons [25]. The photovoltaic effect was first observed by **Alexandre-Edmond Becquerel** in **1839** [26]. After that **Charles Fritts** experimented with the first solar cell in **1884**, which was made of a layer of selenium covered in a thin coating of gold, but it had a very poor efficacy [27]. The first modern solar cell made of silicon was invented by **Russel Ohl** in **1946** [28]. Previously photovoltaic solar cells are made up with thin silicon wafers which convert solar energy into electrical energy. The modern Photovoltaic cells are composed of two layers of semiconductor material (p-type and n-type) that create electron holes when they are exposed to light. In this structure, an electron is ejected from one layer to another when a photon of sufficient energy strikes the junction between the n-type and p-type layers. During this process, an electron and a hole are created, and electrical power is generated [29]. There are various materials which are used for the construction of the PV solar cell, on the basis of these material solar cells are categorized into different groups which are discussed below (also shown in **Figure 1.3**).

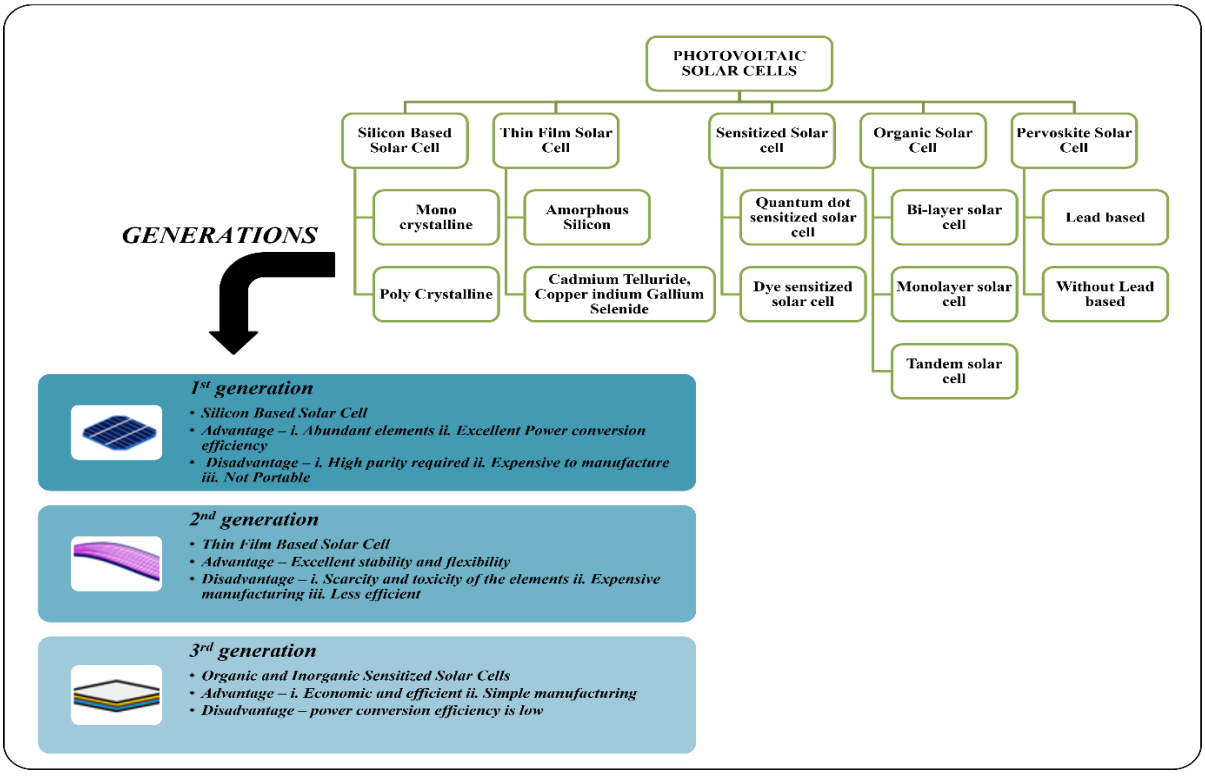


Figure 1.3: Classification of different type photovoltaic solar cells.

1.2.1 First generation solar cells

This is the oldest and most popular PV solar cell due to its high-power conversion efficiency and it is produced by using silicon wafers [30, 31]. This PV solar cell can be further categorized into two subgroups namely –

i. Single or Mono-crystalline silicon solar cell
ii. Poly or Multi-crystalline silicon solar cell

1.2.1.1 Single or mono-crystalline silicon solar cells

Mono-crystalline solar cell, as the name suggests, it is manufactured from single crystals of silicon by a process called Czochralski process [32]. During the manufacturing process, Si crystals are produced from the big sized crystals by slicing. These large single crystal productions require precise processing which make the process more expensive and multi process. The efficiency of mono-crystalline silicon solar cells lies between 17% - 18% [33].

1.2.1.2 Poly or multi-crystalline silicon solar cells

Poly-crystalline silicon solar cells composed of number of different crystals, coupled to one another to develop a single solar cell. The polycrystalline Si solar cells, which are created by cooling a graphite mould filled with molten silicon, are more cost-effective. Presently polycrystalline Si solar cells are the most popular solar cells. Although these solar cells are slightly cheaper to fabricate compared to mono-crystalline silicon solar panels, they are less efficient ~12% - 14% [34].

1.2.2 Second generation solar cells

Most second-generation solar cells, including a-Si and thin film solar cells, are more cost-effective than first generation silicon wafer solar cells. Silicon-wafer cells have light absorbing layers up to 350 μm thick, while thin-film solar cells have very thin light absorbing layers, generally of the order of 1 μm thickness [35]. Second generation solar cells are generally classified based on their composition, such as.

i. Amorphous Silicon Thin Film (a-Si) Solar Cell
ii. Cadmium Telluride (CdTe) Thin Film Solar Cell

1.2.2.1 Amorphous silicon thin film (a-Si) solar cells

The earliest solar cells to be produced industrially were amorphous Si (a-Si) PV solar cells. Low processing temperatures make amorphous (a-Si) solar cells manufacturable, enabling the use of a variety of inexpensive, flexible substrates made of polymers and other materials. These substrates process more efficiently with less energy, which make amorphous Si solar cell cheaper and widely available [36]. The term "amorphous" denotes that the silicon material that makes up the cell lacks a clear arrangement of atoms in the lattice, is not crystalline, or is not highly organized. These are created by covering the substrate/glass plate with a doped silicon substance [37].

The main problem of a-Si solar cell is the poor and almost unstable efficiency. Currently, the efficiencies of a-Si solar cell vary in the range of 4% - 8%. They can be easily operated at elevated temperatures and are suitable for the changing climatic conditions where sun shines for few hours [38].

1.2.2.2 Cadmium telluride (CdTe) thin film solar cells

Cadmium telluride (CdTe) is the first low-cost PV technology and one of the top candidates for the creation of more affordable, commercially viable photovoltaic (PV) devices among thin-film solar cells. It is usually constructed by sandwiching between layers of cadmium sulphide to create a p-n junction diode. CdTe has high chemical stability as well as high optical absorption coefficient and it also has low band gap. These properties make CdTe more attractive material for thin film based solar cell. The efficiency of this type of solar cells usually operates in the range 9% to 11% [39, 40].

There are various environmental issues with this type of solar cells that limit its widespread application. Cadmium is recognized as a heavy metal and potentially dangerous substance that can accumulate in the bodies of people, animals, and plants. The recycling and disposal of dangerous Cd-based materials have the potential to be very expensive and harmful to society and the environment. Therefore, the primary problems with this CdTe technology are a limited supply of cadmium and the environmental risk connected with its use [41].

1.2.3 Third generation solar cells

The 3rd generation solar cell is the newer promising solar cell technology, but these are not commercially developed technologies. The 3rd generation solar cells have lots of advantage over other solar cells, like it is flexible, environment friendly, easy to manufacture and cost effective. Although these solar cells have lots of advantages, it is not used widely because of its poor power conversion efficiency compared with silicon based solar cells [42, 43]. The 3rd generation solar cells are generally categorized into different groups based on their composition and power generation mechanism which are:

i. Sensitized Solar Cell
ii. Organic Solar Cell
iii. Perovskite Solar Cell

1.2.3.1 Sensitized solar cells

Sensitized solar cells are one of the most important types of third generation solar cell which are easy to manufacture and cost effective but have poor power conversion efficiency. The sensitized solar cells are classified into two main groups which are,

i. Quantum-dot-sensitized solar cells (QDSCs)
ii. Dye sensitized solar cells (DSSCs)

Although the mechanism of power generation is same, but they have different composition and structure of solar cell.

1.2.3.1.1 Quantum-dot-sensitized solar cells (QDSCs)

QDSCs are the most promising low-cost alternative of the other existing solar technologies like silicon based solar cell and thin film based solar cell. QDSCs utilize quantum dots (QDs) as light-absorbing materials for generation of solar energies. Quantum dots are nanometre-sized semiconductor particles that exhibit unique electronic and optical properties due to quantum confinement effects [44, 45]. These properties allow QDSSCs to overcome some of the limitations of traditional solar cells, such as low light absorption efficiency and high production costs [44].

The development of QDSCs originated from the pioneering work of Grätzel and O'Regan in 1991, who introduced the concept of dye-sensitized solar cells (DSSCs) [46]. Since then,

extensive research has been conducted to enhance the performance of DSSCs, and the utilization of quantum dots as sensitizers has proven to be a significant breakthrough. The ability to tune the bandgap of quantum dots by controlling their size and composition allows for the efficient harvesting of a broad spectrum of light, including visible and infrared regions, which results in improved overall energy conversion efficiency [47, 48].

Quantum Dot Sensitized Solar Cells (QDSCs) generate power through a process that involves light absorption by quantum dots, creating electron-hole pairs. These pairs are separated, and the electrons flow through a semiconductor film to generate an electric current. Redox mediators regenerate the quantum dots, enhancing overall efficiency [48]. QDSCs offer higher light absorption and improved energy conversion, promising a greener and more efficient solar energy future [44].

1.2.3.1.2 Dye sensitized solar cells (DSSCs)

Dye-sensitized solar cells (DSSCs) represent a breakthrough technology in the field of photovoltaics, offering a cost-effective and efficient alternative to traditional silicon-based solar cells [49]. Conceived in the early 1990s by Michael Grätzel and Brian O'Regan, DSSCs are characterized by their innovative design, utilizing organic dyes to absorb sunlight and convert it into electricity [46]. The unique architecture and fundamental principles of DSSCs have sparked intense research interest and significant advancements in recent years.

DSSCs harness the principles of light absorption and electron transfer to generate electrical power. The core component of a DSSC is a mesoporous semiconductor film, typically composed of titanium dioxide (TiO_2), which acts as an anchoring surface for the light-absorbing dye molecules. These dyes capture photons from sunlight across a wide range of wavelengths, converting the light energy into excited electrons [50]. The process of charge separation occurs when the excited electrons in the dye are injected into the conduction band of the semiconductor material. Simultaneously, the dye cations (positively charged dye molecules) leave behind holes in the valence band, generating electron-hole pairs. Electrons then flow through the semiconductor material to an electrode, creating an electric current. In contrast, the dye cations are regenerated by electron donation from an electrolyte solution, thus closing the circuit and completing the energy conversion cycle [51]. The architecture of DSSCs and material used for its construction is discussed in detail in the **section 1.3**.

The distinguishing features of DSSCs lie in their low-cost fabrication, ease of manufacturing, and their ability to perform efficiently even under low-light conditions. Additionally, the wide variety of dyes available allows for tuning the absorption spectrum, expanding their applications in different environmental settings, and integrating them into flexible and transparent modules [52].

1.2.3.2 Organic solar cells

Organic solar cells come in various types, each utilizing unique materials and architectures to convert sunlight into electricity. Organic photovoltaic cells (OPVs) are generally formed by using thin film of polymers, small molecules or both fabricated by very simple and cost-efficient techniques like spin coating, spray deposition and printing [53]. This type of solar cells is very flexible which allow them to use in different shapes with large surface area. OPVs are generally categorized under small molecule and polymer solar cells [54]. Although these types of solar cells are easy to fabricate, cost effective but still these solar cells have poor efficiencies (approximately 10%) [55].

Organic photovoltaic solar cells are generally comprised of at least 4 layers and also a transparent substrate which is made up with glass, polyesters or any other transparent materials [56]. In some OPVs stainless steel is used as transparent material but, in that case, transparent substrate is designed as backside of the cell. The next layer is comprised of a transparent conductive oxide (TCO) layer which is made up with indium tin oxide (ITO) and it is coated on glass or polyethyleneteraphthalate (PET) material. These two has a dual purpose: it acts as a transparent window for incoming layer and act as an anode to collect the generated holes [55, 56]. The main component of the OPVs is the active layer which contain donor and acceptor organic molecules for the generation of electron-hole pair. The active layer is either applied in layer-by-layer geometry (planar or bilayer heterojunction) or mixed to create blend (bulk heterojunction) [57, 58]. It has been found that the components of the anode get diffused into the active layer and cause its degradation due to the formation of charge trap centres. Another problem arises due to the active layer is in contact with both anode and cathode which may cause flow of electron and hole to the same electrode and cause reduction of the performance of the solar cell [55, 59-61]. A protective layer is used between the active layer and the anode to overcome this problem, this protective layer is made up with poly(3,4-ethylenedioxythiophene)

poly(styrenesulfonate) (PEDOT:PSS). Here, PEDOT:PSS acts as an electron-blocking layer and it also helps to assisting the efficient transfer of photo-generated holes to the anode [62]. The last layer of organic solar cell is the cathode which is generally aluminium, copper, or any other metal which can achieve better electron transport [55].

Here are some notable types of organic solar cells used for power generation:

Single-layer organic solar cells: Single-layer organic solar cells, also known as bulk heterojunction solar cells, are among the earliest and most widely studied organic photovoltaic devices. These cells typically consist of a single layer of an organic semiconductor blend, where donor and acceptor materials are mixed to facilitate efficient charge separation and transport [63].

Bi-layer organic solar cells: Acceptor and donor are brought into close proximity in a bilayer solar cell by subsequent deposition of the two layers. To achieve an ohmic contact to the acceptor and donor, the bilayers are sandwiched between two electrodes. The cathode, which is located on the donor side, should be transparent and have a high work function (ITO is one of the best options). On the acceptor side, the other electrode (anode) with a low work function is deposited. Light penetrates the heterojunction through the transparent cathode [55].

Tandem organic solar cells: Tandem organic solar cells employ a stacked configuration of multiple sub cells with complementary absorption spectra to enhance light harvesting efficiency. This architecture allows the absorption of a broader range of solar wavelengths, boosting overall power conversion efficiency. Tandem structures often use different organic materials in each sub cell to optimize light absorption and electron-hole separation [64-66].

1.2.3.3 Perovskite solar cells:

This type of solar cells contains perovskite structure compound, most commonly hybrid organic-inorganic lead or tin halide-based compounds which act as active layer for light absorption [67]. The perovskites are compounds which are indicated by the formula ABX3 where X represent any halogen (I-, Br-, Cl-) and A B represent the two different cations of different size [68]. In this type of solar cell compound like methylammonium lead halide are used because it is cheap and easy to synthesize. Efficiencies of this type of solar cells which contain this type of perovskite compound have increased from 3.8% in 2009 to 22.7% in 2017; make these fastest

growing solar technologies [69]. This type of solar cells can achieve even higher power generation efficiencies with low production cost. However, the main problem with perovskite solar cell is their stability and durability. The material used for its construction is moisture and oxygen sensitive which leads to its degradation over time and reduce the efficiencies of the solar cell [70]. Researchers are actively working on developing encapsulation techniques and more stable perovskite material to overcome this issue [71, 72]. Another concern of perovskite solar cell is its toxicity, some of the early perovskite materials contain lead which is regarded as heavy metal and cause different disease by accumulating in the human body. While efforts have been made to reduce or eliminate toxic components, ensuring the environmental safety of perovskite solar cells remains a priority for further development [73-75].

1.3 Dye sensitized solar cells (DSSCs)

Dye-Sensitized Solar Cells (DSSC) represent one of the important types of 3rd generation PV solar cells in which different types of organic dyes are used as photosensitizers. Recently, DSSCs have gained considerable amount of attraction due to its advantages over silicon based solar cells, for having limiting weight, low cost of production and environment friendliness [76]. The dye is the main photoactive component of the DSSCs, which generate electrons from photons through series of electrochemical reaction [77, 78]. The DSSCs are prepared by coating the dye molecules onto a wide bandgap semi-conductive oxide substance [79]. The dye is the most important part of the DSSCs because it controls photon harvesting and electron generation [78].

The basic architecture of DSSCs is shown in **Figure 1.4a**. It consists of 7 layers namely a transparent substance (mainly glass or polymer), a transparent conductive oxide (TCO) layer (mainly Fluorine doped tin oxide (FTO), an Indium doped tin oxide (ITO) are used), a blocking layer (ZnO, In_2O_3, MgO, etc.), a semi-conductive oxide (SCO) layer (mainly TiO_2) coated with photoactive dye, electrolyte solution and a counter electrode (Pt). In DSSCs, transparent substance - TCO layer – blocking layer – SCO layer – dye together form a photoanode (PA), and counter electrode – TCO layer – transparent substance is united to form the cathode. Electrolytes like iodide/triiodide (I^-/I_3^-) solution is used for the preparation of DSSCs, where the electrolytes play an important role in the regeneration of dye by redox reaction [77]. A detailed description of all the layers and their functions is represented in the **Table 1.1** [80-84]. In DSSCs, the electrons are generated when the dye undergoes photoexcitation by absorbing radiation coming from the

sun. The electrons are excited to the lowest unoccupied molecular orbital (LUMO) from the highest occupied molecular orbital (HOMO), and subsequently, electrons are transported to the TCO layer by the conduction band of the nanostructured SCO layer. From the TCO layer, the electrons flow through the external circuit and get collected at the platinum counter electrode site. The electrons are then transferred to the HOMO of dyes for their regeneration by the redox reaction of electrolytes which is catalyzed by a platinum counter electrode [77, 78]. The whole process for the generation of electrons is shown in **Figure 1.4b**.

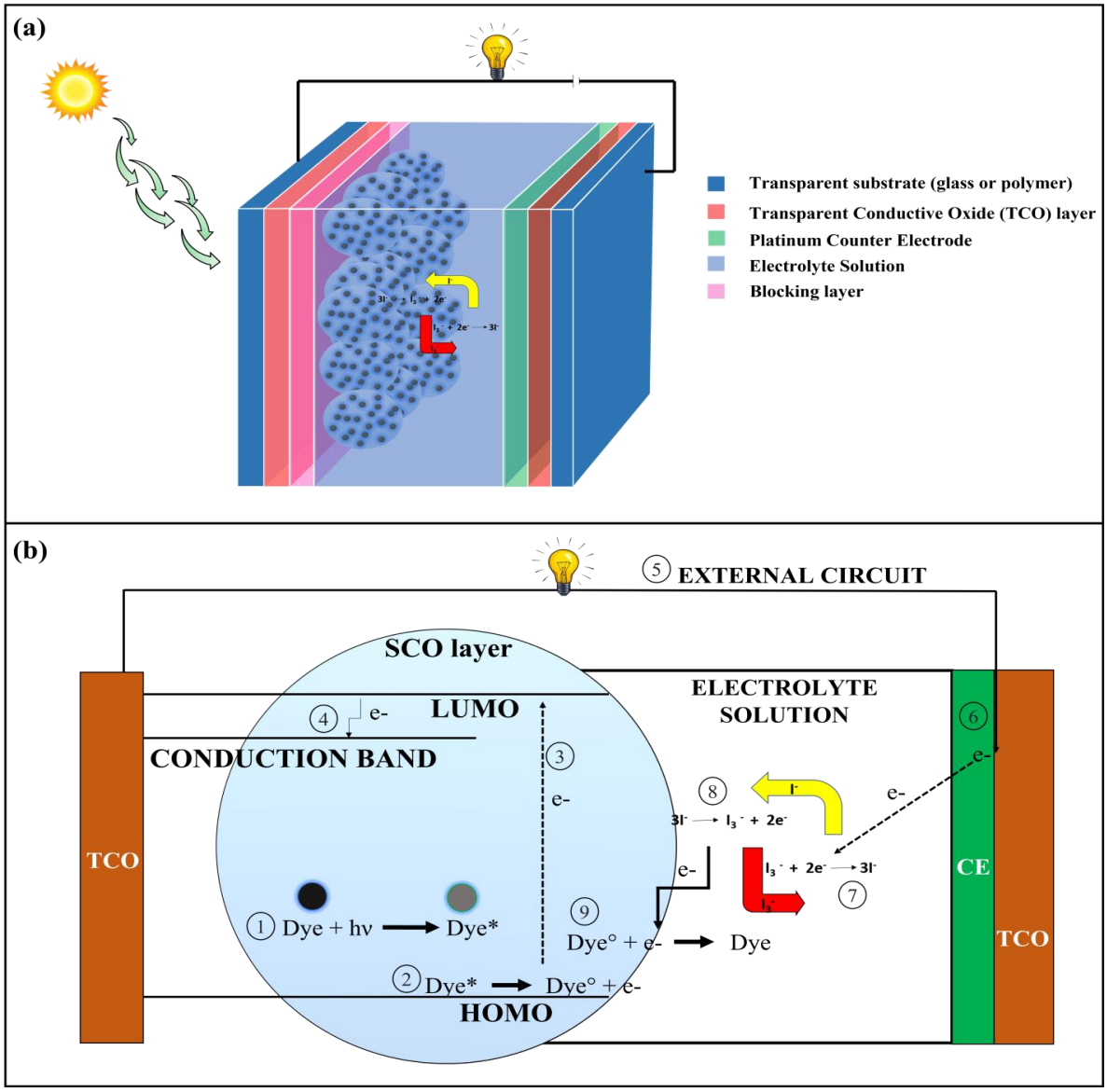


Figure 1.4: (a) Basic structure of Dye-Sensitized Solar cells (DSSCs) (b) Mechanism of electron generation in DSSCs

Table 1.1: Description of all the layers of DSSCs and their functions

Layers of DSSCs	Composition	Function
Transparent Substance [80]	i. Soda-lime glasses ii. Polymer-based substance (**Polyethylene terephthalate (PET)** and **polyethylene naphthalate (PEN)**) ii. PET/Polyamide mesh v. Metal foil-based material (Ti, SS, Zn, W, Fe, Ni)	Act as a transparent window for the incident solar radiation which helps to transfer radiation in the form of visible, infrared and ultraviolet lights to the photochemical cells for maximum participation in photoexcitation.
Transparent conductive oxide (TCO) layer [81]	i. Fluorine-doped tin oxide (FTO) ii. Indium-doped tin oxide (ITO) ii. Aluminium doped ZnO (AZO)	It helps in a significant amount of electron collection, extremely optical transmittance within the visible and infrared solar spectrum for photo participation at Dye, and supports SCO mechanically, chemically, and electrically. It also serves the need for excellent adhesion with transparent substrate and SCO layer.
Blocking layer (BL) [82]	Titanium dioxide, Magnesium oxide, Nickel oxide, Zinc oxide, Niobium pentoxide, and Indium oxide	It provides adhesion between the SCO layer and TCO layer without affecting transmittance and prevents

		recombination of electrons with oxide electrolyte.
Semi-conductive oxide (SCO) layer [83]	i. Nano-structured Titanium dioxide (TiO_2) ii. ZnO, SnO_2, MoS_2	It acts as an anchoring surface for the photoactive dye. It also helps to improve by originating superior light harvesting ability, light scattering ability, wide optical bandgap, efficient electron transport pathway towards TCO, greater charge recombination resistance, etc.
Electrolyte [83]	Iodide-triiodide system	Electrolyte performs the leading protagonist in dye regeneration by transporting electrons from CE to oxidized dye for making it neutral through redox reaction under optimum light absorption.
Counter electrode [84]	Platinum (Pt)	It completes the power generation cycle by collecting electrons from the external circuit and quickly transmitting to the electrolyte through a catalytic reduction reaction.

In DSSCs, the dye is the key element for the generation of solar power, because it controls photon harvesting and electron generation [78]. The dyes used in DSSCs can be classified into two groups namely metal-based inorganic dyes and metal-free organic dyes. The latter types are preferred due to having a low production cost, synthetically feasible, environment friendly, and easy to modify structure [85]. Most of the metal-free organic dyes have Donor – π – Acceptor (D-π-A) type structural configuration in which conjugated π-systems like polyenes and oligothiophenes act as π spacers and have a rod-like configuration for the effective intramolecular charge transfer (ICT) by photoexcitation. The molecular arrangement of these dyes and moieties present in the dye is shown in **Figure 1.5**. The donor units are generally composed of different aromatic moieties like coumarins, triphenylamines, and porphyrins while the acceptor end contains structures like carboxylic acids and cyanoacrylic acids [78, 85]. The organic dyes have lower solar power conversion efficiency (PCE) as compared to the metal-based inorganic dyes due to the poor absorption at red and near-infrared spectrum of solar radiation, charge recombination at semi-conductive oxide layer surface, and aggregation of dyes [85]. In the recent past, different types of structural modifications have been performed to increase the absorption of solar radiation and PCE, like increasing the electron-donating ability of the donor and π-spacer by introducing an electron-donating group or increasing the electron-accepting ability of the acceptor by introducing the electron-withdrawing group or increasing the length of π-spacer [86]. Therefore, by altering the structures, it is possible to generate new dyes with higher PCE values while maintaining the same properties for all other performance-controlling factors. For designing a new dye molecule, a well-known scheme should be developed and checked before the synthesis of the molecule. Different *in-silico* methods are used in the field of material science for designing new dye materials having higher power conversion efficiency compared to the already existing dyes. The *in-silico* methods used for designing new dyes are discussed in detail in the following section.

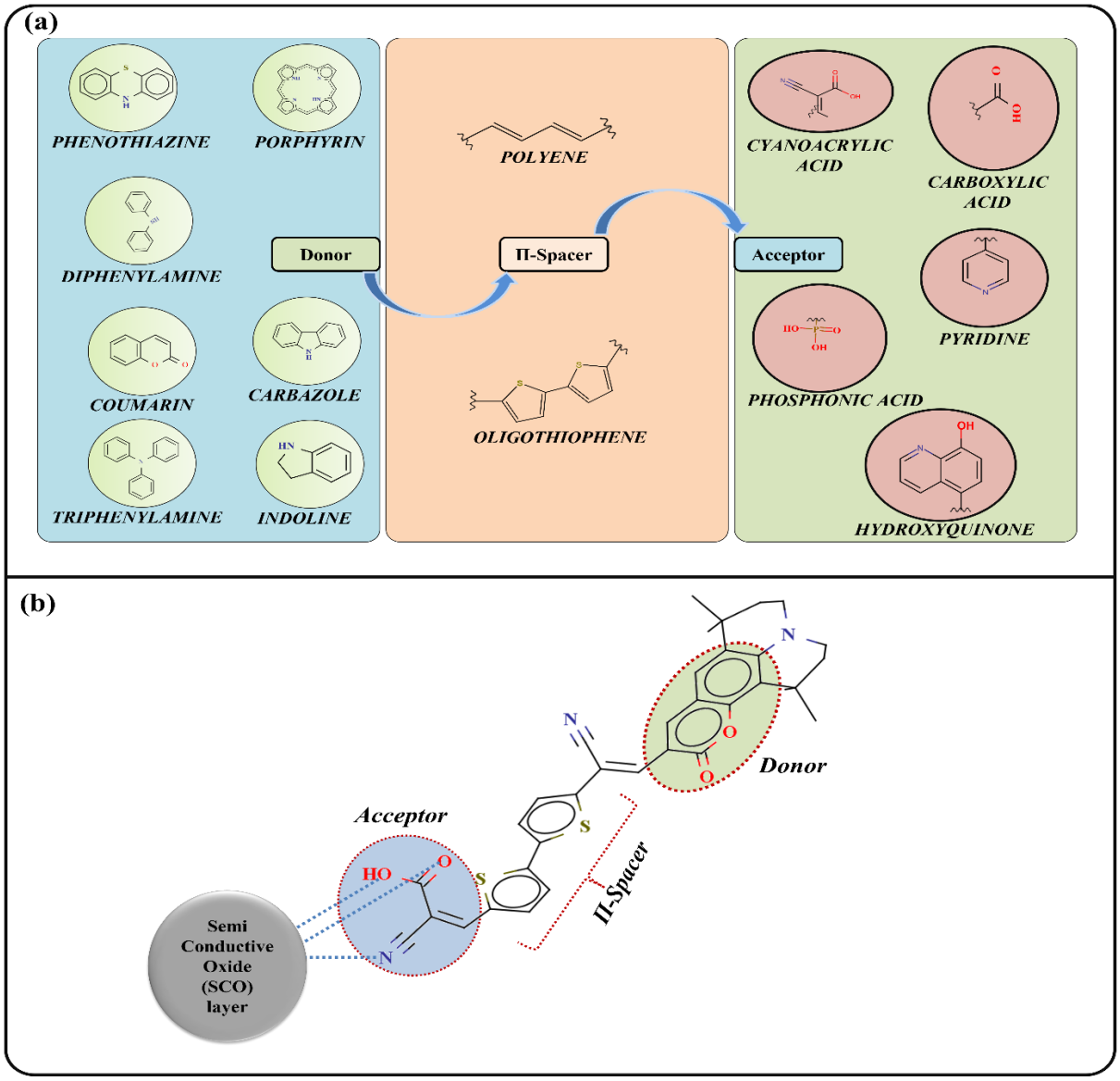


Figure 1.5: (a) Organic moieties present in the structure of the dye (b) Arrangement of the donor, acceptor, and π-spacer moieties in the dye.

1.4 In-silico methods used in the field of materials science

In the recent past, there has been a rise in various *in-silico* methods due to the advancements in computational power and the availability of various software tools. Researchers in the field of materials science have increasingly relied on various computational methods due to the high cost of synthesis of the materials and poor success rate [87]. Materials informatics (MI), a new branch of materials science, has emerged as a result of collaboration between materials science and information science, and this field is expanding with the rapid growth of computer technology

and advancements in mathematical algorithms [88]. MI is a multidisciplinary field that also requires knowledge from various other fields including physics, chemistry, and mathematics. The major goal of MI is to develop a relationship between the properties of materials and molecular structures [89]. One of the prime advantages of MI is the availability of publicly accessible databases that contain information about the properties of materials. By analyzing large databases of material properties, MI researchers try to identify patterns and correlations that can be used to develop predictive models for the behaviour of new materials [90]. MI has the potential to shift the discovery, design, and development of novel materials with defined features for use in several applications, including energy storage [91-94], catalysis [95, 96], and electronics [97, 98]. The traditional methods for materials development involve a lot of trial and error making them time-consuming and expensive. With MI, one can predict the properties of new materials even before their synthesis, which reduces the cost of experimental trials. This not only accelerates the process of discovery and development of materials but also saves time, and resources, and reduces waste [99]. There is a wide array of applications of MI, ranging from improving the property of existing materials to the development of novel materials. Presently, different data-driven methods quantitative structure-property relationship (QSPR), read-across (RA), with or without involvement of machine learning (ML), etc. are being adopted in MI to analyze experimental data and build predictive models. These methods are discussed in detail in the next following section.

1.4.1 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 there is a fundamental relationship between the molecular structure and biological activity. On this assumption, QSAR attempts to establish a correlation between various molecular properties of a set of molecules with 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 [100]. However, the term QSAR can be used in general to denote all these three studies. The QSPR [101] study deals with the molecular features governing their physicochemical properties. The descriptors measure the properties of the molecules that include their hydrophobic, steric, and electronic features in addition to the various structural patterns. The QSTR [102] 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 virtual screening [103] of large numbers of diverse compounds for a definite response parameter. Besides this, the identification of 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 [103] 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 centre of focus for the field of quantitative relationship-based studies. Such correlation represents predictive models derived from the application of statistical tools correlating response data of molecules (including therapeutic activity, property, and toxicity) of chemicals with descriptors representative of molecular structure and/or property [104]. 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 [105]. They showed that biological activity

could be correlated linearly with free-energy-related terms [105], a model referred to as the Linear Free Energy Relationship (LFER) model. The introduction of Hansch's linear and parabolic models made a considerable impact on the understanding of how chemical structures influenced biological activity. The Free-Wilson approach determined the contributions made by various structural fragments to the overall biological activity [106] of the molecules. 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 [107] 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 which activity profile of these molecules. QSAR models are essentially pattern recognition models that identify trends in structural features correlating with the experimental property. QSAR models are useful in several cases like suggesting structural modifications to enhance molecular property. Such quantitative approaches are being applied in many disciplines like risk assessment, toxicity prediction, and regulatory decisions [108] apart from drug discovery and lead optimization [109]. 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 [109].

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. [110]. QSPR models are now regarded as scientifically credible tools for predicting and classifying the properties of untested chemicals. QSPR method has become inevitably an essential tool in different industries, from the discovery of new material with desired properties to the development of that material [111-113]. 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 those

chemicals predicted to have poor outcomes. The generalized workflow for the development of the QSPR model is shown in **Figure 1.6**.

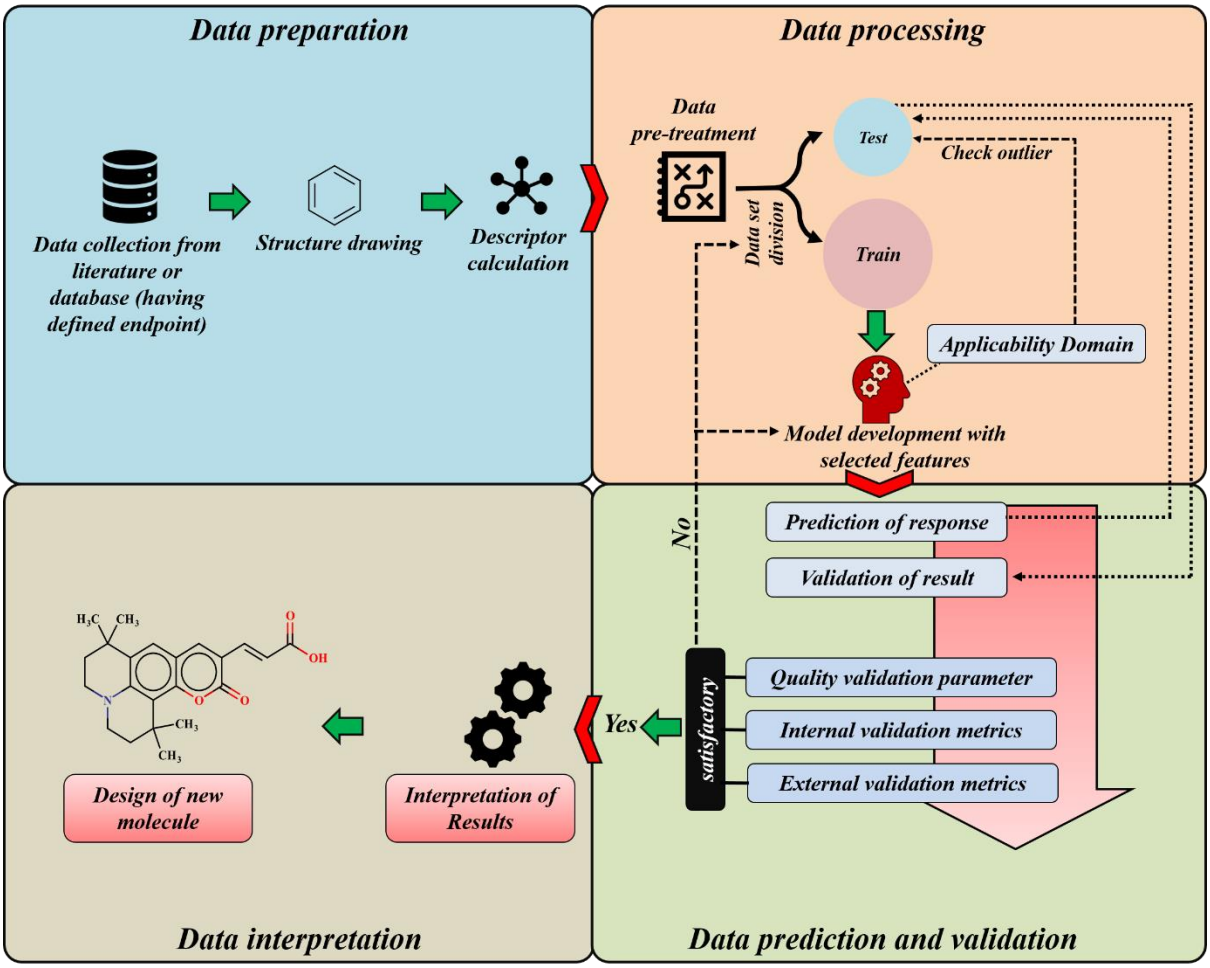


Figure 1.6: General workflow of the QSPR modeling.

1.4.1.1 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 property.

4. Reduction of cost, time, and manpower requirement by the development of more effective compounds using a scientifically less exhaustive approach.

To achieve the aforementioned objectives, 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.

1.4.1.2 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" [114, 115]. 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 the generation of 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 that are 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, certain endpoints may need to take into account 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 [116]. These visualizations are sensible only when descriptor values can be associated with structural features.

1.4.2 Read-Across (RA)

Among the various in-silico approaches, the QSPR method is one of the most popular methods for the development of predictive models. QSPR is a statistical model-building approach that requires a significant amount of data points to build a meaningful model. In addition, the whole dataset needs to be divided into training and test set for validation purpose in order to full fill the requirements as recommended by OECD guideline (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm) [117]. 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 loss of data 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 prediction that involves simple algebraic operations, and no data points are wasted [118].

Read-across is a similarity-based grouping technique that involves simple algebraic operations and uses similarity between two chemical compounds to make predictions [119]. It is a non-experimental data gap-filling method that provide 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 important *in-silico* methods used for data generation, data ,gap-filling and regulatory decision-making [120]. Read-across method can be classified into two groups, one is qualitative read-across and the other is quantitative read-across [121]. 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 [122]. 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 [118], which is calculated by using the following equation:

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

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

1.4.3 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 kind of 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. [123] who developed the classification-based RASAR models whereas Banerjee and Roy [124] were the first to develop

the regression-based quantitative RASAR (q-RASAR) models. The RASPR method is a combined method of read-across and QSPR that encapsulates the advantages of both of these methods and generates enhance predictivity. In this method, selected structural and physicochemical descriptors are used to generate different similarity and error-based measures (known as RASPR descriptors) from the similarity-based read-across approach [125]. 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** and the general working process behind the RASPR is shown in **Figure 1.7**.

Table 1.2: Definition of RASPR descriptors

Descriptors	Description	Mathematical equation
RA function	It is a read-across weighted average prediction value for a target compound that is obtained based on the similarity between selected close source compounds and the target (or query) compound.	$RA\ function = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ $w_i = \frac{S_i}{\sum_{i=1}^{n} S_i}$ w_i = weightage given to individually selected close source compounds S_i= Similarity between individual selected close source compounds and target compounds x_i=observed response value of the selected close source compounds
SD Activity ($S_{weighted}$)	It is the standard deviation of the observed response values of the selected close source compounds for each query compound.	$S_{weighted} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_{wtd}})^2}{\sum_{i=1}^{n} w_i}} \times \frac{n}{n-1}$ $x_{wtd} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ n = effective number of selected close source compound

		w_i = weightage given to individual selected close source compounds
SE	It is the standard error of the observed response value of the selected close source compounds for each query compound.	$SE = \frac{SD\ Activity}{\sqrt{n}}$
CVact ($CV_{activity}$)	It is the coefficient of variation of the observed response (or activity) values of the selected close source compounds for each query compound.	$CV_{activity} = \frac{S_{weighted}}{\overline{x_{wtd}}}$
CVsim ($CV_{similarity}$)	It is the coefficient of variation of the similarity values of the selected close source compounds for each query compound.	$CV_{similarity} = \frac{SD\ Similarity}{\bar{f}}$ $\bar{f}$= average similarity value of the selected close source compounds
MaxPos	Maximum similarity level to the positive close source compounds based on the mean value of training set observed response.	
MaxNeg	Maximum similarity level to the negative close source compounds based on the mean value of training set observed response.	
Abs MaxPos-MaxNeg (Abs.Diff.)	The absolute difference between MaxPos and MaxNeg.	$AbsDiff = \lvert MaxPos - MaxNeg \rvert$
Avg.Sim. (Average Similarity)	It is the mean of similarity values of the selected close	$Avg.Sim.(\bar{f}) = \frac{\sum_{i=1}^{n} f_i}{n}$

	source compounds for each query compound.	f_i = similarity values of the selected close source compounds n=number of selected close source compounds
SD Similarity	It is the standard deviation of the similarity values of selected close source compounds for each query compound.	$SD\ Similarity = \sqrt{\frac{\sum_{i=1}^{n}(f-\bar{f})^2}{n-1}}$ $\bar{f}$= average similarity value of the selected close source compounds n=number of selected close source compounds
g_m *[Banerjee-Roy Coefficient]*	A novel concordance measure	$g_m = (-1)^n \lvert Posfrac - 0.5 \rvert$ n=1 when MaxPos<MaxNeg n=2 when MaxPos>=MaxNeg Posfrac=Fraction of the close source compounds having a response value greater than the training set mean response
g_m**Avg.Sim*	Product of the values of g_m and Average Similarity	
g_m**SD Similarity*	Product of the values of g_m and SD Similarity	
Pos.Avg.Sim	Average similarity value of the selected positive close source compounds based on the mean observed response value of the training set	
Neg.Avg.Sim	Average similarity value of the selected negative close source compounds based on the mean	

	observed response value of the training set	

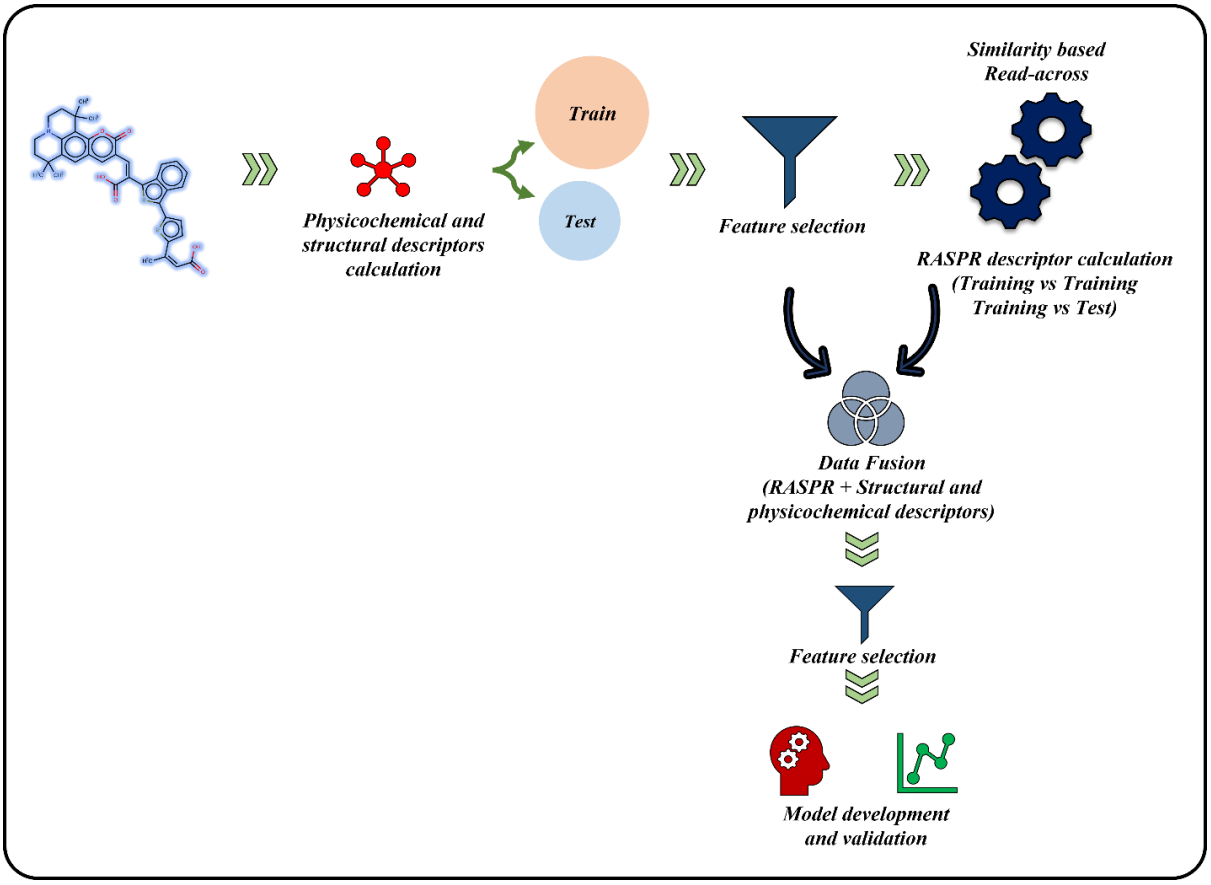


Figure 1.7: Schematic representation of the RASPR algorithm

1.4.4 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 [126]. 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 labeled 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 [127]. Currently, ML algorithms have moved beyond purely theoretical applications to practical applications like the creation of new molecules [128]. 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 [127]. 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 [129]. SHAP or Shapley additive explanation analysis is one of the important methods used for the interpretation of the ML models [130].

Chapter – 2

PRESENT WORK

2. PRESENT WORK

The limited resources and cost involved in experimentation slow down the process of development of new chemicals. The development of a dye to be used in a dye-sensitized solar cell requires lots of trial and error in experimentation resulting in loss of money, effort, and resources. In this scenario, different molecular modeling techniques are being increasingly used as an alternative to experimentation where it is possible. In the last 20 years, the advent of computational tools such as quantitative structure-property relationship (QSPR) has provided significant insight into materials science [130]. These are simply defined as mathematical relationships linking a compound's property with its chemical structure in a qualitative/quantitative manner. All the QSPR models are developed according to the guideline specified by the Organization for Economic Cooperation and Development (OECD) (https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm) [100]. For small data sets, different similarity-based approaches are used for the development of predictive models. Read-across is one of the important similarity-based methods used for data generation and data gap-filling [118]. Another similarity-based approach read-across structure-property relationship (RASPR) is used for modeling purposes. RASPR is a combination of read-across and QSPR where read-across-based similarity and error measures are used as descriptors (also known as RASPR descriptors) [124].

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

In this present study, we have used RASPR descriptors for QSPR modeling of power conversion efficiency (PCE) of organic dyes used in dye-sensitized solar cells (DSSCs). RASPR descriptors are similarity-based read-across similarity and error-based measures. This method uses non-linear similarity functions like Gaussian kernel or Laplacian kernel for the calculation of similarity between two chemicals [118]. Here, we found that the incorporation of RASPR descriptors helps to improve the external predictivity of the developed model where internal validation metrics may or may not improve due to the Leave-same-out (LSO) algorithm of RASPR.

2.1 Study 1

The dye-sensitized solar cell (DSSC) is a 3rd generation photovoltaic solar cell containing photoactive dyes that generate electrons when solar radiation is incident on the solar cell. The dye is the main component of DSSC because it controls photon harvesting and electron generation that depends on the molecular structure of the dyes [52].

In this study, we have adopted a novel quantitative Read-Across Structure-Property Relationship (q-RASPR) approach, which is analogous to the Quantitative Read-Across Structure-Activity Relationship (q-RASAR) first reported by Banerjee and Roy to generate different predictive models. The RASPR method is used for modeling of the power conversion efficiency (PCE) of different classes of organic dyes used in DSSCs. Here we have taken a total of 429 organic dyes distributed across 4 different classes namely – coumarins (56 compounds), diphenylamines (35 compounds), carbazoles (179 compounds), and indolines (159 compounds). The experimental PCE value of the dyes and descriptor values were collected from the work of Krishna et al. [131]. In this study, we have maintained the same composition of the training and test sets to ensure that the chemical compounds span the same chemical space which would help in better comparison with the previous work.

The collected physicochemical and structural descriptors of the dyes were used for the calculation of the RASPR descriptor at different hyper-parameter setting which is obtained by cross-validation of the training set. Compound-specific similarity and error-based measures were used as RASPR descriptors and combined with the initially selected features to generate different predictive models. Different ML approaches in the form of Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (or XGBoosting), Support Vector Machine (SVM), Linear Support Vector Machine (Linear SVM), Ridge Regression (RR) and Partial Least Squares (PLS) models were adopted to predict the PCEs of organic dye based DSSCs. For all the developed models it was found that the models with RASPR descriptor(s) supersede the previously developed QSPR model developed using the physicochemical and structural descriptors.

2.2 Study 2

In this study, we have used the q-RASPR method for the modeling of PCE of organic dyes for three relatively large datasets containing different classes of compounds namely – phenothiazines (215 compounds), porphyrins (300 compounds) and triphenylamines (244 compounds). Here, we have divided the datasets into training and test sets with three different random combinations and developed QSPR and RASPR models initially with the partial least squares (PLS) algorithm. These QSPR and RASPR models were used for the prediction of the test set compounds to show that the external predictivity of the models improves for the RASPR method compared to the QSPR method. We have further modeled these datasets using different ML methods like the RF, GB, adaptive boosting (ADB), XGBoosting, linear SVM, SVM, and RR and compared the performance of the QSPR models with the corresponding RASPR models.

The details of the different datasets are represented in **Table 2.3**.

Table 2.3: Description of the datasets

Dataset	Study	Number of compounds	PCE range	Reference
Coumarins	**1**	56	0.33-7.4	[131]
Diphenylamines		35	0.4-8	
Carbazoles		179	0.038-12.5	
Indolines		159	0.046-9.2	
Phenothiazines	**2**	215	0.12-8.18	
Porphyrins		300	0.0013-12.5	
Triphenylamines		244	0.053-10.1	

Chapter – 3

MATERIALS AND METHODS

3. MATERIALS AND 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 in the following parts:

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

3.1 Details of datasets consisting of PCE data

3.1.1 Coumarin dataset (Study 1)

The dataset consists of 56 compounds, among which 42 compounds are present in the training set and 14 compounds are present in the test set. The detail of this dataset is shows in **Table 3.1**.

Table 3.1: Details of the coumarin dataset

Serial Number	Compound ID	SMILE	Experimental PCE
1	**1**	CCN([C@@H]1C=Cc2c(C1)oc(=O)c(c2)c1ccc(cc1)c1ccc(s1)/C=C(/[C@H]1SC(=S)N(C1=O)CC(=O)O)\[C@H]1SC(=S)N(C1=O)CC(=O)O)CC	**1.39**
2	**2**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(cc1)c1ccc(s1)C/C=C(\C(=O)O)/C#N)CC	**3.62**
3	**3**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(cc1)c1ccc(cc1)/C=C(\[C@H]1SC(=S)N(C1=O)CC(=O)O)/[C@H]1SC(=S)N(C1=O)CC(=O)O)CC	**1.77**
4	**4***	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(cc1)c1ccc(cc1)/C=C(/C(=O)O)\C#N)CC	**2.8**

5	**5***	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(cc1)/C=C(\[C@H]1SC(=S)N(C1=O)C(=O)O)/[C@H]1SC(=S)N(C1=O)CC(=O)O)CC	**1.41**
6	**6**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(cc1)/C=C(\C(=O)O)/C#N)CC	**2.6**
7	**7**	OC(=O)C[n+]1c(/C=C/c2cc3cc4c5c(c3oc2=O)C(C)(C)CCN5CCC4(C)C)sc2c1cccc2	**1.1**
8	**8***	OC(=O)c1ccc2c(c1)C(C)(C)C(=[N+]2C)/C=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C	**0.6**
9	**9**	OC(=O)CN1C(=S)S/C(=C/c2cc3cc4c5c(c3oc2=O)C(C)(C)CCN5CCC4(C)C)/C1=O	**2.6**
10	**10**	OC(=O)C(=C/C=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)C(=O)O	**3.7**
11	**11**	N#C/C(=C/C=C/C=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**3.5**
12	**12**	N#C/C(=C/C=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**5.2**
13	**13***	CCN(c1ccc2c(c1)oc(=O)c(c2)/C=C/C=C(\C(=O)O)/C#N)CC	**4.7**
14	**14**	N#C/C(=C\c1c(=O)oc2c(c1C(F)(F)F)cc1c3c2C(C)(C)CCN3CCC1(C)C)/C(=O)O	**3.1**
15	**15**	N#C/C(=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**4.1**
16	**16**	OC(=O)/C=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C	**3.4**
17	**17**	OC(=O)c1cc2cc3CCCN4c3c(c2oc1=O)CCC4	**0.9**
18	**18***	N#C/C(=C\c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**5.8**
19	**19**	N#C/C(=C/c1ccc(s1)c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**7.4**
20	**20**	N#C/C(=C/c1ccc(s1)c1ccc(s1)c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**6.4**
21	**21**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1csc(n1)c1ccc(s1)C=C(C(=O)O)C#N)CC	**4.78**

22	**22**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1csc(n1)c1ccc(s1)/C=C\1/SC(=S)N(C1=O)CC(=O)O)CC	**0.33**
23	**23**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(cc1)c1ccc(s1)C=C(C(=O)O)C#N)CC	**3.62**
24	**24**	OC(=O)c1ccc(cc1)c1ccccc1c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C	**1.04**
25	**25**	OC(=O)c1ccc(cc1)c1cccc(c1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C	**2.33**
26	**26**	OC(=O)c1ccc(cc1)c1ccc(cc1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C	**2.69**
27	**27**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**4.82**
28	**28**	N#C/C(=C\c1sc(c2c1OCCO2)c1ccc(c2c1nsn2)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**3.93**
29	**29**	N#C/C(=C\c1sc(c2c1OCCO2)c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**6.07**
30	**30***	N#C/C(=C\c1ccc(cc1)c1sc(c2c1OCCO2)c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**3.73**
31	**31**	N#C/C(=C\c1ccc(s1)c1sc(c2c1OCCO2)c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**4.37**
32	**32**	N#C/C(=C/c1ccc(s1)c1ccc(s1)/C(=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C#N)/C(=O)O	**6.5**
33	**33**	N#C/C(=C\c1ccc(s1)/C=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**5.8**
34	**34**	N#C/C(=C\c1ccc(s1)/C(=C/c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C#N)/C(=O)O	**5.3**
35	**35**	N#C/C(=C\1/C=C(/C=C/c2cc3cc4c5c(c3oc2=O)C(C)(C)CCN5CCC4(C)C)CC(C1)(C)C)/C(=O)O	**6.2**
36	**36**	N#C/C(=C/c1ccc(s1)/C=C/c1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**5.5**
37	**37**	CCCCn1c2ccc(cc2c2c1cccc2)c1cc(c(s1)c1cc2cc3c4c(c2oc1=O	**5.53**

		)C(C)(C)CCN4CCC3(C)C)/C=C(\C(=O)O)/C#N	
38	**38**	CCCCC1(CCCC)c2ccccc2c2c1cc(cc2)c1cc(c(s1)c1cc2cc3C(C)(C)CCN4c3c(c2oc1=O)C(C)(C)CC4)/C=C(/C(=O)O)\C#N	**4.02**
39	**39**	N#C/C(=C/c1cc(sc1c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)C)/C(=O)O	**2.74**
40	**40***	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1cc2ccc(cc2oc1=O)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.59**
41	**41***	N#C/C(=C\c1ccc(cc1)c1ccc(s1)c1cc2ccc(cc2oc1=O)N(c1ccccc1)c1ccccc1)/C(=O)O	**6.24**
42	**42***	CCN(c1ccc2c(c1)OC(=O)[C@H](C2)C(=O)O)CC	**1.03**
43	**43***	N#C/C(=C\c1ccc(s1)c1cc2ccc(cc2oc1=O)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.36**
44	**44**	N#C/C(=C/c1ccc(s1)c1ccc(c2c1nsn2)C#Cc1ccc(s1)c1cc2cc3c4c(c2oc1=O)C(C)(C)CCN4CCC3(C)C)/C(=O)O	**1.35**
45	**46**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(c(c1)F)/C=C(\C(=O)O)/C#N)CC	**5.2**
46	**47**	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(c(c1)Cl)/C=C(\C(=O)O)/C#N)CC	**4.1**
47	**48***	CCN(c1ccc2c(c1)oc(=O)c(c2)c1ccc(c(c1)Br)/C=C(\C(=O)O)/C#N)CC	**3.5**
48	**49***	CCO[Si](c1ccc2c(c1)nc(o2)c1cc2ccc(cc2oc1=O)N(CC)CC)(OCC)OCC	**0.72**
49	**50**	N#CC(=Cc1ccc(s1)c1cc2ccc(cc2oc1=O)N(c1ccc2c(c1)Cc1c2ccc1)c1ccc2c(c1)Cc1c2cccc1)C(=O)O	**4.7**
50	**51***	N#CC(=Cc1ccc(s1)c1ccc(s1)c1cc2ccc(cc2oc1=O)N(c1ccc2c(c1)Cc1c2cccc1)c1ccc2c(c1)Cc1c2cccc1)C(=O)O	**5.06**
51	**52**	N#CC(=Cc1ccc(cc1)c1ccc(s1)c1cc2ccc(cc2oc1=O)N(c1ccc2c(c1)Cc1c2cccc1)c1ccc2c(c1)Cc1c2cccc1)C(=O)O	**5.94**
52	**53***	CCN(c1ccc2c(c1)oc(=O)c(c2C)C#CC=C(C(=O)O)C#N)CC	**2.21**
53	**54**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)/C=C/c1ccc(s1)[Si](OCC)(OCC)OCC)c1c(=O)oc2c(c1C)ccc(c2)N(CC)CC	**3.5**

54	**55**	CCCCCCCCCCOc1cc2cc(C#Cc3ccc(cc3)/C=C(/C(=O)O)\C#N)c(=O)oc2cc1OCCCCCCCCCC	**2**
55	**56**	N#CC(=Cc1ccc(cc1)C#Cc1cc2cc(OC)c(cc2oc1=O)OC)C(=O)O	**0.99**
56	**57**	N#CC(=Cc1ccc(cc1)/C=C/c1ccc(cc1)/C=C/c1cc2cc(OC)c(cc2oc1=O)OC)C(=O)O	**1.37**

'*' represent the test set compound

3.1.2 Diphenylamine dataset (Study 1)

This dataset consists of 35 compounds, among which 25 compounds are present in the training set and 10 compounds are present in the test set. The detail of this dataset is shows in **Table 3.2**.

Table 3.2: Details of the diphenylamine dataset

Serial Number	Compound ID	SMILE	Experimental PCE
1	**1***	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(s1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.75**
2	**2**	CCCCCCOc1ccc(cc1)N(c1ncc(s1)c1sc2c(c1)c1cc(sc1c(c2OCCCCCC)OCCCCCC)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC	**5.1**
3	**3**	CCCCCCOc1ccc(cc1)N(c1ccc(s1)c1sc2c(c1)c1cc(sc1c(c2OCCCCCC)OCCCCCC)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC	**5.4**
4	**4***	CCCCCCOc1ccc(cc1)N(c1ncc(s1)c1sc2c(c1)C(c1c2sc(c1)/C=C(/C(=O)O)\C#N)(CCCCCC)CCCCCC)c1ccc(cc1)OCCCCCC	**4.8**
5	**5***	CCCCCCOc1ccc(cc1)N(c1ccc(s1)c1sc2c(c1)C(c1c2sc(c1)/C=C(/C(=O)O)\C#N)(CCCCCC)CCCCCC)c1ccc(cc1)OCCCCCC	**6.6**
6	**6***	N#CC(=Cc1cnc(nc1)c1ccc(s1)N(c1ccccc1)c1ccccc1)C(=O)O	**6.72**
7	**7**	N#CC(=Cc1cnc(nc1)c1ccc(s1)N(c1ccc(cc1)OC)c1ccc(cc1)OC)C(=O)O	**7.05**
8	**8**	CCCCCCOc1ccc(cc1)N(c1ccc(s1)c1ncc(cn1)C=C(C(=O)O)C#N)c1ccc(cc1)OCCCCCC	**7.64**
9	**9***	N#C/C(=C\c1cccc(c1)C#Cc1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N(c1ccccc1)c1ccccc1)/C(=O)O	**0.61**

10	**10**	N#C/C(=C\c1ccc(s1)C#Cc1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N(c1ccccc1)c1ccccc1)/C(=O)O	**1.99**
11	**11***	CCCCCCN(c1ccc(cc1)/C=C(/c1ccc(cc1)c1ccc(cc1)/C=C(\C(=O)O)/C#N)\C#N)c1ccc(cc1)/C=C(/c1ccc(cc1)c1ccc(cc1)/C=C(\C(=O)O)/C#N)\C#N	**1.02**
12	**12***	CCCCCCN(c1ccc(cc1)/C=C(/c1ccc(cc1)c1ccc(cc1)/C=C/1\SC(=S)N(C1=O)CC(=O)O)\C#N)c1ccc(cc1)/C=C(/c1ccc(cc1)c1ccc(cc1)/C=C\1/SC(=S)N(C1=O)CC(=O)O)\C#N	**0.56**
13	**13**	CCCCCCN(c1ccc(cc1)/C=C(/c1ccc(s1)/C=C(/C(=O)O)\C#N)\C#N)c1ccc(cc1)/C=C(/c1ccc(s1)/C=C(/C(=O)O)\C#N)\C#N	**3.16**
14	**14**	N#C/C(=C\c1ccc(cc1)C#Cc1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3**
15	**15**	CCCCCCCCOc1c(OCCCCCCCC)c(c2ccc(s2)N(c2ccccc2)c2ccccc2)c2c(c1c1ccc(cc1)/C=C(\C(=O)O)/C#N)nsn2	**6.66**
16	**16***	CCCCCCCCOc1c(OCCCCCCCC)c(c2ccc(s2)N(c2ccccc2)c2ccccc2)c2c(c1c1ccc(o1)/C=C(\C(=O)O)/C#N)nsn2	**7.16**
17	**17**	CCCCCCCCOc1c(OCCCCCCCC)c(c2ccc(s2)N(c2ccccc2)c2ccccc2)c2c(c1c1ccc(s1)/C=C(\C(=O)O)/C#N)nsn2	**6.19**
18	**18**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(s1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	**3.74**
19	**19**	CCCCCCCCCCCCN(c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.52**
20	**20**	CCCCCCCCCCCCN(c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(cc1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.1**
21	**21**	CCCCCCCCCCCCN(c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(cc1)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.12**
22	**22**	N#CC(=Cc1ccc(cc1)c1ccc(s1)N(c1ccccc1)c1ccccc1)C(=O)O	**5.22**
23	**23**	CCCCCCN(c1ccc(cc1)/C=C(/c1ccc(s1)C=C(C(=O)O)C(=O)O)\C#N)c1ccc(cc1)/C=C(/c1ccc(s1)C=C(C(=O)O)C(=O)O)\C#N	**4.1**
24	**24***	CCCCc1ccc(cc1)N(c1ccc2c(c1)C(c1ccccc1)(c1ccccc1)c1c2ccc(c	**4.26**

		1)c1ccc(cc1)C=C(C(=O)O)C#N)c1ccc(cc1)CCCC	
25	**25**	CCCCCCC1(CCCCCC)c2cc(ccc2c2c1cc(cc2)N(c1ccc(cc1)CCCC)c1ccc(cc1)CCCC)c1ccc(cc1)C=C(C(=O)O)C#N	**5.15**
26	**26**	CCCCCCOc1ccc(cc1)N(c1ccc(c2c1nsn2)c1sc2c(c1)C(c1c2sc(c1)/C=C(/C(=O)O)\C#N)(CCCCCC)CCCCCC)c1ccc(cc1)OCCCCCC	**7.1**
27	**27**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc(c2c1nsn2)N(c1ccc(cc1)c1ccc(cc1OCCCC)OCCCC)c1ccc(cc1)c1ccc(cc1OCCCC)OCCCC)/C=C(/C(=O)O)\C#N	**8**
28	**28**	N#C/C(=C\c1ccc(o1)C1=Nc2c(C1(C)C)cc(cc2)N(c1ccccc1)c1ccccc1)/C(=O)O	**2.8**
29	**29**	N#C/C(=C\c1cnc(s1)C1=Nc2c(C1(C)C)cc(cc2)N(c1ccccc1)c1ccccc1)/C(=O)O	**1.99**
30	**30**	N#C/C(=C\c1ccc(cc1)C1=Nc2c(C1(C)C)cc(cc2)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.09**
31	**31**	CCCCc1ccc(cc1)N(c1ccc2c(c1)Cc1c2ccc(c1)c1ccc(cc1)C=C(C(=O)O)C#N)c1ccc(cc1)CCCC	**4.62**
32	**32***	CCCCCCCCN(c1ccccc1)c1ccc(cc1)C=C1C(=O)NC(=O)NC1=O	**1.53**
33	**33**	CCCCCCCCN(c1ccccc1)c1ccc(cc1)C=C1C(=O)NC(=S)NC1=O	**0.44**
34	**34**	CCCCCCCCC1(CC=C(C=C1)NC(=C1C(=O)NC(=O)NC1=O)c1ccccc1)C=C1C(=O)NC(=O)NC1=O	**1**
35	**35**	CCCCCCCCN(c1ccc(cc1)C=C1C(=O)NC(=S)NC1=O)c1ccc(cc1)C=C1C(=O)NC(=S)NC1=O	**0.4**

'*' represent the test set compound

3.1.3 Carbazole dataset (Study 1)

This dataset consists of 179 compounds, among which 125 compounds are present in the training set and 54 compounds are present in the test set. The detail of this dataset is shows in **Table 3.3**.

Table 3.3: Details of the carbazole dataset

Serial Number	Compound ID	SMILE	Experimental PCE

1	**1***	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/C(=O)O	**2.48**
2	**2**	N#C/C(=C/c1ccc(s1)c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/C(=O)O	**2.39**
3	**3**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc2c(c1)n1c3ccc(cc3c3c1c2cc(c3)C(C)(C)C)C(C)(C)C)/C(=O)O	**3.96**
4	**4**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)n1c3ccc(cc3c3c1c2cc(c3)C(C)(C)C)C(C)(C)C	**2.85**
5	**5**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**5.02**
6	**6**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**5.15**
7	**7**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)c1ccccc1n2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)C(=O)O	**3.87**
8	**8**	N#CC(=Cc1ccc(s1)C=Cc1ccc(s1)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**3.76**
9	**9**	CCCCCCCOc1ccc2c(c1)n(CC)c1c2cc(cc1)c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)/C=C(\C(=O)O)/C#N	**4.8**
10	**10***	CCCCCCCOc1ccc2c(c1)n(CC)c1c2cc(cc1)c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)/C=C(\C(=O)O)/C#N	**4.7**
11	**11**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N)c1sc(cc1CCCCCC)c1ccc2c(c1)c1ccccc1n2CC	**5**
12	**12**	OC(=O)CN1C(=S)SC(=Cc2ccc(s2)c2ccc3c(c2)c2ccccc2n3c2ccc3c(c2)C(C)(C)c2c3cccc2)C1=O	**1.79**
13	**13**	OC(=O)CN1C(=S)SC(=Cc2ccc(s2)c2ccc(s2)c2ccc3c(c2)c2ccccc2n3c2ccc3c(c2)C(C)(C)c2c3cccc2)C1=O	**0.55**
14	**14**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)C=C(C(=O)O)C#N)c1ccc2c(c1)c1ccccc1n2CC	**5.1**
15	**15**	CCCCCCc1cc(sc1c1cc(c(s1)C=C(C(=O)O)C#N)CCCCCC)c1ccc2c(c1)c1ccccc1n2c1ccccc1	**5.3**

16	**16**	CCCCCCc1cc(sc1c1cc(c(s1)C=C(C(=O)O)C#N)CCCCCC)c1ccc2c(c1)c1ccccc1n2c1ccc(cc1)OC	**5.4**
17	**17**	N#CC(=Cc1ccc(s1)c1ccc(c2c1nsn2)c1ccc(cc1)n1c2ccccc2c2c1cccc2)C(=O)O	**1.56**
18	**18**	N#CC(=Cc1ccc(s1)c1ccc(c2c1nsn2)c1ccc2c(c1)c1ccccc1n2c1ccccc1)C(=O)O	**3.29**
19	**19**	CCCCC(Cn1c2ccc(cc2c2c1cccc2)c1ccc(c2c1nsn2)c1ccc(s1)C=C(C(=O)O)C#N)CC	**3.8**
20	**20**	CCCCC(COc1ccc(cc1)n1c2ccc(cc2c2c1cccc2)c1ccc(c2c1nsn2)c1ccc(s1)C=C(C(=O)O)C#N)CC	**2.36**
21	**21***	N#CC(=Cc1ccc(s1)c1ccc(c2c1nsn2)c1ccc2c(c1)n(c1ccccc1)c1c2cccc1)C(=O)O	**1.53**
22	**22***	CCCCC(COc1ccc(cc1)n1c2ccccc2c2c1cc(cc2)c1ccc(c2c1nsn2)c1ccc(s1)C=C(C(=O)O)C#N)CC	**1.2**
23	**23***	N#CC(=Cc1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C(=O)O	**1.65**
24	**24***	N#CC(=Cc1ccc(cc1)n1c2ccccc2c2c1cccc2)C(=O)O	**2.81**
25	**25**	CCCCC(COc1c2c3sc(cc3sc2c2c(c1OCC(CCCC)CC)c1sc(cc1s2)/C=C(\C(=O)O)/C#N)c1ccc(c2c1nc(c1csc(c1)C)c(n2)c1ccc(s1)C)c1ccc2c(c1)c1ccccc1n2CC(CCCC)CC)CC	**4.63**
26	**26**	CCCCC(COc1c2c3sc(cc3sc2c2c(c1OCC(CCCC)CC)c1sc(cc1s2)/C=C(\C(=O)O)/C#N)c1ccc(c2c1nsn2)c1ccc2c(c1)c1ccccc1n2CC(CCCC)CC)CC	**5.46**
27	**27**	CCCCC(COc1c2c3sc(cc3sc2c2c(c1OCC(CCCC)CC)c1sc(cc1s2)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)n(CC(CCCC)CC)c1c2cc(cc1)c1sc2c(c1)sc1c2c(OCC(CCCC)CC)c(OCC(CCCC)CC)c2c1sc1c2sc(c1)/C=C(\C(=O)O)/C#N)CC	**5.34**
28	**28**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)/C=C(/P(=O)(O)O)\C#N)c1sc(cc1CCCCCC)c1ccc2c(c1)c1ccccc1n2CC	**3.4**
29	**29**	CCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/C(=O)O)\C#N)/C=C(/C(=O)O)\C#N	**1.54**

30	**30**	CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccc(cc1)OC)c1ccccc1)N(c1ccc(cc1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)OC	**2.54**
31	**31***	CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccc(cc1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)OC)N(c1ccc(cc1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)OC	**3.23**
32	**32**	N#CC(=Cc1ccc(cc1)C#Cc1ccc(cc1)n1c2ccccc2c2c1cccc2)C(=O)O	**0.695**
33	**33**	CCCCC(COc1c2c3sc(cc3sc2c2c(c1OCC(CCCC)CC)c1sc(cc1s2)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)c1ccccc1n2CC(CCCC)CC)CC	**4.87**
34	**34**	CCCCCCCCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C	**3.64**
35	**35**	CCCCCCCCCCCCn1c2ccc(cc2c2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.8**
36	**36**	CCCCCCCCCCCCn1c2ccc(cc2c2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.69**
37	**37**	CCCCCCCCCCCCn1c2ccc(cc2c2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(s1)c1ccc(s1)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**4.82**
38	**38***	N#C/C(=C/c1ccc(s1)c1ccc(s1)c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/C(=O)O	**4.62**
39	**39**	N#CC(=Cc1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)n1c2ccccc2c2c1cccc2)C(=O)O	**4.3**
40	**40**	N#CC(=Cc1ccc(s1)C=Cc1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C(=O)O	**5.22**
41	**41**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)/C=C(\C(=O)O)/C#N)c1sc(cc1CCCCCC)c1ccc2c(c1)c1ccccc1n2CC	**3.9**
42	**42**	COCCCCc1cc(sc1c1sc(c(c1)CCCCOC)c1sc(c(c1)CCCCOC)/C=C(/C(=O)O)\C#N)c1sc(cc1CCCCOC)c1ccc2c(c1)c1ccccc1n2CC	**2.9**
43	**43**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N)c1sc	**4**

		(cc1CCCCOC)c1sc(cc1CCCCOC)c1ccc2c(c1)c1ccccc1n2CC	
44	**44**	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.45**
45	**45**	N#CC(=Cc1ccc(s1)C#Cc1ccc(cc1)n1c2ccccc2c2c1cccc2)C(=O)O	**0.819**
46	**46**	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.05**
47	**47**	N#CC(=Cc1ccc2c(c1)c1ccccc1n2CC)C(=O)O	**2.43**
48	**48**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)C=C1SC(=S)N(C1=O)CC(=O)O)C=C1SC(=S)N(C1=O)CC(=O)O	**2.81**
49	**49***	CCCCCCc1cc(sc1/C=C(\C(=O)O)/C#N)c1sc2c(c1)c1nn(nc1c1c2sc(c1)c1ccc2c(c1)n(CC(CCCC)CC)c1c2cccc1)CC(CCCC)CC	**8.07**
50	**50**	CCCCC(Cn1nc2c(n1)c1cc(sc1c1c2cc(s1)c1ccc2c(c1)n(CC(CCCC)CC)c1c2cccc1)c1ccc(s1)/C=C(\C(=O)O)/C#N)CC	**7.52**
51	**51***	N#C/C(=C/c1ccc(cc1)/C=C/c1ccc2c(c1)c1ccccc1n2c1ccc(cc1)C)/C(=O)O	**4.5**
52	**52**	CCCCOCc1cc(sc1c1sc(c(c1)CCCCCC)c1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N)c1sc(cc1COCCCC)c1ccc2c(c1)c1ccccc1n2CC	**4**
53	**53**	CCCCn1c2cc(ccc2c2c1cc(cc2)/C=C(/C(=O)O)\C#N)n1c2ccccc2c2c1cccc2	**0.99**
54	**54**	CCCCn1c2cc(ccc2c2c1cc(cc2)n1c2ccccc2c2c1cccc2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.22**
55	**55**	CCCCn1c2cc(ccc2c2c1cc(cc2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.95**
56	**56**	CCCCn1c2cc(ccc2c2c1cc(cc2)n1c2ccccc2c2c1cccc2)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.04**
57	**57**	CCCCn1c2cc(ccc2c2c1cc(cc2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.48**
58	**58***	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)c1ccc2c(c1)c1ccccc1n2CC)c1sc(cc1CCCCCC)c1sc(cc1CCCCCC)/C=C(\C(=O)O)/C#N	**7.7**
59	**59***	N#C/C(=C/c1ccc(s1)c1ccc(s1)c1ccc(s1)c1ccc2c(c1)c1ccccc1n	**4.8**

		2CC)/C(=O)O	
60	**60**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)c1ccc2c(c1)c1ccccc1n2CC)c1sc(cc1CCCCCC)/C=C(\C(=O)O)/C#N	**5.6**
61	**61***	N#CC(=Cc1ccc(cc1)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**2.35**
62	**62**	CCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccncc1)c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1cccs1)c1ccc(cc1)c1cccs1	**1.88**
63	**63**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C(=O)O)C#N)c1ccc(cc1)C=C(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)n1c2ccccc2c2c1cccc2	**5.51**
64	**64***	CCCCCCn1c2ccc(cc2c2c1ccc(c2)C=C(C(=O)O)C#N)c1ccc(cc1)C=C(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)n1c2ccccc2c2c1cccc2	**2.14**
65	**65**	CCCCCCCCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(c2c1nsn2)c1ccc2c(c1)n(CCCCCCCCCCCC)c1c2cc(cc1)/C=C(/C(=O)O)\C#N)/C=C(/C(=O)O)\C#N	**4.35**
66	**66**	CCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1sc(cc1c1ccc(cc1)OC)c1ccc(cc1)OC	**1.44**
67	**67**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)OC	**4.57**
68	**68***	CCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)OC	**2.52**
69	**69**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(cc1)OC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.54**
70	**70***	CCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OC	**1.87**
71	**71***	N#CC(=Cc1ccc(s1)c1ccc(c2c1nsn2)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**4.35**
72	**72***	N#CC(=Cc1ccc(o1)c1ccc(c2c1nsn2)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**4.96**
73	**73***	N#CC(=Cc1ccc(cc1)c1ccc(c2c1nsn2)c1ccc2c(c1)c1ccccc1n2c	**4.05**

		1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	
74	**74**	N#CC(=Cc1ccc(o1)c1ccc2c(c1)c1ccccc1n2c1ccc2c(c1)C(C)(C)c1c2cccc1)C(=O)O	**3.74**
75	**75***	N#CC(=Cc1ccc2c(c1)c1cc(ccc1n2CC)c1ccc(cc1)C=C(c1ccc(cc1)N(C)C)c1ccc(cc1)N(C)C)C(=O)O	**1.9**
76	**76**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/c1ccc(cc1)c1ccc(cc1)/C=C\1/SC(=S)N(C1=O)CC(=O)O)\C#N)C=C(c1ccc(cc1)c1ccc(cc1)/C=C/1\SC(=S)N(C1=O)CC(=O)O)C#N	**4.04**
77	**77**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/c1ccc(cc1)c1ccc(cc1)/C=C(\C(=O)O)/C#N)\C#N)C=C(c1ccc(cc1)c1ccc(cc1)/C=C(\C(=O)O)/C#N)C#N	**2.37**
78	**78***	CCCCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.22**
79	**79**	CCCCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.47**
80	**80**	CCCCCCCCC(n1c2ccc(cc2c2c1ccc(c2)c1sc(cc1CCCCCC)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1sc(cc1CCCCCC)c1ccc(s1)/C=C(\C(=O)O)/C#N)CCCCCCCC	**2.82**
81	**81***	CCCCCCCCC(n1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N)CCCCCCCC	**4.41**
82	**82***	CCCCCCCCC(n1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N)CCCCCCCC	**4.26**
83	**83***	CCCCCCn1c2ccc(cc2c2c1ccc(c2)c1sc(cc1CCCCCC)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1sc(cc1CCCCCC)c1ccc(s1)/C=C(\C(=O)O)/C#N	**4.55**
84	**84***	CCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**3.64**
85	**85***	CCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(s1)/C=C(/C(=	**4.82**

		O)O)\C#N)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N	
86	**86**	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)c1cc(ccc1n2CC)c1ccc(cc1)C=C(c1ccc(cc1)N(C)C)c1ccc(cc1)N(C)C)/C(=O)O	**4**
87	**87**	N#C/C(=C/c1ccc2c(c1)c1ccc(cc1n2CC)c1ccc(s1)c1ccc(s1)c1ccc(cc1)C=C(c1ccc(cc1)N(C)C)c1ccc(cc1)N(C)C)/C(=O)O	**3**
88	**88***	N#C/C(=C/c1ccc2c(c1)c1ccc(cc1n2CC)c1ccc(s1)c1ccc(cc1)C=C(c1ccc(cc1)N(C)C)c1ccc(cc1)N(C)C)/C(=O)O	**3.1**
89	**89**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1)/C=C/1\SC(=S)N(C1=O)CC(=O)O	**1.23**
90	**90***	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1)/C=C(\C(=O)O)/C#N	**1.94**
91	**91**	CCCCCCn1c2ccc(cc2c2c1cccc2)/C=C(\c1ccc(s1)C=C1C(=O)NC(=O)NC1=O)/C#N	**0.19**
92	**92**	CCCCCCn1c2ccc(cc2c2c1cccc2)/C=C(\c1ccc(s1)/C=C\1/SC(=S)N(C1=O)CC(=O)O)/C#N	**0.24**
93	**93***	CCCCCCn1c2ccc(cc2c2c1cccc2)/C=C(\c1ccc(s1)/C=C(/C(=O)O)\C#N)/C#N	**3.55**
94	**94**	CCCCCCC(Cc1cc(sc1c1sc2c(c1)c1nonc1c1c2sc(c1)c1sc(cc1CC(CCCCCC)CCCC)c1ccc2c(c1)c1ccccc1n2CC)/C=C(/C(=O)O)\C#N)CCCC	**7.33**
95	**95**	CCCCOc1ccc(cc1)N(c1ccc2c(c1)n(CCCC)c1c2ccc(c1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCC	**5.76**
96	**96**	CCCCn1c2cc(ccc2c2c1cc(cc2)c1ccc(s1)/C=C(/C(=O)O)\C#N)N(c1ccc(cc1)OCCCC)c1ccc(cc1)OCCCC	**1.36**
97	**97**	CCCCCCCC[N+]1=C(/C=C/c2ccc3c(c2)c2ccccc2n3CCCCCCn2c3ccc(cc3c3c2cccc3)/C=C/C2=[N+](CCCCCCCC)c3c(C2(C)C)cc(cc3)C(=O)O)C(c2c1ccc(c2)C(=O)O)(C)C.[I-].[I-]	**0.0538**
98	**98**	CCCCCCCC[N+]1=C(/C=C/c2ccc3c(c2)c2ccccc2n3CCCCCC)C(c2c1ccc(c2)C(=O)O)(C)C.[I-]	**0.0387**
99	**99**	CCCCCCCCOc1ccc(cc1)c1c2c3ccccc3n(c2c(c2c1n(CCCCCCCC)c1c2cc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCC	**7.58**

		CCC)CCCCCCCC	
100	**100**	CCCCCCCCn1c2ccc(cc2c2c1c(c1ccc(cc1)OCC)c1c(c2c2ccc(cc2)OCC)n(c2c1cccc2)CCCCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.98**
101	**101**	CCCCCCCCn1c2ccc(cc2c2c1c(c1ccc(cc1)N(CC)CC)c1c(c2c2ccc(cc2)N(CC)CC)n(c2c1cccc2)CCCCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**8.09**
102	**102**	CCCCCCCCn1c2ccc(cc2c2c1c(c1ccc(cc1)CC)c1c(c2c2ccc(cc2)CC)n(c2c1cccc2)CCCCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.25**
103	**103**	CCCCCCCCc1cc(sc1c1ccc(s1)c1sc(cc1CCCCCCCC)C=C(C(=O)O)C#N)c1cc2c(c3c1cccc3)n(c1c2cc(OC)cc1)C(CCCCC)CC	**7.54**
104	**104**	CCCCCC(n1c2ccc(cc2c2c1c1ccccc1c(c2)c1ccc(o1)C=C(C(=O)O)C#N)OC)CC	**6.93**
105	**105**	CCCCCC(n1c2ccc(cc2c2c1c1ccccc1c(c2)c1ccc(s1)C=C(C(=O)O)C#N)OC)CC	**6.01**
106	**106***	N#C/C(=C/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C(=O)O	**1.77**
107	**107***	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(s1)n1c2ccc(cc2c2c1ccc(c2)n1c2ccccc2c2c1cccc2)n1c2ccccc2c2c1cccc2)/C(=O)O	**4.86**
108	**108**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(s1)n1c2ccccc2c2c1cccc2)/C(=O)O	**4.3**
109	**109***	N#C/C(=C\c1ccc(s1)c1ccc(s1)n1c2ccccc2c2c1cccc2)/C(=O)O	**2.94**
110	**110**	N#C/C(=C\c1ccc(s1)c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C(=O)O	**2.74**
111	**111***	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1)C=C1C(=O)NC(=S)NC1=O	**0.54**
112	**112**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1)C=C1C(=O)NC(=O)NC1=O	**0.96**
113	**113**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1cc(sc1c1sc(c(c1)c1ccc2c(c1)c1ccccc1n2CCCCCC)c1ccc2c(c1)c1ccccc1n2CCCCCC)/C=C(/C(=O)O)\C#N	**4.24**
114	**114**	CCCCn1c2ccc(cc2c2c1cccc2)c1ccc(s1)c1cc2c(s1)c1c(n2c2ccc3c(c2)C(CCCC)(CCCC)c2c3cccc2)cc(s1)/C=C(\C(=O)O)/C#N	**5.39**

115	**115**	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccc(cc1)c1ccc2c(c1)C(CC)(CC)c1c2cccc1)c1ccc(cc1)c1ccc2c(c1)C(CC)(CC)c1c2cccc1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.2**
116	**116**	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccc(cc1)c1ccc2c(c1)C(CC)(CC)c1c2cccc1)c1ccc(cc1)c1ccc2c(c1)C(CC)(CC)c1c2cccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**1.78**
117	**117***	OC(=O)CCCCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1cnccn1	**1.58**
118	**118**	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1cnccn1	**0.89**
119	**119**	OC(=O)CCCCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1ccncc1	**2.35**
120	**120***	OC(=O)CCCCCCn1c2cc(ccc2c2c1cc(cc2)c1ccncc1)N(c1ccccc1)c1ccccc1	**1.81**
121	**121***	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1ccncc1	**1.84**
122	**122**	c1ccc(cc1)N(c1ccc2c(c1)[nH]c1c2ccc(c1)c1ccc(s1)c1ccncc1)c1ccccc1	**1.89**
123	**123**	CCCCn1c2cc(ccc2c2c1cc(cc2)c1ccncc1)N(c1ccccc1)c1ccccc1	**1.15**
124	**124**	c1ccc(cc1)N(c1ccc2c(c1)[nH]c1c2ccc(c1)c1ccncc1)c1ccccc1	**1.04**
125	**125***	CCCCn1c2cc(ccc2c2c1cc(cc2)N(c1ccccc1)c1ccccc1)c1ccc(cc1)C(=O)O	**0.97**
126	**126**	OC(=O)c1ccc(cc1)c1ccc2c(c1)[nH]c1c2ccc(c1)N(c1ccccc1)c1ccccc1	**0.91**
127	**127***	CCCCCCc1cc(sc1c1sc(cc1CCCCCC)c1sc(c(c1)CCCCCC)/C=C(\C(=O)O)/C#N)c1sc(cc1CCCCCC)c1ccc2c(c1)c1ccccc1n2CC	**3.4**
128	**128***	CCCCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccc(cc1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC)N(c1ccc(cc1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC	**4.9**
129	**129***	CCCCCCCCC(Cn1c2cc(C#Cc3ccc(c4c3nsn4)c3ccc(cc3)C(=O)O)c3c4c2c2c1cc(c1ccc5c(c1)c1ccccc1n5CC(CCCCCCCC)CCCCCC	**10.7**

		)c1c2c(c4ccc3)ccc1)CCCCCC	
130	**130**	CCCCCCCCC(Cn1c2cc(C#Cc3ccc(c4c3nsn4)c3ccc(cc3)C(=O)O)c3c4c2c2c1cc(c1ccc5c(c1)c1ccccc1C5(CC(CCCCCCCC)CCCCCC)CC(CCCCCCCC)CCCCCC)c1c2c(c4ccc3)ccc1)CCCCCC	**9.8**
131	**131**	CCCCCCCCC(Cn1c2ccc3c4c2c2c1cc(C#Cc1ccc(c5c1nsn5)c1ccc(cc1)C(=O)O)c1c2c(c4ccc3)ccc1)CCCCCC	**7.6**
132	**132**	CCCCCCc1cc(sc1c1cc(c(s1)/C=C(/C(=O)Nc1ccc(cc1)[Si](OC)(OC)OC)\C#N)CCCCCC)c1sc(cc1CCCCCC)c1sc(cc1CCCCCC)c1ccc2c(c1)c1ccccc1n2CC	**12.5**
133	**133**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)/C=C(/C(=O)Nc1ccc(cc1)C(=O)O)\C#N)c1sc(cc1CCCCCC)c1sc(cc1CCCCCC)c1ccc2c(c1)c1ccccc1n2CC	**9.32**
134	**134**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc2c(c1)c1cc(ccc1n2CCCC)c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)c1cc(ccc1n2CCCC)c1ccc2c(c1)c1ccccc1n2CCCC	**6.16**
135	**135**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc2c(c1)c1ccccc1n2CCCC)c1ccc2c(c1)c1cc(ccc1n2CCCC)c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N	**6.33**
136	**136***	CCCCn1c2ccc(cc2c2c1ccc(c2)c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)c1ccccc1n2CCCC	**5.87**
137	**137***	N#C/C(=C\c1ccc2c(n1)ccc(c2)n1c2ccc(cc2c2c1ccc(c2)OC)OC)/C(=O)O	**0.75**
138	**138**	CCCCCCCCCCCCn1c2ccccc2c2c1ccc(c2)c1nc2ccc(cc2nc1c1ccc2c(c1)c1ccccc1n2CCCCCCCCCCCC)c1ccc(s1)C=C(C(=O)O)C#N	**2.17**
139	**139**	CCCCCCn1c2ccccc2c2c1ccc(c2)c1nc2ccc(cc2nc1c1ccc2c(c1)c1ccccc1n2CCCCCC)c1ccc(s1)C=C(C(=O)O)C#N	**3.17**
140	**140**	CCCCn1c2ccc(cc2c2c1cccc2)c1nc2cc(ccc2nc1c1ccc2c(c1)c1ccccc1n2CCCC)c1ccc(s1)C=C(C(=O)O)C#N	**4.27**
141	**141**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)nc(c(n2)c1ccc2c(c1)c1ccccc1n2CC)c1ccc2c(c1)c1ccccc1n2CC)C(=O)O	**2.58**

142	**142**	CCCCCCCCn1c2ccccc2c2c1ccc1c2c2c3cc(ccc3n(c2cc1)CCCCCCCC)/C=C(/C(=O)O)\C#N	**3.18**
143	**143***	N#C/C(=C\c1sc2c(c1)C(c1c2sc(c1)c1ccc(s1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)(C)C)/C(=O)O	**4.78**
144	**144***	N#C/C(=C\c1sc2c(c1)C(c1c2sc(c1)c1ccc(s1)c1ccc(s1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)(C)C)/C(=O)O	**4.33**
145	**145***	CCCCCCCCc1cc(sc1c1sc2c(c1)C(c1c2sc(c1)/C=C(/C(=O)O)\C#N)(C)C)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C	**3.38**
146	**146**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1c(/C=C(/C(=O)O)\C#N)sc(c1c1ccc2c(c1)c1ccccc1n2CCCCCC)/C=C(/C(=O)O)\C#N	**3.2**
147	**147***	N#CC(=Cc1ccc(cc1)C#Cc1ccc2c(c1)n(c1ccc(cc1)C(F)(F)F)c1c2ccc(c1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C(=O)O	**2.45**
148	**148**	N#CC(=Cc1ccc(cc1)C#Cc1ccc2c(c1)n(c1ccc(cc1)C(F)(F)F)c1c2ccc(c1)C#Cc1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C(=O)O	**2.24**
149	**149***	N#CC(=Cc1ccc(cc1)C#Cc1ccc(s1)c1ccc(s1)C#Cc1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C(=O)O	**2.23**
150	**150**	N#C/C(=C\c1ccc2c(c1)c1c(n2C)ccc2c1c1c3ccccc3n(c1cc2)C)/C(=O)O	**2.49**
151	**151**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/c1ccc(s1)C=C(C(=O)O)C(=O)O)\C#N)/C=C(/c1ccc(s1)C=C(C(=O)O)C(=O)O)\C#N	**4.7**
152	**152**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(cc1)C(=O)O)c1ccc(s1)c1ccc(cc1)C(=O)O	**1.37**
153	**153**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccncc1)c1ccc(s1)c1ccncc1	**2.04**
154	**154**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccc(c(c1)O)O)c1ccc(s1)c1ccc(c(c1)O)O	**0.07**
155	**155***	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccncc1)c1ccc(s1)c1ccc(cc1)C(=O)O	**1.51**
156	**156**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccncc1)c1ccc(s1)c1ccc(c(c1)O)O	**0.06**

157	**157**	CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccccc1)c1ccccc1)c1ccc(c(c1)O)O	**0.31**
158	**158**	CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1ccncc1	**1.47**
159	**159**	CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccccc1)c1ccccc1)c1ccc(c(c1)O)O.CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1ccncc1	**0.21**
160	**160**	CCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)c1ccncc1)c1ccc(s1)c1ccncc1.CCCCn1c2ccc(cc2c2c1ccc(c2)N(c1ccccc1)c1ccccc1)c1ccc(c(c1)O)O	**0.34**
161	**161***	N#C/C(=C/c1ccc(s1)c1scc(n1)c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C(=O)O	**2.01**
162	**162**	N#C/C(=C/c1ccc(s1)c1scc(n1)c1ccc(cc1)c1ccc2c(c1)n(c1ccccc1)c1c2cccc1)/C(=O)O	**4.72**
163	**163**	CCCCn1c2ccccc2c2c1cc(cc2)c1ccc(cc1)c1csc(n1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**4.47**
164	**164**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)c1cc(ccc1n2c1ccc(cc1)c1ccc(s1)c1ccc(s1)C=C(C(=O)O)C#N)c1ccc(cc1)OCCCCCC	**5.31**
165	**165**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)c1cc(ccc1n2c1ccc2c(c1)C(C)(C)c1c2sc(c1)c1ccc(s1)C=C(C(=O)O)C#N)c1ccc(cc1)OCCCCCC	**5.41**
166	**166**	CCCCCCOc1cc(OCCCCCC)ccc1c1ccc2c(c1)c1cc(ccc1n2c1ccc2c(c1)C(C)(C)c1c2sc(c1)c1ccc(s1)C=C(C(=O)O)C#N)c1ccc(cc1OCCCCCC)OCCCCCC	**5.88**
167	**167**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(o1)/C=C(\C(=O)O)/C#N)c1ccc(o1)C=C(C(=O)O)C#N	**3.2**
168	**168**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1ccc(s1)C=C(C(=O)O)C#N	**3.8**
169	**169***	CCCCCCn1c2ccc(cc2c2c1ccc(c2)c1ccc(cc1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)C=C(C(=O)O)C#N	**3.4**
170	**170***	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1)C=C(C(=O)O)C#N	**3.2**
171	**171**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(c(c1)F)C=C(C(=O)O)C#	**4**

		N	
172	**172**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1F)C=C(C(=O)O)C#N	**2.7**
173	**173**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(c(c1F)F)C=C(C(=O)O)C#N	**3.5**
174	**174***	CCCCn1c2ccc(cc2c2c1ccc(c2)C=C1SC(=O)NC1=O)C=C1SC(=O)NC1=O	**1.44**
175	**175***	CCCCn1c2ccc(cc2c2c1ccc(c2)C=C1C(=O)NC(=O)NC1=O)C=C1C(=O)NC(=O)NC1=O	**0.33**
176	**176**	CCCCn1c2ccc(cc2c2c1cccc2)C=C1C(=O)NC(=O)NC1=O	**1.26**
177	**177***	CCCCCCn1c2ccc(cc2c2c1cccc2)C=C1C(=O)NC(=O)NC1=O	**1.47**
178	**178**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccc2c(c1)c1ccccc1n2c1ccccc1)c1ccc2c(c1)c1ccccc1n2c1ccccc1)c1sc(cc1CCCCCC)/C=C(\C(=O)O)/C#N	**7.43**
179	**179**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccc2c(c1)c1ccccc1n2c1ccccc1)c1ccc2c(c1)c1ccccc1n2c1ccccc1)/C=C(\C(=O)O)/C#N	**5.31**

'*' represent the test set compound

3.1.4 Indoline dataset (Study 1)

This dataset consists of 159 compounds, among which 121 compounds are present in the training set and 38 compounds are present in the test set. The detail of this dataset is shows in **Table 3.4**.

Table 3.4: Details of the indoline dataset

Serial Number	Compound ID	SMILE	Experimental PCE
1	**1**	CCCCCCCCn1c(=O)/c(=c/2\c(=C/c3ccc(s3)c3ccc4c(c3)C3CCCC3N4c3ccc4c(c3)Cc3c4cccc3)\sc(=O)n2CC(=O)O)/sc1=C(C#N)C#N	**5.15**
2	**2**	CCCCCCCCn1c(=O)/c(=c/2\c(=C/c3ccc(s3)c3ccc4c(c3)C3CCCC3N4c3ccc4c(c3)C(C)(C)c3c4cccc3)\sc(=O)n2CC(=O)O)/sc1=C(C#N)C#N	**5.55**
3	**3**	CCCCCCCCn1c(=C(C#N)C#N)s/c(=c\2/c(=C\c3ccc(s3)c3cc	**5.34**

		c4c(c3)C3CCCC3N4c3ccc4c(c3)C(CCCC)(CCCC)c3c4cccc3)/sc(=O)n2CC(=O)O)/c1=O	
4	**4**	CCCCCCCCn1c(=C(C#N)C#N)s/c(=c\2/c(=C\c3ccc(s3)c3ccc4c(c3)C3CCCC3N4c3ccc4c(c3)C(CCCCCCCC)(CCCCCCCC)c3c4cccc3)/sc(=O)n2CC(=O)O)/c1=O	**4.92**
5	**5**	CCCCCCCCN1C(=S)S/C(=c\2/s/c(=C\c3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**4.92**
6	**6**	CCCCCCCCn1c(=O)/c(=c/2\c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)sc(=O)n2CC(=O)O)/sc1=C(C#N)C#N	**5.15**
7	**7**	CCCCC1(CCCC)c2cc(ccc2c2c1cccc2)N1C2CCCC2c2c1ccc(c2)c1ccc(s1)/C=c\1/sc(=O)n(/c1=c/1\sc(=C(C#N)C#N)n(c1=O)CCCCCCCC(=O)O)CC(=O)O	**5.19**
8	**8**	CCCCCc1nc2c3cc(sc3c3c(c2nc1CCCCC)cc(s3)/C=C(/C(=O)O)\C#N)c1ccc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C	**7.12**
9	**9***	CCCCCCCCn1c(=O)s/c(=C/c2ccc(s2)c2ccc3c(c2)C2CCCC2N3c2ccc3c(c2)C(CCCC)(CCCC)c2c3cccc2)/c/1=c\1/sc(=C(C#N)C#N)n(c1=O)CCCCCCCC	**4.49**
10	**10**	CCCCCCn1c2ccc(cc2c2c1cccc2)N1c2ccc(cc2C2C1CCC2)c1ccc(c2c1nsn2)c1ccc(s1)/C=C(\C(=O)O)/C#N	**5**
11	**11***	CCCCCCCCCC1(CCCCCCCC)c2cc(sc2c2c1cc(s2)/C=C(\C(=O)O)/C#N)c1ccc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)c1ccccc1n2CCCCCC	**7.53**
12	**13**	CCCCCCCCOc1cc(OCCCCCCCC)ccc1c1ccc(cc1)C(=Cc1ccc(cc1)N1c2ccc(cc2C2C1CCC2)c1ncc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)c1ccc(cc1OCCCCCCCC)OCCCCCCCC	**8.43**
13	**14**	CCCCCCCCCCCCc1cc(sc1c1sc2c(c1)sc(c2)/C=C(\C(=O)O)/C#N)c1ccc(cc1)N1c2ccc(cc2C2C1C1CCC2C1)C=C(c1ccccc	**3.85**

		1)c1ccccc1	
14	**15***	CCCCCCCCCCCCc1cc(sc1c1sc2c(c1)sc(c2)/C=C(\C(=O)O)/C#N)c1ccc(cc1)N1c2ccc(cc2C2C1C1CCC2C1)c1cccs1	**4.62**
15	**16**	CCn1o/c(=c/2\c(=Cc3ccc(cc3)c3ccc4c(c3)C3CCCC3N4c3ccccc3)sc(=O)n2CC(=O)O)/sc1=S	**4.07**
16	**17***	CCCCCCOc1cc(OCCCCCC)ccc1c1ccc(cc1)N1c2ccc(cc2C2C1CCC2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**8.12**
17	**18**	CCCCCCOc1cc(OCCCCCC)ccc1c1ccc(cc1)C(=Cc1ccc(cc1)N1c2ccc(cc2C2C1CCC2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)c1ccc(cc1OCCCCCC)OCCCCCC	**7.28**
18	**19**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**7.79**
19	**20**	N#Cc1ccc(cc1)N1C2CCCC2c2c1ccc(c2)c1ccc(cc1)C=c1sc(=O)n(/c/1=c\1/on(c(=S)s1)CC)CC(=O)O	**3.34**
20	**21**	Cc1ccc(cc1)N1c2ccc(cc2C2C1CCC2)c1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O	**6.12**
21	**22***	Cc1ccc(cc1)N1c2ccc(cc2C2C1CCC2)C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O	**6.28**
22	**23***	Cc1ccc(cc1)N1C2CCCC2c2c1ccc(c2)c1ccc(c2c1nsn2)C#Cc1ccc(cc1)C(=O)O	**7.13**
23	**24**	N#C/C(=C/c1ccc(s1)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**9.2**
24	**25***	CCCCCCCCC1(CCCCCCCC)c2cc(sc2c2c1cc(s2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C=C(\C(=O)O)/C#N	**9**
25	**26**	CCn1o/c(=c/2\c(=Cc3ccc(cc3)c3ccc4c(c3)C3CCCC3N4c3ccc(cc3)N(C)C)sc(=O)n2CC(=O)O)/sc1=S	**0.22**
26	**28***	COc1ccc(cc1)N1C2CCCC2c2c1ccc(c2)c1ccc(cc1)C=c1sc(=O)n(/c/1=c\1/on(c(=S)s1)CC)CC(=O)O	**4.08**

27	**29**	CCn1o/c(=c/2\c(=Cc3ccc(cc3)c3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C(F)(F)F)sc(=O)n2CC(=O)O)/sc1=S	**3.58**
28	**30**	CCCCn1c2ccc(cc2c2c1nc1ccccc1n2)/C=C(/C(=O)O)\C#N	**0.77**
29	**31***	CCCCn1c2ccc(cc2c2c1nc1ccccc1n2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.07**
30	**32**	CCCCn1c2ccc(cc2c2c1nc1ccccc1n2)C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N	**0.63**
31	**33**	CCCCCCCCCCCCc1cc(sc1c1cc2c(s1)cc(s2)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)C1C3CCC(C1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)C3	**4.38**
32	**34**	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)/C=C(/C(=O)O)\C#N)c1c(c2c2ccc(s2)c2cccc3c2C2C4CCC(C2N3c2ccccc2)C4)nc(c(n1)c1ccccc1)c1ccccc1	**2.7**
33	**35**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)/C=C(\C(=O)O)/C#N)c1c(c2c2ccc(s2)c2cccc3c2C2C4CCC(C2N3c2ccccc2)C4)nc(c(n1)c1ccccc1)c1ccccc1	**1.2**
34	**36**	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)C(=O)O)c1c(c2c2ccc(s2)c2cccc3c2C2C4CCC(C2N3c2ccccc2)C4)nc(c(n1)c1ccccc1)c1ccccc1	**4.8**
35	**37***	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)C(=O)O)c1c(c2c2ccc(s2)c2cccc3c2C2C4CCC(C2N3c2ccccc2)C4)nc(c(n1)c1ccccc1)c1ccccc1	**2.8**
36	**38***	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)/C=C(/C(=O)O)\C#N)c1c(c2c2cccc3c2C2C4CCC(C2N3c2ccccc2)C4)nc(c(n1)c1ccccc1)c1ccccc1	**4.8**
37	**39**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1c(c2c2cccc3c2C2C4CCC(C2N3c2ccccc2)C4)nc(c(n1)c1ccccc1)c1ccccc1	**4.4**
38	**41**	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)/C=C(\C(=O)O)/C#N)c1c(c2c2ccc(s2)c2ccc(cc2)N2C3C4CCC(C3c3c2cccc3)C4)nc(c(n1)c1ccccc1)c1ccccc1	**4.5**

39	**42***	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)/C=C(\C(=O)O)/C#N)c1c(c2c2ccc(s2)c2ccc(cc2)N2C3C4CCC(C3c3c2cccc3)C4)nc(c(n1)c1ccccc1)c1ccccc1	**3.2**
40	**43**	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)C(=O)O)c1c(c2c2ccc(s2)c2ccc(cc2)N2C3C4CCC(C3c3c2cccc3)C4)nc(c(n1)c1ccccc1)c1ccccc1	**5.4**
41	**44**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1ccc2c(c1)cc(n2c1ccc2c(c1)C(C)(C)c1c2cccc1)c1ccccc1)C(=O)O	**6.96**
42	**45***	CCCCCCc1ccc(cc1)c1cc2c(n1c1ccc3c(c1)C(C)(C)c1c3cccc1)ccc(c2)c1ccc(s1)c1ccc(s1)C=C(C(=O)O)C#N	**7.18**
43	**46**	N#C/C(=C\c1cc2C3CCCC3N(c2c2c1cccc2)c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C(=O)O	**2.15**
44	**47***	N#C/C(=C\c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C(=O)O	**2.27**
45	**48**	OC(=O)CN1C(=S)S/C(=C\c2cc3C4CCCC4N(c3c3c2cccc3)c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)/C1=O	**1.66**
46	**49**	OC(=O)CN1C(=S)S/C(=C\c2ccc3c(c2)C2CCCC2N3c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)/C1=O	**3.55**
47	**50**	CCN1C(=S)S/C(=c\2/s/c(=C\c3cc4C5CCCC5N(c4c4c3cccc4)c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**2.65**
48	**51**	CCCCCCCCCCCCc1cc(sc1c1sc2c(c1)sc(c2)/C=C(\C(=O)O)/C#N)c1ccc(cc1)N1c2ccc(cc2C2C1C1CCC2C1)c1ccc(cc1)OC	**4.24**
49	**52**	CCCCCCCCCCCCc1cc(sc1c1sc2c(c1)sc(c2)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)C1C3CCC(C1N2c1ccc(cc1)c1cccs1)C3	**4.23**
50	**53**	CCCCCCCCCCCCc1cc(sc1c1cc2c(s1)cc(s2)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)C1C3CCC(C1N2c1ccc(cc1)c1ccc(cc1)OC)C3	**2.96**
51	**54**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)C(=O)O)c1c(c2c2ccc(s2)c2ccc(cc2)N2C3C4CCC(C3c3c2cccc3)C4)nc(c(n1)c1ccc	**3.1**

		cc1)c1ccccc1	
52	**55**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)N1c3ccccc3C3C1C1CCC3C1)c1c(c2c2ccc(s2)C(=O)O)nc(c(n1)c1ccccc1)c1ccccc1	**3.9**
53	**57**	CCN1C(=S)S/C(=c\2/s/c(=C\c3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**4.11**
54	**58***	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)c1ccccc1n1c2cc2c1cc(cc2)C)/C(=O)O	**6.22**
55	**59**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc2c(c1)c1ccccc1n1c2cc2c1cc(cc2)C)/C(=O)O	**8.34**
56	**60**	CCN1C(=S)S/C(=c\2/s/c(=c\3/c(=Cc4ccc5c(c4)C4CCCC4N5c4ccc(cc4)C=C(c4ccccc4)c4ccccc4)sc(=O)n3CC(=O)O)/c(=O)n2CC)/C1=O	**3.92**
57	**61**	CCCCCCCCn1/c(=C/2\SC(=S)N(C2=O)CCCCCCCC)/s/c(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)sc(=O)n2CC(=O)O)/c1=O	**4.36**
58	**62**	CCCCC(Cn1c(=O)/c(=c/2\c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)sc(=O)n2CC(=O)O)/s/c/1=C\1/SC(=S)N(C1=O)CC(CCCC)CC)CC	**4.12**
59	**63***	OC(=O)Cn1c(=O)sc(=Cc2ccc3c(c2)C2CCCC2N3c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)/c/1=c\1/s/c(=C\2/SC(=S)N(C2=O)c2ccccc2)/n(c1=O)c1ccccc1	**4.09**
60	**64***	CCCCCCCCN1C(=S)S/C(=c\2/s/c(=c\3/c(=Cc4ccc5c(c4)C4CCCC4N5c4ccc(cc4)C=C(c4ccccc4)c4ccccc4)sc(=O)n3CC(=O)O)/c(=O)n2c2ccccc2)/C1=O	**4.35**
61	**65***	CCCCCCCCn1c(=O)/c(=c/2\c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)sc(=O)n2CC(=O)O)/s/c/1=C\1/SC(=S)N(C1=O)c1ccccc1	**4.15**
62	**67**	CCCCCCCCCCCCc1cc(sc1c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N1c2ccccc2C2C1C1CCC2C1	**5.2**
63	**69**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)N1c3ccccc3C3C1C1CCC3C1)c1c(c2c2ccc(cc2)/C=C(/C(=O)O)\C#N)nc(c(n1)c1cc	**4.1**

		ccc1)c1ccccc1	
64	**70**	CCCCCCCCn1c(c2ccc(cc2)OCCCCCCCCC)c(c2c1ccc(c2)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1c2cc(ccc2n(c1c1ccc(cc1)OCCCCCCCC)CCCCCCCC)c1ccc(s1)/C=C(\C(=O)O)/C#N	**6.05**
65	**71**	CCCCCCCCn1c(c2ccc(cc2)OCCCCCCCCC)c(c2c1ccc(c2)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1c2cc(ccc2n(c1c1ccc(cc1)OCCCCCCCCC)CCCCCCCC)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**6.54**
66	**72***	CCCCCCn1c2ccc(cc2c2c1c1ccccc1C2(CCCCCC)CCCCCC)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**6.29**
67	**74***	CCCCCCn1c2ccc(cc2c2c1c1ccccc1C2(CCCCCC)CCCCCC)c1ccc(cc1)/C=c/1\sc(=C(C#N)C#N)[nH]c1=O	**3.6**
68	**75**	CCCCCCn1c2ccc(cc2c2c1c1ccccc1C2(CCCCCC)CCCCCC)c1ccc(s1)/C=c/1\sc(=C(C#N)C#N)[nH]c1=O	**5.41**
69	**76***	CCCCCCn1c2ccc(cc2c2c1c1ccccc1C2(CCCCCC)CCCCCC)C#Cc1ccc(cc1)/C=C(\C(=O)O)/C#N	**5.16**
70	**77**	CCCCCCn1c2ccc(cc2c2c1c1ccccc1C2(CCCCCC)CCCCCC)C#Cc1ccc(s1)/C=C(/C(=O)O)\C#N	**6.86**
71	**78**	CCCCCCn1c2ccc(cc2c2c1c1ccccc1C2(CCCCCC)CCCCCC)C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O	**7.99**
72	**79**	CCCCCCCCCCCCn1c(cc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.55**
73	**80***	CCCCCCCCCCCCn1c(cc2c1ccc(c2)c1ccc(s1)/C=C\1/SC(=S)N(C1=O)CC(=O)O)c1ccc(s1)/C=C/1\SC(=S)N(C1=O)CC(=O)O	**5.13**
74	**81***	CCCCn1c2ccc(cc2c2c1nc1ccccc1n2)C#Cc1cccc(c1)/C=C(/C(=O)O)\C#N	**0.97**
75	**82**	CCCCn1c2ccc(cc2c2c1nc1ccccc1n2)C#Cc1ccc(s1)/C=C(/C(=O)O)\C#N	**2.41**
76	**83***	CCCCn1c2ccc(cc2c2c1nc1ccccc1n2)C#Cc1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**2.3**

77	**84***	CCCCCCCCn1c(c2ccc(cc2)OCCCCCCCCC)c(c2c1ccc(c2)c1ccc(o1)/C=C(\C(=O)O)/C#N)c1c2cc(ccc2n(c1c1ccc(cc1)OCCCCCCCC)CCCCCCCC)c1ccc(o1)/C=C(\C(=O)O)/C#N	**5.87**
78	**85**	CCCCCCCCN1C(=O)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)sc(=O)n2CC(=O)O)/C1=O	**3.38**
79	**86**	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccccc3)sc(=O)n2CC(=O)O)/C1=O	**4.31**
80	**87**	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccccc3C(=O)O)sc(=O)n2CC)/C1=O	**1.92**
81	**88**	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3cccc(c3)C(=O)O)sc(=O)n2CC)/C1=O	**1.6**
82	**89**	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C(=O)O)sc(=O)n2CC)/C1=O	**2.1**
83	**90***	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)CO)sc(=O)n2CC(=O)O)/C1=O	**4.53**
84	**91**	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)CO)sc(=O)n2CCC(=O)O)/C1=O	**3.43**
85	**92***	CCN1C(=S)S/C(=c\2/c(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)CO)sc(=O)n2CCCC(=O)O)/C1=O	**3.69**
86	**93**	CCCCCCn1cc(c2c1cccc2)/C=C(\c1ccc(s1)/C=C/1\SC(=S)N(C1=O)CC(=O)O)/C#N	**0.35**
87	**94**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1cc2CCCC3c2c(c1)C(c1ccc2c(c1)C(C)(C)c1c2cccc1)N3c1ccccc1)C(=O)O	**8.42**
88	**95***	N#CC(=Cc1ccc(s1)c1ccc(s1)c1ccc2c3c1CCCC3N(C2c1ccc2c(c1)C(C)(C)c1c2cccc1)c1ccccc1)C(=O)O	**6.88**
89	**96**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)C=C(C(=O)O)C#N)c1ccc2c(c1)C1CCCC1N2c1ccccc1	**3.4**
90	**97***	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)CCN2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C(=O)O	**3.94**
91	**98**	N#C/C(=C\c1ccc2c(c1)nc(s2)c1ccc2c(c1)CCN2c1ccc(cc1)	**4.02**

		C=C(c1ccccc1)c1ccccc1)/C(=O)O	
92	**99**	CCN1C(=S)S/C(=c\2/s/c(=C/c3ccc(s3)c3ccc4c(c3)N(CC4)c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**1.42**
93	**100**	CCCCCCn1cc(c2c1cccc2)/C=C(\c1ccc(s1)C=C(C(=O)O)C#N)/C#N	**1.04**
94	**101**	CCN1C(=S)S/C(=c\2/s/c(=C\c3ccc4c(c3)nc(s4)c3ccc4c(c3)CCN4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**2.91**
95	**102***	N#CC(=Cc1ccc2c(c1)[C@@H]1CCC[C@H]1N2c1ccc(cc1)C)C(=O)O	**1.79**
96	**103**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)[C@@H]1CCC[C@H]1N2c1ccc(cc1)C)C(=O)O	**4.58**
97	**104**	N#CC(=Cc1ccc(s1)/C=C/c1ccc2c(c1)[C@@H]1CCC[C@H]1N2c1ccc(cc1)C)C(=O)O	**4.46**
98	**105**	N#CC(=Cc1ccc(s1)/C=C/c1ccc(s1)c1ccc2c(c1)[C@@H]1CCC[C@H]1N2c1ccc(cc1)C)C(=O)O	**2.53**
99	**106**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1ccc2c(c1)N(c1ccc(cc1)C)C1C2CCC1)/C(=O)O	**3.86**
100	**107**	CCCCCCc1cc(sc1c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(c2c1nsn2)c1ccc2c(c1)N(c1ccc(cc1)C)C1C2CCC1	**7.63**
101	**108**	CCCCCCn1cc(c2c1ccc(c2)OC)/C=C(\c1ccc(s1)/C=C/1\SC(=S)N(C1=O)CC(=O)O)/C#N	**0.046**
102	**110**	CCCCCCC1(CCCCCC)c2ccccc2c2c1ccc(c2)N1C2CCCC2c2c1cc(/C=C/c1ccc(s1)/C=C(/C(=O)O)\C#N)cc2	**5.09**
103	**111***	N#C/C(=C\c1ccc(s1)/C=C/c1ccc2c(c1)N(c1ccc(cc1)C)C1C2CCC1)/C(=O)O	**3.99**
104	**112**	CCN1C(=S)S/C(=c/2\s/c(=C\c3ccc4c(c3)N(c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)C3C4CCC3)/c(=O)n2CC(=O)O)/C1=O	**6.51**
105	**113***	OC(=O)CN1C(=S)S/C(=c/2\s/c(=C\c3ccc4c(c3)N(c3ccc(cc3)/C=C(\c3ccccc3)/C3=CC=CCC3)C3C4CCC3)/c(=O)n2CC(	**5.5**

		=O)O)/C1=O	
106	**114**	CCN1C(=S)S/C(=c/2\s/c(=C\c3ccc4c(c3)N(c3ccc(cc3)/C=C(\c3ccccc3)/C3=CC=CCC3)C3C4CCC3)/c(=O)n2CCC(=O)O)/C1=O	**5.6**
107	**115**	CCN1C(=S)S/C(=c/2\s/c(=c/3\s/c(=C\c4ccc5c(c4)N(c4ccc(cc4)/C=C(\c4ccccc4)/C4=CC=CCC4)C4C5CCC4)/c(=O)n3CC(=O)O)/c(=O)n2CC)/C1=O	**5.93**
108	**116**	CCCCCCc1cc(sc1c1cc(c(s1)C=C(C(=O)O)C#N)CCCCCC)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)OC	**3**
109	**117**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)c1sc(cc1CCCCCC)/C=C(/C(=O)O)\C#N)c1sc(cc1CCCCCC)c1sc2c(c1)n(c1c2cccc1)CC	**4**
110	**118**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)C=C(C(=O)O)C#N)c1cc2c(s1)c1c(n2CC)cccc1	**7.4**
111	**120**	CCCCCCn1cc(c2c1ccc(c2)OC)/C=C(\c1ccc(s1)/C=N/c1ccc(cc1)C(=O)O)/C#N	**4.12**
112	**121**	CCCCn1c(/C=C/c2ccc(n2c2ccc(cc2)c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C=C(\C(=O)O)/C#N)cc2c1ccc(c2)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**4.31**
113	**122**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(cc1)n1c(ccc1/C=C(\C(=O)O)/C#N)/C=C/c1cc2c(n1CCCC)ccc(c2)c1ccc2c(c1)c1ccccc1n2CCCCCC	**4.31**
114	**123**	CCCCCCc1cc(sc1/C=C(\C(=O)O)/C#N)c1ccc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C	**6.85**
115	**124**	CCCCCCc1cc(sc1c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)c1ccc(c2c1nsn2)c1cc(c(s1)/C=C(\C(=O)O)/C#N)CCCCCC	**4.64**
116	**125**	CCCCCCc1cc(sc1c1sc(c(c1)CCCCCC)C=C(C(=O)O)C#N)c1ccc2c(c1)C1CCCC1N2CC	**3.7**
117	**126**	CCCC1(CCC)c2ccc(cc2c2c1cccc2)N1C2CCCC2c2c1ccc(c2)c1ccc(c2c1nsn2)c1ccc(s1)/C=C(\C(=O)O)/C#N	**6.9**
118	**127**	N#C/C(=C/c1ccc(s1)c1ccc(c2c1nsn2)c1ccc2c(c1)C1CCCC	**5.87**

		1N2c1ccc(cc1)C(C)(C)C)/C(=O)O	
119	**128**	CCCC1(CCC)c2ccc(cc2c2c1cccc2)N1C2CCCC2c2c1ccc(c2)c1ccc(s1)/C=C(\C(=O)O)/C#N	**6.2**
120	**129**	CCCCCCn1cc(c2c1ccc(c2)OC)/C=C(\c1ccc(s1)C=C(C(=O)O)C#N)/C#N	**1.48**
121	**130**	OC(=O)c1ccc2c(c1)C(=O)/C(=C/c1ccc(s1)c1ccc3c(c1)C1CCCC1N3c1ccc3c(c1)C(C)(C)c1c3cccc1)/C2=O	**4.68**
122	**131**	CCCCCCc1cc(sc1/C=C/1\C(=O)c2c(C1=O)cc(cc2)C(=O)O)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)C(C)(C)c1c2cccc1	**6.11**
123	**132**	OC(=O)CN1C(=S)S/C(=C\c2ccc3c(c2)N(c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)C2C3CCC2)/C1=O	**2.35**
124	**133**	CCCCCCCCN1C(=S)SC(=c2sc(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)c(=O)n2CC(=O)O)C1=O	**5.7**
125	**134**	N#C/C(=C\c1ccc(o1)c1ccc(c2c1non2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**6.21**
126	**135**	CCCCCCc1cc(sc1/C=C(/C(=O)O)\C#N)c1ccc(c2c1non2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C	**8.61**
127	**136***	N#CC(=Cc1ccc(s1)c1ccc(s1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)C(=O)O	**5.97**
128	**138**	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**4.42**
129	**139**	CCCCCCCCN1C(=O)c2c(C1=O)c(ccc2c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.11**
130	**140**	COc1ccc(cc1)N1C2CCCC2c2c1ccc(c2)/C=C(/C(=O)O)\C#N	**4.03**
131	**141**	CCCCCCCCn1nc2c(n1)c(ccc2c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C	**8.38**
132	**142**	N#C/C(=C\c1ccc(s1)c1sc(c2c1OCCO2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**4.45**
133	**143**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)n1c2cc(sc2c2c1cc(s2)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)C(	**7.25**

		CCC)(CCC)c1c2cccc1)/C=C(/C(=O)O)\C#N	
134	**144**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)n1c2cc(sc2c2c1cc(s2)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)C(CCC)(CCC)c1c2c2c(c3c1c1ccccc1C3(CCC)CCC)c1c(C2(CCC)CCC)cccc1)/C=C(/C(=O)O)\C#N	**8.78**
135	**145**	CCCCCCN(c1ccc(cc1)n1c2cc(sc2c2c1cc(s2)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)C(CCC)(CCC)c1c2cccc1)/C=C(/C(=O)O)\C#N)CCCCCC	**7.4**
136	**146**	CCCCCCc1cc(sc1/C=C(\C(=O)O)/C#N)c1sc2c(c1)c1nn(nc1c1c2sc(c1)c1cc2c(s1)c1c(n2CC(CCCC)CC)cccc1)CC(CCCC)CC	**6.71**
137	**147**	Cc1ccc(cc1)N1C2CCCC2c2c1ccc(c2)c1ccc2c3c1cccc3c(=O)n(c2=O)CC(=O)O	**1.18**
138	**148**	OC(=O)Cn1c(=O)c2ccc(c3c2c(c1=O)ccc3c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C	**2.7**
139	**149**	N#C/C(=C\1/CCCC(=C1)/C=C/c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**1.92**
140	**150**	N#CC(=Cc1ccc(s1)c1cnc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)C(=O)O	**2.1**
141	**151***	CCCCC(CC1(CC(CCCC)CC)c2cc(sc2c2c1cc(s2)c1ccc(s1)C=C(C(=O)O)C#N)c1cnc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)CC	**5.1**
142	**152**	COc1ccc2c(c1)ccc(c2)N1C2CCCC2c2c1ccc(c2)/C=C/1\SC(=S)N(C1=O)CC(=O)O	**5.5**
143	**153**	COc1ccc(cc1)N1c2ccc(cc2C2C1CCC2)/C=C/1\SC(=S)N(C1=O)CC(=O)O	**5.11**
144	**154**	CCCc1c(/C=C(/C(=O)O)\C#N)sc2c1sc(c2)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)C(CCC)(CCC)c1c2cccc1	**6.95**
145	**155**	CCCCCCOc1ccc(cc1)N1C2CCCC2c2c1ccc(c2)c1cc2c(s1)c(c(s2)/C=C(/C(=O)O)\C#N)CCC	**5.61**
146	**156**	CCCc1c(/C=C(/C(=O)O)\C#N)sc2c1sc(c2)c1ccc2c(c1)C1C	**5.89**

		CCC1N2c1ccc(cc1)OC(C)(C)C	
147	**157**	CCCc1c(/C=C(/C(=O)O)\C#N)sc2c1sc(c2)c1ccc2c(c1)C1CCCC1N2c1ccc2c(c1)C(CCC)(CCC)c1c2c2c(c3c1c1ccccc1C3(CCC)CCC)c1c(C2(CCC)CCC)cccc1	**7.08**
148	**158**	N#C/C(=C\c1ccc(cc1)c1ccc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**5.3**
149	**159***	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C)/C(=O)O	**5.3**
150	**161**	N#C/C(=C/c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C(=O)O	**2.6**
151	**162**	N#C/C(=C/c1ccc(s1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C(=O)O	**3.78**
152	**163***	N#C/C(=C/c1ccc(s1)c1ccc(s1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C(=O)O	**3.19**
153	**164**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1ccc(s1)c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)C(=O)O	**2.08**
154	**165**	CCCCCCc1cc(sc1c1ccc2c(c1)C1CCCC1N2c1ccc(cc1)C=C(c1ccccc1)c1ccccc1)/C=C(\C(=O)O)/C#N	**3.36**
155	**166**	OC(=O)CN1C(=S)S/C(=C/c2ccc(s2)c2ccc3c(c2)C2CCCC2N3c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)/C1=O	**2.83**
156	**167**	CCN1C(=S)S/C(=c\2/s/c(=C/c3ccc(s3)c3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**2.35**
157	**168***	CC(c1cc(cc(c1)C(C)(C)C)CN1C(=S)S/C(=c\2/s/c(=C/c3ccc(s3)c3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O)(C)C	**2.45**
158	**169**	OC(=O)Cn1c(=O)/c(=C\c2ccc(s2)c2ccc(s2)c2ccc3c(c2)C2CCCC2N3c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)/s/c/1=C\1/SC(=S)N(C1=O)Cc1cc(cc(c1)C(C)(C)C)C(C)(C)C	**1.71**
159	**170***	OC(=O)Cn1c(=O)/c(=C\c2ccc(s2)c2ccc(s2)c2ccc(s2)c2ccc3c(c2)C2CCCC2N3c2ccc(cc2)C=C(c2ccccc2)c2ccccc2)/s/c	**1.36**

		/1=C\1/SC(=S)N(C1=O)Cc1cc(cc(c1)C(C)(C)C)C(C)(C)C	

'*' represent the test set compound

3.1.5 Phenothiazine dataset (Study 2)

This dataset consists of 215 compounds, among which 22 compounds are removed from the dataset based on leverage value, and the remaining dataset is divided into three training and test sets with different data combinations. The detail of this dataset is shows in **Table 3.5**.

Table 3.5: Details of the phenothiazine dataset

Serial Number	Compound ID	SMILE	Experimental PCE
1	**1**	N#C/C(=C\c1ccc2c(c1)Oc1c(N2CCCCCCN2c3ccccc3Oc3c2ccc(c3)/C=C(/C(=O)O)\C#N)cccc1)/C(=O)O	**2.71**
2	**2**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1OCCCC)OCCCC)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**7.5**
3	**3**	CCCCCCOc1ccc(cc1)N1c2ccc(cc2Sc2c1ccc(c2)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCCCCCC)cccc1	**6.87**
4	**4**	CCCCCCOc1ccc(cc1)N1c2ccc(cc2Sc2c1cccc2)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**7.78**
5	**5**	CCCCCCOc1ccc(cc1)N1c2ccccc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N	**6.98**
6	**6**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1	**5.05**
7	**7**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**5.44**
8	**8**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2Cc2ccc(cc2)CN2c3ccccc3Sc3c2ccc(c3)/C=C(/C(=O)O)\C#N)cccc1)/C(=O)O	**3.56**
9	**9**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2CCCCN2c3ccccc3Sc3c2ccc(c3)/C=C(/C(=O)O)\C#N)cccc1)/C(=O)O	**4.13**

10	**10**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)c1cc(c(s1)/C=C(/C(=O)O)\C#N)CCCCCC)c1ccc(s1)c1cc(c(s1)/C=C(\C(=O)O)/C#N)CCCCCC)CC	**6.67**
11	**11**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1ccc(c2)c1cc(c(s1)/C=C(/C(=O)O)\C#N)CCCCCC)c1cc(c(s1)/C=C(\C(=O)O)/C#N)CCCCCC)CC	**6.48**
12	**12**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(\C(=O)O)/C#N)CC	**6.29**
13	**13**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)/C=C(\C(=O)O)/C#N)CC	**5.55**
14	**14**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC	**3.78**
15	**15**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**4.41**
16	**16**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)C=C(c1ccccc1)c1ccccc1	**2.48**
17	**17**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2C)ccc(c1)c1ccc(cc1)N(c1ccc2c1cccc2)c1ccccc1)/C(=O)O	**4.6**
18	**18**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1	**5.6**
19	**19**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1	**2.04**
20	**20**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(\C(=O)O)/C#N)c1ccc(s1)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1	**6.22**
21	**21**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2c2ccc(cc2)C(c2ccc(cc2)C)c2ccc(cc2)C)ccc(c1)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	**5.22**
22	**22**	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)N(c2ccc(cc2)C)c2ccc(cc2)C)ccc(c1)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	**4.24**
23	**23**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2c2ccc(cc2)N(c2ccc(cc2)C)c	**4.8**

		2ccc(cc2)C)ccc(c1)c1ccc(s1)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	
24	**24**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)/C=C(/C(=O)O)\C#N	**4.49**
25	**25**	CCCCCCN1c2ccc(cc2Oc2c1cccc2)/C=C(/C(=O)O)\C#N	**3.61**
26	**26**	CCN1c2ccc(cc2Sc2c1cccc2)/C=C(/C(=O)O)\C#N	**2.91**
27	**27**	N#C/C(=C\c1ccc(cc1)N1c2ccccc2Sc2c1cccc2)/C(=O)O	**1.83**
28	**28**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1OCCCC)OCCCC)c1ccc(c2c1nsn2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.4**
29	**29**	CCCCCCCCc1cc(sc1c1ccc(s1)c1sc(cc1CCCCCCCC)/C=C(/C(=O)O)\C#N)c1cc2cc(n3c2c(c1)sc1c3cccc1)c1ccc(cc1)OCCCCCC	**7.54**
30	**30**	CCCCCCOc1ccc(cc1)c1cc2c3n1c1ccccc1sc3cc(c2)c1ccc(o1)/C=C(/C(=O)O)\C#N	**5.6**
31	**31**	CCCCCCOc1ccc(cc1)c1cc2c3n1c1ccccc1sc3cc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.42**
32	**32**	CCCCCC1c2c(C)c(c(n2[B-]([N+]2=C(C(=C(C12)C)c1ccc2c(c1)Sc1c(N2CCC)cccc1)C)(F)F)C)c1ccc(o1)C=C1SC(=S)N(C1=O)CC(=O)O	**1.23**
33	**33**	CCCCCC1c2c(C)c(c(n2[B-]([N+]2=C(C(=C(C12)C)c1ccc2c(c1)Sc1c(N2CCC)cccc1)C)(F)F)C)c1ccc(o1)/C=C(\C(=O)O)/C#N	**5.31**
34	**34**	CCCCCC1c2c(C)c(c(n2[B-]([N+]2=C(C(=C(C12)C)c1ccc2c(c1)Sc1c(N2CCC)cccc1)C)(F)F)C)c1ccc(s1)C=C1SC(=S)N(C1=O)CC(=O)O	**2.26**
35	**35**	CCCCCC1c2c(C)c(c(n2[B-]([N+]2=C(C(=C(C12)C)c1ccc2c(c1)Sc1c(N2CCC)cccc1)C)(F)F)C)c1ccc(s1)/C=C(\C(=O)O)/C#N	**4.61**
36	**36**	CCCCC(CN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)/C=C(/C(=O)O)\C#N)CC	**6.82**
37	**37**	CCCCCCC(Cc1cc(sc1c1sc2c(c1)c1nonc1c1c2sc(c1)c1sc(cc1CC(CCCCCC)CCCC)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)Sc1c(	**5.6**

		N2CC)cccc1)CCCC	
38	**38**	CCCCCC1=C2C(=C(C(=[N+]2[B-](n2c1c(C)c(c2C)c1ccc(s1)/C=C(/C(=O)O)\C#N)(F)F)C)c1ccc2c(c1)Sc1c(N2CCCCCCN2c3ccccc3Sc3c2ccc(c3)C2=C(C)C3=C(CCCCC)c4n([B-]([N+]3=C2C)(F)F)c(c(c4C)/C=C(/C(=O)O)\C#N)C)cccc1)C	**4.89**
39	**39**	CCCCCC1=C2C(=C(C(=[N+]2[B-](n2c1c(C)c(c2C)/C=C(/C(=O)O)\C#N)(F)F)C)c1ccc2c(c1)Sc1c(N2CCCCCCN2c3ccccc3Sc3c2ccc(c3)C2=C(C)C3=C(CCCCC)c4n([B-]([N+]3=C2C)(F)F)c(c(c4C)/C=C(/C(=O)O)\C#N)C)cccc1)C	**3.24**
40	**40**	CCCCCC1=C2C(=C(C(=[N+]2[B-](n2c1c(C)c(c2C)/C=C(/C(=O)O)\C#N)(F)F)C)c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1)C	**2.28**
41	**41**	CCCCCC1=C2C(=C(C(=[N+]2[B-](n2c1c(C)c(c2C)/C=C(/C(=O)O)\C#N)(F)F)C)c1ccc2c(c1)Sc1c(N2CCC)cccc1)C	**2.14**
42	**42**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1cc2c(s1)c1c(n2CCCCCC)ccs1	**5.8**
43	**43**	CCCCCCOc1ccc(cc1)n1c2cc(sc2c2c1ccs2)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**5.7**
44	**44**	CCCCCCCCc1cc(sc1c1ccc(s1)c1sc(cc1CCCCCCCC)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)Sc1c(N2CC(CCCC)CC)ccc(c1)c1ccc(cc1)OCCCC	**7.48**
45	**45**	CCCCCCCCc1cc(sc1c1ccc(s1)c1sc(cc1CCCCCCCC)/C=C(\C(=O)O)/C#N)c1ccc2c(c1)Sc1c(N2CC(CCCC)CC)cccc1	**6.56**
46	**46**	CCCCC(CN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N)CC	**6.14**
47	**47**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(\C(=O)O)/C#N)c1nc(c([nH]1)c1ccc(cc1)OC)c1ccc(cc1)OC	**5.52**
48	**48**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(\C(=O)O)/C#N)c1n	**4.87**

		c(c([nH]1)c1ccccc1)c1ccccc1	
49	**49**	CCCCCCOc1cc(OCCCCCC)ccc1n1c2cc(sc2c2c1ccs2)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**5.75**
50	**50**	CCCCCCOc1cc(OCCCCCC)ccc1n1c2cc(sc2c2c1cc(s2)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1	**7.7**
51	**51**	CCCCN1c2ccc(cc2Sc2c1cccc2)/C=C/c1ccc(cc1)/C=C\1/SC(=S)N(C1=O)CC(=O)O	**2.4**
52	**52**	CCCCN1c2ccc(cc2Sc2c1cccc2)/C=C\1/SC(=S)N(C1=O)CC(=O)O	**1.9**
53	**53**	CCCCN1c2ccc(cc2Sc2c1cccc2)/C=C/c1ccc(cc1)/C=C(/C(=O)O)\C#N	**4.8**
54	**54**	CCCC[C@@H](COc1ccc(cc1)c1ccc2c(c1)Sc1c(N2[C@@H]2CC=C(C=C2)c2ccc(cc2)C(=O)O)ccc(c1)c1ccc(cc1)OC[C@H](CCCC)CC)CC	**0.73**
55	**55**	CCCC[C@@H](COc1ccc(cc1)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)C(=O)O)ccc(c1)c1ccc(cc1)OC[C@H](CCCC)CC)CC	**0.99**
56	**56**	CCCCCCOc1ccc(cc1)N1c2ccc(cc2Sc2c1cccc2)/C=C/C1C/C(=C(/C(=O)O)\C#N)/CC(C1)(C)C	**5.4**
57	**57**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1cccc2)/C=C(/C(=O)O)\C#N)CC	**5.6**
58	**58**	CCCCCCn1c(c2ccc(cc2)c2ccc3c(c2)Sc2c(N3CCCCCC)cccc2)c2c(c1=O)c(n(c2=O)CCCCCC)c1ccc(cc1)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**4.56**
59	**59**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1ccc(c2)/C=C\1/SC(=S)N(C1=O)CC(=O)O)/C=C/1\SC(=S)N(C1=O)CC(=O)O)CC	**1.3**
60	**60**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1cccc2)/C=C\1/SC(=S)N(C1=O)CC(=O)O)CC	**0.4**
61	**61**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)/C=C(/C(=O)O)\C#N)CC	**6.8**
62	**62**	CCCCCCCCCCCC[C@H](CN1c2ccc(cc2Sc2c1ccc(c2)N1CCCC1)/C=C/1\SC(=S)N(C1=O)CC(=O)O)CCCCCCCCCC	**0.5**

63	**63**	CCCCCCCCCCCC[C@H](CN1c2ccc(cc2Sc2c1ccc(c2)n1c2ccccc2c2c1cccc2)/C=C/1\SC(=S)N(C1=O)CC(=O)O)CCCCCCCCCC	**1.3**
64	**64**	CCCCCCCCCCCC[C@H](CN1c2ccc(cc2Sc2c1ccc(c2)N1c2ccccc2Sc2c1cccc2)/C=C\1/SC(=S)N(C1=O)CC(=O)O)CCCCCCCCCC	**1.8**
65	**65**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C/1\SC(=S)N(C1=O)CC(=O)O)c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1	**1.9**
66	**66**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1)C)/C=C/1\SC(=S)N(C1=O)CC(=O)O	**1.2**
67	**67**	CCCCCCCCCCCC[C@H](CN1c2ccc(cc2Sc2c1ccc(c2)C#N)/C=C/1\SC(=S)N(C1=O)CC(=O)O)CCCCCCCCCC	**0.4**
68	**68**	CCCCCCCCCCCC[C@H](CN1c2ccc(cc2Sc2c1ccc(c2)/C=C/1\SC(=S)N(C1=O)CC(=O)O)Br)CCCCCCCCCC	**0.6**
69	**69**	CCCCCCCCCCCC[C@H](CN1c2ccccc2Sc2c1ccc(c2)/C=C/1\SC(=S)N(C1=O)CC(=O)O)CCCCCCCCCC	**0.7**
70	**70**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C(=O)O)C#N)c1ccc(cc1)N(c1ccc2c(c1)C(CCC)(CCC)c1c2cccc1)c1ccc2c(c1)C(CCC)(CCC)c1c2cccc1	**6.49**
71	**71**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C(=O)O)C#N)N(c1ccc(cc1)C(c1ccccc1)(C)C)c1ccc(cc1)C(c1ccccc1)(C)C	**6.14**
72	**72**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C(=O)O)C#N)c1ccc2c(c1)c1ccccc1n2CCCCCC	**7.13**
73	**73**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)C=C(C(=O)O)C#N	**6.63**
74	**74**	CCCCCCc1ccc(cc1)N1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**7.3**
75	**75**	CCCCCCc1ccc(cc1)N1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(\C(=O)O)/C#N)c1ccc(s1)/C=C(\C(=O)O)/C#N	**6.04**
76	**76**	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N1c2ccccc2Sc2c1cccc2)/C(=O)O	**4.47**

77	**77**	CCN1c2ccc(cc2Sc2c1cccc2)/C=C/c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N	**6.72**
78	**78**	CCCCC1(CCCC)c2cc(ccc2c2c1cc(cc2)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)N1c2ccccc2Sc2c1ccc(c2)C=C1c2ccccc2c2c1cccc2	**2.83**
79	**79**	CCCCCCn1c(c2ccc(cc2)c2ccc3c(c2)Sc2c(N3CCCCCC)cccc2)c2c(c1=O)c(n(c2=O)CCCCCC)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**3.26**
80	**80**	CCCCCCn1c(c2ccc(cc2)c2ccc(cc2)/C=C(/C(=O)O)\C#N)c2c(c1=O)c(n(c2=O)CCCCCC)c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1	**5.16**
81	**81**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C\1/SC(=S)N(C1=O)CC(=O)O)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1	**1.74**
82	**82**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)N(c1ccc2c(c1)c1ccccc1n2CCCCCC)c1ccc2c(c1)c1ccccc1n2CCCCCC	**5.93**
83	**83**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1nc2c(n1c1ccccc1)cccc2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.66**
84	**84**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1nc2c(n1c1ccccc1)cccc2)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.6**
85	**85**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1nc2c(n1c1ccccc1)cccc2)c1ccc(cc1)/C=C(\C(=O)O)/C#N	**3.31**
86	**86**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1nc2c(n1c1ccccc1)cccc2)c1ccc(s1)c1ccc(cc1)/C=C(\C(=O)O)/C#N	**3.24**
87	**87**	CCCCN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**1.63**
88	**88**	CCCCCCOc1ccc(cc1)N(c1ccc2c(c1)c1ccccc1n2CCCCCC)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**3.55**
89	**89**	CCCCN1c2ccc(cc2Sc2c1cccc2)/C=C(/C(=O)O)\C#N	**6.3**
90	**90**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C1SC(=S)N(C1=O)CC(=O)O)n1c2ccc(cc2c2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C	**3.31**

		)(C)C)C(C)(C)C)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C	
91	**91**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C1SC(=S)N(C1=O)CC(=O)O)n1c2ccc(cc2c2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C	**3.1**
92	**92**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C=C1SC(=S)N(C1=O)CC(=O)O	**3.58**
93	**93**	CCCCN1c2ccc(cc2Sc2c1ccc(c2)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)C=C1SC(=S)N(C1=O)CC(=O)O	**4.38**
94	**94**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)N(c1ccc2c(c1)c1ccccc1n2CCCCCC)c1ccc2c(c1)scc2	**3.04**
95	**95**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N1c2ccccc2Sc2c1cccc2)/C(=O)O	**4.01**
96	**96**	CCCCC1(CCCC)c2cc(ccc2c2c1cc(cc2)c1ccc(s1)/C=C(/C(=O)O)\C#N)N1c2ccccc2Sc2c1ccc(c2)C=C1c2ccccc2c2c1cccc2	**3.31**
97	**97**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2CCCCCCCCCCN2c3ccccc3Sc3c2ccc(c3)/C=C(/C(=O)O)\C#N)cccc1)/C(=O)O	**4.03**
98	**98**	N#C/C(=C\c1ccc(cc1)N(c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N1c2ccccc2Sc2c1cccc2)/C(=O)O	**2.66**
99	**99**	OC(=O)CN1C(=S)S/C(=C\c2ccc(cc2)N(c2ccc(cc2)N2c3ccccc3Sc3c2cccc3)c2ccc(cc2)/C=C/2\SC(=S)N(C2=O)CC(=O)O)/C1=O	**1.12**
100	**100**	CCCCCCSC1=C(SCCCCCC)SC(=Cc2ccc(cc2)c2ccc(cc2)c2ccc3c(c2)N(CCCC)c2c(S3)ccc(c2)/C=C(/C(=O)O)\C#N)S1	**6.59**
101	**101**	CCCCCCOc1cc(ccc1C=C1SC(=C(S1)SCCCCCC)SCCCCCC)c1ccc2c(c1)N(CCCC)c1c(S2)ccc(c1)/C=C(/C(=O)O)\C#N	**5.07**
102	**102**	CCCCN1c2cc(ccc2Sc2c1cc(cc2)/C=C(/C(=O)O)\C#N)c1ccc(cc1F)C=C1SC(=C(S1)SCCCCCC)SCCCCCC	**5.66**
103	**103**	CCCCN1c2cc(ccc2Sc2c1cc(cc2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)C=C1SC(=C(S1)SCCCCCC)SCCCCCC	**4.22**
104	**104**	CCCCN1c2cc(ccc2Sc2c1cc(cc2)/C=C(/C(=O)O)\C#N)C=C1SC(=C(S1)SCCCCCC)SCCCCCC	**4.56**

105	**105**	CCCCN1c2cc(ccc2Sc2c1cccc2)/C=C(/C(=O)O)\C#N	**4.39**
106	**106**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2CCCCCCCCN2c3ccccc3Sc3c2ccc(c3)/C=C(/C(=O)O)\C#N)cccc1)/C(=O)O	**4**
107	**107**	CCCCN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)c1cc2c(s1)c1c(n2c2ccc3c(c2)C(CCCC)(CCCC)c2c3cccc2)cc(s1)/C=C(\C(=O)O)/C#N	**4.87**
108	**108**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1	**5.24**
109	**109**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1	**7.45**
110	**110**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C\1/SC(=S)N(C1=O)CC(=O)O)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1	**2.57**
111	**111**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C#N)C#N)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1	**0.7**
112	**112**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C#N)C#N)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1	**1.06**
113	**113**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)Sc1c(N2C2CCCCC2)cccc1)C(=O)O	**3.2**
114	**114**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)c1ccccc1n2CC)C(=O)O	**1.12**
115	**115**	N#CC(=Cc1ccc(s1)c1ccc2c(c1)Sc1c(N2CCCCCCn2c3ccc(cc3c3c2cccc3)c2ccc(s2)C=C(C(=O)O)C#N)cccc1)C(=O)O	**1.61**
116	**116**	C(#N)/C(/C(=O)O)=C\C=1SC(=CC1)C=1C=CC=2N(C3=CC=C(C=C3SC2C1)C1=CC=C(C=C1)OC)C1=CC=C(C=C1)OC	**4.52**
117	**117**	CSC1=CC=C(C=C1)C=1C=C2SC=3C=C(C=CC3N(C2=CC1)CCCCCCOC1=CC=C(C=C1)N1C2=CC=C(C=C2SC=2C=C(C=CC12)C1=CC=C(S1)/C=C(/C(=O)O)\C#N)C1=CC=C(C=C1)OCCCCCCN1C2=CC=C(C=C2SC=2C=C(C=CC12)C=1SC(=CC1)\C=C(/C#N)\C(=O)O)C1=CC=C(C=C1)SC)C=1SC(=CC1)\C=C(/C#N)\C(=O)O	**6.09**
118	**118**	COC1=CC=C(C=C1)C=1C=C2SC=3C=C(C=CC3N(C2=CC1)CCCCCCOC1=CC=C(C=C1)N1C2=CC=C(C=C2SC=2C=C(C=CC1	**5.65**

		2)C1=CC=C(S1)/C=C(/C(=O)O)\C#N)C1=CC=C(C=C1)OCCCCCCCN1C2=CC=C(C=C2SC=2C=C(C=CC12)C=1SC(=CC1)\C=C(/C#N)\C(=O)O)C1=CC=C(C=C1)OC)C=1SC(=CC1)\C=C(/C#N)\C(=O)O	
119	**119**	CC1=CC=C(C=C1)C=1C=C2SC=3C=C(C=CC3N(C2=CC1)CCCCCCCOC1=CC=C(C=C1)N1C2=CC=C(C=C2SC=2C=C(C=CC12)C1=CC=C(S1)/C=C(/C(=O)O)\C#N)C1=CC=C(C=C1)OCCCCCCCN1C2=CC=C(C=C2SC=2C=C(C=CC12)C=1SC(=CC1)\C=C(/C#N)\C(=O)O)C1=CC=C(C=C1)C)C=1SC(=CC1)\C=C(/C#N)\C(=O)O	**5.21**
120	**120**	CCCCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2CC)ccc(c1)/C=C(/C(=O)O)\C#N	**6.32**
121	**121**	CCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)O	**5.23**
122	**122**	CCCCCCCCOc1ccc(cc1)N1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.82**
123	**123**	CCCCCCCCOc1ccc(cc1)N1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1cccs1	**5.12**
124	**124**	CCCCCCCCOc1ccc(cc1)N1c2ccccc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.34**
125	**125**	CCCCN1c2ccc(cc2Sc2c1cc(SC)cc2)/C=C(/C(=O)O)\C#N	**6.53**
126	**126**	CCCCN1c2ccc(cc2Sc2c1cc(SC)cc2)/C=C/c1ccc(s1)/C=C(/C(=O)O)\C#N	**7.44**
127	**127**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2CC)ccc(c1)/C=C(/C(=O)O)\C#N	**6.87**
128	**128**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**8.18**
129	**129**	CCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC	**8.08**
130	**130**	CCCCCCCCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC	**7.57**

131	**131**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)c1ccccc1n2CCCCCC	**6.72**
132	**132**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc2c(c1)c1ccccc1C2(CCCCCC)CCCCCC	**6.13**
133	**133**	CCCCCCn1c2ccccc2c2c1cc(cc2)N1c2ccccc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N	**5.12**
134	**134**	CCCCCCC1(CCCCCC)c2cc(ccc2c2c1cccc2)N1c2ccccc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N	**3.98**
135	**135**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)C=C(c1ccc(cc1)N1c2ccccc2Sc2c1cccc2)c1ccc(cc1)N1c2ccccc2Sc2c1cccc2	**6.55**
136	**136**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)C=C(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)n1c2ccccc2c2c1cccc2	**5.51**
137	**137**	CCCCCCn1c2ccc(cc2c2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)C=C(c1ccc(cc1)N1c2ccccc2Sc2c1cccc2)c1ccc(cc1)N1c2ccccc2Sc2c1cccc2	**2.69**
138	**138**	CCCCCCc1sc(c2c1OCCO2)c1sc(c2c1OCCO2)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**2.24**
139	**139**	CCCCCCc1sc(c2c1OCCO2)c1ccc(s1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**5.62**
140	**140**	CCCCCCc1ccc(s1)c1ccc(s1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**7.87**
141	**141**	CCCCCCc1sc(c2c1OCCO2)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**7.98**
142	**142**	CCCCCCc1ccc(s1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**8.07**
143	**143**	CCCCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCCCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**5.73**
144	**144**	CCCCCCCCCCCCOc1ccc(cc1)N1c2ccccc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N	**4.54**

145	**145**	CCCCCCCCOc1ccc(cc1)N1c2ccccc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N	**4.79**
146	**146**	COc1ccc(cc1)N1c2ccc(cc2Sc2c1cccc2)/C=C(/C(=O)O)\C#N	**4.53**
147	**147**	CCCCCCCCCCCCC1(CCCCCCCCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)Sc1c(N2CCCC)cccc1)/C=C(/C(=O)O)\C#N	**7**
148	**148**	N#C/C(=C/c1ccc(s1)/C=C/c1ccc2c(c1)Sc1c(N2)cc(cc1)SC)/C(=O)O	**7.4**
149	**149**	CCCCCCOc1c2cc(ccc2c(c2c1ccc(c2)/C=C(/C(=O)O)\C#N)OCCCCC)N(c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1)c1ccccc1	**2.77**
150	**150**	CCCCCCOc1c2cc(ccc2c(c2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)OCCCCCC)N(c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1)c1ccccc1	**4.64**
151	**151**	CCCCCCOc1c2cc(ccc2c(c2c1cccc2)OCCCCCC)N(c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N)c1ccccc1	**7.13**
152	**152**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(\C(=O)O)/C#N)N(c1ccc2c(c1)c(OCCCCCC)c1c(c2OCCCCCC)cccc1)c1ccccc1	**5.73**
153	**153**	CCCCCCOc1c2cc(ccc2c(c2c1cccc2)OCCCCCC)N(c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(\C(=O)O)/C#N)c1ccccc1	**4.09**
154	**154**	CCCCC(CN1c2ccc(cc2Sc2c1ccc(c2)/C=C\C(=O)O)/C=C/C1=CC(=C2C(=O)N(CC)C(=S)N(C2=O)CC)C=C(O1)/C=C\c1ccc2c(c1)Sc1c(N2CC(CCCC)CC)ccc(c1)/C=C/C(=O)O)CC	**0.93**
155	**155**	CCCCC(CN1c2ccc(cc2Sc2c1ccc(c2)/C=C/C(=O)O)/C=C\C1=CC(=C(C#N)C#N)C=C(O1)/C=C/c1ccc2c(c1)Sc1c(N2CC(CCCC)CC)ccc(c1)/C=N\C(=O)O)CC	**2.82**
156	**156**	CCCCC(CN1c2ccc(cc2Sc2c1ccc(c2)/C=C\C(=O)O)/C=C/C1=C/C(=C\2/C(=O)c3c(C2=C(C#N)C#N)cccc3)/C=C(O1)/C=C\c1ccc2c(c1)Sc1c(N2CC(CCCC)CC)ccc(c1)/C=C/C(=O)O)CC	**0.12**
157	**157**	CCCCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCCCCCCC)ccc(c1)c1ccc(o1)/C=C(/C(=O)O)\C#N	**7.33**
158	**158**	CCCCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCC	**6.44**

		CCCCCC)ccc(c1)c1ccc(s1)/C=C(/C(=O)O)\C#N	
159	**159**	CCCCCCc1ccc(s1)c1ccc(s1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)C=C(C(=O)O)C#N	**5.4**
160	**160**	CCCCCCc1ccc(s1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)C=C(C(=O)O)C#N	**4.3**
161	**161**	OC(=O)CN1C(=S)S/C(=C/c2ccc3c(c2)Sc2c(N3CCCCCn3c4ccc(cc4c4c3ccc(c4)C(C)(C)C)C(C)(C)C)cccc2)/C1=O	**2.5**
162	**162**	N#C/C(=C/c1ccc2c(c1)Sc1c(N2CCCCCn2c3ccc(cc3c3c2ccc(c3)C(C)(C)C)C(C)(C)C)cccc1)/C(=O)O	**6.3**
163	**163**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)/C=C\1/SC(=S)N(C1=O)C(=O)O	**5.6**
164	**164**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)/C=C(\C(=O)O)/C#N	**5.8**
165	**165**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1)OC)c1ccc(o1)/C=C(/C(=O)O)\C#N	**3.97**
166	**166**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccccc1)c1ccc(o1)/C=C(/C(=O)O)\C#N	**4.34**
167	**167**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1)OC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.03**
168	**168**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.87**
169	**169**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1)OC)c1cc(F)c(c(c1)F)/C=C(/C(=O)O)\C#N	**2.96**
170	**170**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccccc1)c1cc(F)c(c(c1)F)/C=C(/C(=O)O)\C#N	**2.79**
171	**171**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1)OC)c1ccc(c(c1)F)/C=C(/C(=O)O)\C#N	**3.41**
172	**172**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccccc1)c1ccc(c(c1)F)/C=C(/C(=O)O)\C#N	**3.43**
173	**173**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(cc1)OC)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**3.46**
174	**174**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccccc1)c1ccc(cc1)/C=	**3.53**

		C(/C(=O)O)\C#N	
175	**175**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OC	**4.15**
176	**176**	CC1(C)c2cc(ccc2c2c1cccc2)N1c2ccc(cc2Sc2c1ccc(c2)c1ccc2c(c1)C(C)(C)c1c2cccc1)c1ccc(s1)C#Cc1cc(cc(c1)c1cccnc1)c1cccnc1	**0.45**
177	**177**	OC(=O)c1cccc(c1)c1cc(C#Cc2ccc(s2)c2ccc3c(c2)Sc2c(N3c3ccc4c(c3)C(C)(C)c3c4cccc3)ccc(c2)c2ccc3c(c2)C(C)(C)c2c3cccc2)cc(c1)c1cccc(c1)C(=O)O	**4.55**
178	**178**	OC(=O)c1ccc(cc1)C#Cc1ccc(s1)c1ccc2c(c1)Sc1c(N2c2ccc3c(c2)C(C)(C)c2c3cccc2)ccc(c1)c1ccc2c(c1)C(C)(C)c1c2cccc1	**2.72**
179	**179**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(\C(=O)O)/C#N)C#Cc1ccc(cc1)C#Cc1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(\C(=O)O)/C#N	**7.44**
180	**180**	CCCCn1c2cc(C#Cc3ccc4c(c3)Sc3c(N4CCCCCC)ccc(c3)/C=C(\C(=O)O)/C#N)ccc2c2c1cc(C#Cc1ccc3c(c1)Sc1c(N3CCCCCC)ccc(c1)/C=C(\C(=O)O)/C#N)cc2	**6.64**
181	**181**	CCCCC1(CCCC)c2cc(ccc2c2c1cc(cc2)C#Cc1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(\C(=O)O)/C#N)C#Cc1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(\C(=O)O)/C#N	**6.72**
182	**182**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(\C(=O)O)/C#N)C#Cc1ccc(cc1)c1ccc(cc1)C#Cc1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(\C(=O)O)/C#N	**6.06**
183	**183**	CCC(CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)C1=CC=C(C1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)c1ccc2c(c1)ccc(c2c1c(OCCCCC(CC)CC)ccc2c1cccc2)OCCCCC(CC)CC)CC	**2.74**
184	**184**	CCC(CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)C1=CC=C(C1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)c1ccc2c(c1)ccc(c2)OCCCCC(CC)CC)CC	**5.32**

185	**185**	CCC(CCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)C1=CC=C(C1)/C=C(\C(=O)O)/C#N)c1ccc(cc1)OCCCCC(CC)CC)CC	**5.64**
186	**186**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)c1ccc(cc1)c1n(CCCCCC)c(=O)c2c1c(=O)n(c2c1ccc(cc1)C#Cc1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N)CCCCCC	**7.94**
187	**187**	N#C/C(=C\c1ccc(cc1)N(c1ccccc1)CCCCCCN1c2ccccc2Sc2c1ccc(c2)/C=C(/c1ccc(s1)/C=C(/C(=O)O)\C#N)\C#N)/c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.19**
188	**188**	N#C/C(=C\c1ccc(cc1)N(c1ccccc1)CCCCCCN1c2ccccc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)/C(=O)O	**4.66**
189	**189**	N#C/C(=C\c1ccc2c(c1)c1ccccc1n2CCCCCCc1ccc(c(c1)Nc1ccccc1)/C=C(/C(=O)O)\C#N)/C(=O)O	**2.82**
190	**190**	CCCCCCN(c1ccccc1)c1ccc(cc1)/C=C(/c1ccc(s1)/C=C(/C(=O)O)\C#N)\C#N	**3.8**
191	**191**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1sc(cc1CCCCCC)c1cc[n+](cc1)C(=O)O)c1ccc(cc1OCCCCCC)OCCCCCC	**7.1**
192	**192**	CCCCCCc1cc(sc1c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1)c1cc[n+](cc1)C(=O)O	**6.9**
193	**193**	CCSc1cc2c(cc1C=C(C(=O)O)C#N)Sc1c(N2c2ccc(cc2)OC)cccc1	**5.74**
194	**194**	CCCCN1c2cc(SCC)c(cc2Sc2c1cccc2)C=C(C(=O)O)C#N	**5.7**
195	**195**	COc1ccc(cc1)N1c2ccc(cc2Sc2c1cccc2)C=C(C(=O)O)C#N	**6.02**
196	**196**	CCCCN1c2ccc(cc2Sc2c1cccc2)C=C(C(=O)O)C#N	**5.78**
197	**197**	CCCCN1c2ccc(cc2Sc2c1cccc2)c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N	**6.09**
198	**198**	CCCCN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.32**
199	**199**	CCCCN1c2ccc(cc2Sc2c1cccc2)c1ccc(o1)/C=C(/C(=O)O)\C#N	**6.58**
200	**200**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)OC[C@H](CCCC)CC)c1cc	**2**

		c(s1)[C@@H]1[C@H]([O-])/C(=C/[C@@H]2[NH+](CCCCCCCC)c3c(C2(C)C)cc(cc3)C(=O)O)/C1=O	
201	**201**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C=C\c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)/C=C(/C(=O)O)\C#N	**5.3**
202	**202**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)/C=C(/C(=O)O)\C#N)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**6.1**
203	**203**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.2**
204	**204**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.2**
205	**205**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)/C=C\c1cc(c2ccc(s2)/C=C(\C(=O)O)/C#N)c(cc1/C=C/c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1)c1ccc(s1)C=C(C(=O)O)C#N	**4.5**
206	**206**	CCCCCCN1c2ccc(cc2Sc2c1cccc2)/C=C/c1cc(c(cc1c1ccc(s1)/C=C(/C(=O)O)\C#N)/C=C/c1ccc2c(c1)Sc1c(N2CCCCCC)cccc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**1.5**
207	**207**	CC(CCOc1ccc(cc1)N1c2ccc(cc2Sc2c1cccc2)/C=C/c1nc2cc(ccc2nc1/C=C/c1ccc2c(c1)Sc1c(N2c2ccc(cc2)OCCC(C)C)cccc1)C(=O)O)C	**4.36**
208	**208**	N#C/C(=C/c1ccc2c(c1)Sc1c(N2CCCCCCN2c3ccccc3Sc3c2ccc(c3)/C=C(/C(=O)O)\C#N)cccc1)/C(=O)O	**3.91**
209	**209**	N#C/C(=C\c1ccc2c(c1)Sc1c(N2CCCCCOc2ccc(cc2)[C@@](c2ccc(cc2)OCCCCCCN2c3ccccc3Sc3c2ccc(c3)/C=C(\C(=O)O)/C#N)(c2ccc(cc2)OCCCCCCN2c3ccccc3Sc3c2ccc(c3)/C=C(/C(=O)O)\C#N)C)cccc1)/C(=O)O	**4.9**
210	**210**	CCCCN1c2ccc(cc2Sc2c1cccc2)/C=C/c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.4**
211	**211**	CCN1c2ccc(cc2Sc2c1cccc2)C=CC(=O)O	**3.46**
212	**212**	CCN1c2ccc(cc2Sc2c1cccc2)C=C(C(=O)O)C#N	**5.53**

213	**213**	CCCCCCCCC1(CCCCCCCC)c2cc(ccc2c2c1cccc2)c1ccc2c(c1)Sc1c(N2c2ccc3c(c2)sc2c3sc3c2ccc(c3)c2cc(c(s2)C=C(C(=O)O)C#N)C)ccc(c1)c1ccc2c(c1)C(CCCCCCCC)(CCCCCCCC)c1c2cccc1	**4.54**
214	**214**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C(C(=O)O)C#N)C=C1SC(=C(S1)SC)SC	**8.01**
215	**215**	CCCCCCN1c2ccc(cc2Sc2c1ccc(c2)C=C1SC(=C(S1)SC)SC)c1ccc(cc1)C=C(C(=O)O)C#N	**8.12**
Test set compounds of Division 1 (compound ids) – 3,8,13,14,15,17,18,27,28,29,34,43,53,56,62,71,73,75,77,86,89,101,108,110,114,125,128,134,141,146,151,153,155,160,172,179,196,204,206			
Test set compounds of Division 2 (compound ids) – 2,6,22,30,41,49,55,58,59,60,64,70,83,88,97,100,105,113,116,121,127,131,132,140,145,148,149,159,161,162,165,171,187,191,194,195,210,212,215			
Test set compounds of Division 3 (compound ids) – 7,9,23,31,35,40,46,61,65,68,72,76,78,79,92,98,99,106,115,126,130,133,135,139,152,163,164,166,180,182,190,192,198,201,203,211,213,214			
Compounds removed based on leverage value – 1,32,38,39,90,91,111,112,117,118,138,154,156,169,170,176,177,183,186,189,200,209			

3.1.6 Porphyrin dataset (Study 2)

This dataset consists of 300 compounds, among which 30 compounds are removed from the dataset based on leverage value, and the remaining dataset is divided into three training and test sets with different data combinations. The detail of this dataset is shows in **Table 3.6**.

Table 3.6: Details of the porphyrin dataset

Serial Number	Compound ID	SMILE	Experimental PCE
1	**1**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(F)(F)F)/c1ccc(cc1)C(F)(F)F)/c1ccc(cc1)C(F)(F)F	**3**

2	**2**	CCc1cc(CC)c(c(c1)CC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(CC)cc(cc1CC)CC)/c1ccc(cc1)C(=O)O)/c1c(CC)cc(cc1CC)CC	**1.4**
3	**3**	OC(=O)c1ccc2c(c1)ccc1c2nc2c3/C(=C\4/C=CC(=N4)/C(=c/4\n5[Zn]n3c(c2n1)/C(=C\1/C=CC(=N1)/C(=c\5/cc4)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**5.1**
4	**4**	N#Cc1nc2cc3c(cc2nc1C(=O)O)c1n2c3/C(=C\3/C=CC(=N3)/C(=c/3\n([Zn]2)/c(=C(\C2=N/C(=C\1/c1c(C)cc(cc1C)C)/C=C2)/c1c(C)cc(cc1C)C)/cc3)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C	**0.8**
5	**5**	Cc1cc(C)c(c(c1)C)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)c1c5cc2c(c1)nc(cn2)C(=O)O)/c1c(C)cc(cc1C)C)/C=C4)\c1c(C)cc(cc1C)C	**6.3**
6	**6**	Cc1ccc(cc1)N(/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/N(c1ccc(cc1)C)c1ccc(cc1)C)cc5)/c1c(C)cc(cc1C)C)/C=C4)c1ccc(cc1)C	**10.1**
7	**7**	CCCCCCc1ccc(cc1)N(/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)cc5)/C#Cc1ccc(cc1)C(=O)O)/C=C4)\c1cc(cc(c1)C(C)(C)C)C(C)(C)C)c1ccc(cc1)CCCCCC	**9.1**
8	**8**	Cc1ccc(cc1)N(/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/N(c1ccc(cc1)C)c1ccc(cc1)C)cc5)/c1c(C)cc(cc1C)C)/C=C4)\c1ccc(cc1)C(=O)O)c1ccc(cc1)C	**8.3**
9	**9**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccccc1)/c1ccccc1)/c1ccccc1	**1.2**
10	**10**	Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C	**3**
11	**11**	COc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OC)/c1ccc(cc1)C(	**3**

		=O)O)/c1ccc(cc1)OC	
12	**12**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1ccc(cc1)C(=O)O)/c1c(C)cc(cc1C)C	**3.4**
13	**13**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1ccc(cc1)c1ccc(cc1)C(=O)O)/c1c(C)cc(cc1C)C	**2.4**
14	**14**	OC(=O)c1ccc(cc1)C#Cc1cc2n3c1/C=C/1\N=C(C=C1C#Cc1ccc(cc1)C(=O)O)/C(=c/1\n([Zn]3)/c(=C(\C3=N/C(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C3)/N(c2ccc(cc2)C(C)(C)C)c2ccc(cc2)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**4**
15	**15**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**6.2**
16	**16**	OC(=O)/C=C/C1=CC2=N/C/1=C\c1ccc3n1[Zn]n1/c(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc/c/1=C(/C1=N/C(=C\3/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C1)\N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C	**3.2**
17	**17**	OC(=O)/C=C/C=C/C1=CC2=N/C/1=C\c1ccc3n1[Zn]n1/c(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc/c/1=C(/C1=N/C(=C\3/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C1)\N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C	**4.22**
18	**18**	OC(=O)/C=C/C=C/c1cc2n3c1/C=C/1\N=C(C=C1/C=C/C=C/C(=O)O)/C(=c/1\n([Zn]3)/c(=C(\C3=N/C(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C3)/N(c2ccc(cc2)C(C)(C)C)c2ccc(cc2)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**4.36**
19	**19**	N#C/C(=C/C1=CC2=N/C/1=C(/c1ccc(cc1)C(C)(C)C)\c1ccc3n1[Zn]n1/c(=C\2/c2ccc(cc2)C(C)(C)C)/cc/c/1=C(\c1ccc(cc1)C(C)(C)C)/C1=N/C(=C\3/c2ccc(cc2)C(C)(C)C)/C(=C1c1ccccc1)c1ccccc1)/C(=O)O	**1.72**

20	**20**	N#C/C(=C/C1=CC2=N/C/1=C(/c1c(C)cc(cc1C)C)\c1ccc3n1[Zn]n1/c(=C\2/c2ccc(cc2)C(C)(C)C)/cc/c/1=C(/C1=N/C(=C\3/c2c(C)cc(cc2C)C)/C=C1)\c1c(C)cc(cc1C)C)/C(=O)O	**5.35**
21	**21**	Cc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc(cc1)C)c1ccc(cc1)C)/c1c(C)cc(cc1C)C)/c1ccc(cc1)C(=O)O)c1ccc(cc1)C	**3.8**
22	**22**	OC(=O)c1ccc(cc1)C#Cc1cc2n3c1/C=C\1/C=CC(=N1)/C(=c/1\n([Zn]3)/c(=C(\C3=N/C(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C3)/N(c2ccc(cc2)C(C)(C)C)c2ccc(cc2)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**5.9**
23	**23**	Cc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O)/N(c1ccc(cc1)C)c1ccc(cc1)C)c1ccc(cc1)C	**5.5**
24	**24**	Cc1ccc(cc1)N(/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)C(=O)O)/C=C4)\c1c(C)cc(cc1C)C)c1ccc(cc1)C	**6.5**
25	**25**	Cc1cc(C)c(c(c1)C)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1c(C)cc(cc1C)C)/C=C4	**4.4**
26	**26**	CCCCO/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C=C/C(=O)O)/c1ccccc1)/c1ccccc1	**6.04**
27	**27**	CCCCO/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C=C(/C(=O)O)\C#N)/c1ccccc1)/c1ccccc1	**3.08**
28	**28**	CCCCO/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C=C/C=C(/C(=O)O)\C#N)/c1ccccc1)/c1ccccc1	**1.44**
29	**29**	CCn1c2ccccc2c2c1c1c3cc(ccc3n(c1c1c2n(CC)c2c1cccc2)CC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**4.83**

30	**30**	N#C/C(=C\c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)cc5)/c1ccc2c(c1)c1c(n2CC)c2c3ccccc3n(c2c2c1n(CC)c1c2cccc1)CC)/C=C4)\c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/C(=O)O	**6.05**
31	**31**	OC(=O)c1ccc(cc1)C#Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**4.01**
32	**32**	OC(=O)c1ccc(cc1)C#Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C#Cc1ccc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**2.55**
33	**33**	OC(=O)c1ccc(cc1)C#Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1ccc(cc1)C#Cc1ccc(cc1)C(=O)O)/c1ccc(cc1)C#Cc1ccc(cc1)C(=O)O	**0.58**
34	**34**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccnc(c1)C(=O)O)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)OCCCCCCCCCCCC	**8.3**
35	**35**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccnc(c1)O)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)OCCCCCCCCCCCC	**8.5**
36	**36**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccncc1)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)OCCCCCCCCCCCC	**8.2**
37	**37**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc(cc1)C)c1ccc(cc1)C)/N(c1ccc(cc1)C)c1ccc(cc1)C)/CCc1ccc(cc1)C(=O)O)CO	**9.1**

		CCCCCCCC	
38	**38**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc(cc1)C)c1ccc(cc1)C)/N(c1ccc(cc1)C)c1ccc(cc1)C)/CCc1c(OCCCC)cc(cc1OCCCC)C(=O)O)COCCCCCCCC	**7.2**
39	**39**	CCCCCCCCOCc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc(cc1)C)c1ccc(cc1)C)/N(c1ccc(cc1)C)c1ccc(cc1)C)/CCc1c(OCCCCCCCC)cc(cc1OCCCCCCCC)C(=O)O)OCCCCCCCC	**6.9**
40	**40**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/CCc1ccc(cc1)C(=O)O)OCCCCCCCC	**8.6**
41	**41**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O	**0.8**
42	**42**	Oc1c(O)cccc1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cccc(c1O)O)/c1cccc(c1O)O)/c1cccc(c1O)O	**0.6**
43	**43**	Oc1cc(ccc1O)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(c(c1)O)O)/c1ccc(c(c1)O)O)/c1ccc(c(c1)O)O	**0.23**
44	**44**	Oc1ccc(c(c1)O)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1O)O)/c1ccc(cc1O)O)/c1ccc(cc1O)O	**0.03**
45	**45**	Oc1ccc(c(c1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(O)ccc1O)/c1cc(O)ccc1O)/c1cc(O)ccc1O)O	**0.02**
46	**46**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=O)O)/c1ccc(cc1)N(c1ccc(cc1)OC	**3.4**

		CCCCC)c1ccc(cc1)OCCCCCC)/c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	
47	**47**	OC(=O)c1ccc(cc1)C#C/C/1=c/2\cc/c/3=C(\c4ccc(cc4)N(c4ccccc4)c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)cc5)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C=C4	**2.06**
48	**48**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**8.42**
49	**49**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(c5c4nsn5)c4ccc(cc4)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**12.5**
50	**50**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(s1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)\N(c1ccc(cc1)c1ccc(s1)CCCCCC)c1ccc(cc1)c1ccc(s1)CCCCCC)OCCCCCCCC	**5.62**
51	**51**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)N(c4ccc(cc4)c4ccc(s4)CCCCCC)c4ccc(cc4)c4ccc(s4)CCCCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(s1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**6.4**
52	**52**	OC(c1cccc(c1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)N(c4ccccc4)c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)cc5)/c1cccc(c1)C(=O)O)/C=C4)O	**1.81**
53	**53**	OC(=O)c1cccc(c1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/c1cccc(c1)C(=O)O)/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**3.37**
54	**54**	OC(=O)c1cccc(c1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)N(c4ccccc4)c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c	**3.72**

		1cccc(c1)C(=O)O)cc5)/c1cccc(c1)C(=O)O)/C=C4	
55	**55**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**2.88**
56	**56**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**2.84**
57	**57**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**4.02**
58	**58**	OC(=O)c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C(=O)O)cc5)/c1ccc(cc1)C(=O)O)/C=C4)\c1ccc(cc1)N(c1ccccc1)c1ccccc1	**5.36**
59	**59**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cccs1)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O	**3**
60	**60**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cccs1)/c1cccs1)/c1ccc(cc1)C(=O)O	**2.5**
61	**61**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(=O)O)/c1ccc s1)/c1cccs1	**1.8**
62	**62**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cccs1)/c1cccs1)/c1ccc s1	**0.2**
63	**63**	Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c\2/cc/c(=C(/C3=N/C(=C(\c4[nH]c1cc4)/c1ccc(cc1)C(=O)O)/C=C3)\c1ccc(cc1)C)/[nH]2)/c1ccc(cc1)C	**0.42**
64	**64**	Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c\2/cc/c(=C(/C3=N/C(=C(\c4[nH]c1cc4)/c1cccc(c1)C(=O)O)/C=C3)\c1ccc(cc1)C)/[nH	**0.71**

		]2)/c1ccc(cc1)C	
65	**65**	Cc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C)/c1cccc(c1)C(=O)O)/c1ccc(cc1)C	**4.17**
66	**66**	CCCCCCc1c(C#C/C/2=c/3\cc/c/4=C(\c5ccc(cc5)N(c5ccccc5)c5ccccc5)/C5=N/C(=C(\c6n([Zn]n34)c(/C(=C/3\N=C2C=C3)/c2ccc(cc2)N(c2ccccc2)c2ccccc2)cc6)/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C=C5)sc(c1CCCCCC)/C=C(/C(=O)O)\C#N	**6**
67	**67**	CCCCCCc1c(C#C/C/2=c/3\cc/c/4=C(\c5ccc(cc5)N(c5ccccc5)c5ccccc5)/C5=N/C(=C(\c6n([Zn]n34)c(/C(=C/3\N=C2C=C3)/c2ccc(cc2)N(c2ccccc2)c2ccccc2)cc6)/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C=C5)sc(c1CCCCCC)/C=C/c1sc(c(c1CCCCCC)CCCCCC)/C=C(/C(=O)O)\C#N	**4.1**
68	**68**	CC(c1cc(cc(c1)C(C)(C)C)/C/1=c/2\cc/c/3=C/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(cc1)C(=O)O)cc5)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/C=C4)(C)C	**0.22**
69	**69**	Cc1cc(C)c(c(c1)C)Oc1cccc2c1cc1c(cccc1c2/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)cc5)/C#Cc1ccc(cc1)C(=O)O)/C=C4)\c1cc(cc(c1)C(C)(C)C)C(C)(C)C)Oc1c(C)cc(cc1C)C	**0.66**
70	**70**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**6.36**
71	**71**	CCCCCCc1ccc(cc1)N(c1ccc(cc1)CCCCCC)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**5.6**
72	**72**	OC(=O)C(=C/C=C/C1=C/C/2=C(\c3ccccc3)/c3ccc4n3[Zn]n3/c(=C(\C1=N2)/c1ccccc1)/cc/c/3=C(/C1=N/C(=C\4/c2ccccc2)/C=C1)\c1ccccc1)C(=O)O	**5.1**

73	**73**	Cc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C(=C2)/C=C/C=C(C(=O)O)C(=O)O)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4	**7.1**
74	**74**	CCc1ccc(cc1)/C/1=C\2/N=C(C(=C2)/C=C/C=C(C(=O)O)C(=O)O)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)CC)/c1ccc(cc1)CC)/c1ccc(cc1)CC	**5.8**
75	**75**	CCCCc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=C(C(=N1)/C(=c\3/cc2)/c1ccc(cc1)CCCC)/C=C/C=C(C(=O)O)C(=O)O)/c1ccc(cc1)CCCC)/c1ccc(cc1)CCCC	**6.4**
76	**76**	CCCCCCCCc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=C(C(=N1)/C(=c\3/cc2)/c1ccc(cc1)CCCCCCCC)/C=C/C=C(C(=O)O)C(=O)O)/c1ccc(cc1)CCCCCCCC)/c1ccc(cc1)CCCCCCCC	**5.3**
77	**77**	Cc1cc(C)cc(c1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C(=C2)/C=C/C=C(C(=O)O)C(=O)O)/c1cc(C)cc(c1)C)cc5)/c1cc(C)cc(c1)C)/C=C4)\c1cc(C)cc(c1)C	**6.1**
78	**78**	CCCCCCCCCCCCOc1cc(cc(c1OCCCCCCCCCCCC)OCCCCCCCCCCCC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OCCCCCCCCCCCC)c(c(c1)OCCCCCCCCCCC)OCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O	**4.78**
79	**79**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCC	**7.43**
80	**80**	CN(c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)C	**9.34**
81	**81**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#	**10.17**

		Cc1ccc(cc1)N(C)C)OCCCCCCCCCCCC	
82	**82**	CCCCCCCCCCCCOc1ccc(cc1)C1=C2CCC3N2[Zn]24n5c1ccc5C(=C1N2C(CC1)C(=c1n4c(=C3c2ccc(cc2)OCCCCCCCCCCCC)cc1)c1ccc(cc1)C(=O)O)c1ccc(cc1)OCCCCCCCCCCCC	**4.97**
83	**83**	CCCCCCCCCCCCOc1cc(ccc1OCCCCCCCCCCCC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/CCC(/C(=c\3/cc2)/c2ccc(cc2)C(=O)O)N1)/c1ccc(c(c1)OCCCCCCCCCCCC)OCCCCCCCCCCC)/c1ccc(c(c1)OCCCCCCCCCCCC)OCCCCCCCCCCC	**5.98**
84	**84**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)c4nc(nc(n4)c4ccccc4)c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**0.97**
85	**85**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc2c(c1)c1ccccc1n2CC)OCCCCCCCC	**4.4**
86	**86**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)OCCCCCCCC	**5**
87	**87**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)N(c4ccc(cc4)C)c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**5.58**
88	**88**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1cccc(c1)C(=O)O)/C#Cc1ccc2c(c1)c1ccccc1n2CC)OCCCCCCCC	**3.75**
89	**89**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)N(c4ccc(cc4)C)c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1cccc(c1)C(=O)O)cc5)/c1c(OCCCCCCCC)cc	**4.3**

		cc1OCCCCCCCC)/C=C4)OCCCCCCCC	
90	**90**	COCCOCCOCCOc1ccc(cc1)/C/1=c/2\cc/c(=C(\c3ccc(cc3)OCCOCCOCOC)/C3=N/C(=C(\c4[nH]c(/C(=C/5\N=C1C=C5)/c1ccc(cc1)OCCOCCOCCOC)cc4)/c1ccc(cc1)C(=O)O)/C=C3)/[nH]2	**0.65**
91	**91**	COCCOCCOCCOc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)OCCOCCOCOC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)OCCOCCOCCOC)cc5)/c1ccc(cc1)C(=O)O)/C=C4	**2.42**
92	**92**	COCCOCCOCCOc1ccc(cc1)/C/1=c/2\cc/c(=C(\c3ccc(cc3)OCCOCCOCOC)/C3=N/C(=C(\c4[nH]c(/C(=C/5\N=C1C=C5)/c1ccc(cc1)OCCOCCOCCOC)cc4)/c1cccc(c1)C(=O)O)/C=C3)/[nH]2	**0.65**
93	**93**	COCCOCCOCCOc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)OCCOCCOCOC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)OCCOCCOCCOC)cc5)/c1cccc(c1)C(=O)O)/C=C4	**2.45**
94	**94**	CCCCCCO/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C=c\3/cc2)/C(=O)O)/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**3.64**
95	**95**	CCCCCCO/C/1=c/2\cc/c/3=C(\c4ccc(cc4)N(c4ccccc4)c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C(=O)O)cc5)/OCCCCCC)/C=C4	**5.77**
96	**96**	CCCCCCOc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OCCCCCC)/c1ccc(cc1)N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)/c1cccc(c1)C(=O)O)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OCCCCCC)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)/c1ccc(cc1)C(=O)O	**2.3**
97	**97**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)N(CCCC)CCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**7.56**

98	**98**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)N(c1ccc(cc1)CCCCCC)CCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**9**
99	**99**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**9.19**
100	**100**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**7.93**
101	**101**	CCCc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)CCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)CCC)cc5)/c1ccc(cc1)C(=O)O)/C=C4	**2.09**
102	**102**	CCCc1ccc(cc1)/C/1=c/2\cc/c(=C(/c3ccc(cc3)CCC)\C3=N/C(=C(\c4[nH]c(/C(=C\5/N=C1C=C5)/c1ccc(cc1)CCC)cc4)/c1cccc(c1)C(=O)O)/C=C3)/[nH]2	**0.3**
103	**103**	CCCc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)CCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)CCC)cc5)/c1cccc(c1)C(=O)O)/C=C4	**1.99**
104	**104**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc1c4/c(=C\4/C=CC(=N4)/C(=c\3/cc2)/c2c(C)cc(cc2C)C)/c2c3c1cccc3c(cc2)C(=O)O)/c1c(C)cc(cc1C)C	**4.1**
105	**105**	COC1CCC2C3C1CCCC3C1=C/C/3=C(\C4C(C)CC(CC4C)C)/C4N5[Zn]N6/C(=C/2\C1=N3)/CC/C/6=C(/C1=N/C(=C(\C5CC4)/C2CCC(CC2)C(=O)O)/C=C1)\C1C(C)CC(CC1C)C	**1.1**
106	**106**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1ccc(c2c1cccc2)C(=O)O)/c1c(C)cc(cc1C)C	**2.8**
107	**107**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4	**8.79**

		c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/N(c1ccccc1)c1ccccc1)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	
108	**108**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4 c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/N(c1ccc(cc1)C)c1ccc(cc1)C)/C#Cc1ccc(cc1)C (=O)O)OCCCCCCCC	**8.92**
109	**109**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4 c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/N(c1ccc2c(c1)cccc2)c1ccc2c(c1)cccc2)/C#Cc 1ccc(cc1)C(=O)O)OCCCCCCCC	**7.58**
110	**110**	COC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1 ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C(=O)OC)/c 1ccc(cc1)C#Cc1cccnc1)/c1ccc(cc1)C(=O)OC	**3.36**
111	**111**	COC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1 ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccncc1)/c1ccc(cc1)C (=O)OC)/c1ccncc1	**2.27**
112	**112**	COC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1 ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccncc1)/c1ccc(cc1)C (=O)OC)/c1ccc(cc1)C(=O)OC	**3.35**
113	**113**	n1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(= C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccncc1)/c1ccncc1)/c1ccncc1	**2.46**
114	**114**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc c4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1ccc(s 1)C(=O)O)/c1c(C)cc(cc1C)C	**3.1**
115	**115**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc c4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1c(C)c c(cc1C)C)/c1ccc(o1)C(=O)O	**2.3**
116	**116**	Cc1cc(C)c(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1cc c4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1scc(c 1)C(=O)O)/c1c(C)cc(cc1C)C	**1.8**

117	**117**	Cc1cc(C)c(c(c1)C)/C/1=C\2/N=C(c3c2nc2cc(ccc2n3)C(=O)O)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C	**5.2**
118	**118**	Cc1cc(C)c(c(c1)C)/C/1=C\2/N=C(c3c2nc2cc(C(=O)O)c(cc2n3)C(=O)O)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C	**4**
119	**119**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/n1c2ccc(cc2c2c1ccc(c2)CCCCCC)CCCCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	**3.99**
120	**120**	CCCCCCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)c4n(CC(CCCC)CC)c(=O)c5c4c(=O)n(c5c4ccc(cc4)c4ccc(s4)/C=C(/C(=O)O)\C#N)CC(CCCC)CC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)OC)cc5)/c1c(OCCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C=C4)OCCCCCCCCCCCC	**2.48**
121	**121**	CCCCCCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)c4n(CC(CCCC)CC)c(=O)c5c4c(=O)n(c5c4ccc(cc4)c4ccc(s4)/C=C(/C(=O)O)\C#N)CC(CCCC)CC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/n1c2ccc(cc2c2c1ccc(c2)CCCCCC)CCCCCC)cc5)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C=C4)OCCCCCCCCCCCC	**2.58**
122	**122**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\c4ccc(cc4)N(c4ccc(cc4)c4ccc(s4)CCCCCC)c4ccc(cc4)c4ccc(s4)CCCCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(s1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**5.09**
123	**123**	Cc1ccc(cc1)/C/1=c/2\cc/c(=C(\c3ccc(cc3)C)/C3=N/C(=C(\c4[nH]c(/C(=C/5\N=C1C=C5)/c1ccc(cc1)C)cc4)/c1ccccc1C(=O)O)/C=C3)/[nH]2	**0.11**
124	**124**	Cc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C)cc5)/c1ccccc1C(=O)O)/C=C4	**0.37**

125	**125**	COc1ccc(c2c1cccc2)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)C(=O)O)/C=C4)\c1c(C)cc(cc1C)C	**3.4**
126	**126**	Oc1ccc(c(c1)C(=O)O)/N=N/c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)/N=N/c1ccc(cc1C(=O)O)O)/c1ccc(cc1)/N=N/c1ccc(cc1C(=O)O)O)/c1ccc(cc1)/N=N\c1ccc(cc1C(=O)O)O	**0.56**
127	**127**	OC(=O)c1cc(/N=N\c2ccc(cc2)/C/2=C/3\C=CC(=N3)/C(=c/3\n4[Zn]n5c2ccc5/C(=C\2/C=CC(=N2)/C(=c\4/cc3)/c2ccc(cc2)/N=N/c2ccc(c(c2)C(=O)O)O)/c2ccc(cc2)/N=N\c2ccc(c(c2)C(=O)O)O)/c2ccc(cc2)/N=N\c2ccc(c(c2)C(=O)O)O)ccc1O	**2.04**
128	**128**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C=c\3/cc2)/c1ccccc1)/c1ccccc1	**2.26**
129	**129**	OC(=O)c1cccc(c1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C=c\3/cc2)/c1ccccc1)/c1ccccc1	**0.64**
130	**130**	OC(=O)c1ccc(cc1O)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C=c\3/cc2)/c1ccccc1)/c1ccccc1	**4.55**
131	**131**	OC(=O)c1cc(ccc1O)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C=c\3/cc2)/c1ccccc1)/c1ccccc1	**1.22**
132	**132**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C=c\3/cc2)/c1ccccc1)/c1ccccc1	**3.01**
133	**133**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/c1ccccc1)/c1ccccc1	**3.47**
134	**134**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N1c2ccccc2CCc2c1cccc2)/c1ccccc1)/c1ccccc1	**3.62**
135	**135**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N1c2ccccc2C	**2.55**

		=Cc2c1cccc2)/c1ccccc1)/c1ccccc1	
136	**136**	N#C/C(=C\c1sc(cc1C)c1ccc(c2c1nsn2)c1sc(c(c1)C)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C)/c1ccc(cc1)C)/c1ccc(cc1)C)/C(=O)O	**5.46**
137	**137**	CCCCCCc1cc(sc1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C)/c1ccc(cc1)C)/c1ccc(cc1)C)c1ccc(c2c1nsn2)c1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N	**4.67**
138	**138**	CCCCCCc1ccc(s1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(s1)CCCCCC)cc5)/c1ccc(s1)/C=C(/C(=O)O)\C#N)/C=C4)\c1ccc(s1)CCCCCC	**4.44**
139	**139**	N#C/C(=C\c1ccc(s1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(s1)C)/c1ccc(s1)C)/c1ccc(s1)C)/C(=O)O	**5.71**
140	**140**	CCCCCCc1cc(sc1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(s1)C)/c1ccc(s1)C)/c1ccc(s1)C)/C=C(/C(=O)O)\C#N	**4.54**
141	**141**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1cc(OCCCCCC)c(cc1OCCCCCC)n1c2ccccc2c2c1cccc2)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCC	**8**
142	**142**	N#C/C(=C\c1sc(c2c1OCCO2)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(s1)C)/c1ccc(s1)C)/c1ccc(s1)C)/C(=O)O	**4.54**
143	**143**	N#C/C(=C\c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)C)/C=C4)\c1c(C)cc(cc1C)C)/C(=O)O	**2.61**
144	**144**	N#C/C(=C\c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)/C=C/c1ccc(cc1)/C=C/c1ccc2c(c1)c1cc(ccc1n2CCC(CCCC(	**3.26**

		C)C)C)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C=C4)\c1c(C)cc(cc1C)C)/C(=O)O	
145	**145**	N#C/C(=C\c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)/C=C/c1ccc(cc1)/C=C/c1ccc2c(c1)c1cc(ccc1n2CCC(CCCC(C)C)C)c1ccc(cc1)C(F)(F)F)/C=C4)\c1c(C)cc(cc1C)C)/C(=O)O	**3.22**
146	**146**	N#C/C(=C\c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)/C=C/c1ccc2c(c1)c1cc(ccc1n2CC[C@@H](CCCC(C)C)C)c1ccc(cc1)C)/C=C4)\c1c(C)cc(cc1C)C)/C(=O)O	**3.28**
147	**147**	COc1ccc(cc1)c1ccc2c(c1)c1cc(/C=C/c3ccc(cc3)/C=C/c3ccc(cc3)/C/3=C/4\C=CC(=N4)/C(=c/4\n5[Zn]n6c3ccc6/C(=C\3/C=CC(=N3)/C(=c\5/cc4)/c3ccc(cc3)/C=C(/C(=O)O)\C#N)/c3c(C)cc(cc3C)C)/c3c(C)cc(cc3C)C)ccc1n2CCC(CCCC(C)C)C	**3.35**
148	**148**	N#C/C(=C\c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)/C=C/c1ccc(cc1)/C=C/c1ccc2c(c1)c1cc(ccc1n2CCC(CCCC(C)C)C)c1cccc(c1)C(F)(F)F)/C=C4)\c1c(C)cc(cc1C)C)/C(=O)O	**3.42**
149	**149**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/c1ccc(cc1)N(C)C)/C#Cc1cc2c(s1)c(c(s2)C(=O)O)C)OCCCCCCCC	**4.53**
150	**150**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C#Cc1ccc(s1)C(=O)O)OCCCCCCCC	**3.98**
151	**151**	CCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC)cccc1OCCCCCC)/c1ccc(cc1)C(=O)O)OCCCCCC	**3.94**
152	**152**	CCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC)cccc1OC	**2.93**

		CCCCC)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O)OCCCCCC	
153	**153**	CCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC)cccc1OCCCCC)/C=C/C(=O)O)OCCCCCC	**4.16**
154	**154**	OC(=O)CNc1nc(Nc2ccc(cc2)/C/2=C/3\C=CC(=N3)/C(=c/3\n4[Zn]n5c2ccc5/C(=C\2/C=CC(=N2)/C(=c\4/cc3)/c2ccccc2)/c2ccccc2)/c2ccccc2)nc(n1)Nc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/c1ccccc1)/C=C4	**3.61**
155	**155**	OC(=O)CNc1nc(Nc2ccc(cc2)/C/2=C/3\C=CC(=N3)/C(=c\3/cc/c(=C(/C4=N/C(=C(\c5[nH]c2cc5)/c2ccccc2)/C=C4)\c2ccccc2)/[nH]3)/c2ccccc2)nc(n1)Nc1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/c1ccccc1)/C=C4)\c1ccccc1	**4.46**
156	**156**	N#C/C(=C\C1=CC2=N/C/1=C(/c1ccccc1)\c1ccc3n1[Zn]n1/c(=C\2/c2ccccc2)/cc/c/1=C(\c1ccccc1)/C1=N/C(=C\3/c2ccccc2)/C=C1)/C(=O)O	**3.12**
157	**157**	N#C/C(=C\C1=CC2=N/C/1=C(/c1ccc(cc1)C)\c1ccc3n1[Zn]n1/c(=C\2/c2ccc(cc2)C)/cc/c/1=C(/C1=N/C(=C\3/c2ccc(cc2)C)/C=C1)\c1ccc(cc1)C)/C(=O)O	**2.12**
158	**158**	N#C/C(=C\C1=CC2=N/C/1=C(/c1ccc(cc1)CC(C)C)\c1ccc3n1[Zn]n1/c(=C\2/c2ccc(cc2)CC(C)C)/cc/c/1=C(\c1ccc(cc1)CC(C)C)/C1=N/C(=C\3/c2ccc(cc2)CC(C)C)/C=C1)/C(=O)O	**1.56**
159	**159**	N#C/C(=C\C1=CC2=N/C/1=C(/c1ccc(cc1)C(C)(C)C)\c1ccc3n1[Zn]n1/c(=C\2/c2ccc(cc2)C(C)(C)C)/cc/c/1=C(/C1=N/C(=C\3/c2ccc(cc2)C(C)(C)C)/C=C1)\c1ccc(cc1)C(C)(C)C)/C(=O)O	**1.01**
160	**160**	OC(=O)c1ccc(cc1)C#CC1=CC2=N/C/1=C(/c1ccccc1)\c1ccc3n1[Zn]n1/c(=C\2/c2ccccc2)/cc/c/1=C(\c1ccccc1)/C1=N/C(=C\3/c2ccccc2)/C=C1	**3.55**
161	**161**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc	**2.1**

		c1OCCCCCCCCC)/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C#Cc1cc2c(s1)c(c(s2)C(=O)O)C)OCCCCCCCCC	
162	**162**	Cc1ccc(cc1)/C/1=C\2/N=C(C=C2C#Cc2ccc(cc2)C(=O)O)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)C)/c1ccc(cc1)C)/c1ccc(cc1)C	**4.18**
163	**163**	CC(Cc1ccc(cc1)/C/1=C\2/N=C(C=C2C#Cc2ccc(cc2)C(=O)O)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)CC(C)C)/c1ccc(cc1)CC(C)C)/c1ccc(cc1)CC(C)C)C	**4.79**
164	**164**	OC(=O)c1ccc(cc1)C#CC1=CC2=N/C/1=C(/c1ccc(cc1)C(C)(C)C)\c1ccc3n1[Zn]n1/c(=C\2/c2ccc(cc2)C(C)(C)C)/cc/c/1=C(/C1=N/C(=C\3/c2ccc(cc2)C(C)(C)C)/C=C1)\c1ccc(cc1)C(C)(C)C	**5.08**
165	**165**	OC(=O)c1ccc(cc1)C#C/C/1=C\2/N=C(c3c2cccc3)/C(=c/2\n3[Zn]n4c1c1ccccc1c4/C(=C/1\N=C(/C(=c\3/c3c2cccc3)/c2ccccc2)c2c1cccc2)/c1ccccc1)/c1ccccc1	**0.95**
166	**166**	CCCCCCCCCCCCOc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OCCCCCCCCCCC)/C#Cc1ccc(cc1)N(CCCCCCCCCCCC)CCCCCCCCCCCC)/c1ccc(cc1)C(=O)O	**1.89**
167	**167**	CCCCCCCCCCCCOc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OCCCCCCCCCCC)/C(=C(C#N)C#N)C(=C(C#N)C#N)c1ccc(cc1)N(CCCCCCCCCCCC)CCCCCCCCCCCC)/c1ccc(cc1)C(=O)O	**0.38**
168	**168**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCC	**7.09**
169	**169**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/c1ccc(cc1)C(=O)O)OCCCCCCCCCCC	**4.77**
170	**170**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[	**5.92**

		Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/c1ccc(cc1)/C=C(/C(=O)O)\C#N)OCCCCCCCCCCCC	
171	**171**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N)OCCCCCCCCCCCC	**5.75**
172	**172**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C#Cc1ccc(c(c1)C#N)C(=O)O)OCCCCCCCCCCC	**7.59**
173	**173**	CCCCCCCCCCCCOc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc(cc1)n1c2ccccc2c2c1cccc2	**2.22**
174	**174**	CCCCCCCCCCCCOc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)OCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/c1ccc(cc1)n1c2ccccc2c2c1cccc2	**5.51**
175	**175**	N#C/C(=C/c1sc(cc1C)c1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4)/C(=O)O	**5.55**
176	**176**	CCCCCCc1cc(sc1/C=C(\C(=O)O)/C#N)c1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4	**5.8**
177	**177**	N#C/C(=C/c1sc(c2c1OCCO2)c1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4)/C(=O)O	**6.47**
178	**178**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1c2ccccc2c(c2c1cccc2	**9.73**

		)C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc(cc1)N(C)C)OCCCCCCCCCCCC	
179	**179**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1c2ccccc2c(c2c1cccc2)C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1)OCCCCCCCCCCC	**6.97**
180	**180**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1c2ccccc2c(c2c1cccc2)C#Cc1ccc(cc1)N(C)C)OCCCCCCCCCCCCC	**9.51**
181	**181**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1c2ccccc2c(c2c1cccc2)C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1)OCCCCCCCCCCC	**6.7**
182	**182**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1ccc(cc1)N(C)C)/C#CC(=O)O)OCCCCCCCCCCCC	**10.03**
183	**183**	CCCCCCc1ccc(cc1)N(/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)cc5)/N(c1cccc(c1)CCCCCC)c1cccc(c1)CCCCCC)/C=C4)c1ccc(cc1)CCCCCC	**6.8**
184	**184**	CCCCCCc1cccc(c1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(cc1)N(c1cccc(c1)OC)c1ccc(cc1)OC)/C#Cc1ccc(cc1)C(=O)O)c1cccc(c1)CCCCCC	**4.2**
185	**185**	CCCCCCCCc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=	**5.5**

		O)O)/c1ccc(cc1)N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)/c1ccc(cc1)N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)c1ccc(cc1)CCCCCCCC	
186	**186**	CCCCCCCCc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=O)O)/c1cc(c(c(c1)C(C)(C)C)OC)C(C)(C)C)/c1cc(c(c(c1)C(C)(C)C)OC)C(C)(C)C)c1ccc(cc1)CCCCCCCC	**7**
187	**187**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)c1ccccc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**6.3**
188	**188**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc2c(c1)Cc1c2cccc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**8.1**
189	**189**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc2c3c1cccc3CC2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**7.4**
190	**190**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1cncnc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**5.4**
191	**191**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ncccn1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**4.1**
192	**192**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/c1ccc(s1)N(c1ccccc1)c1ccccc1)OCCCCCCCCCCCC	**6.01**
193	**193**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC	**4.23**

		CCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N)/c1ccc(s1)N(c1ccccc1)c1ccccc1)OCCCCCCCCCCCC	
194	**194**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccc(s1)/C=C(\C(=O)O)/C#C)/c1ccc(s1)N(c1ccccc1)c1ccccc1)OCCCCCCCCCCCC	**4.38**
195	**195**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1sc(c(c1)CCCCCC)/C=C(\C(=O)O)/C#C)/c1ccc(s1)N(c1ccccc1)c1ccccc1)OCCCCCCCCCCCC	**5**
196	**196**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccc(cc1)N(C)C)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	**9.25**
197	**197**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c\2/cc/c(=C(/C3=N/C(=C(\c4[nH]c1cc4)/C#Cc1ccc(cc1)N(C)C)/C=C3)\c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/[nH]2)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	**5.62**
198	**198**	CCCCCCCCN(c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)CCCCCCCC	**8.77**
199	**199**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)N(C)C)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**6.39**
200	**200**	CCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCC)cccc1OCCCC)/C#Cc1ccc(cc1)N(C)C)/C#Cc1ccc(cc1)C(=O)O)OCCCC	**5.91**
201	**201**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[	**10.24**

		Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1ccc(cc1)N(CCCCCCCC)CCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	
202	**202**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccccc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**5.11**
203	**203**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1cccc2c1cccc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**7.83**
204	**204**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1c2ccccc2cc2c1cccc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**6.62**
205	**205**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1cc2ccccc2c2c1cccc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**8.26**
206	**206**	OC(=O)c1ccc(cc1)C#C/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc2c3c1ccc1c3c(cc2)ccc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**10.06**
207	**207**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1ccc(cc1)n1c2ccccc2c2c1cccc2)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	**6.84**
208	**208**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCCC)/C#Cc1cc(OCCCCCC)c(cc1OCCCCCC)n1c2ccccc2c2c1cccc2)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCC	**7.94**
209	**209**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[	**8.59**

		Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC CCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1cc(OC)c(cc1OC)n1c2 ccccc2c2c1cccc2)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	
210	**210**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1c cc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N(c1ccccc1)c 1ccccc1)/c1ccccc1)/c1ccccc1	**2.09**
211	**211**	COc1cc(cc(c1OC)OC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c 1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OC)c(c(c1)OC)O C)/c1ccc(cc1)C(=O)O)/c1cc(OC)c(c(c1)OC)OC	**1.75**
212	**212**	COc1cc(cc(c1OC)OC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c 1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OC)c(c(c1)OC)O C)/c1ccc(cc1)/N=C/c1ccc(cc1)C(=O)O)/c1cc(OC)c(c(c1)OC)O C	**1.06**
213	**213**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC CCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C# Cc1cc(OCCCCCC)c(cc1OCCCCCC)N(c1ccccc1)c1ccccc1)OCCC CCCCCCCCC	**8.6**
214	**214**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCC CCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1cc(OCCCCCC)c(cc1O C)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OC)/C#Cc1ccc(cc1)C(=O) O)OCCCCCCCCCCC	**8.7**
215	**215**	CCCCCCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4cc(OC CCCCC)c(cc4OCCCCCC)N(c4ccc(cc4)OCCCCCC)c4ccc(cc4)OCC CCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc 1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCC CCCCCC)/C=C4)OCCCCCCCCCCC	**9.1**
216	**216**	CCCCCCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4cc(OC CCCCC)c(cc4OCCCCCC)N(c4ccc(cc4)OCCCCCC)c4ccc(cc4)OCC CCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc	**9.5**

		1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C=C4)OCCCCCCCCCCCC	
217	**217**	CCCCCCc1cc(sc1/C=C(/C(=O)O)\C#N)c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C=C4)\c1ccccc1.CCCCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCCCC)c1ccc(cc1)c1ccc(s1)c1ccc(c2c1nccn2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**7.88**
218	**218**	CCCCCCc1cc(sc1/C=C(/C(=O)O)\C#N)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C=C4)\c1ccccc1.CCCCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCCCC)c1ccc(cc1)c1ccc(s1)c1ccc(c2c1nccn2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**8.14**
219	**219**	Cc1cc(C)c(c(c1)C)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)N(c4ccccc4)c4ccccc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C(=O)O)cc5)/c1c(C)cc(cc1C)C)/C=C4	**2.74**
220	**220**	CCCCCCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4c5ccccc5c(c5c4cccc5)C#Cc4ccc(cc4)N(CCCCCCCC)CCCCCCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCC)/C=C4)OCCCCCCCCCCCC	**6.5**
221	**221**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)/C#Cc1c2ccccc2c(c2c1cccc2)N(CCCCCCCC)CCCCCCCC)OCCCCCCCCCCCC	**5.5**
222	**222**	CCCCCCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4c5ccccc5c(c5c4cccc5)N(c4ccc(cc4)CCCCCCCC)c4ccc(cc4)CCCCCCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCC)/C=C4)OCCCCCCCCCCCC	**10.3**

223	**223**	CCCCCCCCCCCCC#CC#CCCCCCCCCCNC(=O)c1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C(=O)NCCCCCCCCCC#CC#CCCCCCCCCCCCC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C(=O)NCCCCCCCCCC#CC#CCCCCCCCCCCCC)cc5)/c1ccc(cc1)C(=O)O)/C=C4	**2.3**
224	**224**	Br/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/Br)/c1ccc(cc1)C(C)(C)C)/c1ccncc1	**3.57**
225	**225**	Br/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C(C)(C)C)cc5)/Br)/C=C4)\c1ccnc(c1)c1ccccn1	**5.08**
226	**226**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc5c6c4cc(c4ccc(cc4OCCCCCC)OCCCCCC)c4n6c(c(c5)c5ccc(cc5OCCCCCC)OCCCCCC)cc4)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**7.35**
227	**227**	N#C/C(=C\c1ccc(s1)c1ccc(s1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(s1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/c1ccc(s1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/c1ccc(s1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/C(=O)O	**6.25**
228	**228**	N#C/C(=C\c1ccc(s1)c1ccc(s1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/C(=O)O	**5.09**
229	**229**	N#C/C(=C\c1ccc(s1)c1ccc(s1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/c1ccc(cc1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/C(=O)O	**3.5**
230	**230**	CCCCC(Cc1ccc(cc1)N(c1ccc(cc1)CC(CCCC)CC)c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=	**3.36**

		C2)/c1c(C)cc(cc1C)C)cc5)/c1ccc(cc1)C(=O)O)/C=C4)\c1c(C)cc(cc1C)C)CC	
231	**231**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)N(c4ccc(cc4)OC)c4ccc(cc4)OC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(cc1)C(=O)O)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**8.1**
232	**232**	CCCCCCCCOc1cccc(c1/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)N(c4ccc(cc4)OC)c4ccc(cc4)OC)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N)cc5)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C=C4)OCCCCCCCC	**5.5**
233	**233**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc(cc1)N(CCCC)CCCC)OCCCCCCCC	**6.6**
234	**234**	CCCCCOCc1c(OC)cc(cc1OC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OC)c(c(c1)OC)OCCCCCC)/c1cc2cccc3c2c2c1cccc2cc3)/C#Cc1ccc(s1)/C=C(\C(=O)O)/C#N	**1.35**
235	**235**	CCCCCOCc1c(OC)cc(cc1OC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OC)c(c(c1)OC)OCCCCCC)/c1cc2cccc3c2c2c1cccc2cc3)/C#Cc1ccc(s1)C=C(C(=O)O)C(=O)O	**3.14**
236	**236**	CCCCCOCc1c(OC)cc(cc1OC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OC)c(c(c1)OC)OCCCCCC)/c1ccc2c(c1)C(CCCCCC)(CCCCCC)c1c2cccc1)/C#Cc1ccc(s1)/C=C(\C(=O)O)/C#N	**0.92**
237	**237**	COc1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C/1\N=C(/C(=c\3/cc2)/c2ccc(cc2)OC)C=C1/C=C(\C(=O)O)/C#N)/c1ccc(cc1)OC)/c1ccc(cc1)OC	**2.27**
238	**238**	CCCCCOCc1c(OC)cc(cc1OC)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(OC)c(c(c1	**1.28**

		)OC)OCCCCCC)/c1ccc2c(c1)C(CCCCCC)(CCCCCC)c1c2cccc1)/C#Cc1ccc(s1)C=C(C(=O)O)C(=O)O	
239	**239**	CCCCCCCCc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(c2c1cccc2)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)c1ccc(cc1)CCCCCCCC	**7.23**
240	**240**	CCCCCCCCc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1c2ccccc2c(c2c1cccc2)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)c1ccc(cc1)CCCCCCCC	**2.57**
241	**241**	CCCCCCCCc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(c2c1cccc2)/C=C(/C(=O)O)\C#N)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)c1ccc(cc1)CCCCCCCC	**4.28**
242	**242**	CCCCCCCCc1ccc(cc1)N(/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1c2ccccc2c(c2c1cccc2)/C=C(/C(=O)O)\C#N)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)c1ccc(cc1)CCCCCCCC	**2.86**
243	**243**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(c2c1cccc2)C(=O)O)OCCCCCCCC	**8.04**
244	**244**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1c2ccccc2c(c2c1cccc2)C(=O)O)OCCCCCCCC	**6.03**
245	**245**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(c2c1cccc2)/C=C(/C(=O)O)\C#N)OCCCCCCCC	**5.8**
246	**246**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4	**3.64**

		c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C# Cc1c2ccccc2c(c2c1cccc2)/C=C(/C(=O)O)\C#N)OCCCCCCCC	
247	**247**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4 c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/C#Cc1c2ccccc2c(c2c1cccc2)N(c1ccc(cc1)OC) c1ccc(cc1)OC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**7.7**
248	**248**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4 c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/C#CC1=CC=C(CC1)n1c2ccc(cc2c2c1ccc(c2)C(C)(C)C)C(C)(C)C)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCC	**4.6**
249	**249**	CCCCCCc1cc(sc1/C=C(/C(=O)O)\C#N)c1ccc(cc1)/C/1=c/2\cc/ c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1cc ccc1)cc5)/c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C=C4)\c1ccccc1	**4.9**
250	**250**	CCCCCCc1cc(sc1/C=C(/C(=O)O)\C#N)c1ccc(c2c1nc(c1ccccc1) c(n2)c1ccccc1)c1ccc(cc1)/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c 5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/c1ccc(cc1) n1c2ccccc2c2c1cccc2)/C=C4)\c1ccccc1	**6.02**
251	**251**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4 c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)ccc c1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C# Cc1ccc(=O)c(cc1)O)OCCCCCCCC	**7.7**
252	**252**	OC(=O)c1ccc(cc1)c1ccc2c(c1)C1=N/C/2=N\c2n3[Zn]n4/c(=N \1)/c1cc(Oc5ccc(cc5)C(CC(C)(C)C)(C)C)c(cc1/c/4=N/C1=N/C(=N\c3c3c2cc(Oc2ccc(cc2)C(CC(C)(C)C)(C)C)c(c3)Oc2ccc(cc2) C(CC(C)(C)C)(C)C)/c2c1cc(c(c2)Oc1ccc(cc1)C(CC(C)(C)C)(C)C) Oc1ccc(cc1)C(CC(C)(C)C)(C)C)Oc1ccc(cc1)C(CC(C)(C)C)(C)C	**0.73**
253	**253**	OC(=O)c1ccc2c(c1)C1=N/C/2=N\c2c3cc(ccc3c3n2[Zn]n2/c(= N\1)/c1ccc(cc1/c/2=N/C1=N/C(=N\3)/c2c1ccc(c2)C(C)(C)C)C (C)(C)C)C(C)(C)C	**1.54**
254	**254**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]N	**8.32**

		4C1C=CC4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)/C#Cc1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)c1ccc(cc1)OCCCCCC)OCCCCCCCC	
255	**255**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)C#C/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\C5N([Zn]n23)C(/C(=C/2\N=C1C=C2)/c1c(cccc1OCC(CCCC)CC)OCC(CCCC)CC)C=C5)/C#Cc1ccc(cc1)C(=O)O)/C=C4)\c1c(OCCC(CCC)CC)cccc1OCC(CCCC)CC	**8.28**
256	**256**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]N4C1C=CC4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)/C#Cc1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)c1ccc(cc1)OCCCCCC)OCCCCCCCC	**6.46**
257	**257**	CCCCCCOc1ccc(cc1)c1ccc2c(c1)Sc1c(N2CCCCCC)ccc(c1)C#C/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\C5N([Zn]n23)C(/C(=C/2\N=C1C=C2)/c1c(cccc1OCC(CCCC)CC)OCC(CCCC)CC)C=C5)/C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)/C=C4)\c1c(OCCC(CCC)CC)cccc1OCC(CCCC)CC	**7.5**
258	**258**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1cccc2)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)/C=C(\C(=O)O)/C#N)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)CC	**5.21**
259	**259**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1cccc2)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(s1)/C=C(\C(=O)O)/C#N)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)CC	**5.8**
260	**260**	CCCC[C@@H](CN1c2ccc(cc2Sc2c1cccc2)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)CC	**6.25**
261	**261**	Cc1ccc(cc1)N1c2ccc(cc2C2C1CCC2)/C/1=C/2\C=CC(=N2)/C(	**5.59**

		=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C#Cc1ccc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	
262	**262**	CCCCCCc1c(/C=C/c2ccc(cc2)/C/2=C/3\C=CC(=N3)/C(=c/3\n4[Zn]n5c2ccc5/C(=C\2/C=CC(=N2)/C(=c\4/cc3)/c2c(C)cc(cc2C)C)/c2c(C)cc(cc2C)C)/c2c(C)cc(cc2C)C)sc(c1CCCCCC)/C=C/c1sc(c(c1CCCCCC)CCCCCC)/C=C(/C(=O)O)\C#N	**4.77**
263	**263**	N#C/C(=C\c1sc(c2c1OCCO2)/C=C/c1sc(c2c1OCCO2)/C=C/c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C)/c1c(C)cc(cc1C)C)/C(=O)O	**1.64**
264	**264**	OC(=O)CCC/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/CCCC(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**0.14**
265	**265**	OC(=O)/C=C/C1=CC2=N/C/1=C\c1[nH]c(cc1/C=C/C(=O)O)/C(=C\1/C=CC(=N1)/C(=c/1\[nH]/c(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**0.63**
266	**266**	OC(=O)/C=C/c1cc2n3c1/C=C/1\N=C(C=C1/C=C/C(=O)O)/C(=c/1\n([Zn]3)/c(=C(\C3=N/C(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C3)/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**1.48**
267	**267**	OC(=O)/C=C/C=C/C1=CC2=N/C/1=C\c1ccc3n1[Zn]n1/c(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc/c/1=C(/C1=N/C(=C\3/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C1)\c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**2.08**
268	**268**	OC(=O)/C=C/C=C/c1cc2n3c1/C=C/1\N=C(C=C1/C=C/C=C/C(=O)O)/C(=c/1\n([Zn]3)/c(=C(\C3=N/C(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C3)/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**2.37**
269	**269**	CCCC[C@H](CN1c2ccc(cc2Sc2c1cccc2)/C/1=C/2\C=CC(=N2)/	**4.01**

		C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc1)C(=O)O)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)CC	
270	**270**	OC(=O)C(=C/C=C/c1cc2n3c1/C=C/1\N=C(C=C1/C=C/C=C(C(=O)O)C(=O)O)/C(=c/1\n([Zn]3)/c(=C(\C3=N/C(=C\2/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/C=C3)/c2cc(cc(c2)C(C)(C)C)C(C)(C)C)/cc1)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C)C(=O)O	**3.03**
271	**271**	CN(c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c\2/cc/c(=C(/C3=N/C(=C(\c4[nH]c1cc4)/c1ccc(cc1)C(=O)O)/C=C3)\c1ccc(cc1)N(C)C)/[nH]2)/c1ccc(cc1)C(=O)O)C	**3.8**
272	**272**	CN(c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccc(cc1)N(C)C)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O)C	**4.83**
273	**273**	CCCCOc1ccccc1/C/1=C\2/N=C(C=C2C#Cc2ccc(cc2)/C=C(/C(=O)O)\C#N)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccccc1OCCCC)/c1ccccc1OCCCC)/c1ccccc1OCCCC	**5.42**
274	**274**	CCCCCCOc1ccccc1/C/1=C\2/N=C(C=C2C#Cc2ccc(cc2)/C=C(/C(=O)O)\C#N)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccccc1OCCCCCC)/c1ccccc1OCCCCCC)/c1ccccc1OCCCCCC	**6.09**
275	**275**	CCCCCCCCOc1ccccc1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C/1\N=C(/C(=c\3/cc2)/c2ccccc2OCCCCCCCC)C=C1C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N)/c1ccccc1OCCCCCCCC)/c1ccccc1OCCCCCCCC	**6.14**
276	**276**	CCCCCCCCCCCCOc1ccccc1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C/1\N=C(/C(=c\3/cc2)/c2ccccc2OCCCCCCCCCCCC)C=C1C#Cc1ccc(cc1)/C=C(/C(=O)O)\C#N)/c1ccccc1OCCCCCCCCCC)/c1ccccc1OCCCCCCCCCCCC	**6.32**
277	**277**	CCCCC(COc1ccccc1/C/1=C\2/N=C(C=C2C#Cc2ccc(cc2)/C=C(/C(=O)O)\C#N)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1ccccc1OCC(CCCC)CC)/c1ccccc1OCC(CCCC)CC)/	**5.72**

		c1ccccc1OCC(CCCC)CC)CC	
278	**278**	N#C/C(=C\c1ccc(cc1)C#CC1=CC2=N/C/1=C(/c1ccccc1OC1CCCCC1)\c1ccc3n1[Zn]n1/c(=C\2/c2ccccc2OC2CCCCC2)/cc/c/1=C(/C1=N/C(=C\3/c2ccccc2OC2CCCCC2)/C=C1)\c1ccccc1OC1CCCCC1)/C(=O)O	**5.59**
279	**279**	CCCCCCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCCCCC)cccc1OCCCCCCCCCCCC)/C#Cc1c2ccccc2c(c2c1cccc2)C#CN(CCCCCCCC)CCCCCCCC)/C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCC	**9.95**
280	**280**	OC(=O)C(c1ccc2c(c1)C1=N/C/2=N\c2n3[Zn]n4/c(=N\1)/c1cc(ccc1/c/4=N/C1=N/C(=N\c3c3c2ccc(c3)C(C)(C)C)/c2c1cc(cc2)C(C)(C)C)C(C)(C)C)C(=O)O	**3.05**
281	**281**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/N=c/2\n3[Zn]N4C1C=CC4/N=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C	**1.56**
282	**282**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/N=c/2\n3[Zn]N4C1C=CC4/N=C\1/C=CC(=N1)/C(=c\3/cc2)/c1cc(cc(c1)C(C)(C)C)C(C)(C)C.CCC1=C(C)C2=C(c3ccc(cc3)C#CCn3ccnc3)c3n([B-]([N+]2=C1C)(F)F)c(c(c3C)CC)C	**3.82**
283	**283**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]N4C1C=CC4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(c2c1nsn2)c1ccc(cc1)C(=O)O)/C#Cc1ccc(cc1)N(CCCCCCCC)CCCCCCCC)OCCCCCCCC	**7.37**
284	**284**	Cc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C(=C2)c1ccc(o1)C=C(C(=O)O)C(=O)O)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4	**0.102**
285	**285**	N#C/C(=C\c1ccc(o1)c1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)C)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4)/C(=O)O	**0.106**
286	**286**	Cc1ccc(cc1)/C/1=c/2\cc/c/3=C(\c4ccc(cc4)c4ccc(o4)C=C(C(=	**0.0013**

		O)O)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccc(cc1)C)cc5)/c1ccc(cc1)C)/C=C4	
287	**287**	N#C/C(=C\c1ccc(cc1)/C=C\C\1=C\2/C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/C=C/c1ccc(cc1)/C=C(/c1ccc(cc1)[N+](=O)[O-])\C#N)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)C(=O)O)/c1ccc(cc1)[N+](=O)[O-]	**2.9**
288	**288**	Cc1cc(C)c(c(c1)C)/C/1=c/2\cc/c/3=C(\C#Cc4c(F)c(F)c(c(c4F)F)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)cc5)/c1c(C)cc(cc1C)C)/C=C4	**4.6**
289	**289**	Cc1cc(C)c(c(c1)C)/C/1=c/2\cc/c/3=C(\C#Cc4ccc(cc4)C(=O)O)/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/N(c1ccc(cc1)C(C)(C)C)c1ccc(cc1)C(C)(C)C)cc5)/c1c(C)cc(cc1C)C)/C=C4	**6.9**
290	**290**	N#C/C(=C/c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc2c(c1)C(C)(C)c1c2cccc1)c1ccc2c(c1)C(C)(C)c1c2cccc1)/c1ccccc1)/c1ccccc1)/C(=O)O	**4.37**
291	**291**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(c2c1nsn2)C(=O)O)OCCCCCCCC	**2.52**
292	**292**	Oc1cc2/C/3=N/c4c5ccc(cc5c5n4[Zn]n4/c(=N\C(=N3)c2cc1O)/c1cc(ccc1/c/4=N/C1=N/C(=N\5)/c2ccc(cc12)C(C)(C)C)C(C)(C)C)C(C)(C)C	**0.92**
293	**293**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]N4C1C=CC4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/C#Cc1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(cc1)C(=O)O)/C#Cc1ccc(cc1)N(CCCCCCCC)CCCCCCCC)OCCCCCCCC	**6.46**
294	**294**	N#C/C(=C\c1ccc(o1)C1=C/C/2=C(\c3ccc(cc3)C)/c3ccc4n3[Zn	**0.133**

		]n3/c(=C(\C1=N2)/c1ccc(cc1)C)/cc/c/3=C(\c1ccc(cc1)C)/C1=N/C(=C\4/c2ccc(cc2)C)/C=C1)/C(=O)O	
295	**295**	OC(=O)c1ccc(cc1)/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/N(c1ccc2c(c1)C(C)(C)c1c2cccc1)c1ccc2c(c1)C(C)(C)c1c2cccc1)/c1ccccc1)/c1ccccc1	**3.16**
296	**296**	OC(=O)c1ccc(cc1)C#C/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/N(c1ccccc1)c1ccccc1)/C=C4)\c1ccccc1	**3.6**
297	**297**	OC(=O)c1ccc(cc1O)C#C/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/N(c1ccccc1)c1ccccc1)/C=C4)\c1ccccc1	**5.8**
298	**298**	N#C/C(=C/c1ccc(cc1)C#C/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/N(c1ccccc1)c1ccccc1)/C=C4)\c1ccccc1)/C(=O)O	**2.9**
299	**299**	OC(=O)C(=Cc1ccc(cc1)C#C/C/1=c/2\cc/c/3=C(/C4=N/C(=C(\c5n([Zn]n23)c(/C(=C/2\N=C1C=C2)/c1ccccc1)cc5)/N(c1ccccc1)c1ccccc1)/C=C4)\c1ccccc1)C(=O)O	**1.4**
300	**300**	CCCCCCCCOc1cccc(c1/C/1=C/2\C=CC(=N2)/C(=c/2\n3[Zn]n4c1ccc4/C(=C\1/C=CC(=N1)/C(=c\3/cc2)/c1c(OCCCCCCCC)cccc1OCCCCCCCC)/N(c1ccc(cc1)CCCCCC)c1ccc(cc1)CCCCCC)/C#Cc1ccc(cc1)/C=C(\C(=O)O)/O)OCCCCCCCC	**8.5**
Test set compounds of Division 1 (compound ids) - **5,10,15,20,25,30,35,40,47,52,57,62,67,72,77,82,87,98,103,109,115,122,129,134,141,146,152,159,164,169,174,179,184,189,194,199,204,209,214,219,226,232,237,242,247,254,259,264,269,274,279,287,294,299**			
Test set compounds of Division 2 (compound ids) – **11,12,22,25,28,40,48,55,60,62,69,72,75,76,89,97,98,103,111,114,118,130,131,134,143,145,161,162,163,165,167,175,183,186,187,191,192,195,196,197,201,203,205,211,220,226,230,232,243,271,274,279,284,297**			
Test set compounds of Division 3 (compound ids) – **4,6,7,8,15,23,26,31,32,38,39,52,54,56,70,79,83,102,107,117,125,128,142,147,150,152,153,160,169,176,180,190,193,194,204,208,209,210,212,218,228,249,251,256,261,263,265,268,270,277,286,294,295,299**			
Compounds removed based on leverage value –			

1,42,45,90,91,92,93,95,96,105,113,120,121,126,127,136,137,148,154,155,224,225,227,252,253,280,281, 282,288,292

3.1.7 Triphenylamine dataset (Study 2)

This dataset consists of 244 compounds, among which 25 compounds are removed from the dataset based on leverage value, and the remaining dataset is divided into three training and test sets with different data combinations. The detail of this dataset is shows in **Table 3.7**.

Table 3.7: Details of the triphenylamine dataset

Serial Number	Compound ID	SMILE	Experimental PCE
1	**1**	CCCCCCC(Cc1cc(sc1c1cc2c(s1)c1sc(cc1c1c2non1)c1sc(cc1CC(CCCCCC)CCCC)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1)CCCC	**6.18**
2	**2**	N#C/C(=C\c1ccc(cc1)/C=C/c1cccc(c1)N(c1ccccc1)c1ccccc1)/C(=O)O	**2.23**
3	**3**	N#CC(=Cc1ccc(cc1)C=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**4.12**
4	**4**	N#CC(=Cc1cccc(c1)N(c1ccccc1)c1ccccc1)C(=O)O	**1.27**
5	**5**	N#CC(=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**2.92**
6	**6**	OC(=O)c1ccc(cc1)/C=C/c1cccc(c1)N(c1ccccc1)c1ccccc1	**0.48**
7	**7**	OC(=O)c1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**1.78**
8	**8**	OC(=O)c1cccc(c1)N(c1ccccc1)c1ccccc1	**0.08**
9	**9**	OC(=O)Cc1ccc(cc1)N(c1ccccc1)c1ccccc1	**0.83**
10	**10**	N#C/C(=C\1/C=C(/C=C/c2ccc(cc2)N(c2ccc(cc2)n2c3ccccc3c3c2cccc3)c2ccc(cc2)n2c3ccccc3c3c2cccc3)CC(C1)(C)C)/C(=O)O	**6.02**
11	**11**	N#C/C(=C\1/C=C(/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)CC(C1)(C)C)/C(=O)O	**5.77**
12	**12**	CCCCCCOc1c(OCCCCCC)c(c2ccc(s2)/C=C(/C(=O)O)\C#N)c2c	**5.57**

		(c1c1ccc(cc1)N(c1ccccc1)c1ccccc1)nsn2	
13	**13**	CCCCCCOc1c(OCCCCCC)c(c2ccc(s2)/C=C(/C(=O)O)\C#N)c2c(c1c1ccc3c(c1)C(CC)(CC)c1c3ccc(c1)N(c1ccccc1)c1ccccc1)nsn2	**5.34**
14	**14**	CCCCCCOc1c(OCCCCCC)c(c2ccc(s2)/C=C(/C(=O)O)\C#N)c2c(c1C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)nsn2	**6.22**
15	**15**	CCCCCCOc1c(OCCCCCC)c(c2ccc(s2)/C=C(/C(=O)O)\C#N)c2c(c1c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)nsn2	**6.72**
16	**16**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.01**
17	**17**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccccc1)c1ccccc1)C=C(C(=O)O)C#N	**6.62**
18	**18**	CCCCCCc1cc(sc1c1ccc(s1)c1sc(cc1CCCCCC)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1sc(c(c1)CCCCCC)c1ccc(s1)c1sc(cc1CCCCCC)/C=C(\C(=O)O)/C#N	**4.81**
19	**19**	CCCCCCc1cc(sc1c1ccc(s1)c1sc(cc1CCCCCC)C=C(C(=O)O)C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**5.96**
20	**20**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**3.92**
21	**21**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**2.75**
22	**22**	OC(=O)CN1C(=S)S/C(=C/c2cc(c3cccs3)c(cc2c2cccs2)/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**1.19**
23	**23**	OC(=O)CN1C(=S)S/C(=C/c2cc(Cl)c(cc2Cl)/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**1.29**
24	**24**	OC(=O)CN1C(=S)S/C(=C/c2cc(Br)c(cc2Br)/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**1.14**
25	**25**	OC(=O)CN1C(=S)S/C(=C/c2ccc(cc2)/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**3.27**
26	**26**	N#C/C(=C\c1cc(C#N)c(cc1C#N)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**0.44**

27	**27**	N#C/C(=C\c1cc(c2cccs2)c(cc1c1cccs1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.49**
28	**28**	N#C/C(=C\c1cc(Cl)c(cc1Cl)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.86**
29	**29**	N#C/C(=C\c1cc(Br)c(cc1Br)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.36**
30	**30**	N#C/C(=C\c1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.33**
31	**31**	OC(=O)CN1C(=S)S/C(=C/c2ccc(cc2)N(c2ccccc2)c2ccc(cc2)/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**2.3**
32	**32**	CCCCCCCCCCCCOc1cc(C#Cc2ccc(cc2)N(c2ccccc2)c2ccccc2)c(cc1C#Cc1ccc(cc1)C=C(C(=O)O)C#N)OCCCCCCCCCCCC	**3.61**
33	**33**	CCCCCCCCCCCCOc1cc(C#Cc2ccc(cc2)N(c2ccccc2)c2ccccc2)c(cc1C#Cc1ccc(cc1)C(=O)O)OCCCCCCCCCCCC	**3.06**
34	**34**	N#CC(=Cc1ccc(cc1)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**2.89**
35	**35**	OC(=O)c1ccc(cc1)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1	**2**
36	**36**	N#C/C(=C\c1cc2c(s1)c1c(s2)cc(s1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**2.76**
37	**37**	OC(=O)C(=Cc1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**3.33**
38	**38**	CCOC(=O)/C(=C\c1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**2.43**
39	**39**	N#C/C(=C\c1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/S(=O)(=O)O	**1.49**
40	**40**	O/N=C/C(=C/c1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C#N	**0.44**
41	**41**	CCCCCCCCN1C(=S)S/C(=c\2/s/c(=C\c3ccc(cc3)N(c3ccc(cc3)C)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**4.59**
42	**42**	CCCCCCCCN1C(=S)S/C(=c\2/s/c(=C\c3ccc(cc3)N(c3ccc(cc3)/C=c/3\s/c(=C\4/SC(=S)N(C4=O)CCCCCCCC)/n(c3=O)CC(=O)	**5.3**

		O)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	
43	**43**	CCCCCCCCN1C(=S)S/C(=c\2/s/c(=C/c3ccc(cc3)N(c3ccc(cc3)/C=c/3\s/c(=C\4/SC(=S)N(C4=O)CCCCCCCC)/n(c3=O)CC(=O)O)c3ccc(cc3)/C=c/3\s/c(=C\4/SC(=S)N(C4=O)CCCCCCCC)/n(c3=O)CC(=O)O)/c(=O)n2CC(=O)O)/C1=O	**2.61**
44	**44**	CCCCCCn1nc2c(n1)c(ccc2c1ccc(o1)C=C(C(=O)O)C#N)c1ccc(o1)c1ccc(cc1)N(c1ccccc1)C(=CC)C=C	**4.37**
45	**45**	CCCCCCn1nc2c(n1)c(ccc2c1ccc(s1)/C=C(\C(=O)O)/C#N)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**6.2**
46	**46**	CCCCCCn1nc2c(n1)c(ccc2c1ccc([se]1)C=C(C(=O)O)C#N)c1ccc([se]1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**5.66**
47	**47**	CCCCCCn1nc2c(n1)c(ccc2c1sc(c(c1)CCCCCC)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1sc(c(c1)CCCCCC)C=C(C(=O)O)C#N	**5.89**
48	**48**	N#CC(=Cc1ccc(s1)c1c(F)c(F)c(c2c1n[nH]n2)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**6.01**
49	**49**	CCCCCCCCCc1nc(sc1c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1sc(c(n1)CCCCCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**2.98**
50	**50**	N#C/C(=C\c1ccc(cc1)N(c1ccccc1)c1ccc(cc1)c1cc2CCCN3c2c(c1)CCC3)/C(=O)O	**2**
51	**51**	OC(=O)CN1C(=S)S/C(=C\c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**1.6**
52	**52**	OC(=O)CN1C(=S)S/C(=C\c2ccc(cc2)N(c2ccc(cc2)c2cc3CCCN4c3c(c2)CCC4)c2ccccc2)/C1=O	**3.2**
53	**53**	N#C/C(=C\c1c(F)cc(cc1F)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.62**
54	**54**	N#C/C(=C\c1c(F)cc(cc1F)/C=C/c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	**4.01**
55	**55**	N#C/C(=C\c1ccc(cc1F)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.19**
56	**56**	N#C/C(=C\c1ccc(cc1)/C=C/c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	**4.58**

57	**57**	N#C/C(=C\c1ccc(cc1F)/C=C/c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1)/C(=O)O	**4.73**
58	**58**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(cn1)/C=C(/C(=O)O)\C#N	**6.21**
59	**59**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)OCCCCCC	**7.09**
60	**60**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**5.18**
61	**61**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(c2c1nsn2)c1ccc(s1)/C=C(/C(=O)O)\C#N	**7.12**
62	**62**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)nc(c(c2)/C=C(\C(=O)O)/C#N)Cl)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**1.3**
63	**63**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)[nH]c(=O)c(c2)/C=C(\C(=O)O)/C#N)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**0.77**
64	**64**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)[nH]c(=O)c(c2)C=C(C#N)C#N)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**3.23**
65	**65**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)[nH]c(=O)c(c2)C(=O)O)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**3.08**
66	**66**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)nc(c(c2)C(=O)O)Cl)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**0.27**
67	**67**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc2c(c1)[nH]c(=O)c(c2)C=O)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**0.45**
68	**68**	OC(=O)c1cc2C(=O)C(=Cc3ccc(cc3)N(c3ccccc3)c3ccccc3)C(=O)c2cc1C(=O)O	**0.95**
69	**69**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)C=C1	**1.1**

		C(=O)c2c(C1=O)cc(c(c2)C(=O)O)C(=O)O	
70	**70**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)c1cc(cc(n1)c1cc(cc(n1)c1cc2c(s1)c1c(C2(CCCCCC)CCCCCC)cc(s1)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)C(=O)O)C(=O)O	**3.3**
71	**71**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1sc2c(c1)C(c1c2sc(c1)c1ccc(c(n1)c1ccc(c(n1)c1sc2c(c1)C(c1c2sc(c1)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)(CCCCCC)CCCCCC)C(=O)O)C(=O)O)(CCCCCC)CCCCCC	**3.9**
72	**72**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(s1)c1nc(c(n1CCCCCC)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.34**
73	**73**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(s1)c1nc2c(n1CCCCCC)c1cc(sc1c1c2cc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.54**
74	**74**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1[nH]c(c(n1)c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**4.11**
75	**75**	N#CC(=c1s/c(=C\c2ccc(cc2)N(c2ccccc2)c2ccccc2)/c(=O)[nH]1)C#N	**2.92**
76	**76**	N#CC(=c1[nH]c(=O)/c(=C/c2ccc(cc2)N(c2ccc(cc2)/C=c/2\sc(=C(C#N)C#N)[nH]c2=O)c2ccccc2)/s1)C#N	**3.78**
77	**77**	N#CC(=c1s/c(=C/c2ccc(cc2)N(c2ccc(cc2)/C=c/2\sc(=C(C#N)C#N)[nH]c2=O)c2ccc(cc2)/C=c/2\sc(=C(C#N)C#N)[nH]c2=O)/c(=O)[nH]1)C#N	**3.44**
78	**78**	N#CC(=c1[nH]c(=O)/c(=C/c2ccc(s2)c2ccc(cc2)N(c2ccccc2)c2ccccc2)/s1)C#N	**2.11**
79	**79**	N#CC(=c1[nH]c(=O)/c(=C/c2ccc(s2)c2ccc(cc2)N(c2ccc(cc2)c2ccc(s2)/C=c/2\sc(=C(C#N)C#N)[nH]c2=O)c2ccccc2)/s1)C#N	**2.37**
80	**80**	C=Cc1ccc(cc1)N(c1ccccc1)c1ccc(cc1)/C=C\1/SC(=S)N(C1=O)	**5.84**

		CC(=O)O	
81	**81**	C=Cc1ccc(cc1)N(c1ccccc1)c1ccc(cc1)/C=C/C=C\1/SC(=S)N(C1=O)CC(=O)O	**1.27**
82	**82**	OC(=O)CN1C(=S)S/C(=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**3.64**
83	**83**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.81**
84	**84**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1ccc(s1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.37**
85	**85**	N#C/C(=C\c1ccc(s1)c1ccc(c2c1nsn2)c1ccc(s1)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.55**
86	**86**	CCCC[C@@H](COc1ccc(cc1)N(c1ccc(cc1)OC[C@H](CCCC)CC)c1ccc(cc1)c1ccc(s1)c1ccc(c2c1nsn2)c1ccc(s1)/C=C(/C(=O)O)\C#N)CC	**7.3**
87	**87**	OC(=O)CN1C(=S)S/C(=C/C=C/c2ccc(cc2)N(c2ccccc2)c2ccccc2)/C1=O	**0.63**
88	**88**	N#CC(=Cc1sc2c(c1)[Si](c1c2sc(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)(c1ccccc1)c1ccccc1)C(=O)O	**6.65**
89	**89**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1sc2c(c1)[Si](c1c2sc(c1)C=C(C(=O)O)C#N)(c1ccccc1)c1ccccc1	**7.6**
90	**90**	CCCCCCCCC1(CCCCCCCC)c2cc(C#CC#Cc3ccc(cc3)N(c3ccccc3)c3ccccc3)ccc2c2c1cc(cc2)C=C(C(=O)O)C#N	**2.52**
91	**91**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1sc(c(n1)OC)c1ccc(cc1)N(c1ccc(cc1)C)c1ccc(cc1)C)C(=O)O	**1.02**
92	**92**	N#CC(=Cc1ccc(s1)c1nc(c(s1)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC)OC)C(=O)O	**1.08**
93	**93**	N#CC(=Cc1ccc(s1)c1nc(c(s1)c1ccc(cc1)N(c1ccc(cc1)C)c1ccc(cc1)C)OC)C(=O)O	**1.7**
94	**94**	CCCCCCc1cc(sc1/C=C(\C(=O)O)/C#N)c1ccc2c(c1)C1(c3c2ccc(c3)c2ccc(cc2)N(c2ccccc2)c2ccccc2)c2ccccc2c2c1cccc2	**4.12**
95	**95**	CCCCCCc1cc(sc1/C=C(\C(=O)O)/C#N)c1ccc2c(c1)n(CC)c1c2	**6.51**

		ccc(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	
96	**96**	N#C/C(=C/c1ccc(s1)c1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**7.03**
97	**97**	CCCCCCc1cc(sc1C=Cc1ccc(cc1)N(c1ccc(cc1)C)c1ccc(cc1)C)C=C(C(=O)O)C#N	**7.83**
98	**98**	N#CC(=Cc1cc(c(s1)C=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C)C(=O)O	**7.62**
99	**99**	N#CC(=Cc1cc(c(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C)C(=O)O	**8.27**
100	**100**	CCCCCCCCC1(CCCCCCCC)c2cc(ccc2c2c1cc(cc2)C=C(C(=O)O)C#N)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1	**3.5**
101	**101**	CCCCCCCCC1(CCCCCCCC)c2cc(ccc2c2c1cc(cc2)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C#Cc1ccc2c(c1)C(CCCCCCCC)(CCCCCCCC)c1c2ccc(c1)C=C(C(=O)O)C#N	**1.8**
102	**102**	OC(=O)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**1.66**
103	**103**	OC(=O)CN1C(=S)SC(=Cc2ccc(cc2)N(c2ccccc2)c2ccccc2)C1=O	**1.12**
104	**104**	C=Cc1ccc(cc1)N(c1ccccc1)c1ccc(cc1)C=C(C(=O)O)C#N	**2.86**
105	**105**	C=Cc1ccc(cc1)N(c1ccccc1)c1ccc(cc1)C=CC=C(C(=O)O)C#N	**4.82**
106	**106**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc2c(c1)C(=O)c1c2ccc(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.71**
107	**107**	CCCCCCOc1cc(OCCCCCC)ccc1n1c2cc(sc2c2c1cc(s2)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**7.67**
108	**108**	CCCCCCOc1cc(OCCCCCC)ccc1n1c2cc(sc2c2c1cc(s2)c1sc2c(c1)n(c1c2sc(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1cc(OCCCCCC)cc(c1)OCCCCCC)/C=C(/C(=O)O)\C#N	**4.19**
109	**109**	N#C/C(=C/c1ccc(cc1)c1ccc(cc1)N(c1ccc2c(c1)C(C)(C)c1c2ccc1)c1ccc(cc1)c1ccccc1)/C(=O)O	**2.28**
110	**110**	N#C/C(=C/c1ccc(s1)c1ccc(cc1)N(c1ccc2c(c1)C(C)(C)c1c2ccc1)c1ccc(cc1)c1ccccc1)/C(=O)O	**4.83**
111	**111**	N#C/C(=C/c1ccc(o1)c1ccc(cc1)N(c1ccc2c(c1)C(C)(C)c1c2ccc	**5.81**

		c1)c1ccc(cc1)c1ccccc1)/C(=O)O	
112	**112**	N#C/C(=C\c1cc(c(s1)c1ccc(s1)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**8.08**
113	**113**	N#C/C(=C\c1cc(c(s1)c1sc(cc1c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**7.21**
114	**114**	N#C/C(=C\c1cc(c(s1)c1sc(cc1c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1sc(c(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.79**
115	**115**	CCN(c1ccc(cc1)C=C[N+](=O)[O-])CC	**0.6**
116	**116**	OC(=O)C=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1	**1.7**
117	**117**	[O-][N+](=O)C=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1	**0.7**
118	**118**	OC(=O)C=Cc1ccc(cc1)N(c1ccc(cc1)C=CC(=O)O)c1ccccc1	**3.9**
119	**119**	[O-][N+](=O)C=Cc1ccc(cc1)N(c1ccc(cc1)C=C[N+](=O)[O-])c1ccccc1	**1.1**
120	**120**	COc1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)c1ccc(c2c1nsn2)c1ccc(s1)C=C(C(=O)O)C#N	**7.66**
121	**121**	CCCCCCCCOc1cc2c3cc(sc3c3c(c2cc1OCCCCCCCC)cc(s3)C=C(C(=O)O)C#N)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC	**5.2**
122	**122**	CCCCCCCCOc1cc2c3cc(sc3c3c(c2cc1OCCCCCCCC)cc(s3)C=C(C(=O)O)C#N)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC	**5.5**
123	**123**	CCCCCCCCOc1cc2c3cc(sc3c3c(c2cc1OCCCCCCCC)cc(s3)CC(C(O)O)CN)c1ccc(c2c1nsn2)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC	**8.2**
124	**124**	CCCCCCc1cc(sc1c1cnc(c2c1nsn2)c1ccc(cc1)N(c1ccc(cc1)OCCCCC)c1ccc(cc1)OCCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**2.86**
125	**125**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ncc(c2c1nsn2)c1sc(cc1CCCCCC)c1sc(c(c1)CCCCCC)/C=C(/C(=O)O)\C#N	**3.84**
126	**126**	CCCCCCc1cc(sc1c1cnc(c2c1nsn2)c1ccc(cc1)N(c1ccc(cc1)OC	**4.46**

		CCCCC)c1ccc(cc1)OCCCCCC)c1ccc(cc1)/C=C(/C(=O)O)\C#N	
127	**127**	CCCCCCc1cc(sc1c1ncc(c2c1nsn2)c1sc(cc1CCCCCC)c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC	**6.3**
128	**128**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(s1)c1ncc(c2c1nsn2)c1ccc(s1)c1ccc(cc1)/C=C(/C(=O)O)\C#N	**1.32**
129	**129**	N#C/C(=C\c1cc(c(s1)c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccccc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.87**
130	**130**	N#C/C(=C\c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.56**
131	**131**	N#C/C(=C\c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.97**
132	**132**	c1ccc(cc1)N(c1ccccc1)c1ccc(cc1)c1cccs1.OC(=O)C(=C)C#N	**2.75**
133	**133**	c1ccc(cc1)N(c1ccccc1)c1ccc(cc1)C1(CC=CS1)C=Cc1cccs1.OC(=O)C(=C)C#N	**2.73**
134	**134**	c1ccc(cc1)N(c1ccccc1)c1ccc(cc1)C=CC1(CC=CS1)C=Cc1cccs1.OC(=O)C(=C)C#N	**1.7**
135	**135**	N#CC(=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**1.55**
136	**136**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)c1ccc(c2c1nsn2)c1sc(c(c1)CCCCCC)c1ccc(cc1)C=C(C(=O)O)C#N	**3.16**
137	**137**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)c1ccc(c2c1nsn2)c1sc(c(c1)CCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.81**
138	**138**	N#CC(=Cc1ccc(c(c1)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C#C)C(=O)O	**0.82**
139	**139**	N#CC(=Cc1ccc(c(c1)C#C)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**3.26**
140	**140**	N#CC(=Cc1ccc(c(c1)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)C#C	**5.41**

		c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	
141	**141**	COc1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**6.28**
142	**142**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(s1)c1ccc(s1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**3.4**
143	**143**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(s1)c1sc(c(c1)c1ccc(s1)c1cc(sc1c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)c1ccc(s1)C=C(C(=O)O)C#N)c1ccc(s1)C=C(C(=O)O)C#N	**6.6**
144	**144**	CCN1C(=S)S/C(=c\2/s/c(=C/c3ccc(s3)c3cn4cc(nc4c(c3)c3ccc(cc3)N(c3ccccc3)c3ccccc3)C)/n(c2=O)CC(=O)O)/C1=O	**0.18**
145	**145**	N#C/C(=C\c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)I)/C(=O)O	**5.48**
146	**146**	N#C/C(=C\c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C(=O)O	**5.55**
147	**147**	N#C/C(=C\c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccc(cc1)n1c2ccccc2c2c1cccc2)c1ccc(cc1)n1c2ccccc2c2c1cccc2)/C(=O)O	**6.15**
148	**148**	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)/C=C(/C(=O)O)\C#N)c1c(c2c2ccc(s2)c2ccc(cc2)N(c2ccccc2)c2ccccc2)nc(c(n1)c1ccccc1)c1ccccc1	**2.6**
149	**149**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1c(c2c2ccc(s2)c2ccc(cc2)N(c2ccccc2)c2ccccc2)nc(c(n1)c1ccccc1)c1ccccc1	**2.6**
150	**150**	CCCCCc1nc2c3cc(sc3c3c(c2nc1CCCCC)cc(s3)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**6.37**
151	**151**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1sc2c(c1)c1nc(CCCCC)c(nc1c1c2sc(c1)/C=C(/C(=O)O)\C#N)CCCCC	**4.03**
152	**152**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1sc(c2c1OCCO2)c1sc2c(c1)c1nc(CCCCC)c(nc1c1c2sc(c1)/C=C(/C(=O)O)\C#N)CCCCC	**4.01**

153	**153**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(c2c1nsn2)c1sc2c(c1)c1nc(CCCCC)c(nc1c1c2sc(c1)/C=C(/C(=O)O)\C#N)CCCCC	**6.78**
154	**154**	CCCCCCCCn1nc2c(n1)c(c1ccc(s1)C(=O)O)c1c(c2c2ccc(s2)c2ccc(cc2)N(c2ccccc2)c2ccccc2)nc(c(n1)c1ccccc1)c1ccccc1	**4.4**
155	**155**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)C(=O)O)c1c(c2c2ccc(s2)c2ccc(cc2)N(c2ccccc2)c2ccccc2)nc(c(n1)c1ccccc1)c1ccccc1	**4.2**
156	**156**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1c(c2c2ccc(s2)/C=C(/C(=O)O)\C#N)nc(c(n1)c1ccccc1)c1ccccc1	**2.2**
157	**157**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1c(c2c2ccc(cc2)C=C(C(=O)O)C#N)nc(c(n1)c1ccccc1)c1ccccc1	**3.4**
158	**158**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1c(c2c2ccc(s2)C(=O)O)nc(c(n1)c1ccccc1)c1ccccc1	**3.3**
159	**159**	CCCCCCCCn1nc2c(n1)c(c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1c(c2c2ccc(cc2)C(=O)O)nc(c(n1)c1ccccc1)c1ccccc1	**0.3**
160	**160**	COC(=O)c1c(C)nc2n1cc(cc2c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(s1)/C=C(\C(=O)O)/C#N	**0.64**
161	**161**	N#C/C(=C/c1ccc(s1)c1cn2cc(nc2c(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C)/C(=O)O	**0.45**
162	**162**	CCCCCCCCc1nc2c(ccc(c2nc1CCCCCCC)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**7.78**
163	**163**	CCCCCCCCc1ccc(s1)c1nc2c(nc1c1ccc(s1)CCCCCCCCCC)c(ccc2c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**7.29**
164	**164**	CCCCCCCCc1sc2c(c1)c1cc(sc1c1c2nc2c(ccc(c2n1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)CCCCCCCC	**4.03**
165	**165**	CCCCCCC1(CCCCCC)c2cc(ccc2c2c1cc(cc2)N(c1ccc(cc1)C(=O)O)c1ccc(cc1)C(=O)O)c1ccc(s1)C=C1C(=O)N(CC)C(=S)N(C1=O)CC	**0.092**
166	**166**	CCCCCCC1(CCCCCC)c2cc(sc2c2c1cc(s2)C=C(C#N)C#N)c1ccc	**0.087**

		(cc1)N(c1ccc(cc1)C(=O)O)c1ccc(cc1)C(=O)O	
167	**167**	N#CC(=Cc1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1ccc(s1)C=C(C#N)C#N)c1ccc(cc1)C(=O)O)C#N	**0.101**
168	**168**	CCCCCCC1(CCCCCC)c2cc(ccc2c2c1cc(cc2)c1ccc(s1)C=C(C#N)C#N)N(c1ccc(cc1)C(=O)O)c1ccccc1	**0.06**
169	**169**	CCCCCCC1(CCCCCC)c2cc(ccc2c2c1cc(cc2)N(c1ccc(cc1)C(=O)O)c1ccccc1)c1ccc(s1)c1ccc(s1)C=C(C#N)C#N	**0.058**
170	**170**	CCCCCCC1(CCCCCC)c2cc(ccc2c2c1cc(cc2)N(c1ccc(cc1)C(=O)O)c1ccccc1)c1ccc(s1)C=C1C(=O)N(CC)C(=S)N(C1=O)CC	**0.053**
171	**171**	N#CC(=Cc1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)C(=O)O)c1ccc(cc1)C(=O)O)C#N	**0.093**
172	**172**	N#C/C(=C\c1ccc(s1)c1sc(cc1c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**7.02**
173	**173**	N#C/C(=C\c1ccc(cc1)c1sc(cc1c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.56**
174	**174**	N#C/C(=C\c1ccc(cc1)c1sc(cc1c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**2.6**
175	**175**	CCCCCCn1c2ccc(cc2c2c1cccc2)c1ccc(s1)c1sc(c(c1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(s1)/C=C(/C(=O)O)\C#N	**7.28**
176	**176**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.45**
177	**177**	N#C/C(=C\c1ccc(s1)c1sc(cc1c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**6.95**
178	**178**	O=C1NC(=O)C(=Cc2ccc(s2)c2ccc(c3c2nsn3)c2ccc(cc2)N(c2ccccc2)c2ccccc2)N1	**4.9**
179	**179**	COc1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)c1ccc(c2c1nsn2)c1ccc(s1)C=C1NC(=O)NC1=O	**7.58**
180	**180**	CCN1C(=S)SC(=c2sc(=Cc3ccc4c(c3)C3CCCC3N4c3ccc(cc3)C=C(c3ccccc3)c3ccccc3)c(=O)n2CC(=O)O)C1=O	**1.4**
181	**181**	CCCCCCCCCCCCC1(CCCCCCCCCCCC)c2cc(sc2c2c1cc(s2)c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)C=C(C(=O	**8.8**

		)O)C#N	
182	**182**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1sc2c(c1)c1cc(sc1c(c2OCCCCCC)OCCCCCC)/C=C(/C(=O)O)\C#N	**5.6**
183	**183**	N#CC(=Cc1ccc(s1)c1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**3.09**
184	**184**	N#CC(=Cc1ccc(cc1)C=Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**5.14**
185	**185**	N#CC(=Cc1ccc(cc1)c1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**6.78**
186	**186**	N#CC(=Cc1ccc(s1)C=Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**4.86**
187	**187**	N#C/C(=C\c1ccc(s1)C#Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.5**
188	**188**	N#C/C(=C\c1ccc(cc1)C#Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC)/C(=O)O	**7.03**
189	**189**	N#C/C(=C\c1ccc(s1)C#Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC)/C(=O)O	**6.82**
190	**190**	N#C/C(=C\c1ccc(cc1)c1c2ccccc2c(c2c1cccc2)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**3.93**
191	**191**	N#C/C(=C\c1ccc(cc1)c1c2ccccc2c(c2c1cccc2)C#Cc1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC)/C(=O)O	**5.66**
192	**192**	N#C/C(=C\c1ccc(cc1)C=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.73**
193	**193**	N#C/C(=C\c1ccc(cc1)C#Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.3**
194	**194**	N#C/C(=C\c1ccc(cc1)C=Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**5.14**
195	**195**	N#C/C(=C\c1ccc(cc1)C#Cc1c2ccccc2c(c2c1cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**6.78**
196	**196**	N#C/C(=C\c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**4.7**

197	**197**	N#CC(=Cc1ccc(s1)c1ccc(cc1)N(CCO[Si](C(C)(C)C)(C)C)C(=O)O	**5.5**
198	**198**	CN(c1ccc(cc1)/C=C/c1ccc(s1)C=C(C(=O)O)C#N)CCO[Si](C(C)(C)C)(C)C	**6.4**
199	**199**	N#CC(=Cc1ccc(c2c1nsn2)c1ccc(cc1)N(CCO[Si](C(C)(C)C)(C)C)C(=O)O	**2.4**
200	**200**	N#CC(=Cc1ccc(cc1)c1ccc(c2c1nsn2)c1ccc(cc1)N(CCO[Si](C(C)(C)C)(C)C)C(=O)O	**4.3**
201	**201**	CCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCC)c1ccc(cc1)c1sc(c2c1OCCO2)c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N	**5.66**
202	**202**	CCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCC)c1ccc(cc1)c1ccc(s1)c1sc(c2c1OCCO2)/C=C(/C(=O)O)\C#N	**4.66**
203	**203**	CCCCC(COc1cc(OCC(CCCC)CC)ccc1c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)c1sc(c2c1OCCO2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(cc1)C=C(C(=O)O)C#N)CC	**8.37**
204	**204**	CCCCC(CC1(CCC(CCC)CC)c2cc(sc2c2c1cc(s2)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(cc1)C=C(C(=O)O)C#N)c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)CC	**8.06**
205	**205**	CCCCC(COc1cc(OCC(CCCC)CC)ccc1c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)c1ccc(s1)c1ccc(c2c1nc(c1ccccc1)c(n2)c1ccccc1)c1ccc(cc1)C=C(C(=O)O)C#N)CC	**6.57**
206	**206**	N#CC(=Cc1ccc2c(c1)C(C)(C)c1c2ccc(c1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**3.1**
207	**207**	N#CC(=Cc1ccc2c(c1)C(C)(C)c1c2ccc(c1)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**4.6**
208	**208**	N#CC(=Cc1ccc2c(c1)C(C)(C)c1c2ccc(c1)c1ccc(s1)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**5.6**
209	**209**	CCCCCCCCc1cc(sc1c1sc(c(c1)CCCCCCCC)c1ccc(cc1)N(c1cccc	**5.4**

		c1)c1ccccc1)c1ccc2c(c1)C(C)(C)c1c2ccc(c1)C=C(C(=O)O)C#N	
210	**210**	N#CC(=Cc1ccc2c(c1)C(C)(C)c1c2ccc(c1)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**5.8**
211	**211**	N#CC(=Cc1ccc(s1)c1ccc(s1)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**4.4**
212	**212**	N#CC(=Cc1ccc2c(c1)C(C)(C)c1c2ccc(c1)c1ccc(s1)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)C(=O)O	**4.9**
213	**213**	N#C/C(=C\c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1cccs1)c1ccc(cc1)c1cccs1)/C(=O)O	**4.9**
214	**214**	N#C/C(=C/c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)c1cccs1)/C(=O)O	**6.1**
215	**215**	N#C/C(=C/c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)/C(=O)O	**6.3**
216	**216**	CCCCCCc1nc(sc1c1ccc(cc1)N(c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1nc(c(s1)c1ccc(cc1)N(c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)/C=C(/C(=O)O)\C#N)CCCCC	**4.9**
217	**217**	CCCCCCOc1ccc(cc1)c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1ccc(c2c1nsn2)c1ccc(s1)/C=C\1/SC(=S)C(C1=O)CC(=O)O)c1ccc(cc1)c1ccc(s1)c1ccc(cc1)OCCCCCC	**2.03**
218	**218**	CCCCCCOc1ccc(cc1)c1ccc(s1)c1ccc(cc1)N(c1ccc(cc1)c1sc(c2c1OCCO2)/C=C\1/SC(=S)N(C1=O)CC(=O)O)c1ccc(cc1)c1ccc(s1)c1ccc(cc1)OCCCCCC	**2.68**
219	**219**	CCCCOc1ccc(c(c1)OCCCC)c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCCCC)OCCCC)c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**6.74**
220	**220**	CCCCOc1cc(OCCCC)ccc1c1ccc(cc1)N(c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N)c1ccc(cc1)c1ccc(s1)/C=C(/C(=O)O)\C#N	**5.51**
221	**221**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)OCCCCCC)c1cc[n+](cc1)CC(=O)O	**5.9**

222	**222**	OC(=O)C[n+]1ccc(cc1)/C=C/c1ccc(cc1)N(c1ccccc1)c1ccccc1	**2.6**
223	**223**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1cc[n+](cc1)CC(=O)O	**3.1**
224	**224**	CCCCCCc1cc(sc1c1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)OC)c1cc[n+](cc1)CC(=O)O	**5.6**
225	**225**	COc1ccc(cc1)N(c1ccc(cc1)OC)c1ccc(cc1)/C=C/c1cc[n+](cc1)CC(=O)O	**5.9**
226	**226**	N#C/C(=C\c1ccc(s1)c1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N(c1ccc2c(c1)C(CC)(CC)c1c2cccc1)c1ccc2c(c1)C(CC)(CC)c1c2cccc1)/C(=O)O	**4.65**
227	**227**	N#C/C(=C\c1ccc(s1)c1ccc(s1)c1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N(c1ccc2c(c1)C(CC)(CC)c1c2cccc1)c1ccc2c(c1)C(CC)(CC)c1c2cccc1)/C(=O)O	**5.8**
228	**228**	N#C/C(=C/c1ccc(cc1)N(c1ccc(cc1)/C=C(/C(=O)O)\C#N)c1ccccc1)/C(=O)O	**2.86**
229	**229**	OC(=O)CN1C(=S)S/C(=C\c2ccc(cc2)N(c2ccc(cc2)/C=C\2/SC(=S)N(C2=O)CC(=O)O)c2ccccc2)/C1=O	**3.15**
230	**230**	CCCCCCCCN1C(=S)S/C(=c\2/s/c(=C/c3ccc(cc3)N(c3ccc(cc3)/C=c/3\s/c(=C\4/SC(=S)N(C4=O)CCCCCCCC)/n(c3=O)CC(=O)O)c3ccccc3)/c(=O)n2CC(=O)O)/C1=O	**4.64**
231	**231**	CCCCCCOc1ccc(cc1)N(c1ccc(cc1)OCCCCCC)c1ccc(cc1)c1ccc(s1)c1n(CC(CCCC)CC)c(=O)c2c1c(=O)n(c2c1ccc(cc1)c1ccc(cc1)C=C(C(=O)O)C#N)CC(CCCC)CC	**7.67**
232	**232**	CCCCC(Cn1c(c2ccc(s2)c2ccc(cc2)N(c2ccc(cc2)OCCCCCC)c2ccc(cc2)OCCCCCC)c2c(c1=O)c(n(c2=O)CC(CCCC)CC)c1ccc(cc1)c1ccc(o1)C=C(C(=O)O)C#N)CC	**7.6**
233	**233**	N#C/C(=C\c1ccc(cc1F)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**8.22**
234	**234**	N#C/C(=C\c1ccc(cc1)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccccc1)c1ccccc1)/C(=O)O	**7.25**
235	**235**	N#C/C(=C\c1ccc(c(c1)F)c1sc(c2c1OCCO2)c1ccc(cc1)N(c1ccc	**7**

		cc1)c1ccccc1)/C(=O)O	
236	**236**	CCCCCCCCCCCCCCCCc1c(C=C(C(=O)O)C#N)sc2c1sc1c2sc2c1sc(c2CCCCCCCCCCCCCCCC)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**4.76**
237	**237**	CCCCCCCCCCCCCCCCc1c(C=C(C(=O)O)C#N)sc2c1sc1c2sc2c1sc(c2CCCCCCCCCCCCCCCC)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**6.15**
238	**238**	CCCCCCCCCCCCCCCCc1c(sc2c1sc1c2sc2c1sc(c2CCCCCCCCCCCCCCCC)c1ccc(s1)C=C(C(=O)O)C#N)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**10.1**
239	**239**	CCCCCCCCCCCCCCCCc1c(sc2c1sc1c2sc2c1sc(c2CCCCCCCCCCCCCCCC)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(s1)C=C(C(=O)O)C#N	**6.91**
240	**240**	Oc1nc2ccc(cc2nc1O)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**1.18**
241	**241**	Oc1nc2ccc(cc2nc1O)c1ccc(s1)c1ccc(cc1)N(c1ccccc1)c1ccccc1	**1.58**
242	**242**	CCCCC(COc1cc(OCC(CCCC)CC)ccc1c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)c1sc2c(c1)C(c1c2sc(c1)/C=C(/C(=O)O)\C#N)(CC(CCCC)CC)CC(CCCC)CC)CC	**5.87**
243	**243**	CCCCC(CC1(CC(CCCC)CC)c2cc(sc2c2c1cc(s2)c1ccc(c2c1nsn2)c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)/C=C(/C(=O)O)\C#N)CC	**6.69**
244	**244**	CCCCC(CC1(CC(CCCC)CC)c2cc(sc2c2c1cc(s2)c1ccc(cc1)N(c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(cc1)c1ccc(cc1OCC(CCCC)CC)OCC(CCCC)CC)c1ccc(c2c1nsn2)c1ccc(s1)/C=C(/C(=O)O)\C#N)CC	**5.5**
Test set compounds of Division 1 (compound ids) - 3,4,7,9,11,16,38,40,44,49,57,59,73,78,80,83,84,87,97,98,103,121,124,139,140,141,147,148,152,163,169,171,173,182,186,187,203,205,206,219,221,230,239,241			
Test set compounds of Division 2 (compound ids) - 3,5,7,10,16,22,24,30,33,35,36,46,51,59,65,73,81,89,95,106,109,110,116,125,126,129,145,148,152,155,17			

5,176,178,190,196,207,213,215,219,223,231
Test set compounds of Division 3 (compound ids) - **6,7,15,16,19,20,21,31,32,60,66,81,83,84,85,90,97,102,120,121,129,141,150,155,164,168,173,175,178,181,182,193,196,209,210,215,218,220,221,226,229,236,237,242**
Compounds removed from dataset based on leverage value - **23,28,43,48,53,54,62,70,71,77,79,115,119,123,132,133,134,143,197,198,199,200,216,235,244**

3.2 Study-wise specific description of methodologies utilized in each study

3.2.1 Study -1

3.2.1.1 Data collection

Study 1 deals with q-RASPR modeling of different organic dyes used in DSSCs for which experimental PCE and descriptor values are collected from the previous literature [131]. To demonstrate the performance of the q-RASPR approach, in comparison to the previous QSPR models using the same level of chemical information, we have collected a total of 429 compounds distributed across 4 representative datasets, i.e. coumarins, carbazoles, indolines, and diphenylamines. The original training and test sets were retained for the q-RASPR modeling analysis aiming at comparing with the previously developed QSPR model by Krishna et al [131]. We used the same combination of structural and physicochemical descriptors as used in the original analysis. The dataset in each case was prepared by combining the descriptors of five individual models of the previous work and removing the common descriptors of these models. The logarithmic conversion of the endpoint parameter PCE was not required as it represents the performance of the solar cell. The datasets contain 56 coumarin dyes (42 training and 14 test compounds), 35 diphenylamine dyes (25 training and 10 test compounds), 179 carbazole dyes (125 training and 54 test compounds), and 159 indoline dyes (121 training and 38 test compounds). These datasets were used as input files for the calculation of the RASPR descriptors.

3.2.1.2 Read-Across hyperparameter optimization

Read-across is a data-gap filling method in which information of one or more chemicals is/are used to predict the endpoint of a target chemical. In read-across, structural similarity between a source compound (or a training compound) and a target compound (or a test compound) is used for the calculation of the endpoint value [118, 125, 132]. A java-based tool Read-Across-v4.1 (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) was used for the computation of Read-Across-based predictions. This tool utilizes 3 different similarity-based methods, i.e., Euclidean distance (ED)-based similarity; Gaussian kernel (GK)-based similarity and the Laplacian kernel (LK)-based similarity methods for the computation of the predictions for the query compound(s) [118]. In compliance with the theory associated with Machine Learning, there is a need for the optimization of the hyperparameters associated with the three different similarity-based approaches. The Euclidean Distance-based predictions require the optimum number of close source compounds, the Gaussian Kernel-based predictions require the optimum value of σ and the number of close source compounds while the Laplacian Kernel-based predictions require the optimum value of γ and the number of close source compounds.

To select the optimum values for σ, and γ, the training set is randomly divided into 5 sub-training and sub-test sets by the sorted response-based division algorithm [133]. Read-Across-based predictions and validation metrics were calculated using these 5 sub-training and sub-test sets for each value of σ, γ, and CTC; and the average of external validation metrics for the subtest sets of 5 divisions was taken. The selection of the optimum σ and γ depends on where the Q^2_{F1} value (subtest set) is maximum for the GK and LK methods, respectively and then the same values of σ and γ were applied for the original training and test sets at each value of close source compounds (CTC) between 2 to 10. The CTC value which corresponds to the maximum Q^2_{F1} value (subtest set) in the ED approach is selected. These optimized settings of the σ, γ and CTC values were used for the computation of the RASPR descriptors. Note that the subtest sets are derived from the training set itself and different from the actual test set.

3.2.1.3 RASPR descriptor calculation

As defined by Todeschini and Consonni, a descriptor is "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiments" [134]. In the present work, we have used different structural and physicochemical descriptors along with RASPR descriptors to develop predictive models. Different similarity and error-based measures obtained from the read-across-based predictions were used as RASPR descriptors [124]. We have used RASAR-Desc-Calc-v2.0 (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) for RASPR descriptor calculation for both training and test sets. The calculation of the RASPR descriptors for the training set involves the Leave-same-out (LSO) algorithm which does not take the identical compound among the list of close source compounds. The optimized similarity-based method and the associated optimized hyperparameters were used for the computation of the RASPR descriptors. These RASPR descriptors were then combined with the initial structural and physicochemical descriptors and after subsequent feature selection based on cross-validation, q-RASPR models were generated. The optimized hyperparameter settings along with the similarity-based approach used for the calculation of the RASPR descriptors are shown in **Table 3.8**.

Table 3.8. Optimized hyperparameter settings and methods used for RASPR descriptor calculation.

Datasets	Method	σ	γ	CTC
Coumarins	**GK**	1.75	-	8
Carbazoles	**GK**	2	-	5
Indoliness	**LK**	-	0.5	5
Diphenylamine	**LK**	-	0.5	2

GK = Gaussian Kernel, LK = Laplacian Kernel, CTC = Close Training Compound

3.2.1.4 Feature selection and PLS model development

In the present work, we have used the Best Subset Selection (BSS) method to identify the significant descriptors or features that can influence the performance of DSSCs. Feature

selection is important to reduce model noise and identify the significant descriptors for PCE in order to lower the risk of overtraining or overfitting [135, 136]. Significant descriptors for the models were identified by using the tool Best Subset Selection v2.1 (available from https://dtclab.webs.com/software-tools) using the cross-validation statistics.

Best Subset Selection (BSS) is an algorithm that helps to identify the best descriptor combinations by developing models using a specific number of descriptor subset of input descriptors. This algorithm generates models using every possible combination of the descriptors, and the best combination is selected based on different internal. The number of descriptors in the models was selected based on the cross-validation Q^2_{LOO} score, and after that, we have developed several individual models for each data set, and the best models showing acceptable internal and external validation statistics are reported. For the present work, we have used the final models with 8 descriptors for coumarins, 5 descriptors for diphenylamines, 6 descriptors for indolines and 8 descriptors for carbazoles.

Partial Least Squares (PLS) is a generalized form of the multiple linear regression (MLR) that can be applied for collinear, correlated and noisy data containing multiple X variables (or descriptors) and one or more Y variable(s) (or endpoint(s)). The main idea behind PLS is to derive latent variables (LVs) T (or X-scores) and U (or Y-scores) from descriptors and response variables, respectively. These X-scores are then used to predict Y-scores which in turn are used to calculate the response [133]. Here, we have used PLS_SingleY_1.0_14May2020 tool (available from https://dtclab.webs.com/software-tools) for the development of PLS models of selected descriptors. We have also compared the derived PLS models with other machine-learning models obtained using the same feature combinations.

For the purpose of interpretation and explanation of individual descriptors, different PLS plots were developed using SIMCA-P v10.0 software (https://www.sartorius.com/). We have developed the score plots (individual compounds are defined in the LV space and show their distribution and similarity among compounds), the loading plots (loading of all descriptors among the plotted first two LVs, and distance from the origin denote the importance of these descriptors), the Y-randomization plot (plot developed by plotting R^2 and Q^2 value of random model (Y axis) vs. correlation coefficient between observed PCE and permuted PCE), scatter

plots (plots of predicted PCE (Y axis) vs observed PCE (X axis)) and the variable importance plots (in the form of bubble plot).

3.2.1.5 Machine-Learning (ML) models

Machine learning (ML) is a part of artificial intelligence which enables machines to learn from its past data and improve performance based on past experience for the future aspect [137]. In ML, machines are trained with a large amount of data and a suitable algorithm to accomplish a job. This trained algorithm is then applied to a query data set for reliable and accurate predictions. ML can be classified into 3 main groups: supervised (the labeled data is used to train the machine), unsupervised (the training data is not labeled) and reinforcement (feedback-based method, where the learning agents get a reward or penalty based on its action) [138, 139]. In the present study, we have performed regression analysis using different supervised learning methods namely ridge regression (RR), and linear support vector machine (LSVM), support vector machine (SVM), random forest (RF), gradient boosting (GB), and extreme gradient boost or XGBoost (XGB). Some details of these methods are discussed below.

a) Ridge regression: It is a popular method, which is used to address the multicollinearity problem of MLR models without removing any independent variables. In this method, a small amount of bias (penalty) is added to get better predictions. It is an important regularization technique that helps to reduce the complexity of a model, and it is known as Tikhonov regularization. The generalized equation of ridge regression is: $L(x, y) = Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n}(w_i)^2)$ where w_i is weightage of individual feature and λ is penalty term [140, 141].

b) Linear Support Vector Machine (LSVM): LSVM is a form of SVM algorithm, where the data domain can be classified linearly without any kind of transformation. The main steps in LSVM are mapping of data domain into response data, and then division of the data domain. The generalized equation for LSVM is: $\hat{y} = w^T X + b$ [142].

c) Support Vector Machine (SVM): It is a classification machine-learning algorithm, but it can also be used for regression problems, as known as Support Vector Regression (SVR). The main idea behind SVM is to draw a decision boundary between observations to perform its predictions. For nonlinear SVM, we have to transform the data into a feature

space (using a kernel function) before mapping with the response, and the generalized equation for SVM (non-linear) is represented as follows: $\hat{y} = w^T \phi(X) + b$, where $\hat{y}$ is predictions, w is the vector of weights, X is a vector of input features, ϕ is a kernel function and b is bias. This decision boundary is known as a plane for a three-dimensional space and known as a hyperplane for higher order space where a large number of features are present. The SVM method considers both margins which are formed by the area between the decision boundary and the closest training compound and the hyperplane for predictions. The margin is represented by the following equation: $margin = \frac{1}{w^T w}$. SVM tries to maximize the distance between the two closest training compounds on either side of decision boundary [143].

d) Random forest (RF): The RF algorithm builds multiple decision tree models and combines their outcome for more accurate and stable prediction. This method helps to overcome overfitting problem of a decision tree models. The RF algorithm is based on an ensemble learning method known as Bagging (bootstrap aggregating) which is a resampling technique applied to a dataset. In bootstrapping, observations are selected by random sampling with replacement, and random feature subsets are selected. In bagging, a large number of datasets are created by bootstrapping the original dataset, multiple decision tree models are formed using these datasets, and finally average of predictions are taken [143].

e) Gradient boosting (GB): Boosting is also an ensemble method that helps to form a strong learner by combining many weak learners. It is also a tree-based method in which decision trees are generated sequentially, and every tree tries to correct its predecessor. Gradient boosting (GB) is one of the boosting methods in which it tries to fit the current predictor with residual error made by the previous predictor [144].

f) XGBoost: It stands for Extreme Gradient Boosting and was first implemented by the researchers of the University of Washington. This method was built based on the same algorithm of GB, but the main drawback of GB is that it searches for minimizing the loss function across all possible splits to create a new branch of a decision tree. Thus, GB method becomes time-consuming when thousands of features are present because there are thousands of possibilities to split the node of a decision tree. XGBoost handles these drawbacks by taking information of feature distribution across all data points in a single

leaf node and by this way it reduces the search space. This method cannot generate multiple decision trees in parallel but can generate multiple branches of a decision tree in parallel [144].

All the above-mentioned machine-learning models were developed using Anaconda Navigator software (version 2022.05) in Jupyter Notebook IDE (version 6.4.8) [145] with python 3.10.4 64-bit. Different python-based modules were used such as numpy (version 1.23.5), pandas (version 1.5.2), Scikit-learn (version 1.2.0), matplotlib (version 3.5.1) and xgboost (version 1.7.1) for model development. For all the machine-learning models, we have used the same inputs as used for the PLS model development and optimized all the hyper-parameters by the cross-validation method using GridSearchCV function of Scikit-learn. For ML modeling, we standardized the descriptors and endpoints values based on the training set mean and standard deviation which were then used as the input.

In this work, we developed machine learning models using moderate-sized data to train the model and a sufficient number of compounds for the validation of the models. Here, we optimize the hyper-parameters using the GridSearchCV method which is basically a cross-validation method in which the training set is divided into five-folds, and four folds are used to build the models each time when the remaining fold is used to validate the model. After building machine learning models, we calculate various cross-validation metrics to check that the models are not overfitted. The hyperparameter setting is chosen based on the best cross-validation statistics from the five-fold CV data. Again, a small difference between Mean Absolute Error (MAE) values for the training and test sets also indicates that the generated models are not overfitted.

3.2.1.6 Statistical Quality and Validation Metrics

Validation of a model is important to justify the model quality and to determine its further application [133]. There are different statistical metrics available to check the model quality, goodness-of-fit, robustness, reliability and predictivity. The model quality is checked by the determination coefficient (R^2) and the explained variance ($R^2_{(adj)}$). The model quality is increased when its value become closer to 1. Model validation metrics can be classified into two groups – the internal validation metrics and the external validation metrics. Internal validation is performed only on the training set to check the goodness-of-fit and robustness of the model while the predictivity of a model is checked by an external validation performed on the test set.

The robustness and goodness-of-fit of a model are checked by the internal validation metrics Q^2_{LOO} (Leave-one-out correlation coefficient) performed on the training set. The original dataset is divided into a training set and a test set; the test set is used to check the predictivity of a model by calculating external validation metrics like Q^2_{F1}, Q^2_{F2} and MAE_{test} [133]. The validation metrics which demonstrated the quality of our PLS and ML models are shown in Eqs. 3.1-3.5.

$$R^2 = 1 - \frac{\sum(Y_{obs(train)} - Y_{cal(train)})^2}{\sum(Y_{obs(train)} - \bar{Y}_{train})^2} \quad (3.1)$$

$$Q^2_{LOO} = 1 - \frac{\sum(Y_{obs(train)} - Y_{cal(train)})^2}{\sum(Y_{obs(train)} - \bar{Y}_{train})^2} \quad (3.2)$$

$$Q^2_{F1} = 1 - \frac{\sum(Y_{obs(test)} - Y_{cal(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{train})^2} \quad (3.3)$$

$$Q^2_{F2} = 1 - \frac{\sum(Y_{obs(test)} - Y_{cal(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{test})^2} \quad (3.4)$$

$$MAE_{test} = \frac{\sum|Y_{obs(test)} - Y_{cal(test)}|}{n_{test}} \quad (3.5)$$

Here, $Y_{obs(train)}$ = observed response value of training set, $Y_{cal(train)}$ = calculated response value of training set, $\bar{Y}_{train}$ = mean observe response value of the training set, $Y_{obs(test)}$ = observed response value of the test set, $Y_{cal(test)}$ = calculated response value of the test set, $\bar{Y}_{test}$ = mean observed response value of the test set, n_{test} = the number of observations in the test set. A model is considered to be well predictive if the values of ${Q_{F1}}^2$ and ${Q_{F2}}^2$ cross the threshold limit of 0.5 and MAE_{test} attains a minimum value [133].

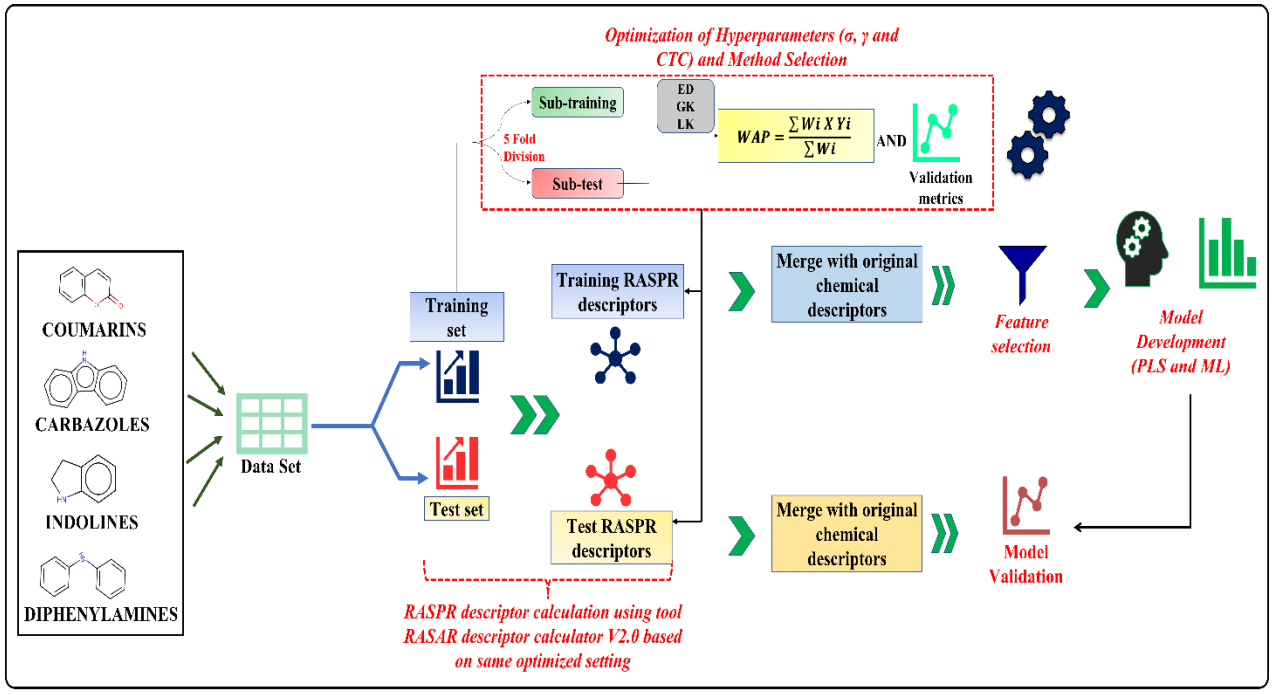
3.2.1.7 SHAP (SHapley Additive exPlanation) analysis

The SHAP analysis is performed to identify the global and local contributions of each feature or descriptor for the predictions. By using SHAP, we can determine the feature's contribution in case of complex machine-learning models. SHAP uses the Shapley values to determine feature contribution, the concept coming from the fair distribution in a cooperative game based on the

player's importance [146]. The Shapley values determine the contributions of different features (by the magnitude of the Shapley values) and direction (sign). The positive sign indicates a positive contribution to the predictions and the negative sign indicates a negative contribution to the predictions. The Shapley value for each feature is calculated using the following formula:

$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$ where ϕ_i is the Shapley value for each feature, $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ is the model output for a subset of features including a particular feature, $f_S(x_S)$ is the model output for the subset of features without that feature, F is the number of input features and S is the number of features in a subset [146, 147, 148].

The complete workflow for the current work is shown in the **Figure 3.1**.



RASPR = Read-across structure-property relationships, PLS = Partial least squares, ML = Machine learning, CTC = Close training compound, ED = Euclidean distance, GK = Gaussian kernel, LK = Laplacian kernel, WAP = Weighted average predictions

Figure 3.1: Schematic representation of complete work for q-RASPR and ML model development

3.2.2 Study 2

3.2.2.1 Data Collection

The experimental PCE values for the organic dyes were collected from the work of Krishna et al., [131]. In this study, we have used three relatively large datasets namely phenothiazines, porphyrins, and triphenylamines for modeling. These three data sets consist of 215 phenothiazine dyes (PCE range: 8.18 to 0.12), 300 porphyrin dyes (PCE range: 12.5 to 0.0013), and 244 triphenylamine dyes (PCE range: 10.1 to 0.053). The logarithmic transformation of PCE was not performed because it just represents the performance of solar cells. The SMILES of each compound were converted to 2D chemical structure (.mol file format) using Marvin Sketch software (https://chemaxon.com/marvin) and each molecule is named using the java-based tool Assign Molecule Names In MOL Files (available from https://dtclab.webs.com/software-tools). After that, a single .sdf file is generated from the .mol files by using OpenBabelGUI software (https://openbabel.org/wiki/Category:Installation) for all three datasets. The compounds present in the salt form were manually curated by removing the salt portion from the molecular structure.

3.2.2.2 Descriptor Calculation

The descriptor is defined as "*The molecular descriptor is the ultimate result of a logical and mathematical operation which transfigures the chemical information encrypted within a symbolic representation of a molecule into the result of some regularized (standardized) experiments*" [149]. The .sdf file was used for the calculation of 2D structural and physicochemical descriptors by using alvaDesc software [150]. The descriptors calculated using the alvaDesc software belong to the classes of constitutional, functional group count, 2D atom pairs, atom-centered fragments, atom type E-states, ring descriptors, connectivity index, molecular properties, and ETA indices. The columns with constant and near-constant values in the calculated descriptor matrix were removed using the alvaDesc software. In this work, we have only used 2D descriptors to avoid the conformational complexity of a molecule and for easy interpretation and identification of structural fragments and physicochemical properties.

3.2.2.3 Data Curation

The calculated descriptor matrix (along with the response values) was used for principal component analysis (PCA), [151] and the calculated principal components along with the PCE values were used for the leverage calculation. The leverage value of a chemical substance is calculated based on the HAT matrix by using the following equation: $H = (X^T(X^TX)^{-1}X)$ where *X* represents the descriptor matrix and *H* represents a square matrix which orthogonally projects vectors into the space spanned by the column of *X*. The leverage threshold value (h^*) to define the applicability domain (AD) is calculated by the formula $h^* = \frac{3(p+1)}{n}$ where p is the number of features and n represents the number of observations [133,152]. The compounds which have a leverage value higher than h* are defined as out-of-AD compounds. Normally, only the final model descriptors are used for leverage calculation. However, we have used here the response values also to understand the outlier behavior of the compounds with respect to both structural and response similarity. In this present work, we have calculated the leverage values using a java based tool Hi_Calculator-v2.0 (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). Compounds with high leverage value were removed from each dataset (approximately 10 percent) before modeling, which were later used as an external validation set for prediction after model development.

3.2.2.4 Data Pre-treatment and dataset division

The whole descriptor matrix in a dataset contains noise which can affect the performance of the model since many of the descriptors are incomplete, possess missing values and are highly inter-correlated. Thus, it is necessary to filter this noisy data before modeling. To reduce the noise in the data, the features with missing value, features with a constant value (variance <0.0001), or inter-correlated (|r| >0.95) features were removed from the descriptor matrix. This process of removal of noisy values is known as data pre-treatment. The datasets, which were generated after the removal of the high-leverage compounds, were pre-treated using a Java-based in-house tool DataPreTreatmentGUI 1.2 (available from https://dtclab.webs.com/software-tools). These pre-treated datasets were divided into training and test set in a 3:1 ratio, where the training set was used for the model development and the test set was used for model validation purpose. Here, we have randomly divided our datasets into 3 different combinations and the divided training and

test sets were further pre-treated by using the tool dataPreTreatmentTrainTest1.0 (available from https://dtclab.webs.com/software-tools). This final pre-treated training and test sets were used for further analysis. The training and test sets composition for respective datasets is shown in **Figure 3.2**.

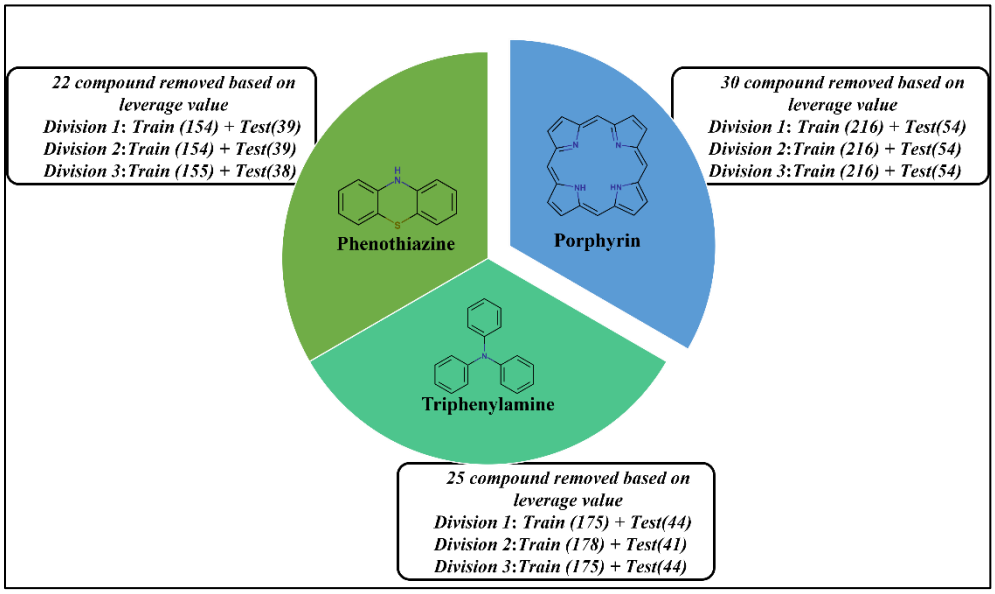


Figure 3.2: The composition of training and test sets for respective datasets

3.2.2.5 Feature selection

In order to employ only the descriptors that are most significant to the model and decrease the noise coming from the descriptors, it is crucial to carefully choose the features [135, 136]. Feature selection reduces the higher dimensional feature space to lower dimensional space by eliminating the noisy and insignificant descriptors, which helps in better interpretability of the model [153]. In the present work, the reduction of the initial feature pool was carried out by genetic algorithm (GA), [154] by using a java based in house tool GeneticAlgorithm_v4.1 (available from https://dtclab.webs.com/software-tools). The GA method of feature selection consists of two steps where in the first step it produces a population of N descriptors and calculates the fitness score of each descriptor combination in the first generation. Based on a scaled fitness score, the following phase produces a portion of the features of the next generation by crossing over (crossover children) and the remainder through mutation (mutation children) from the parents. In the new generation, the offspring contains information about their parents.

This process of crossover and mutation is continued until all member of the *N* population is replaced by their children. The fitness score of new members is again calculated in the next generation, and this cycle is repeated until 80% of the generations have the same fitness score [133]. Here, we have used only the training set of all three divisions independently for the selection of the feature.

Then the reduced pool of features obtained from GA is subjected to the best subset selection method to find out the best possible combination of descriptors for modeling. In the best subset feature selection method, a subset of the input feature pool is selected, and a model is developed with these selected features. This process of feature selection and model development is continued for all possible combinations. In this work, we have used a java-based tool "Best Subset Selection (BSS) v2.1" (available from http://dtclab.webs.com/software-tools) to find out the best subset of descriptors or features for modeling. Here, we have developed several different multiple linear regression (MLR) models with different feature combinations and the best feature combination is selected based on the model quality. In this present work, we have developed all models with 10 descriptors for all the respective datasets and divisions.

3.2.2.6 QSPR model development

The selected descriptors along with the response data were standardized with respect to the training set mean and standard deviation. A partial least squares (PLS) model was developed from the selected descriptors by using the PLS_SingleY_1.0 tool (available from http://dtclab.webs.com/software-tools). PLS is a generalized form of multiple linear regression (MLR) technique, used when there is inter-correlation and multi-collinearity between descriptors is present [155]. This descriptor combination is used for further analysis through machine learning modeling. In this study, we have developed machine learning models like random forest (RF), adaboost (ADB), gradient boost (GB), extreme gradient boost (XGB), support vector machine (SVM), linear support vector machine (LSVM), and ridge regression (RR) [143]. For the development of all ML models, we have used an in-house tool RSL v2.1 (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis) which builds regression-based ML models. In this work, all ML models were developed using the default settings of the associated hyper-parameter.

The internal and external validation metrics of the developed models were checked to determine their robustness, quality, and predictivity. The internal validation metrics like the determination coefficient (R^2), mean absolute error of the training set ($MAE_{training}$), and root mean square error of calibration (RMSEC) are calculated on the training set. To check the predictivity of the models we have calculated the metrics like the predictive R^2 (Q^2_{F1}), Q^2_{F2}, Q^2_{F3}, mean absolute error of test set (MAE_{test}) and root mean square error of prediction (RMSEP), using the test set. We have also calculated the different cross-validation metrics like leave-one-out cross-validated R^2 (Q^2_{LOO}), leave-one-out mean absolute error (MAE_{LOO}), and leave-one-out mean squared error (MSE_{LOO}), using training sets to see whether the generated models are overfitted or not [133]. The mathematical formula for all these metrics is shown in equations no. 3.1 to 3.11.

$$MAE_{training} = \frac{\sum|Y_{obs(training)} - Y_{cal(training)}|}{n_{training}} \quad (3.6)$$

$$RMSEC = \sqrt{\frac{\sum(Y_{obs(training)} - Y_{cal(training)})^2}{n_{training}}} \quad (3.7)$$

$$MAE_{LOO} = \frac{\sum|Y_{obs(training)} - Y_{loo\ predicted(training)}|}{n_{training}} \quad (3.8)$$

$$MSE_{LOO} = \frac{\sum(Y_{obs(training)} - Y_{loo\ predicted(training)})^2}{n_{training}} \quad (3.9)$$

$$Q^2_{F3} = 1 - \frac{[\sum(Y_{obs(test)} - Y_{pred(test)})^2]/n_{test}}{\left[\sum(Y_{obs(training)} - \bar{Y}_{training})^2\right]/n_{training}} \quad (3.10)$$

$$RMSEP = \sqrt{\frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{n_{test}}} \quad (3.11)$$

Here, $Y_{obs(trainining)}$ is the observed response value of the training set, $Y_{cal(training)}$ is the calculated response value of the training set, $\bar{Y}_{training}$ is the mean observed response value of the training set, $Y_{loo\ predicted(training)}$ is the leave one out predicted response value of the training set, $n_{training}$ is the number of observation in the training set, $Y_{obs(test)}$ is the observed response value of the test set, $Y_{pred(test)}$ is the predicted response value of the test set, $\bar{Y}_{test}$ is the mean observed response value of the test set, n_{test} is the number of observation in the test set.

The developed models were then used for the prediction of compounds that were removed from the datasets based on the leverage values. These removed compounds are analyzed by each ML model and the error of prediction in terms of MAE and RMSEP is calculated.

3.2.2.7 RASPR descriptor calculation and q-RASPR model development

RASPR descriptors are different similarity and error measures obtained from the similarity-based read-across predictions [124]. The final pooled structural and physicochemical descriptors were used to calculate the RASPR descriptors for both training and test sets. The RASPR descriptors 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 RASPR descriptor calculation of the training set, the training set itself is used as the query set (training vs training) but for the computation of the test set RASPR descriptors, the test set is used as the query set (training vs test). The calculation of the RASPR descriptors for the training set is based on the Leave-same-out (LSO) method where identical compounds are not considered among the list of close source compounds. In the present work, we have calculated the RASPR descriptor for both training and test sets at the default settings of the associated hyper-parameters (σ=1, γ=1, number of close training compounds (CTC)=10). There are three similarity-based functions provided by this tool for the calculation of the RASPR descriptor, but we have used only two methods for RASPR descriptor calculation i.e., the Gaussian Kernel (GK)-based similarity and the Laplacian Kernel (LK)-based similarity functions. These calculated RASPR descriptors are then merged with initial physicochemical and structural descriptors to build the q-RASPR model after subsequent feature selection. Here, we develop models with both GK and LK methods and among them; only the best model is selected based on the predictive performance of the developed models.

The final dataset of RASPR descriptors merged with the structural and physicochemical descriptors was made to undergo feature selection by the best subset selection method using the "Best Subset Selection (BSS) v2.1" tool (available from http://dtclab.webs.com/software-tools). From the results of the best subset selection, we have selected four feature combinations based on the internal validation metrics like R^2, $Q^2_{LOO,}$ and $MAE_{training}$. These four feature combination were used for the development of PLS models using the PLS_SingleY_1.0 tool (available from

http://dtclab.webs.com/software-tools). Among the four models, the best model was reported based on the productive quality of the model toward the test set. This descriptor combination was also used for the development of ML models like RF, ADB, GB, XGB, SVM, LSVM, and RR by using the tool RSL v2.1, as stated above. For all the developed models we have calculated the same set of internal (R^2, $MAE_{training}$, RMSEC), external (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, MAE_{test}, RMSEP), and cross-validation (Q^2_{LOO}, MAE_{LOO}, MSE_{LOO}) metrics to check the models' quality, robustness, predictivity and whether the models were overfitted or not.

The developed q-RASPR models were also used for the prediction of compounds initially removed from the datasets based on the leverage values. The RASPR descriptors of these compounds were calculated by taking the training set of developed models and considering these removed compounds as a query set. These RASPR descriptors were also calculated at the default hyper-parameter setting that is $\sigma = 1$, $\gamma = 1$, CTC = 10. Here, we have also calculated the prediction error in terms of MAE and RMSEP for comparison with QSPR models, and to show that the RASPR method enhances the predictive performance.

3.2.2.8 PLS plots [155, 156]

Different types of PLS plots were generated for q-RASPR PLS models using Simca-P v10.0, available from https://www.sartorius.com/. The **Score Plot** shows the compounds in latent variable (LV) space, where the axis of this plot is formed by the first two LVs. The compounds which are located in the ellipse of the plot and present closer to each other are considered "similar compounds" but if compounds are lying outside of the ellipse, they are called "outliers". The **DModX** approach is used to check the applicability domain of the models. The **Loading plot** represents the relationship between the dependent variable (Y) and the independent variable (X) of the models. The **Bubble plot** represents the importance of the descriptors to the response in the form of a bubble size, where the x-axis represents descriptors, and the y-axis of the plot represent the standardized regression coefficient value. The importance of the descriptors is also represented by their location along x axis, and colour difference of the bubbles represent the direction of contribution (i.e., positive contribution is represented by green colour and negative contribution is represented by red colour). A **y-Randomization plot** tests whether the developed models are obtained by chance or not. To study the deviation of the predicted PCE value from

the observed PCE value, we developed a **Scatter Plot** by plotting the predicted PCE value along the y-axis and the observed PCE along the x-axis.

3.2.2.9 SHAP plots

Shapley additive explanations (SHAP) is a game theoretic approach used to explain the feature importance toward the predictions made by an ML model [146]. The SHAP method was developed based on two main principles Shapley value [157] and linear LIME (Local Interpretable Model-agonistic Explanations) [158]. The Shapley value is a mean marginal contribution of a feature for a particular prediction and LIME is used to develop a model which tries to approximate the main machine learning model locally. The Shapley value not only determines the contribution of a feature (magnitude) but also indicates the direction (negative or positive). By using the SHAP method we can find out the importance of a feature globally (overall importance of a feature) and locally (importance of a feature for a particular prediction) but in this study, we have used SHAP only to explain the global importance of the feature [147, 148]. Here we have used a python-based module SHAP (version 0.41.0) to find out feature importance toward predictions in the form of a summary plot. In the summary plot, the features are arranged along the vertical axis according to the feature importance in descending order and the horizontal axis indicates the SHAP value distribution. In this plot, each data point is represented by a dot which is colored based on its feature value and helps to determine the direction of feature contribution (i.e., the high value of the feature increase SHAP value or not).

The complete working process of this present study is shown in **Figure 3.3.**

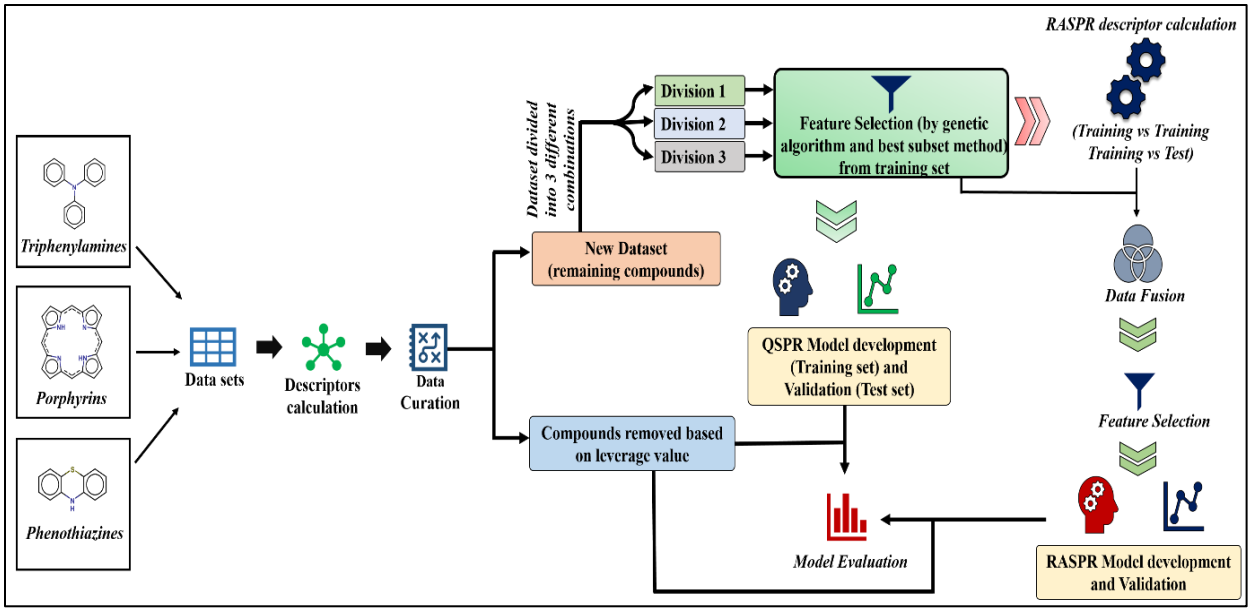


Figure 3.3: The schematic representation of the complete working process of q-RASPR model development

Chapter – 4

RESULT AND DISCUSSION

4. RESULT AND DISCUSSION

4.1 Study 1: Machine learning - based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells

In this study, we initially developed PLS models for each of the considered data sets and then compared the quality of these models to ML-derived model predictions.

4.1.1 Partial Least Squares (PLS) models and interpretation of modeled descriptors

Most significant and statistically robust PLS models for different categories of dyes along with their quality in terms of different internal and external validation metrics are shown in **Table 4.1**. These models were developed with 8, 8, 6 and 5 descriptors for coumarins, carbazoles, indolines and diphenylamines, respectively, and all these models consist of both RASPR and 2D structural descriptors. The models were developed using 7, 3, 3 and 3 latent variables (LVs), respectively based on the leave-one-out Q^2 values. All the developed models satisfy the threshold limit required to become robust, reliable and good predictivity [159].

Table 4.1: Developed q-RASPR PLS models for different type of organic dye used in DSSCs.

Types of organic dye	PLS Models	Training set metrics		Test set metrics		
		R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}	MAE_{test} (95%)

Coumarins (LV = 7)	$PCE = -1.71195 + 0.60957 \times \text{SD Activity(GK)} + 0.94835 \times \text{MaxPos(GK)} - 0.75671 \times \text{nRCN} + 1.07215 \times \text{nThiophenes} - 1.01363 \times \text{nRC} - +1.18272 \times \text{nR} = \text{Ct} - 0.12347 \times \text{T(S..S)} + 0.76919 \times \text{C} - 040$	0.75	0.63	0.72	0.70	0.75
Carbazoles (LV = 3)	$PCE = -0.23418 + 1.26064 \times \text{Avg. Sim(GK)} - 1.51529 \times \text{F06[N} - \text{N]} + 0.92434 \times \text{nR10} + 0.19841 \times \text{F04[C} - \text{N]} + 2.12686 \times \text{B04[N} - \text{O]} - 0.4133 \times \text{N\%} + 2.35133 \times \text{F06[N} - \text{O]} + 1.42348 \times \text{B02[C} - \text{S]}$	0.71	0.66	0.77	0.76	0.61

Indolines (LV = 3)	$PCE = 1.52408 + 0.88535 \times \text{RA function(LK)} - 0.89273 \times \text{CVsim(LK)} - 0.92139 \times \text{Neg. Avg. Sim} + 0.01956 \times \text{F04[C} - \text{N]} + 0.70912 \times \text{B09[O} - \text{S]} - 0.05307 \times \text{nCconj}$	0.63	0.59	0.81	0.81	0.55
Diphenylamines (LV = 3)	$PCE = 1.28039 + 0.8856 \times \text{RA function(LK)} + 1.53133 \times \text{SD similarity(LK)} - 0.14367 \times \text{F01[C} - \text{N]} - 0.15417 \times \text{StsC} + 0.35804 \times \text{F04[N} - \text{S]}$	0.83	0.73	0.90	0.90	0.62

LV = Latent variable

The importance of the features toward the PCE is represented in the form of the bubble plot (**Figure 4.1**), in which variable importance scores and coefficient scores are calculated by using SIMCA-P v10.0 software (https://www.sartorius.com/). The importance of these descriptors is represented by the diameter of the bubbles and their relative position along the y-axis whereas color difference denotes positive and negative contribution.

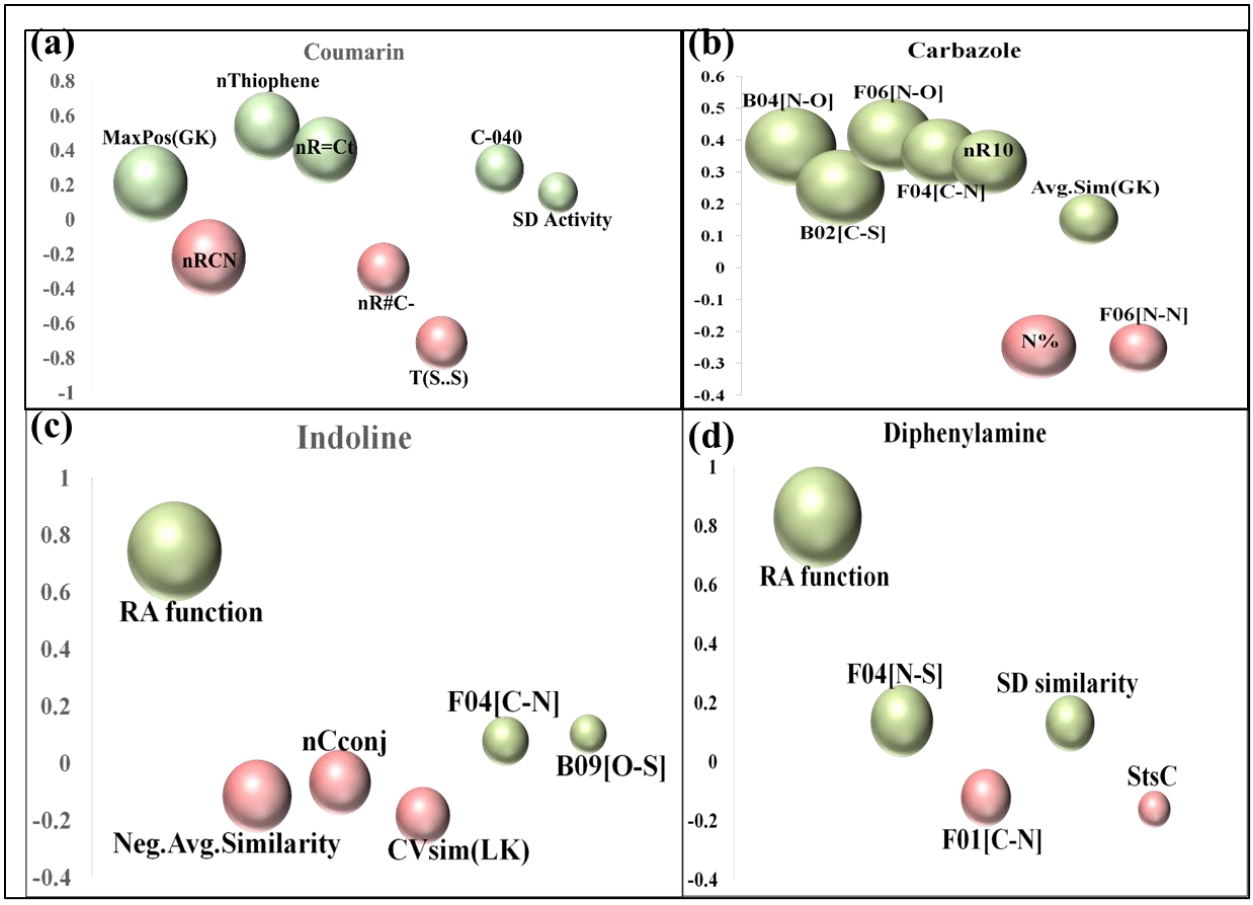


Figure 4.1: Bubble plot showing regression coefficients for individual descriptors in the PLS model (size of the bubble shows variable importance score)

Modelling of PCE of coumarin dyes

The most significant descriptors in the form of a mathematical equation are shown in **Table 4.1** and the contribution of these descriptors towards PCE in the form of a bubble plot in **Figure 4.1a**. The mechanistic interpretation of these descriptors is discussed below:

MaxPos (GK) is a RASPR descriptor that represents the similarity value to the nearest positive close source compound based on training set mean, obtained by Gaussian kernel similarity-based method [124]. From the bubble plot, it was found that this descriptor has the highest contribution to the PCE, as shown in **Figure 4.1a.** *MaxPos(GK)* shows a positive contribution as reflected in following example: **19** (*MaxPos(GK)* = 1, PCE = 7.4), **20** (*MaxPos(GK)* = 1, PCE = 6.4) and vice versa for the dyes **56** (*MaxPos(GK)* = 0.014, PCE = 0.99), **22**(*MaxPos(GK)* = 0.004, PCE = 0.33).

The Functional group count descriptor nThiophene denoting the number of thiophene rings in the coumarin dyes contributes positively to the PCE. Therefore, presence of such functional group in the dye increases the performance of DSSCs as represented by the following examples: **19** (nThiophene = 2, PCE = 7.4), **32** (nThiophene = 2, PCE = 6.5). The PCE value may reduce for the compounds where no such functional group is present as shown in the following examples: **56** (nThiophene = 0, PCE = 0.99), **17** (nThiophene = 0, PCE = 0.9). Thiophene groups are the part of the π-spacer which not only improves light absorption and dipole moment but also decreases the dihedral angle between donor/acceptor and π-spacer plane for better orbital overlap which in turn improve electron injection to TiO_2 [160].

Two other functional group count descriptors nR=Ct (number of an aliphatic tertiary carbon atom with the 'sp2' hybridization) and nR#C- (number of a non-terminal carbon atom with the 'sp' hybridization) have positive and negative contributions to the PCE, respectively. The presence of aliphatic tertiary 'sp2' hybridized C atom and the absence of non-terminal 'sp' hybridized C atom frequency of 's' is responsible for the enhancement of absorption [161]. The contribution of the descriptor nR=Ct is represented by the following examples: **35** (nR=Ct = 4, PCE = 6.2), **32** (nR=Ct = 3, PCE = 6.5 **17** (nR=Ct = 0, PCE = 0.9), **22** (nR=Ct = 0, PCE = 0.33); and the following examples represent the contribution of nR#C- **44** (nR#C- = 2, PCE = 1.35), **56** (nR#C- = 2, PCE = 0.99) **19** (nR#C- = 0, PCE = 7.4), **32** (nR#C- = 0, PCE = 6.5).

nRCN is a functional group count descriptor denoting the number of aliphatic nitriles in the dye which contributes negatively to the PCE of coumarin dyes. Therefore, with the increasing number of nitrile groups, the performance of DSSCs is reduced as indicated by the following examples: **44** (nRCN = 1, PCE = 1.35), **56** (nRCN = 1, PCE = 0.99) and *vice versa* for the dyes **10** (nRCN = 0, PCE = 3.7), **54** (nRCN = 0, PCE = 3.5) where no nitrile group is present. Anchoring groups are a part of the dye which involves adsorption on TiO_2 surface that determines electron injection ability and optoelectrical property of the dye. Nitrile groups are generally a part of this anchoring group which may increase adsorption stability when CN group itself is involved in the binding. On the other hand, nitrile groups may reduce the photovoltaic property when it is not involved in binding [162].

T(S..S) is a 2D atom pair descriptor that indicates the sum of the topological distance between two sulfur atoms where they are part of two thiophene rings. The negative contribution of this descriptor signifies that with the increasing distance between sulfur atoms, the performance of

the DSSCs may decrease as represented by the following examples: **22** (T(S..S) = 31, PCE = 0.33), **3** (T(S..S) = 28, PCE = 1.77), **44** (T(S..S) = 21, PCE = 1.35) and *vice versa* for the dyes **19** (T(S..S) = 3, PCE = 7.4), **32** (T(S..S) = 3, PCE = 6.5), **29** (T(S..S) = 3, PCE = 6.07). The possible reason for this may be due to the disruption of the planar structure of the π-spacer and increase of dihedral angle between adjacent donor/acceptor and π-spacer. Another reason is that a hole is created in the π-spacer after injection of an electron to the TiO_2; this hole is transferred to the donor part and prevents charge recombination. Therefore, with the increasing length of π-spacer, the possibility of this hole transfer is reduced, and this may cause back transfer of electron and reduce DSSCs performances [160].

C-040 is an atom-centered fragments descriptor that represents fragments like R–C(=X)-X/R–C#X/X = C = X (R: any group linked through carbon; X: any electronegative atom like N, S, P, O, halogen; #: triple bond) in the dye which contributes positively to the PCE. The positive contribution of this descriptor signifies that the presence of such fragments in the dye may increase the performance of the dye as shown in the following examples: **19** (C-040 = 3, PCE = 7.4), **20** (C-040 = 3, PCE = 6.4), **35** (C-040 = 3, PCE = 6.2) and *vice versa* for the dyes **24** (C-040 = 2, PCE = 1.04), **17** (C-040 = 2, PCE = 0.9). These fragments are generally parts of the anchoring group containing carboxylic acid or cyanoacrylic acid as a binder which increases the stability of adsorption on TiO_2 surface and helps in efficient electron transfer [162].

SD Activity (GK) is a RASPR descriptor that denotes the weighted standard deviation of the response value of the selected close source compound for each query compound. The positive contribution of this descriptor signifies that with an increasing numerical value of this descriptor PCE of DSSCs may increase [124]. The positive contribution is represented by the following examples: **29** (*SD Activity (GK)* = 1.53777, PCE = 6.07), **36** (*SD Activity (GK)* = 1.40648, PCE = 5.5), **7** (*SD Activity (GK)* = 1.04559, PCE = 1.1), **17** (*SD Activity (GK)* = 0.9868, PCE = 0.9).

The mechanistic interpretation of the 2D structural descriptor of the q-RASPR PLS model for the coumarin dyes is schematically represented in **Figure 4.2.**

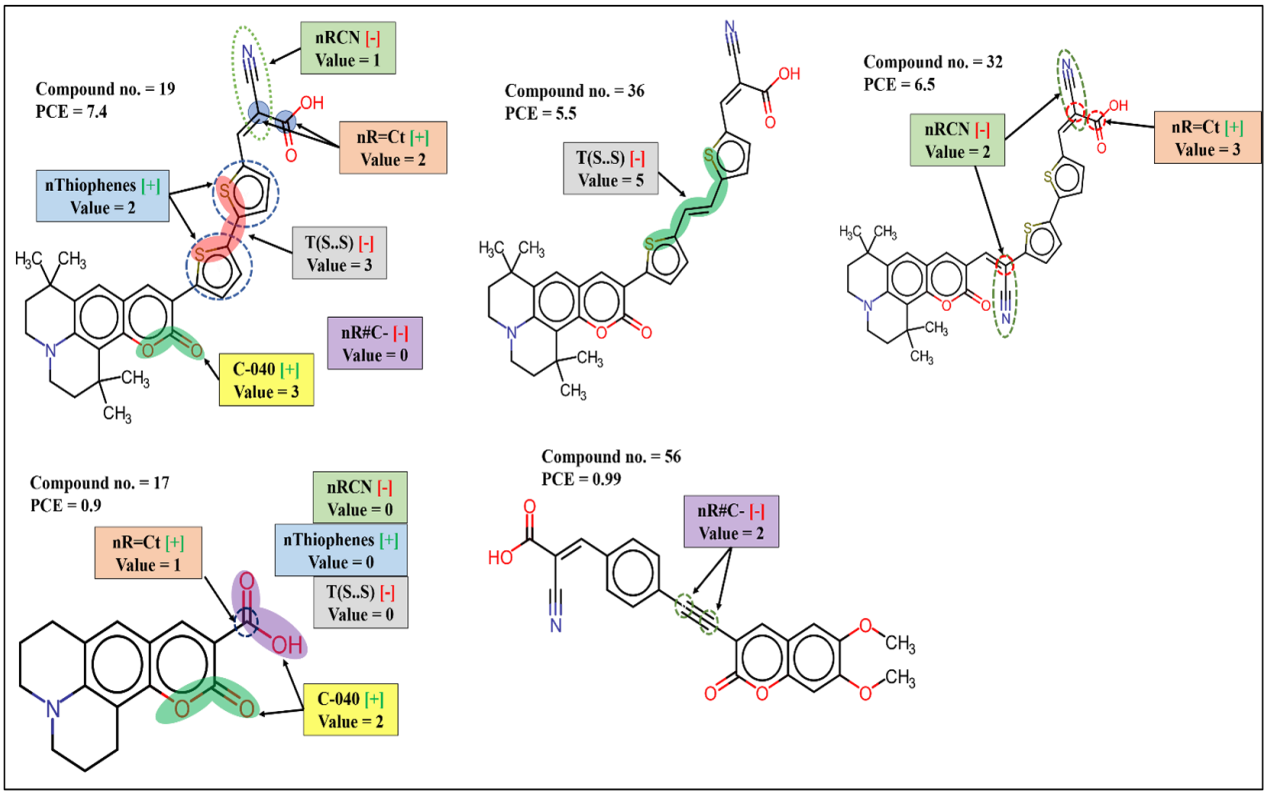


Figure 4.2: Mechanistic interpretation of 2D structural descriptors of q-RASPR PLS model for the coumarin dataset

Modeling of PCE of Carbazoles

The PLS q-RASPR model related to the carbazole dyes consisting of 8 descriptors has been shown in **Table 4.1,** and the contribution of the descriptors in the form of a bubble plot is shown in **Figure 4.1b.** The mechanistic interpretation of these descriptors and their influence on PCE is discussed below:

The 2D atom pair descriptor B04[N-O] indicates the presence or absence of nitrogen and oxygen atoms at the topological distance 4, and this descriptor contributes positively to the PCE. This fragment is part of an anchoring group cyanoacrylic acid. Cyanoacrylic acid is one of the most common anchoring groups for metal oxide (TiO_2) surfaces because of its dual character of a strong adsorber and a good acceptor. Its strong binding with TiO_2 provides stability to the adsorbed dyes which in turn helps in the efficient transfer of an electron to TiO_2. This cyanoacrylic acid also has a strong electron-withdrawing ability which helps in intramolecular charge transfer from donor to metal oxide [163]. Therefore, when such fragments are present in the dye, the performance of DSSCs will increase as represented by the following examples: **132**

(B04[N-O] =1, PCE= 12.5), **133** (B04[N-O] =1, PCE = 9.32), **101** (B04[N-O] =1, PCE = 8.09) and *vice versa* for the dyes **160** (B04[N-O] =0, PCE = 0.34), **157** B04[N-O] =0, (PCE = 0.31), **159** (B04[N-O] =0, PCE = 0.21).

B02[C-S] is a 2D atom pair descriptor that indicates the presence or absence of carbon and sulfur at the topological distance 2. The positive contribution of this descriptor indicates that the presence of such fragment increases the performance of DSSCs. This fragment is a part of the thiophene group that acts as a π-spacer present between donor and acceptor moiety. This electron-rich π-spacer is responsible for the enhancement absorption of photons which in turn increases PCE of carbazole dye [160]. The positive contribution of this descriptor is represented by the following examples: **132** (B02[C-S] =1, PCE= 12.5), **133** (B02[C-S] =1, PCE = 9.32), **101** (B02[C-S] =1, PCE = 8.09) and *vice versa* for the dyes **160** (B02[C-S] =0, PCE = 0.34), **157** (B02[C-S] =0, PCE = 0.31).

F06[N-O] is another 2D atom pair descriptor that indicates the frequency of nitrogen and oxygen atoms at the topological distance 6, and it contributes positively toward the PCE. This fragment is present in the dye either as a part of the phenyl moiety between the acceptor (cyanoacrylic acid) and adsorber or as a part of the linker between the donor and π-spacer (furan or enedioxythiophene). A dye containing this fragment between acceptor and adsorber will have an improved performance by its diode like effect (which prevents the back transfer of electrons from TiO_2 to the dye) [163]. It helps in an efficient intramolecular charge transfer for the dyes containing this fragment as a linker between the donor moiety and π-spacer [164]. The positive contribution of this descriptor is represented by the following examples: **132** (F06[N-O] = 3, PCE = 12.5), **133** (F06[N-O] = 2, PCE = 9.32) and vice versa for the dyes where no such fragment is present, **160** (F06[N-O] = 0, PCE = 0.34), **157** (F06[N-O] = 0, PCE = 0.31).

F04[C-N] is a 2D atom pair descriptor that denotes the frequency of carbon and nitrogen atoms at the topological distance 4, and this descriptor contributes positively to the PCE. This carbon to nitrogen fragment increases the electron donating ability of the dyes which in turn increases the performance of the DSSCs as indicated by the positive contribution of this descriptor [165]. The PCE value increases in the presence of such fragments in the dyes as indicated by the following examples: **50** (F04[C-N] = 22, PCE = 7.52), **101** (F04[C-N] = 17, PCE = 8.09), **130** (F04[C-N]

= 15, PCE = 9.8) and *vice versa* for the dyes **97** (F04[C-N] = 0, PCE = 0.0538), **98** (F04[C-N] = 0, PCE = 0.0387) where no such fragment is present.

nR10 is a ring descriptor that indicates the number of 10 membered rings in a dye which contributes positively to the PCE. The presence of 10-membered rings in the carbazole dyes increases the electron-donating ability of the dyes and is thus responsible for the high PCE values [166]. Therefore, performances of DSSCs should increase when such ring is present in the structures, which is indicated by the following examples: **130** (nR10 = 6, PCE = 9.8), **131** (nR10 = 6, PCE = 7.6) and *vice versa* for the dyes **159** (nR10 = 0, PCE = 0.21), **91** (nR10 = 0, PCE = 0.19), **154** (nR10 = 0, PCE = 0.07) where no 10 membered rings are present.

The negative contribution of the constitutional descriptor N% (percentage of the nitrogen atoms in the dye) and 2D atom pair descriptor F06[N-N] (frequency of two nitrogen atoms at the topological distance 6) indicates that the presence of such fragments hinders the performance of DSSCs. Higher numerical values of these descriptors of a dye may decrease the PCE value which is represented by the following examples: **91** (N% = 6.25, PCE = 0.19), **118** (N% = 5.6338, PCE = 0.89), **53** (N% = 4.83871, PCE = 0.99), **112** (N% = 4.83871, PCE = 0.96) for the descriptor N% ; **141** (F06[N-N] = 2, PCE = 2.58), **138** (F06[N-N] = 2, PCE = 2.17) for F06[N-N]. On the other hand, dyes with lower numerical value of this descriptor may have higher PCE values as shown in following examples: **99** (N% = 1.81818, PCE = 7.58), **103** (N% = 1.50376, PCE = 7.54), **130** (N% = 1.34529, PCE = 9.8) for N%, **132** (F06[N-N] = 0, PCE = 12.5), **130** (F06[N-N] = 0, PCE = 9.8), **133** (F06[N-N] = 0, PCE = 8.09) for the F06[N-N].

Avg. Sim (GK) is a RASPR descriptor that denotes the mean similarity value of the selected close source compounds for each query compound based on the Gaussian kernel similarity-based method. The positive contribution of this descriptor indicates a molecule having a higher *Avg. Sim* value may have a higher PCE value as represented by the following examples: **94** (*Avg. Sim (GK)* = 0.92848, PCE = 7.33), **56** (*Avg. Sim (GK)* = 0.89553, PCE = 6.04), **99** (*Avg. Sim (GK)* = 0.75399, PCE = 7.58) and *vice versa* for the dyes **156** (*Avg. Sim (GK)* = 0.24768, PCE = 0.06), **154** (*Avg. Sim (GK)* = 0.04227, PCE = 0.07).

The interpretation of 2D structural descriptors for the carbazole dyes is represented schematically in **Figure 4.3**.

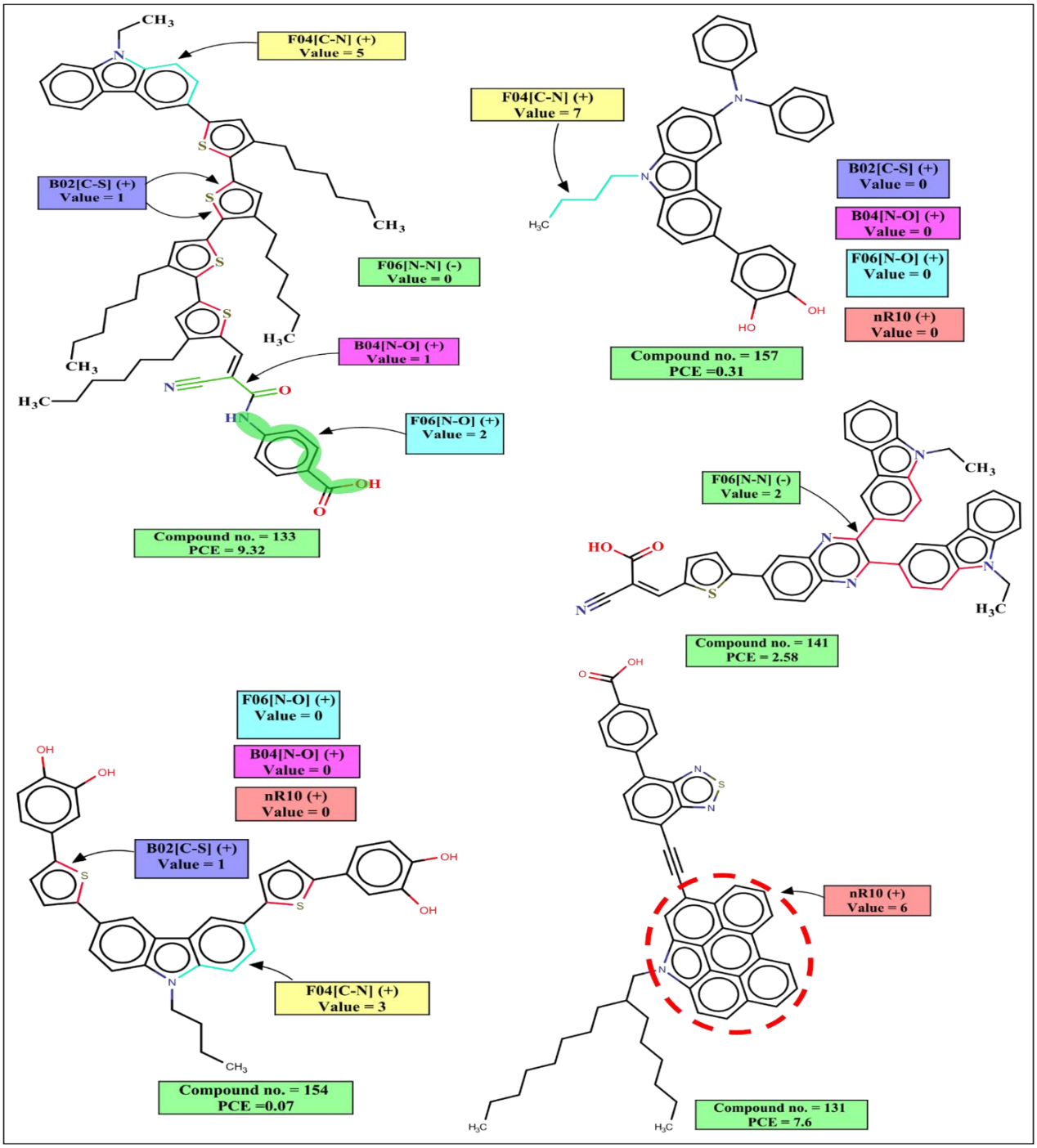


Figure 4.3: Mechanistic interpretation of the 2D structural descriptors of the q-RASPR PLS model for the carbazole dataset

Modeling of PCE of Indolines

The PLS q-RASPR model of indoline dyes consisting of 6 descriptors has been presented in **Table 4.1,** and the contribution of the descriptors in the form of a bubble plot is shown in **Figure 4.1c**. The meaning of these descriptors and their influence on PCE are discussed below:

RA function is a Read-Across-derived RASPR descriptor which encodes information of all the selected structural and physicochemical descriptors [167]. It contributes positively to the PCE as indicated by the following examples: **141** (*RA function* = 8.1741, PCE = 8.38), **8** (*RA function* = 7.9697, PCE = 7.12), **24** (*RA function* = 7.88, PCE = 9.2) and *vice versa* for the dye **129** (*RA function* = 1.8131, PCE = 1.48), **32** (*RA function* = 1.5372, PCE = 0.63), **30** (*RA function* = 1.4248, PCE = 0.77).

Both RASPR descriptors *Neg.Avg.Sim* (denoting the mean of the similarity values of the negative close source compounds for a particular query compound) and *CVsim (LK)* (coefficient of variation of the similarity values of the selected close source compound for each query compound) contribute negatively to the PCE. The negative contribution of these descriptors signifies that a higher value of the descriptors may decrease the PCE value as represented by the following examples: **93** (*Neg.Avg.Similaritry* = 0.2883, PCE = 0.35), **108** (*Neg.Avg.Similaritry* = 0.2883, PCE = 0.046) for *Neg.Avg.Similaritry*; **93** (*CVsim (LK)* = 1.404, PCE = 0.35), **108** (*CVsim (LK)* = 1.404, PCE = 0.046) for *CVsim (LK)*; and *vice versa* for the dye **144** (*Neg.Avg.Similaritry* = 0, PCE = 8.78), **24** (*Neg.Avg.Similaritry* = 0, PCE = 9.2), **135** (*Neg.Avg.Similaritry* = 0, PCE = 8.61) for *Neg.Avg.Similaritry*; **135** (*CVsim (LK)* = 0.4427, PCE = 8.61), **78** (*CVsim (LK)* = 0.3868, PCE = 7.99) for *CVsim (LK)*.

The functional group count descriptor nCconj denotes the number of non-aromatic conjugated sp2 hybridized carbon atoms that contributes negatively to the PCE. The negative contribution of this descriptor signifies that the PCE value may decrease when the number of non-aromatic conjugated sp2 carbon increases as represented by the following examples: **11** (nCconj = 11, PCE = 2.65), **10** (nCconj = 10, PCE = 2.7) and *vice versa* for the dyes with a low numerical value of nCconj like **155** (nCconj = 1, PCE = 5.61) and **152** (nCconj = 1, PCE = 5.5).

F04[C-N] is a 2D atom pair descriptor that indicates the frequency of carbon and nitrogen atoms at the topological distance 4 in the dye, and this descriptor contributes positively to the PCE. It

was found that if the donor group is present with a non-planar orientation with other groups, it may increase the PCE value. Although this fragment is present as a part of the dye in a non-planar structure, it may increase the performance of the DSSCs as indicated by its positive contribution to the PCE [165, 168]. Therefore, the presence of such fragment increases the performance of DSSCs as shown by the following examples: **21** (F04[C-N] = 21, PCE = 8.43), **8** (F04[C-N] = 20, PCE = 7.12), **24** (F04[C-N] = 18, PCE = 9.2) and *vice versa* for the dyes **105** (F04[C-N] = 3, PCE = 2.53), **164** (F04[C-N] = 3, PCE = 2.08).

Another 2D atom pair descriptor B09[O-S] indicates the presence or absence of oxygen and sulfur atoms at the topological distance 9, and this descriptor contributes positively to the PCE. Oxygen and sulfur atoms control electron density delocalization which helps in π bond conjugation. As a result, the molar extinction coefficient of the dye increases which may lead to shifting of the absorption maxima [169]. Dyes containing this type of fragment may increase PCE values as represented by the following examples: **78** (B09[O-S] = 1, PCE = 7.99), **21** (B09[O-S] = 1, PCE = 6.12), **131** (B09[O-S] = 1, PCE = 6.11) and *vice versa* for the dyes **30** (B09[O-S] = 0, PCE = 0.77), **32** (B09[O-S] = 0, PCE = 0.63), **93** (B09[O-S] = 0, PCE = 0.35).The mechanistic interpretation of the relevant descriptors for the indoline dataset is schematically represented in **Figure 4.4.**

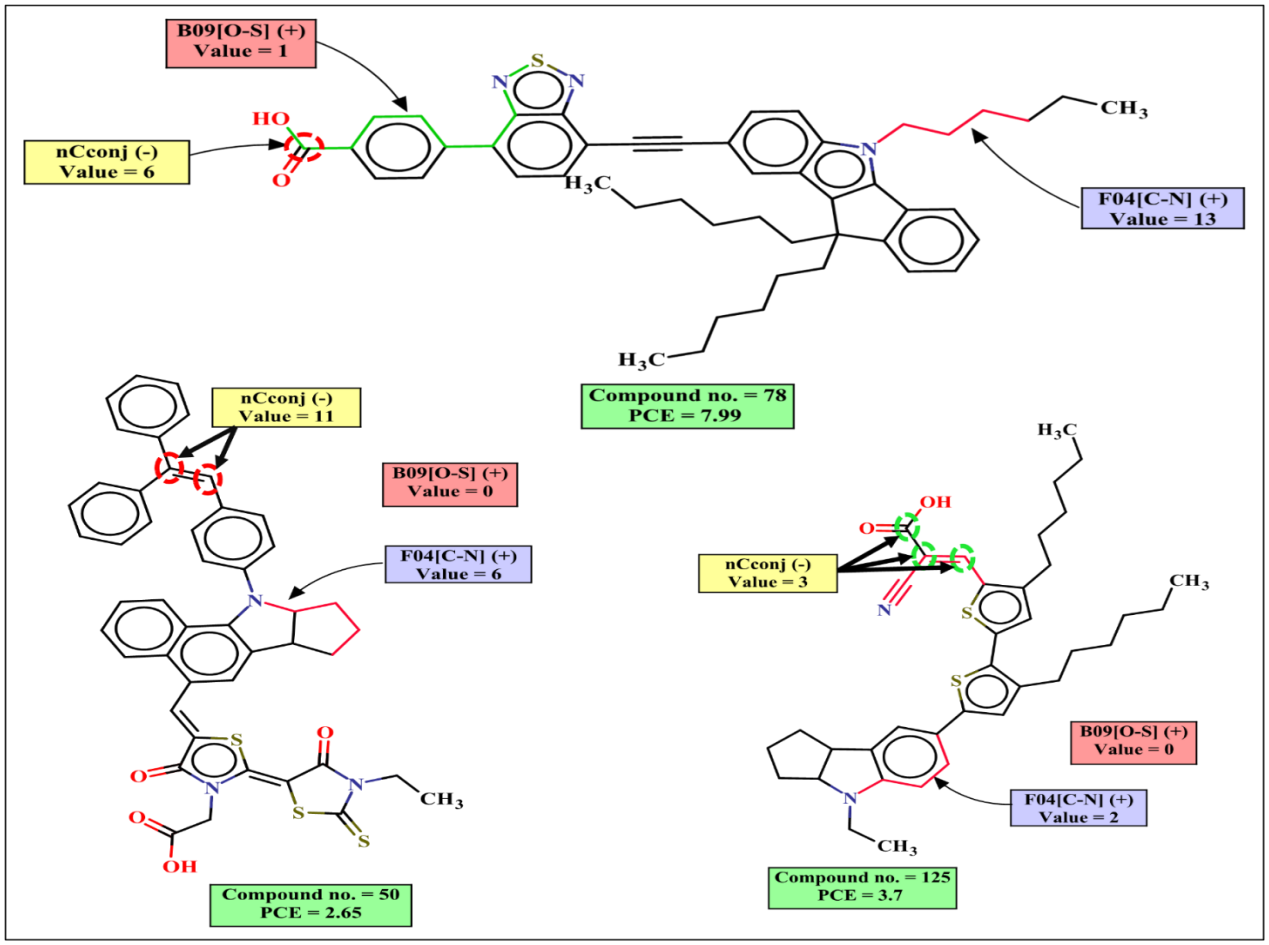


Figure 4.4: Mechanistic interpretation of the 2D structural descriptors of q-RASPR PLS model for the indoline dataset

Modelling of PCE of Diphenylamines

The PLS q-RASPR model consisting of 5 descriptors has been shown in **Table 4.1,** and the contribution of the descriptors in the form of a bubble plot is shown in **Figure 4.1d.** The mechanistic interpretation of these descriptors is represented below:

F01[C-N] is a 2D atom pair descriptor that indicates the frequency of carbon and nitrogen atoms at the topological distance of 1, and this descriptor contributes negatively to the PCE. This fragment affects the polarity of the dye which leads to its aggregation on the surface of the semiconductor layer and decreases photo absorption [170]. Therefore, the presence of such fragment reduces the performance of the DSSCs as represented by the following examples: **35** (F01[C-N] = 11, PCE = 0.4), **34** (F01[C-N] = 10, PCE = 1) and *vice versa* for the dyes **3** (F01[C-N] = 4, PCE = 5.4), **22** (F01[C-N] = 4, PCE = 5.22), where no such fragment is present.

Another 2D atom pair descriptor F04[N-S] denotes the frequency of nitrogen and sulfur atoms at the topological distance 4, and this descriptor contributes positively to the PCE. Therefore, the presence of such fragments increases the performance of the DSSCs as represented by the following examples: **27** (F04[N-S] = 2, PCE = 8), **26** (F04[N-S] = 2, PCE = 7.1), **17** (F04[N-S] = 2, PCE = 6.19) and *vice versa* for the dyes **34** (F04[N-S] = 0, PCE = 1), **33** (F04[N-S] = 0, PCE = 0.44), **35** (F04[N-S] = 0, PCE = 0.4) where no such fragment is present.

SD_similarity is a RASPR descriptor that denotes the standard deviation of the similarity values of close source compounds for each query compound. The positive contribution of this descriptor signifies that a high numerical value of the descriptor may increase PCE value as shown in the following examples: **7** (*SD similarity* = 0.33695, PCE = 7.05), **8** (*SD similarity* = 0.33274, PCE = 7.64) and *vice versa* for the dyes **35** (*SD similarity* = 0.177502, PCE = 0.4), **33** (*SD similarity* = 0.016381, PCE = 0.44).

StsC is an atom type E-state descriptor that indicates the sum of tsC E-states (≡C—), which contributes negatively to the PCE property of the DSSCs. Therefore, increasing the values of this descriptor may decrease the PCE value as observed for the dyes **10** (StsC = 8.292574, PCE = 1.99), **13** (StsC = 7.76829, PCE = 3.16) and *vice versa* for the dyes **8** (StsC = 1.671126, PCE = 7.64), **7** (StsC = 1.644351, PCE = 7.05).

The mechanistic interpretation of the significant 2D-structural descriptors for the diphenylamine dataset is schematically represented in **Figure 4.5**.

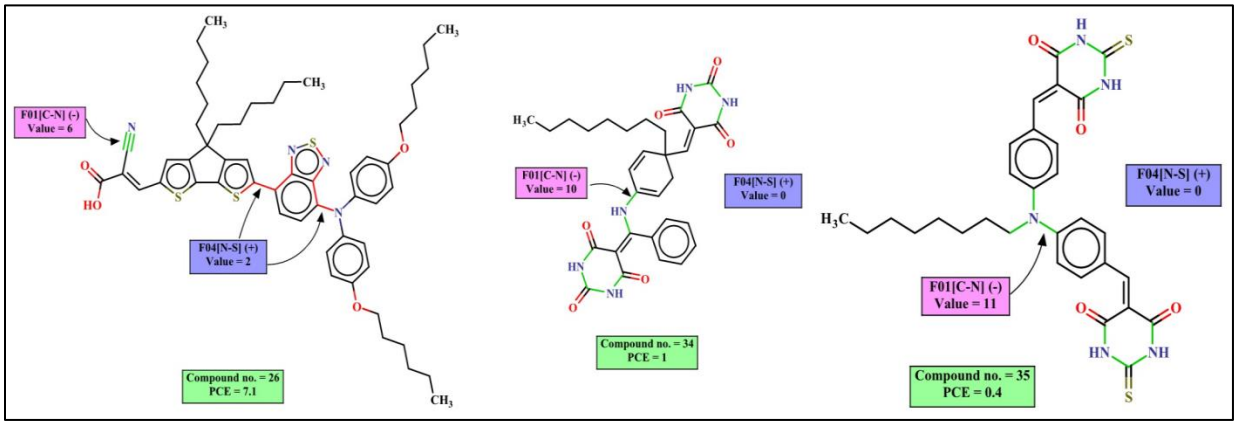


Figure 4.5: Mechanistic interpretation of the 2D structural descriptors of q-RASPR PLS model for the diphenylamine dataset

For all the datasets, the PLS **Scatter plots** (**Figure 4.6**) show that there is not so much difference between observed and predicted PCE indicating the good quality of the test set predictions. The **Loading plots** (**Figure 4.7**) show that the descriptors *MaxPos(GK)* (for coumarins), BO4[N-O] (for carbazoles) and *RA function* (for both indoline and diphenylamines) have the highest contributions to the PCE because they are present closest to the response variable (PCE). The **Score plots** (**Figure 4.8**) show that there are 2 coumarin (**3, 54**), 5 carbazole (**132, 138, 139, 140, and 141**) and 1 indoline (**18**) molecules which are present outside the applicability domain of the corresponding models. The **Y-Randomization plots** (**Figure 4.9**) show that all the models have R^2 and Q^2 intercept values within their threshold limits (0.3 for R^2 and 0.05 for Q^2), indicating that our models are not obtained by chance.

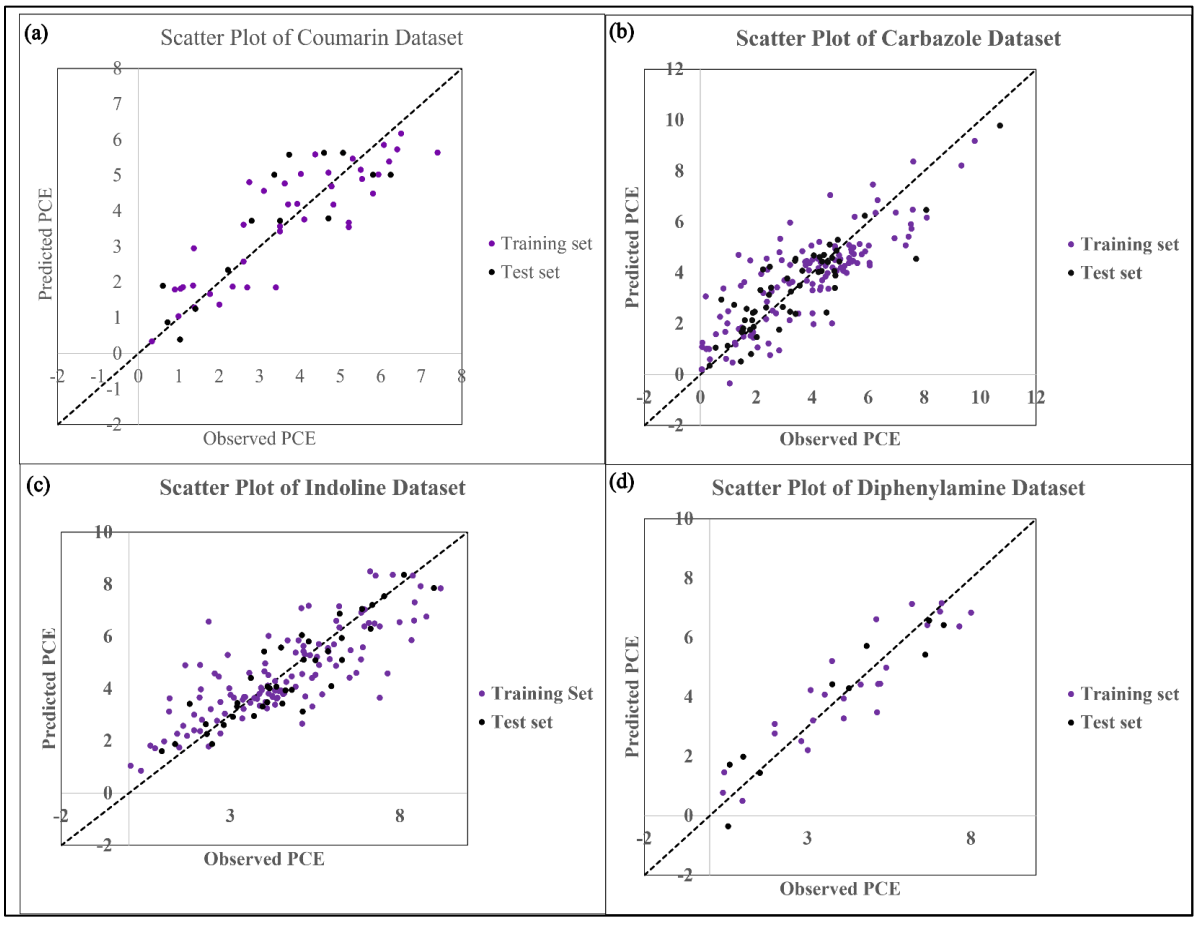


Figure 4.6: Scatter Plots indicating the prediction quality of the developed q-RASPR PLS models

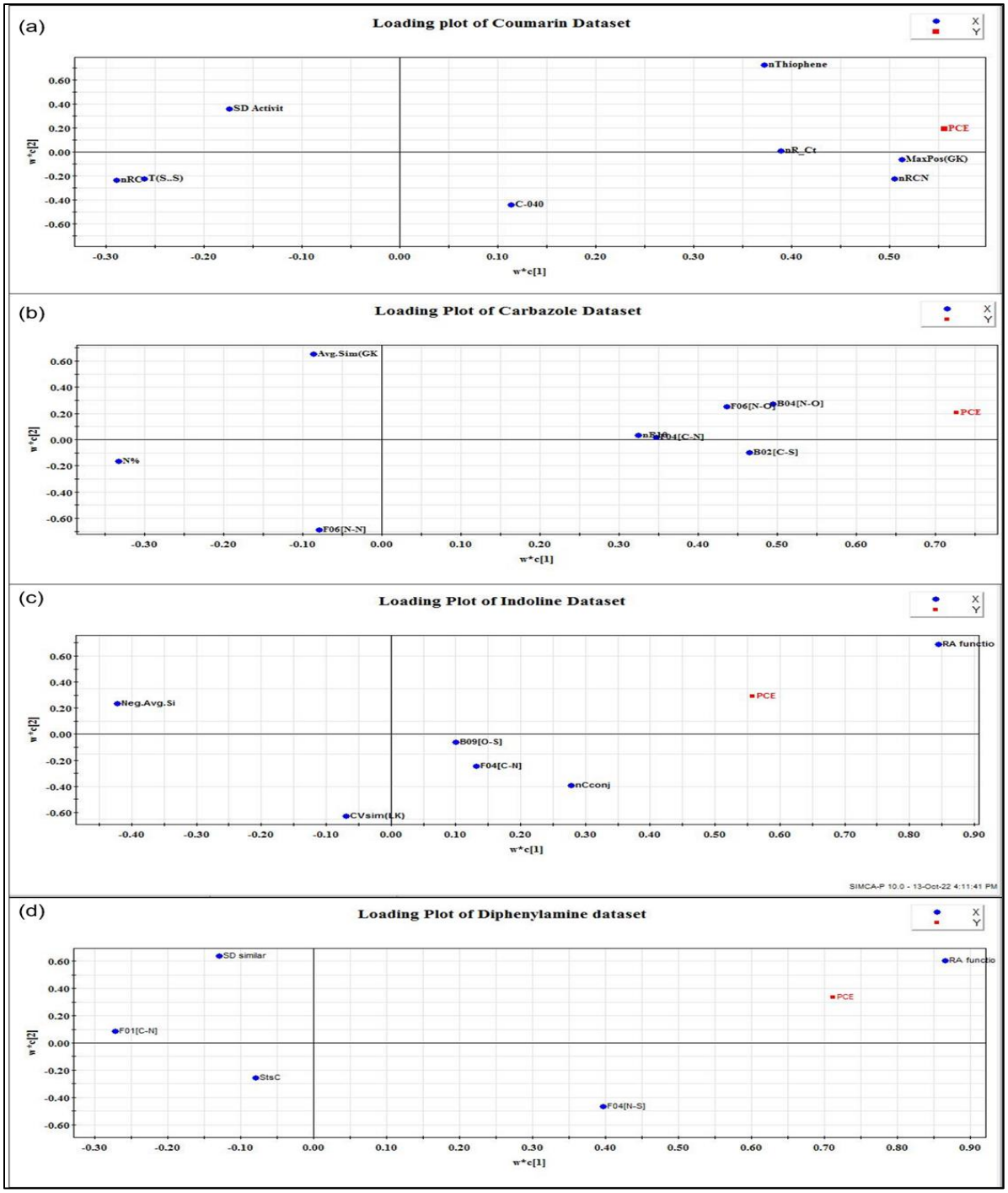


Figure 4.7: Loading Plots showing significance of the descriptors to PCE

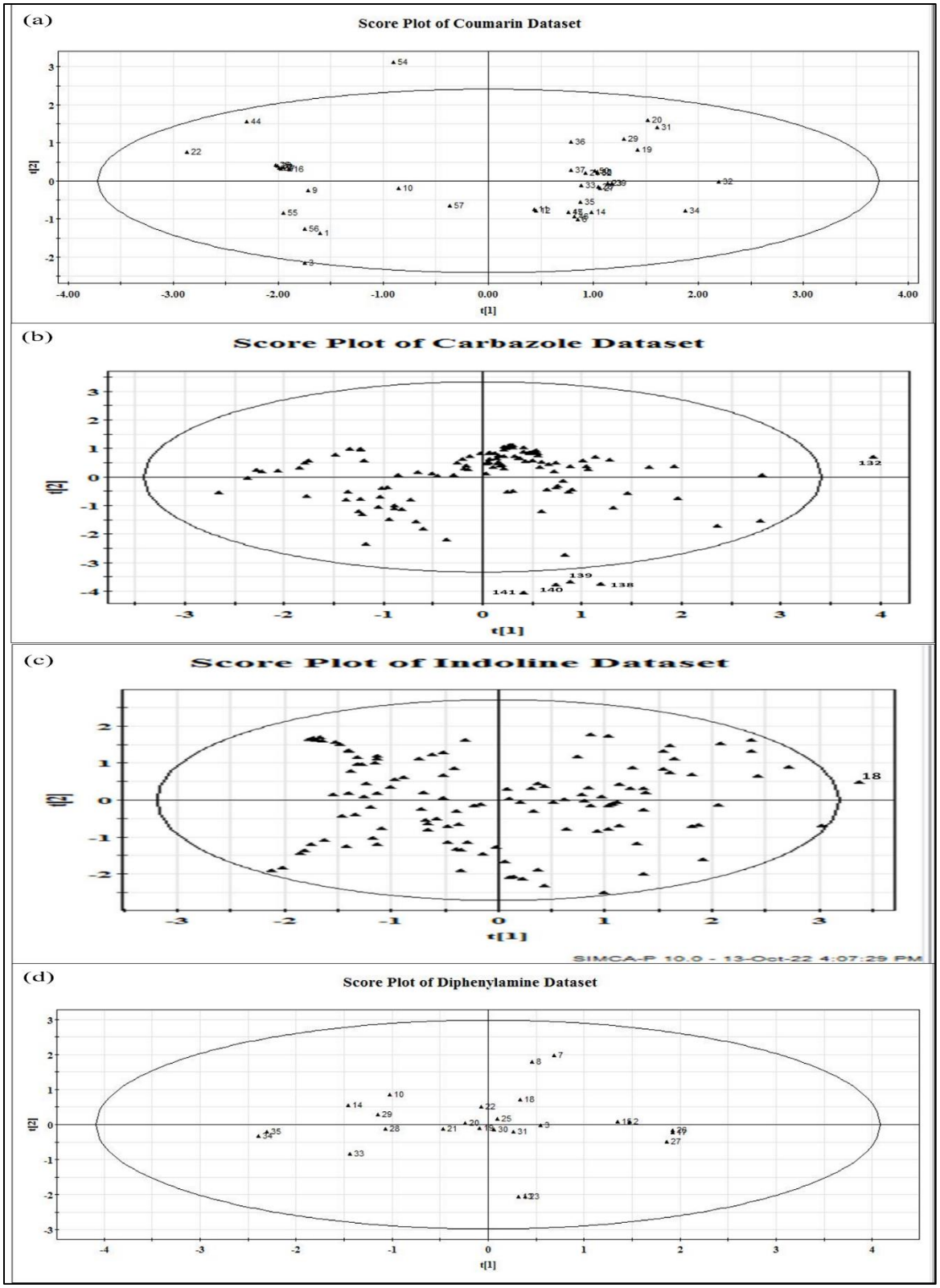


Figure 4.8: Score Plots indicating the applicability domain of PLS models

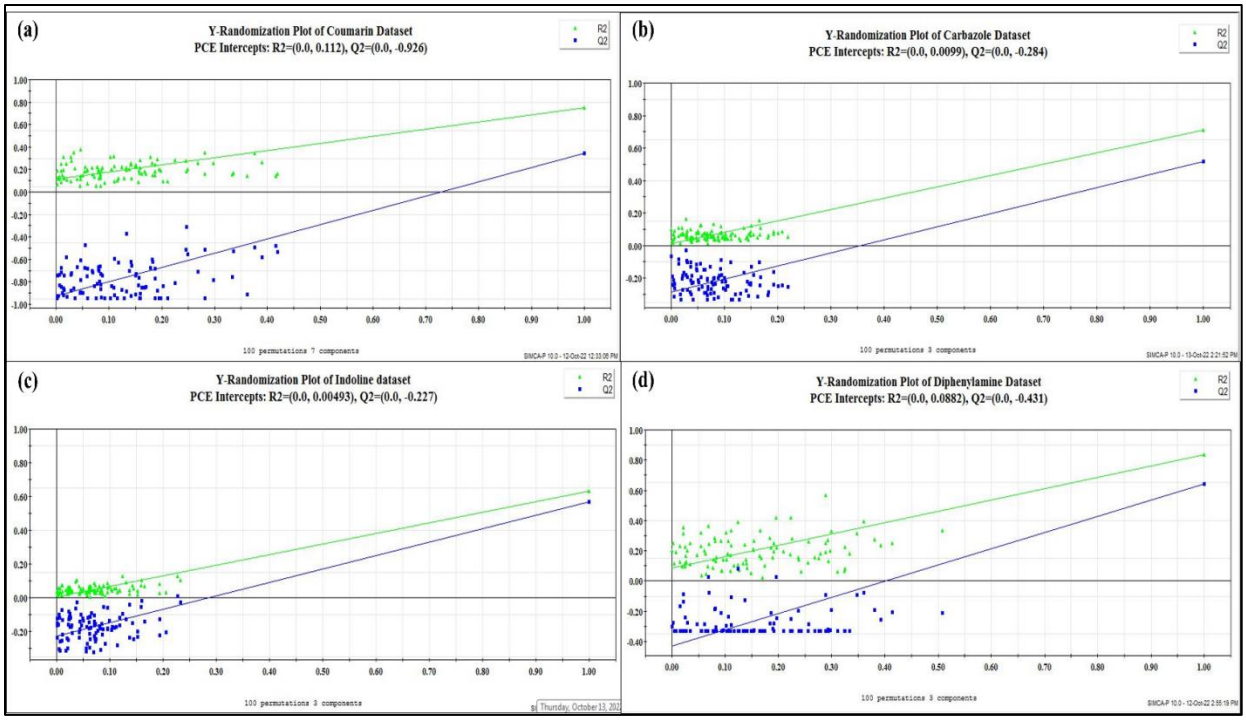


Figure 4.9: PLS randomization plots

4.1.2 Machine-Learning (ML) models

The qualities of the Machine-Learning (ML) models and the different validation metrics for all 4 datasets are shown in **Table 4.2** along with their different optimized hyperparameter settings. To determine the models' quality and predictability, we calculated various quality and validation metrics on the training and test sets. For the purpose of comparison, we have also included PLS q-RASAR models for which the same method was used for calculation of the internal, external and cross-validation metrics. For all 4 datasets, ridge regression, XGBoost and PLS models show almost similar Q^2_{F1}, MAE_{test} and R^2 score (of training set) values.

Table 4.2: Comparison of the quality of PLS and other machine learning models

Datasets	Methods	Training set Metrics						Test set Metrics		Optimized Hyperparameters
		Model Statistics	Cross-Validation Statistics					Prediction Statistics		
		R^2	MAE_{LOO}	MAE ± SEM (20 times 5-fold CV)	r2 ± SEM (20 times 5-fold CV)	MAE ± SEM (1000 times ShuffleSplit CV)	r2 ± SEM (1000 times ShuffleSplit CV)	MAE_{test}	Q^2_{F1}	
Coumarins	PLS	0.75	0.49	0.54±0.015	0.41±0.053	0.56±0.004	0.44±0.011	0.45	0.72	n_components:7
	RR	0.74	0.51	0.54±0.011	0.46±0.036	0.57±0.003	0.48±0.007	0.44	0.73	'alpha': 1.0
	LSVM	0.72	0.52	0.60±0.017	0.27±0.077	0.62±0.005	0.32±0.015	0.49	0.67	'C': 1.0, 'max_iter': 1000
	SVM	0.74	0.66	0.67±0.013	0.17±0.053	0.67±0.004	0.27±0.009	0.5	0.68	'C': 1.0, 'degree': 2, 'gamma': 'auto'

	RF	0.7	0.58	0.61±0.013	0.23±0.062	0.61±0.004	0.35±0.009	0.47	0.68	'max_depth':2, 'min_samples_leaf':2, 'min_samples_split':2, 'n_estimators':200
	GB	0.85	0.57	0.62±0.014	0.18±0.065	0.62±0.003	0.32±0.009	0.55	0.58	'max_depth':2, 'min_samples_leaf':3, 'min_samples_split':3, 'n_estimators':50
	XGB	0.75	0.49	0.54±0.014	0.42±0.050	0.56±0.004	0.44±0.011	0.45	0.72	'booster':'gblinear', 'learning_rate':1.0, 'max_depth': None, 'n_estimators':90
Carbazoles	PLS	0.71	0.47	0.47±0.006	0.62±0.012	0.48±0.002	0.62±0.004	0.31	0.77	n_components:3
	RR	0.71	0.47	0.47±0.006	0.62±0.011	0.48±0.002	0.62±0.003	0.32	0.77	'alpha': 0.5
	LSVM	0.6	0.53	0.51±0.008	0.53±0.016	0.51±0.002	0.54±0.005	0.43	0.64	'C': 5.0, 'max_iter': 100
	SVM	0.84	0.53	0.51±0.009	0.50±0.014	0.52±0.002	0.48±0.004	0.41	0.63	'C': 25.0, 'degree': 2, 'gamma': 'auto'
	RF	0.81	0.56	0.55±0.009	0.44±0.015	0.56±0.002	0.41±0.005	0.45	0.49	'max_depth':6, 'min_samples_leaf':1, 'min_samples_split':4, 'n_estimators':70
	GB	0.82	0.51	0.55±0.009	0.44±0.017	0.56±0.002	0.42±0.005	0.41	0.53	'max_depth':2, 'min_samples_leaf':5, 'min_samples_split':2, 'n_estimators':90
	XGB	0.71	0.47	0.47±0.006	0.62±0.012	0.48±0.002	0.62±0.003	0.32	0.77	'booster': 'gblinear', 'learning_rate':0.1, 'max_depth': None, 'n_estimators':90
Indolines	PLS	0.63	0.48	0.49±0.008	0.55±0.017	0.49±0.002	0.56±0.004	0.3	0.81	n_components:3
	RR	0.63	0.49	0.49±0.007	0.55±0.016	0.49±0.002	0.56±0.004	0.3	0.82	'alpha': 1.0

	LSVM	0.58	0.47	0.51±0.008	0.51±0.021	0.51±0.002	0.53±0.004	0.36	0.73	'C': 5.0, 'max_iter': 100
	SVM	0.71	0.5	0.51±0.008	0.51±0.018	0.52±0.002	0.51±0.004	0.34	0.76	'C': 1.0, 'degree': 2, 'gamma': 'scale'
	RF	0.83	0.45	0.48±0.007	0.54±0.016	0.49±0.002	0.55±0.004	0.42	0.65	'max_depth':5, 'min_samples_leaf':3, 'min_samples_split':2, 'n_estimators':80
	GB	0.81	0.49	0.51±0.007	0.48±0.017	0.51±0.002	0.5±0.004	0.36	0.73	'max_depth':2, 'min_samples_leaf':1, 'min_samples_split':5, 'n_estimators':50
	XGB	0.63	0.48	0.49±0.008	0.55±0.017	0.49±0.002	0.56±0.004	0.3	0.81	'booster': 'gblinear', 'learning_rate':1.0, 'max_depth': None, 'n_estimators':120
Diphenylamines	PLS	0.83	0.44	0.47±0.013	0.39±0.072	0.48±0.005	0.49±0.031	0.31	0.9	n_components:3
	RR	0.83	0.44	0.49±0.012	0.37±0.071	0.5±0.004	0.53±0.013	0.31	0.91	'alpha': 0.5
	LSVM	0.77	0.47	0.51±0.016	0.18±0.133	0.54±0.005	0.38±0.018	0.34	0.87	'C': 15.0, 'max_iter': 100
	SVM	0.87	0.48	0.55±0.019	0.32±0.066	0.57±0.005	0.42±0.014	0.64	0.61	C': 1.0, 'degree': 2, 'gamma': 'auto'
	RF	0.88	0.48	0.51±0.01	0.41±0.059	0.51±0.004	0.55±0.01	0.4	0.8	'max_depth':2, 'min_samples_leaf':1, 'min_samples_split':3, 'n_estimators':120
	GB	1	0.56	0.56±0.014	0.13±0.109	0.54±0.004	0.43±0.013	0.43	0.78	'max_depth':3, 'min_samples_leaf':1, 'min_samples_split':5, 'n_estimators':60
	XGB	0.84	0.44	0.47±0.014	0.37±0.08	0.49±0.005	0.5±0.02	0.3	0.91	'booster': 'gblinear', 'learning_rate':0.1, 'max_depth': None, 'n_estimators':90

To evaluate the quality of developed models, we have also performed 20 times 5-fold repetitive cross-validation, and 1000 times shuffle-split cross-validation with 30% data holding in the validation set. The result of cross-validation for all the datasets are shown in the **Figure 4.10 to 4.13**. For the coumarin dataset, the mean R^2 value for both the repetitive CV and Shuffle-split CV indicates that the Ridge Regression method is the best model among all models while the PLS and XGBoost models show comparable results. For the carbazole and indoline datasets, the ridge regression, PLS and XGBoost models show comparable results, as shown in **Figure 4.11** and **Figure 4.12**. For the diphenylamine dataset, the random forest model shows the highest mean R^2 value in both repetitive cross-validation and shuffle-split cross-validation method but ridge regression, XGBoost and PLS models show comparable results, as shown in **Figure 4.13**.

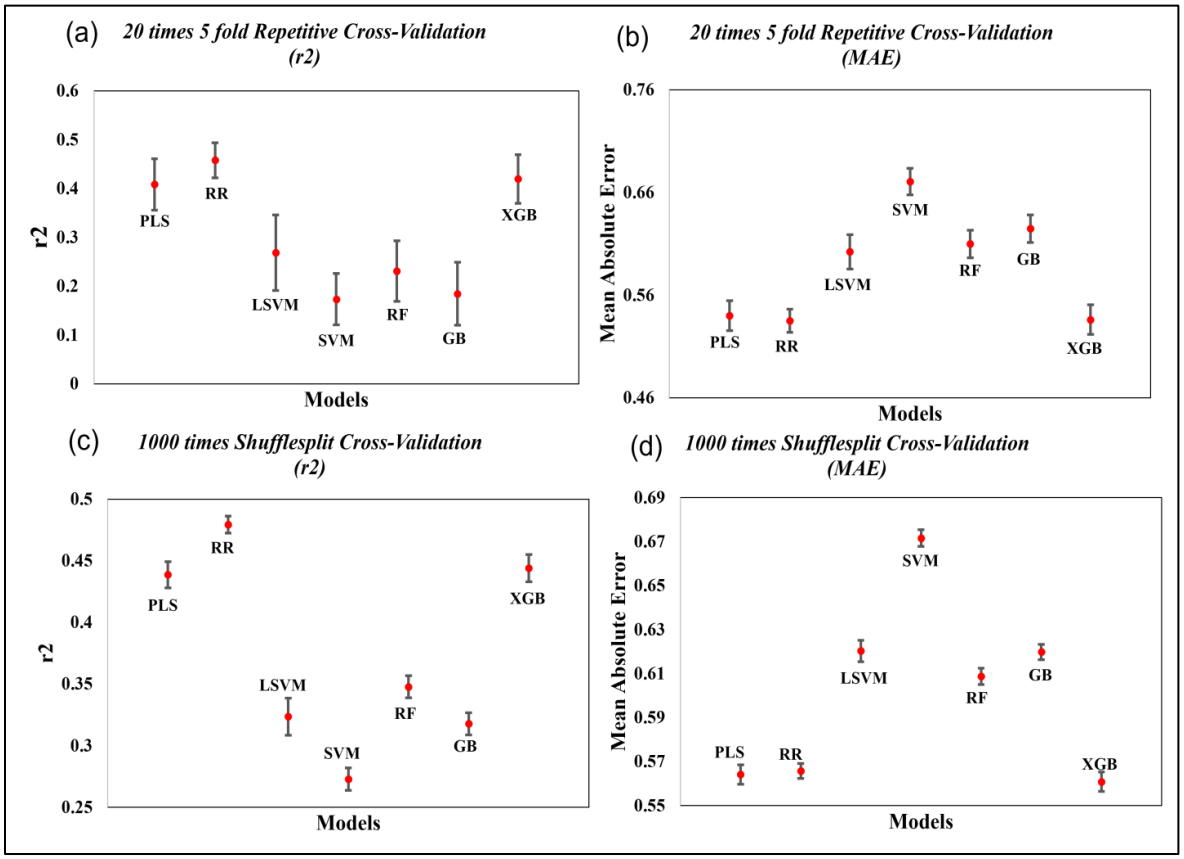


Figure 4.10: Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the coumarin dataset

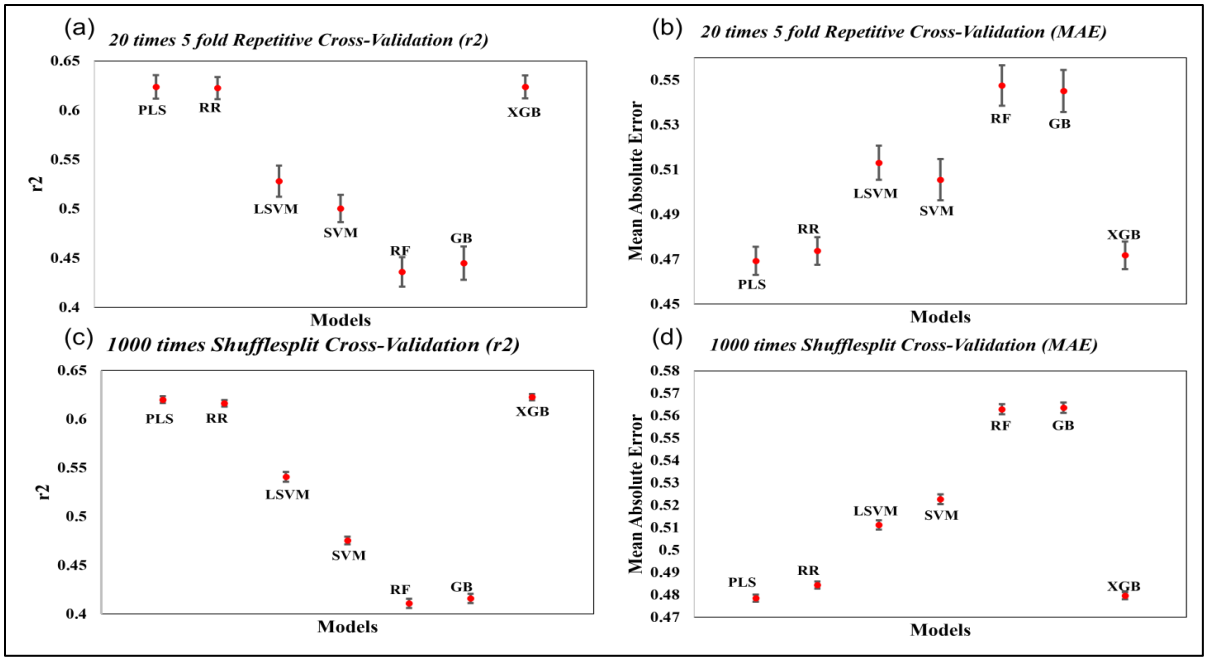


Figure 4.11: Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the carbazole dataset

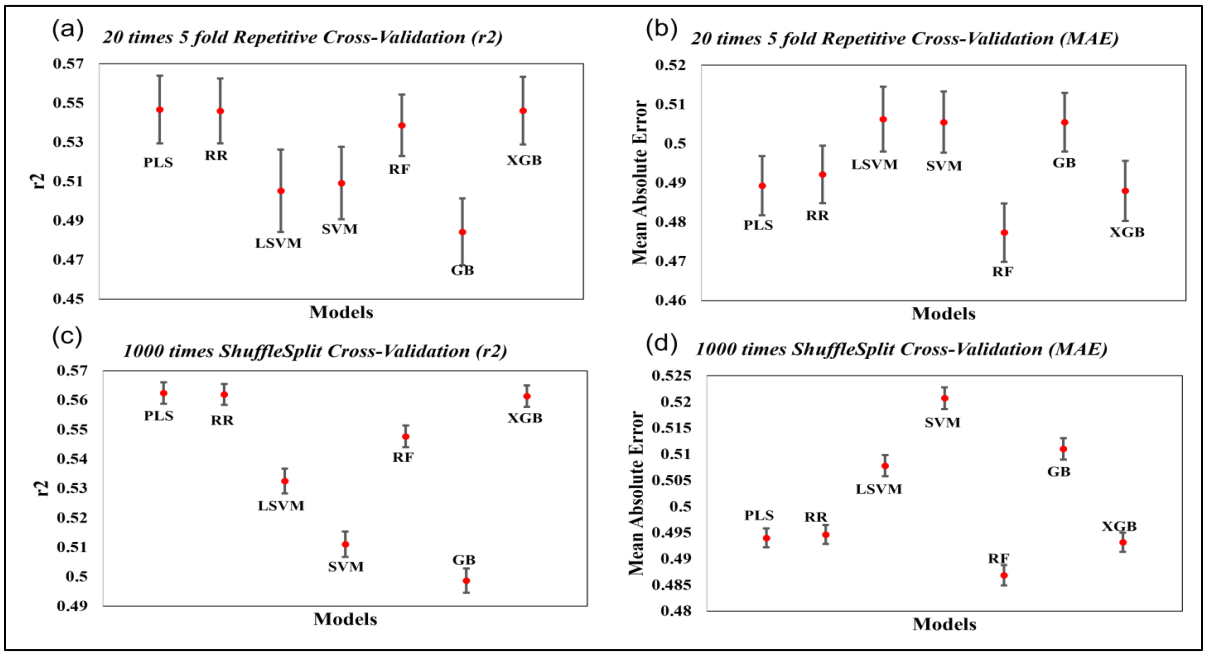


Figure 4.12: Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the indoline dataset

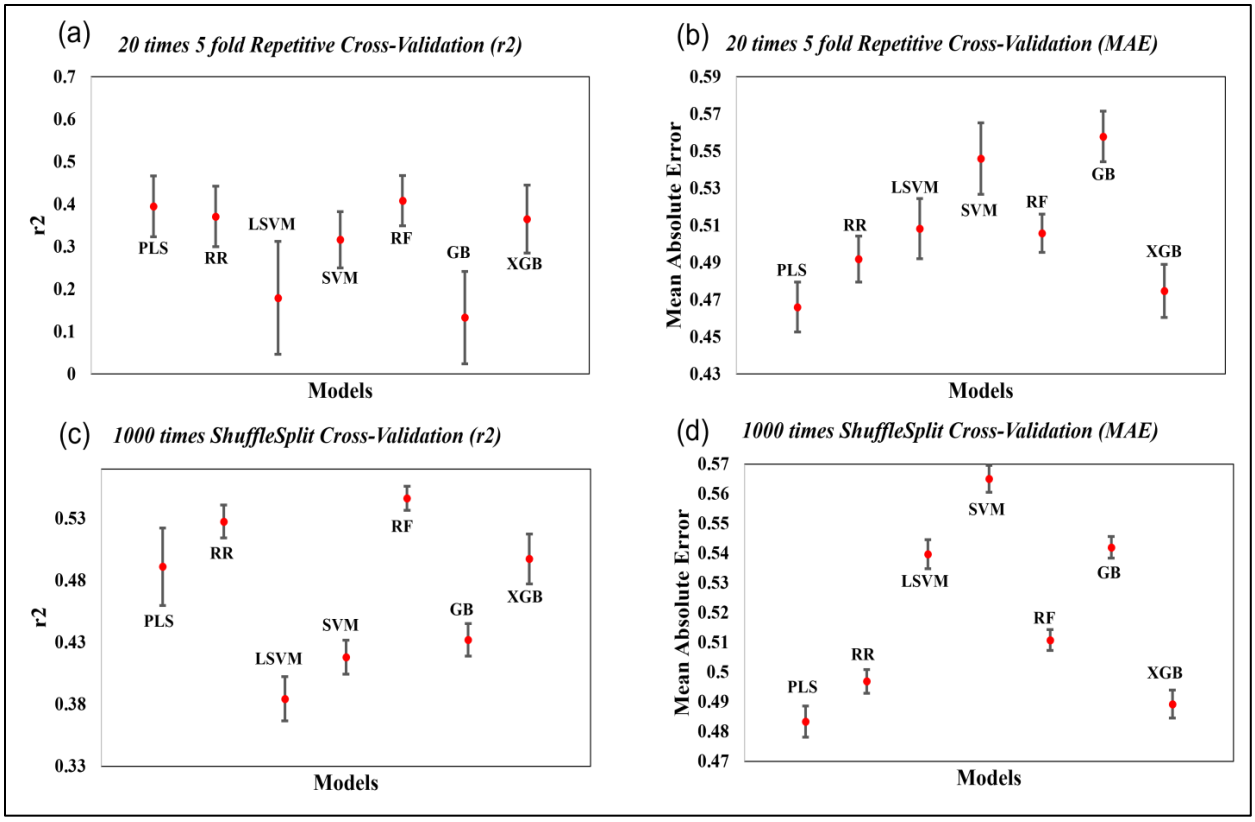


Figure 4.13: Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (Mean ± SEM) for the diphenylamine dataset

To evaluate the importance of descriptors in the machine-learning models, we have performed SHAP analysis on the training set data. We have represented the importance of descriptors in the form of heatmap plots of SHAP as shown in **Figure 4.14**. The PLS, ridge regression and XGBoost models are considered here, and the plots of the remaining models are shown in the **Figure 4.15** to **4.18.** On the Y-axis of the heatmap plot, the features are arranged based on their mean absolute SHAP values, which in turn denotes their importance to the predictions. From the heatmap plot, we can also obtain how the model's prediction changes over every instance which is denoted by the wavy line above the plot. The colour difference in the plot indicates how the SHAP value of the features changes over every instance and how it affects the model's output.

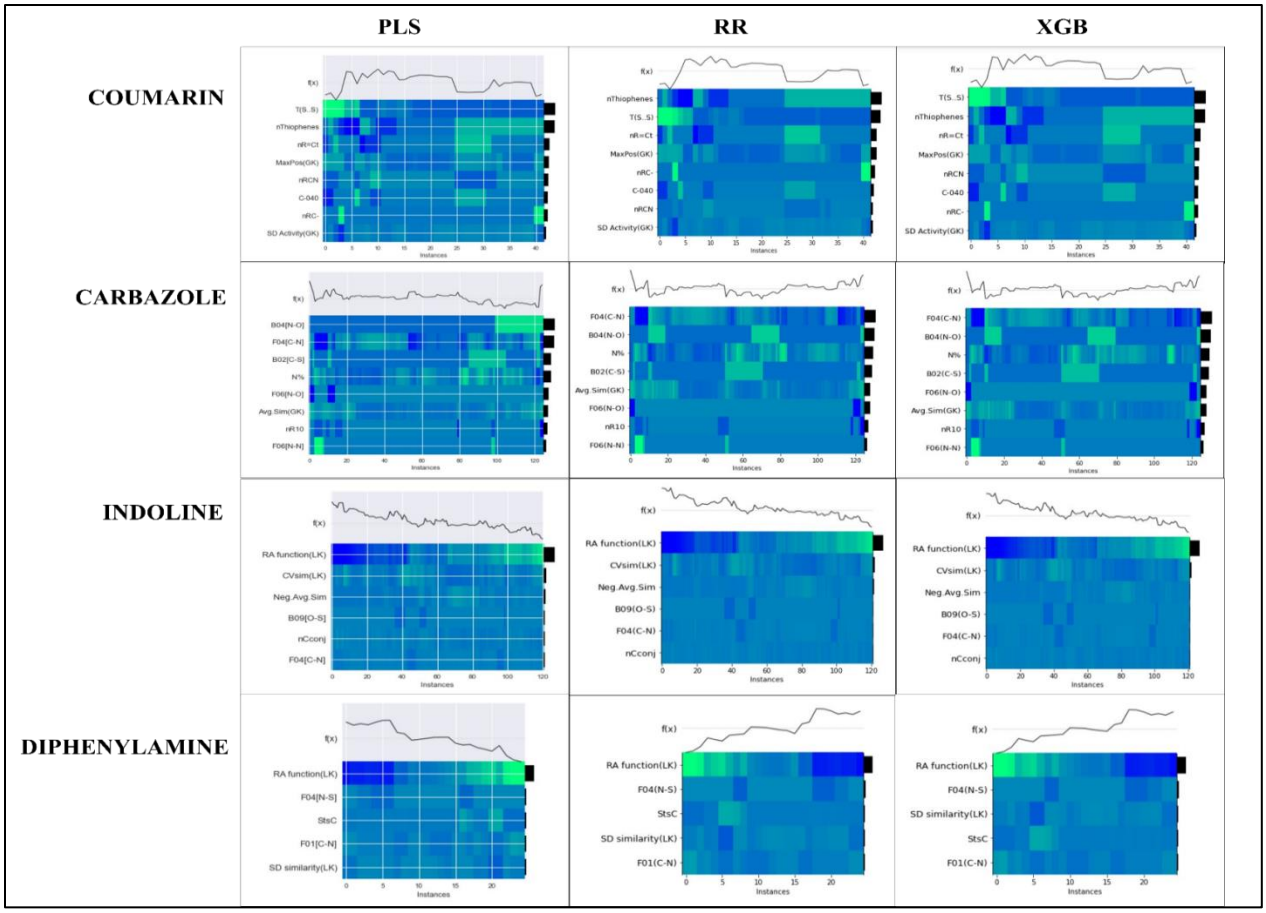


Figure 4.14: Heatmap plots for the PLS, Ridge Regression and XGBoost models for all datasets, indicating relative importance of descriptors

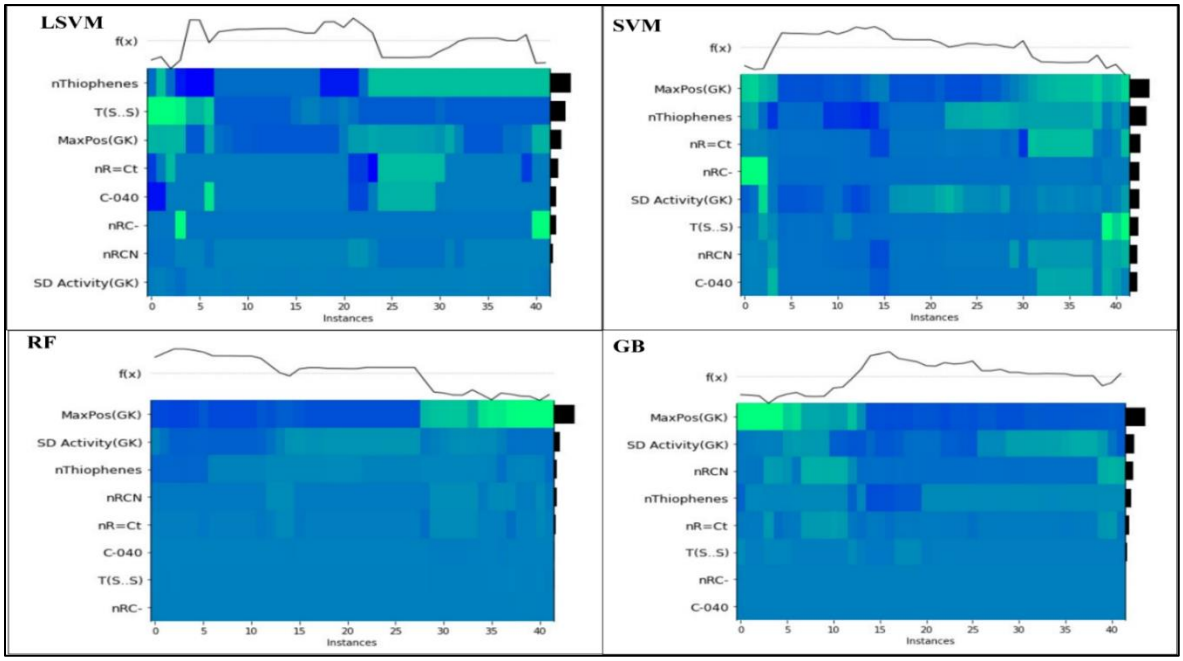


Figure 4.15: Heatmap plots of LSVM, SVM, RF and GB models for the coumarin dataset, indicating the relative importance of descriptors

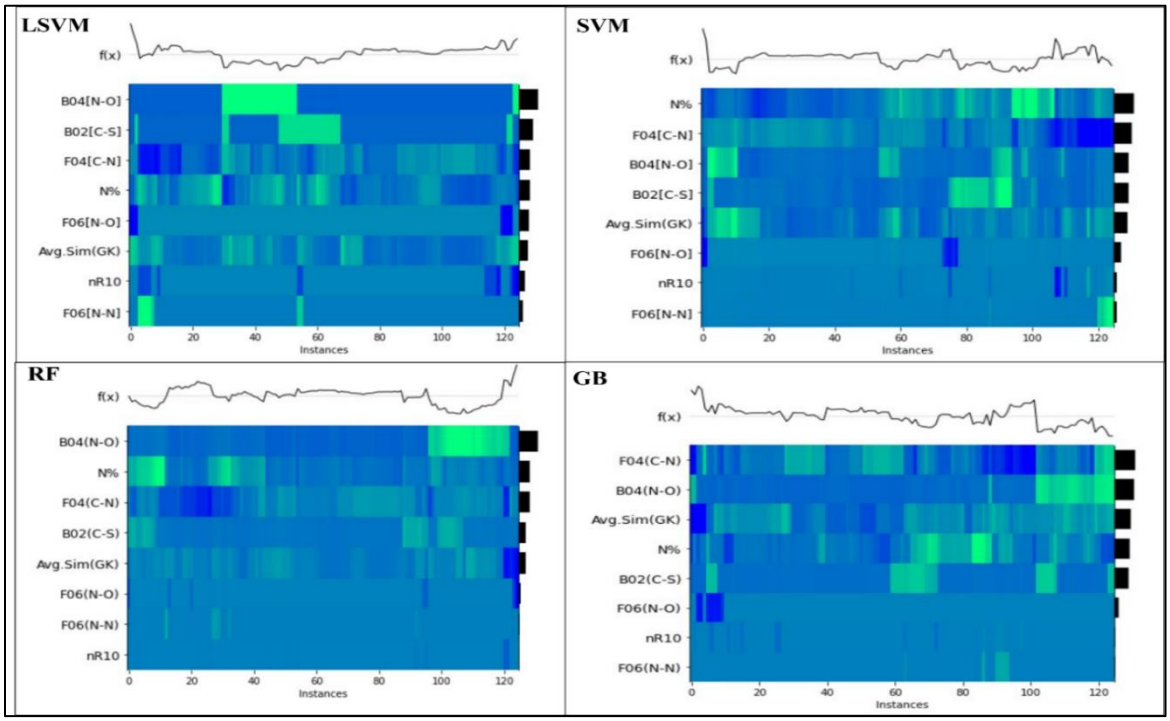


Figure 4.16: Heatmap plots of LSVM, SVM, RF and GB models for the carbazole dataset, indicating the relative importance of descriptors

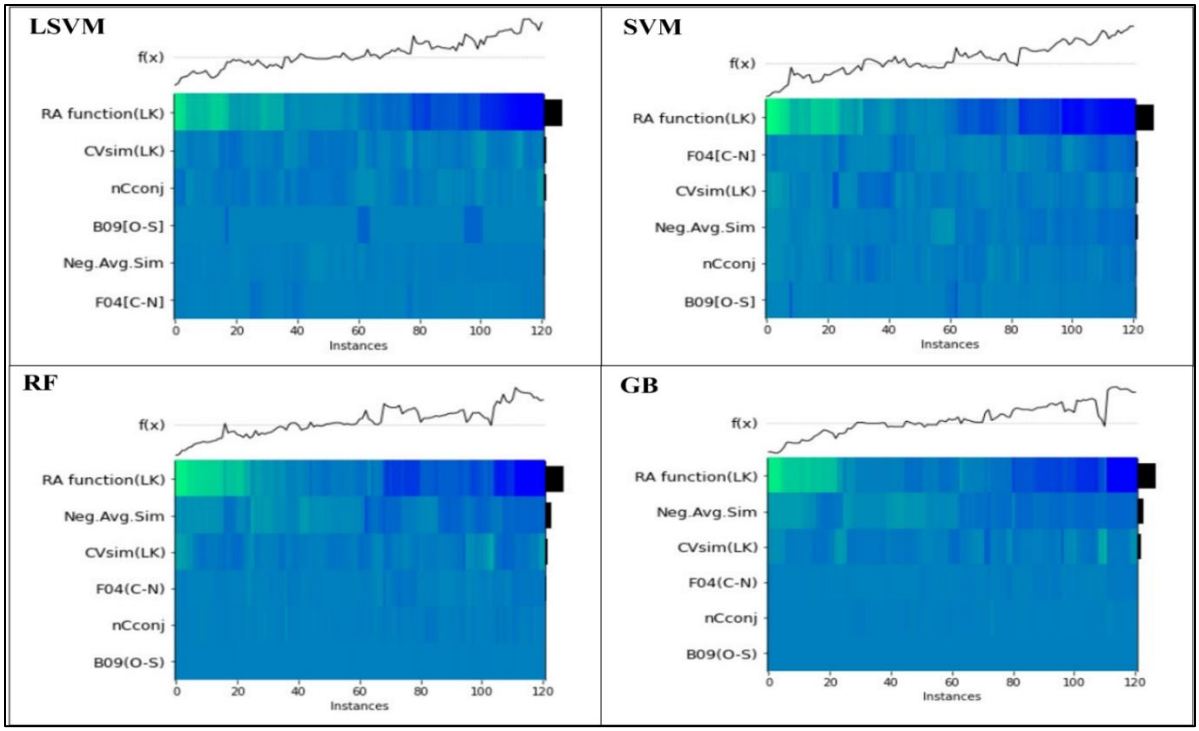


Figure 4.17: Heatmap plots of LSVM, SVM, RF and GB models for the indoline dataset, indicating the relative importance of descriptors

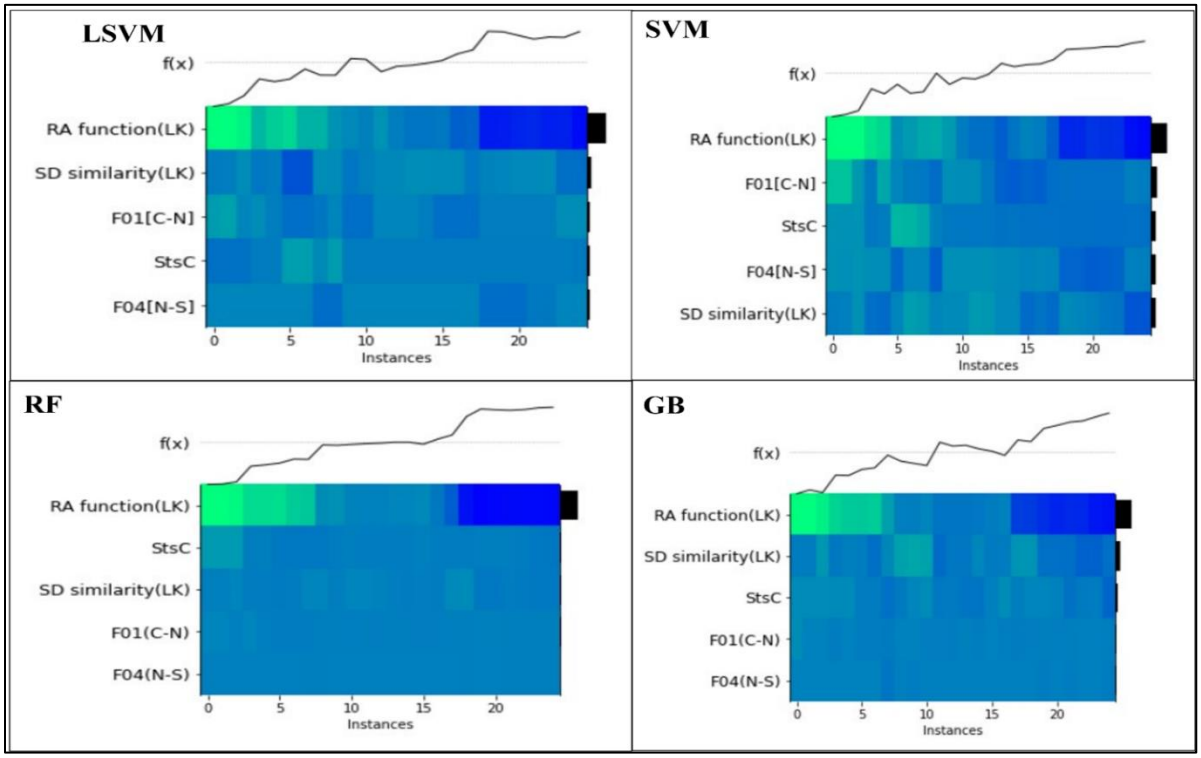


Figure 4.18: Heatmap plots of LSVM, SVM, RF and GB models for the diphenylamine dataset, indicating the relative importance of descriptors

4.1.3 Design of new dyes with improved PCEs

We can design new dyes with improved power conversion efficiency (PCE) by incorporating structural fragments that have positive contributions to the PCE of solar cells or by removing fragments that contribute negatively to the PCE of solar cells. The favorable fragments improve PCE either by increasing intramolecular charge transfer or by increasing the anchoring stability of the dye on the TiO2 semiconductor oxide surface. The descriptors which encode the structural information help to identify the necessary modifications that must be made to improve the PCE. In this study, we use the PLS variable importance plot to identify the relative importance of the positive and negatively contributing descriptors. Based on this, we have designed new dyes with improved predicted PCE values as shown in **Table 4.3**.

For carbazole dyes, the incorporation of a cyano acrylic acid group increases the value of B04[N-O] (presence or absence of nitrogen and oxygen atoms at the topological distance 4) while a para-aminobenzoic acid group increases the value of F06[N-O] (count of nitrogen and oxygen at the topological distance 6). For example, the F06[N-O] value increases when one incorporates 4-(2-cyanoprop-2-enamido)benzoic acid (as shown in NCA1, NCA2 and NCA3) and 2-cyano-N-[4-(trimethoxysilyl)phenyl]prop-2-enamide (as shown in NCA4 and NCA5) moieties in the carbazole structure. These fragments are generally a part of the anchoring group which increases the stability of the binding of the dye with the TiO2 surface. We can increase the value of B02[C-S] (presence or absence of carbon and sulfur atoms at the topological distance of 2) by incorporating a thiophene group, increase the value of F04[C-N] (frequency of carbon and nitrogen atoms at the topological distance of 4) by attaching a long aliphatic chain to the nitrogen atoms and increase the value of nR10 (the number of 10-member rings) by incorporating 10 membered rings in the structure. All these fragments are responsible for the generation of electrons which are transferred to the TiO2 surface.

For coumarin dyes, one can increase the value of positively contributing descriptors like nThiophene (number of thiophene group), nR=Ct (number of aliphatic tertiary C atom with sp2 hybridization) and C-040 (R-C(=X)-X / R-C#X / X=C=X). These descriptors are generally responsible for electron generation and intramolecular charge transfer. One can also try removing the negatively contributing descriptor nRCN (number of aliphatic nitrile groups). The

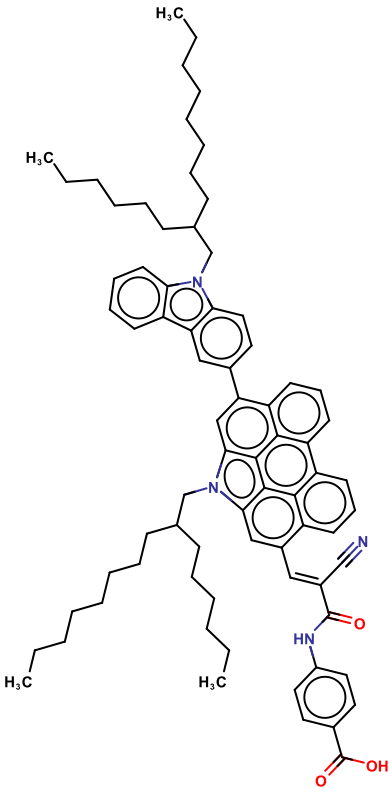
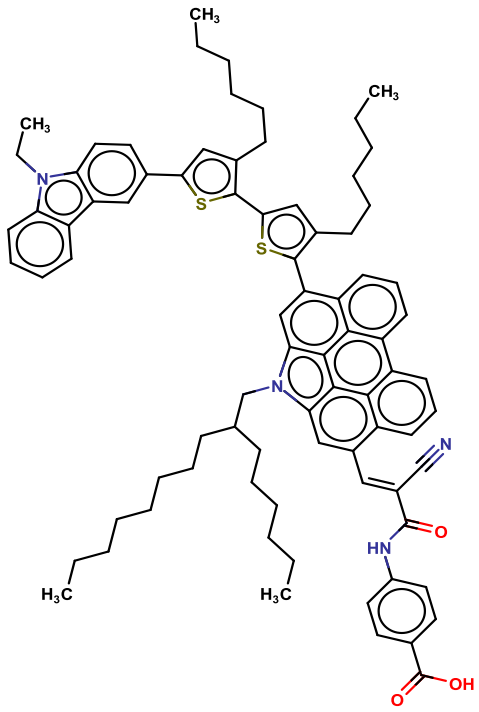
nitrile group is a part of the anchoring group cyanoacrylic acid; therefore, one can try using other anchoring groups like carboxylic acid, pyridine, etc.

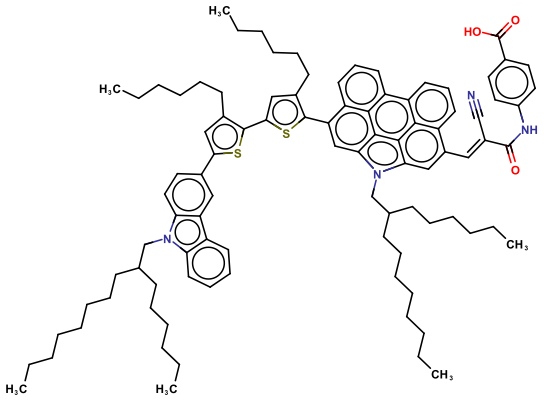
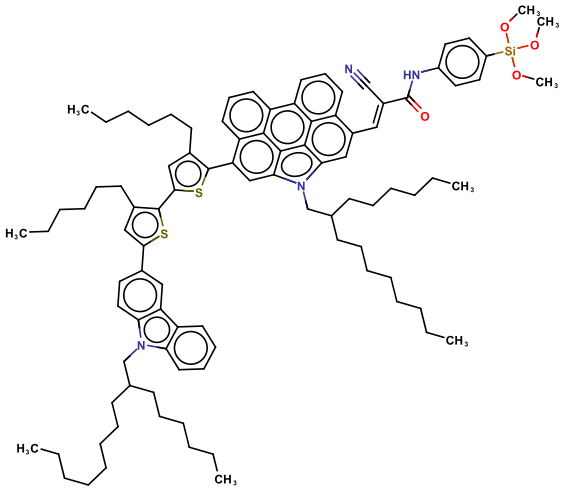
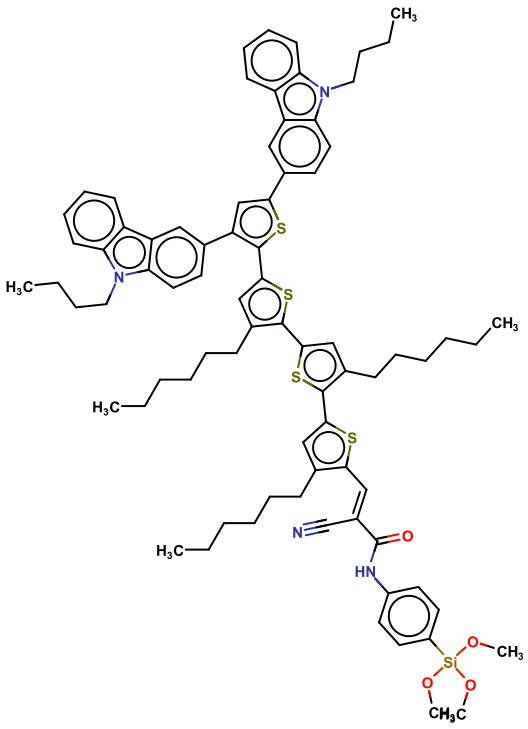
For diphenylamine dyes, one can increase the value of positively contributing descriptor F04[N-S] (frequency of nitrogen and sulfur atoms at the topological distance 4) by incorporating groups like a pyrimidine ring adjacent to a thiophene ring (as shown below in NDI1), 2,1,3-benzothiadiazole and 1,2,3-benzodithiazole groups adjacent to the thiophene and pyrimidine rings respectively (as shown below in NDI5 and NDI3). These fragments are generally a part of the linker between the donor part and the acceptor part which helps to improve performance by increasing intramolecular charge transfer.

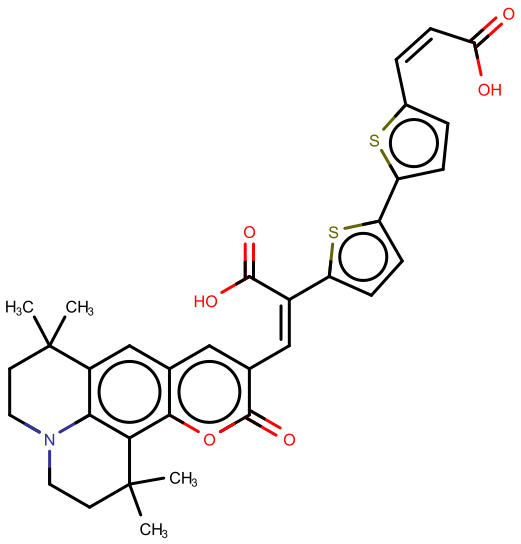
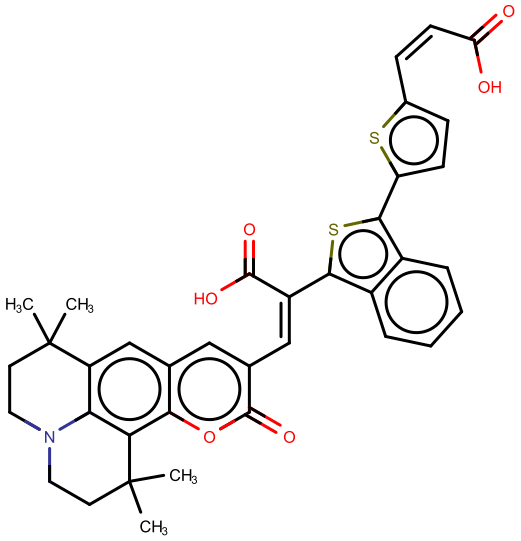
For indoline dyes, one can increase the value of positively contributing descriptors like F04[C-N] (frequency of carbon and nitrogen atoms at the topological distance of 4) by attaching different aliphatic and aromatic groups to the nitrogen atoms, and B09[O-S] (presence of oxygen and sulfur atoms at the topological distance 9) by increasing the length of the π-spacer (for example, compound NIN1 is formed by incorporating a butylene group between the thiophene ring and the cyanoacrylic acid). These fragments help in the generation of electrons and improve intramolecular charge transfer.

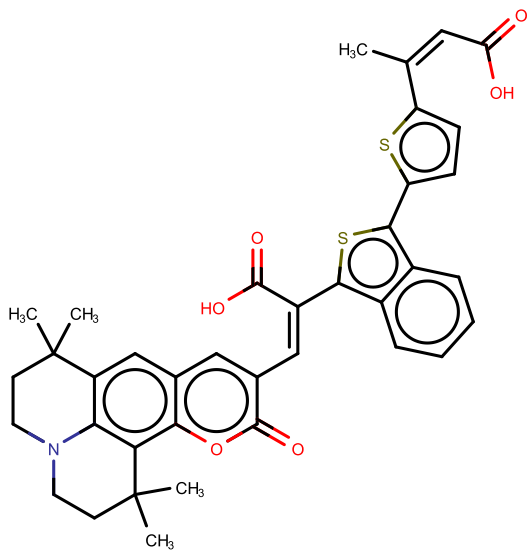
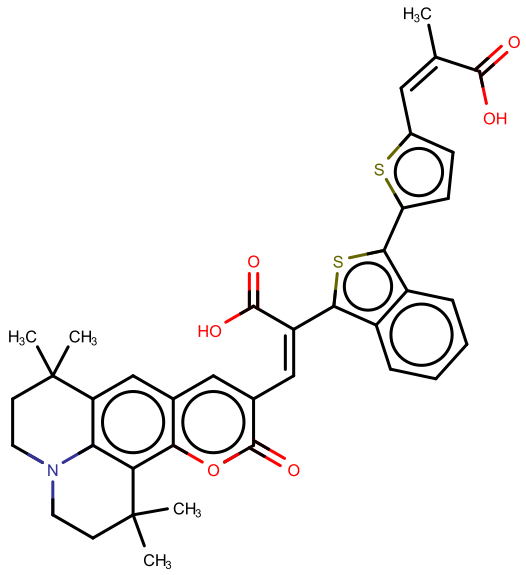
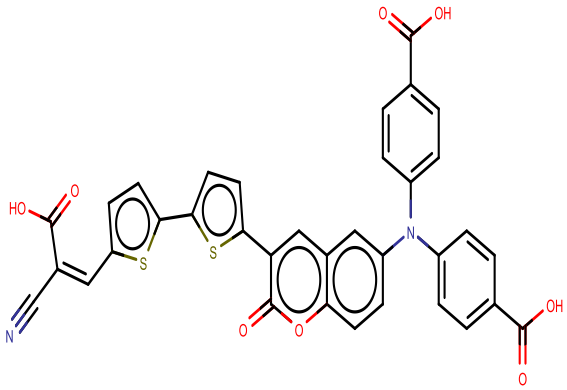
Table 4.3: List of designed dyes and their predicted PCE values

Compounds	**Structure**	**Predicted PCE Values**		
		PLS	**Ridge Regression**	**XGBoost**
Carbazoles				

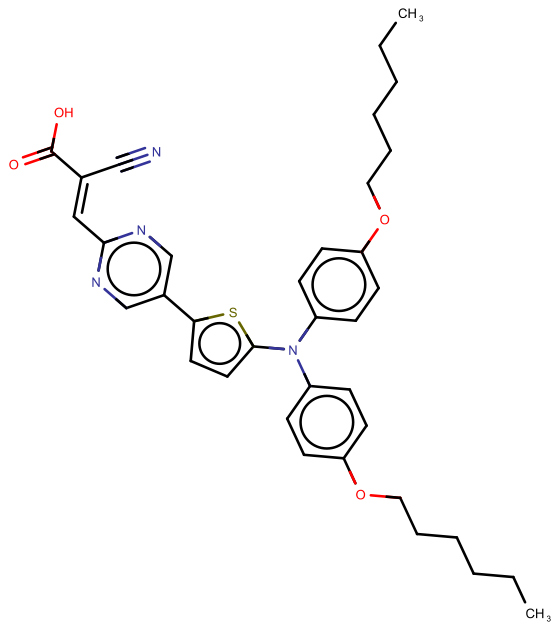
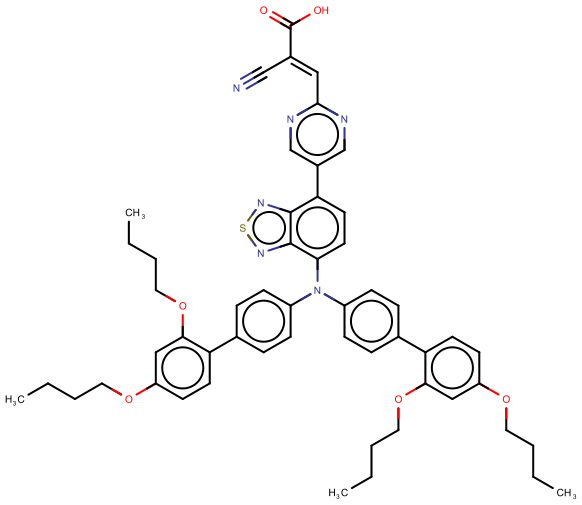
NCA1		**14.15**	**14.02**	**13.97**
NCA2		**15.22**	**15.09**	**15.06**

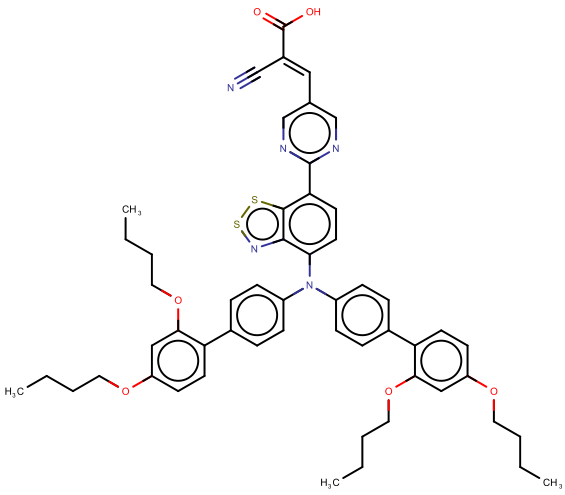
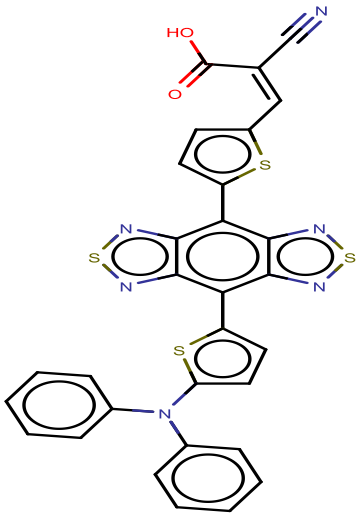
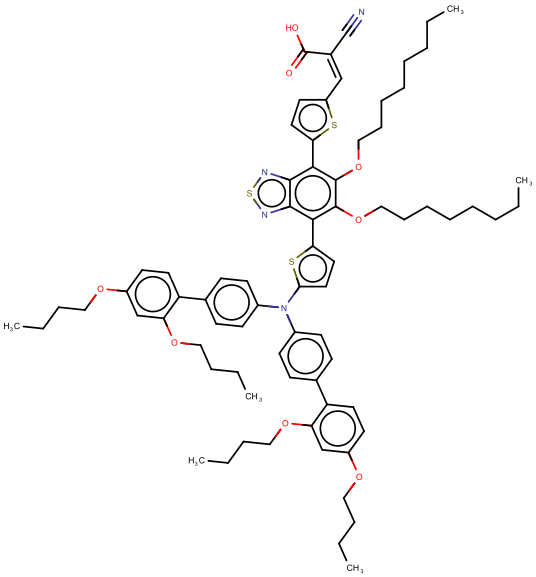
NCA3		**15.79**	**15.73**	**15.69**
NCA4		**18.18**	**18.33**	**18.24**
NCA5		**11.47**	**11.89**	**11.81**

Coumarins				
NCO1		**5.67**	**5.1**	**5.67**
NCO2		**5.67**	**5.1**	**5.67**

NCO3		**6.74**	**6.03**	**6.74**
NCO4		**6.74**	**6.03**	**6.74**
NCO5		**6.6**	**5.94**	**6.6**

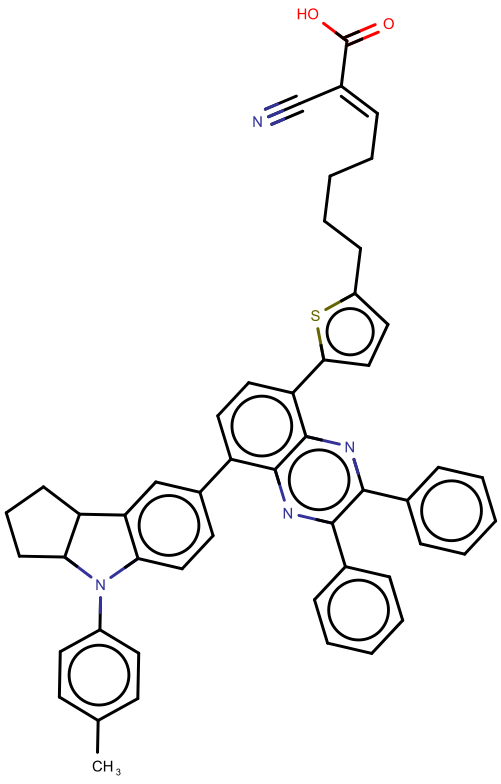
Diphenylamines

NDI1		**7.06**	**7.05**	**7.1**
NDI2		**6.41**	**6.33**	**6.4**

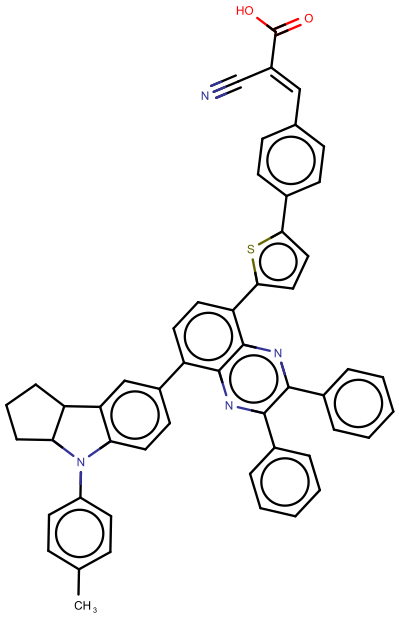
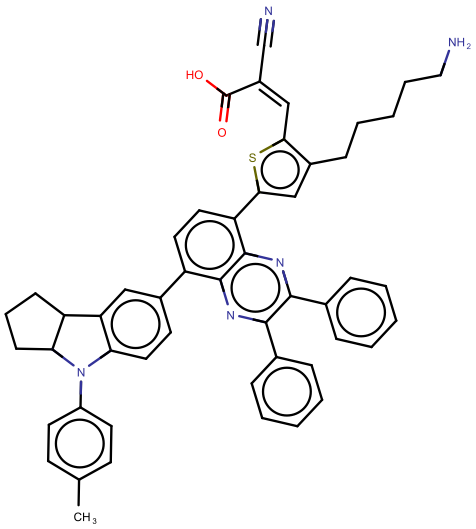
NDI3		**7.62**	**7.75**	**7.77**
NDI4		**7.07**	**7.22**	**7.22**
NDI5		**7.16**	**7.16**	**7.19**

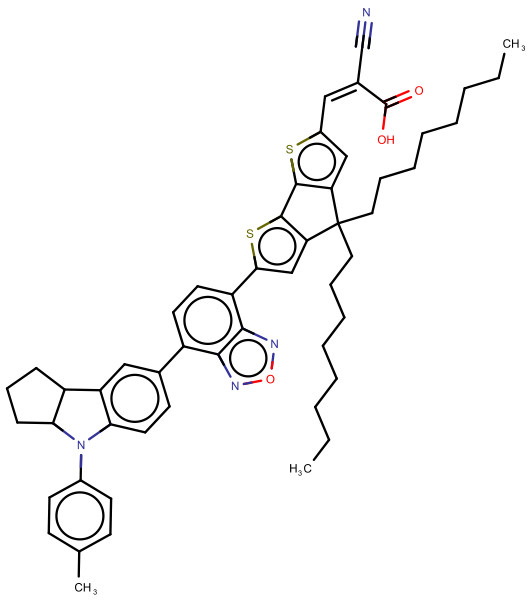
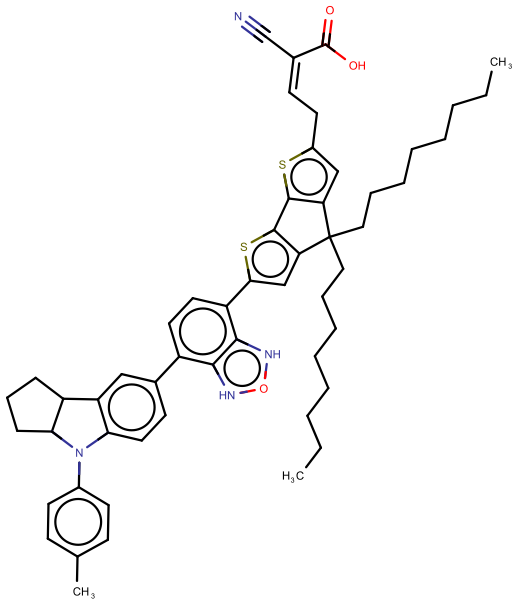
Indolines

NIN1



6.39 **6.34** **6.36**

NIN2		**6.39**	**6.34**	**6.36**
NIN3		**8.22**	**8.15**	**8.19**

NIN4		**8.29**	**8.2**	**8.23**
NIN5		**8.49**	**8.41**	**8.45**

4.1.4 Comparison with the previous work

In 2020, Krishna et al. worked on the prediction of PCE value of metal free organic DSSCs by PLS regression using 2D structural descriptors [131]. The models reported by them were statistically sound with good quality validation metrics. However, our q-RASPR PLS models outperformed them with better prediction quality with the same level of chemical information used, also explaining the importance of RASPR descriptor. The advantage of RASPR descriptors

is that they are simple, easy to calculate and transferable. In comparison to the previous models, the present models have higher Q^2_{F1} scores and lower MAE_{test} values. The comparison between the previous PLS models and the present q-RASPR PLS models are shown in the form of column plots in **Figure 4.19**. The previous study developed five individual models and consensus models of datasets, but here we have considered only the best individual models for comparison. It is clear from the column plots that our newly developed models are much more predictive. This study shows that inclusion of RASPR descriptors along with structural and physicochemical descriptors enhance the predictivity of the models.

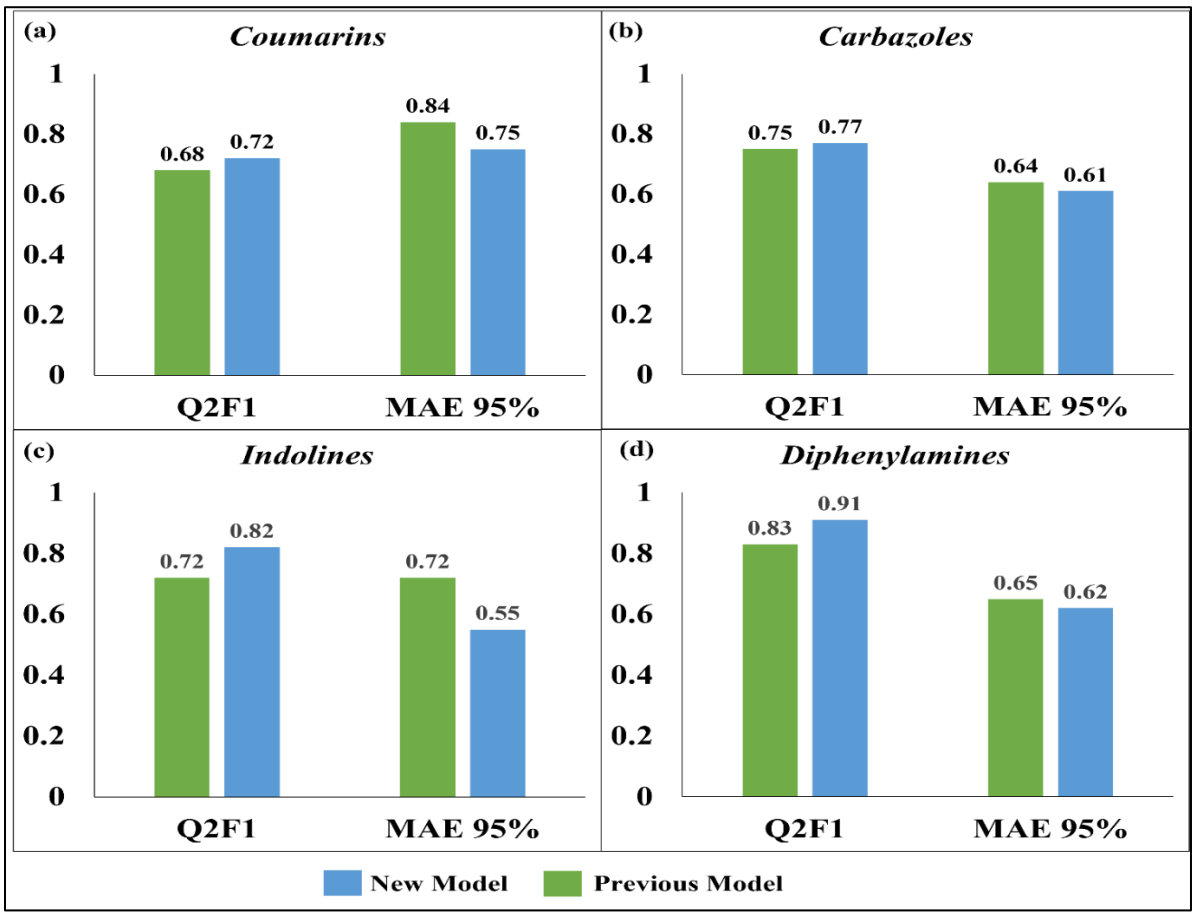


Figure 4.19: Comparison between the previous QSPR PLS models and the newly developed q-RASPR PLS models

4.2 Study 2: Application of machine learning-based read-across structure-property relationship (RASPR) as a new tool for predictive modelling: prediction of power conversion efficiency (PCE) for selected classes of organic dyes in dye-sensitized solar cells (DSSCs)

In this study, we have first d eveloped QSPR and q-RASPR PLS models for all the data sets with three different training and test sets, and then the derived nine models were used for external test set prediction. We further modeled these datasets through other machine learning methods with the same set of descriptors from PLS models and performed SHAP analysis to determine importance of a descriptor to the PCE.

4.2.1 QSPR and q-RASPR PLS models:

The results of statistically significant and robust QSPR and q-RASPR PLS models of all three divisions for the respective datasets, in terms of various internal and external validation metrics, are shown in **Table 4.4**. All these models were developed with 10 descriptors and the components (LV) of the PLS models were selected based on the optimum Q^2_{LOO} value. All the models show acceptable value for internal and external validation metrics to become reliable, robust, and good predictive models. Here we can see that for all the datasets, the external validation metrics were significantly improved for q-RASPR PLS models compared with QSPR PLS models which signify the importance of the RASPR descriptors.

Table 4.4: Results of QSPR and RASPR PLS models

Dataset	Division	Method (PLS Model)	LVs	Training set Metrics						Test set Metrics				
				R^2	Q^2_{LOO}	MAE_{train}	MAE_{LOO}	MSE_{LOO}	RMSEC	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	RMSEP	MAE_{test}
Phenothiazine	1	QSPR	5	0.60	0.55	0.96	1.02	1.61	1.20	0.57	0.57	0.57	1.25	0.98
		RASPR LK	7	0.61	0.56	0.93	0.99	1.59	1.18	0.62	0.62	0.61	1.18	0.96
	2	QSPR	5	0.59	0.54	0.99	1.06	1.67	1.21	0.64	0.64	0.63	1.16	0.95
		RASPR	5	0.59	0.53	0.97	1.04	1.67	1.21	0.67	0.67	0.65	1.12	0.97

		GK												
	3	**QSPR**	4	0.59	0.53	0.99	1.07	1.72	1.23	0.60	0.60	0.62	1.18	0.91
		RASPR GK	4	0.59	0.54	1.00	1.07	1.70	1.23	0.67	0.66	0.69	1.07	0.81
Porphyrin	1	**QSPR**	3	0.56	0.52	1.37	1.44	3.41	1.77	0.50	0.50	0.55	1.79	1.48
		RASPR LK	5	0.57	0.52	1.33	1.40	3.39	1.75	0.64	0.64	0.68	1.51	1.25
	2	**QSPR**	3	0.54	0.50	1.39	1.46	3.47	1.78	0.61	0.61	0.62	1.63	1.21
		RASPR LK	3	0.57	0.53	1.35	1.41	3.26	1.73	0.66	0.66	0.67	1.53	1.14
	3	**QSPR**	3	0.55	0.51	1.39	1.45	3.43	1.77	0.62	0.62	0.64	1.60	1.27
		RASPR LK	3	0.56	0.51	1.37	1.43	3.40	1.76	0.69	0.69	0.70	1.45	1.09
Triphenylamine	1	**QSPR**	4	0.55	0.50	1.25	1.33	2.60	1.53	0.53	0.53	0.54	1.54	1.24
		RASPR GK	2	0.55	0.50	1.26	1.33	2.55	1.53	0.57	0.57	0.58	1.47	1.23
	2	**QSPR**	2	0.53	0.48	1.30	1.37	2.88	1.61	0.52	0.52	0.71	1.27	1.05
		RASPR LK	4	0.55	0.50	1.27	1.35	2.78	1.58	0.57	0.57	0.74	1.20	1.05
	3	**QSPR**	4	0.57	0.51	1.23	1.30	2.61	1.52	0.49	0.47	0.59	1.48	1.26
		RASPR GK	3	0.60	0.55	1.18	1.25	2.44	1.47	0.62	0.61	0.70	1.27	1.04

LV= the number of latent variables

4.2.1.1 Modeling of PCE of phenothiazine dyes

The descriptors used for the modeling of the PCE of phenothiazine dyes are shown in the **Box-1**, in the form of mathematical equations. The initial pool of physicochemical and structural

features used for the calculation of RASPR descriptors and present in the q-RASPR PLS models of the respective divisions are shown in the **Table 4.5**. The description of the modeled RASPR descriptors of the q-RASPR PLS models are shown in **Table 4.6** along with their influence on the PCE. The contribution of the descriptors of the q-RASPR PLS models in the form of bubble plot and correlation between observed and predicted PCE in the form of a scatter plot is shown in **Figure 4.20**. The importance of the descriptors is represented by size and position of the bubble along the x axis, and the color represents the direction of their contributions, i.e., the green color represents a positive contribution and the red color represent a negative contribution. The PLS plots for all three divisions are shown in **Figure 4.21 to 4.24.**

Box-1

Phenothiazine dataset

QSPR Models

M1 (Division 1) $PCE = -11.071 + 25.373 \times X0Av + 0.307 \times nR\#C - -3.274 \times nArCOOH - 0.686 \times SsssCH + 0.394 \times MaxsCH3 - 0.634 \times MaxdssC - 4.182 \times B02[N-S] + 1.851 \times B07[N-S] - 1.187 \times B08[N-O] - 1.304 \times B09[S-S]$

M2 (Division 2) $PCE = -19.785 + 39.822 \times X0Av + 0.432 \times nR\#C - -3.356 \times nArCOOH - 0.084 \times NssCH2 - 0.613 \times MaxdssC - 2.188 \times B04[O-S] + 1.138 \times B07[N-S] + 0.351 \times F01[C-O] - 1.079 \times F03[N-O] - 0.809 \times F08[N-F]$

M3 (Division 3) $PCE = -10.396 + 23.485 \times X0Av - 3.222 \times nArCOOH - 1.259 \times C-033 + 0.570 \times MaxsCH3 - 0.526 \times MaxdssC - 4.575 \times B02[N-S] - 0.865 \times B06[N-N] + 1.009 \times B07[N-S] - 1.100 \times F08[N-F] - 1.229 \times F09[S-S]$

RASPR Models

M4 (Division 1) $PCE = -6.702 + 0.675 \times SD\ Activity\ (LK) - 2.790 \times CVact(LK) - 0.392 \times CVsim(LK) + 0.460 \times g_m(LK) + 19.482 \times X0Av + 0.381 \times nR\#C - -2.701 \times B02[N-S] + 1.639 \times B07[N-S] - 1.113 \times B08[N-O] - 0.801 \times B09[S-S]$

M5 (Division 2) $PCE = -16.599 + 4.131 \times g_m * SD\ Similarity + 34.722 \times X0Av + 0.366 \times nR\#C - -3.041 \times nArCOOH - 0.089 \times NssCH2 - 1.460 \times B04[O-S] + 1.047 \times B07[N-$

	$S] + 0.322 \times F01[C-O] - 0.931 \times F08[N-O] - 0.945 \times F03[N-F]$
M6 (Division 3)	$PCE = -6.601 + 1.511 \times SD\ Activity\ (GK) - 1.189 \times SE(GK) - 0.024 \times g_m(GK) + 3.937 \times g_m * SD\ Similarity + 16.219 \times X0Av - 2.120 \times nArCOOH - 1.555 \times C - 033 - 0.341 \times MaxdssC - 3.309 \times B02[N-S] + 1.409 \times B07[N-S]$

Table 4.5: The description of the modelled structural and physicochemical descriptors of QSPR and q-RASPR PLS models of phenothiazine dyes.

Features	Meaning	Model
X0Av	It is a connectivity indices descriptor represent the average valence connectivity of order 0.	M1, M2, M3, M4, M5, M6
nR#C-	It is functional group count descriptor which indicate the number of non-terminal sp hybridized carbon atoms present in a chemical structure.	M1, M2, M4, M5
nArCOOH	It is functional group count descriptor that represent the number of aromatic carboxylic acid present in the molecular structure.	M1, M2, M3, M5, M6
SsssCH	It is an atom type E-state indices descriptor that represent the sum of sssCH (single, single and single bond) E-states.	M1
MaxsCH3	It is an atom type E-state indices descriptor that represent Maximum single bond CH3 E-state.	M1, M3
MaxdssC	It is an atom type E-state indices descriptor that represent Maximum double bond, single bond and single bond carbon atom (=<) E-state.	M1, M2, M3, M6
B02[N-S]	It is a 2D atom pair descriptor that represent the presence or absence of nitrogen and sulfur atoms at a topological distance of 2.	M1, M3, M4, M6
B07[N-S]	It is a 2D atom pair descriptor that represent the presence or absence of nitrogen and sulfur atoms at a topological distance of 7.	M1, M2, M3, M4, M5, M6
B08[N-O]	It is a 2D atom pair descriptor that represent the presence or absence of nitrogen and oxygen atoms at a topological distance of 8.	M1, M4
B09[S-S]	It is a 2D atom pair descriptor that represent the presence or absence of two sulfur atoms at a topological distance of 9.	M1, M4

NssCH2	It is an atom type E-state indices that represent the number of atoms of single bond, single bond CH2 (-CH2-).	M2, M5
B04[O-S]	It is a 2D atom pair descriptor that represent the presence or absence of oxygen and sulfur atoms at a topological distance of 4.	M2, M5
F01[C-O]	It is a 2D atom pair descriptor that represent the frequency of carbon and oxygen atoms at a topological distance of 1.	M2, M5
F03[N-O]	It is a 2D atom pair descriptor that represent the frequency of nitrogen and oxygen atoms at a topological distance of 3.	M2, M5
F08[N-F]	It is a 2D atom pair descriptor that represent the frequency of nitrogen and fluorine atoms at a topological distance of 8.	M2, M3, M5
C-033	It is an atom-centered fragment descriptor that represent the fragment R-CH..X (R: any group attached to the carbon atom; X: any electronegative atom like O, N, S, P, Se, halogens; ..: represent aromatic single bond).	M3, M6
B06[N-N]	It is a 2D atom pair descriptor that represent the presence or absence of two nitrogen atoms at a topological distance of 6.	M3
F09[S-S]	It is a 2D atom pair descriptor that represent the frequency of two sulfur atoms at a topological distance of 9.	M3

Table 4.6: Descriptors of the q-RASPR PLS models of phenothiazine dataset

Descriptor	Model	Description
SD Activity ($S_{weighted}$)	M4, M6	It is a RASPR descriptor that represents the weighted standard deviation of the observed response values of the selected close source compounds for each query compound. It is calculated by following equation: $$S_{weighted} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_{wtd}})^2}{\sum_{i=1}^{n} w_i}} \times \frac{n}{n-1}$$ $$\overline{x_{wtd}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$ n = number of close source compounds w_i = weightage given to individual selected close source compound This descriptor shows a positive contribution for both the models M4 and M6, as shown in the **Figure 4.20**.

CV activity (or *CVact*)	M4	This descriptor represents the coefficient of variation of the observed response (or activity) values of the selected close source compounds for each query compounds. It is calculated by using the following equation: $$CV_{activity} = \frac{S_{weighted}}{\overline{x_{wtd}}}$$ This descriptor shows a negative contribution to the PCE in model M4.
CV similarity (or *CVsim*)	M4	It is the coefficient of variation of the similarity values of the selected close source compounds for each query compounds. This descriptor is calculated by using following equation: $$SD\ Similarity = \sqrt{\frac{\sum_{i=1}^{n}(f-\bar{f})^2}{n-1}}$$ $$CV_{similarity} = \frac{SD\ Similarity}{\bar{f}}$$ f = similarity value of the selected close source compounds $\bar{f}$= average similarity value of the selected close source compounds n=number of selected close source compounds This descriptor shows negative contribution in the model M4.
g_m [Banerjee-Roy Coefficient]	M4, M6	It is a concordance measure calculated by using following equation: $$g_m = (-1)^n \lvert Posfrac - 0.5 \rvert$$ n = 1 when MaxPos<MaxNeg n = 2 when MaxPos>=MaxNeg *Posfrac* = Fraction of the close source compounds having a response value greater than the training set mean response This descriptor shows the highest contribution to the PCE for both the models M4 and M6 but for M4 it shows a positive contribution and for model M6 it represents negative contribution.
g_m*SD Similarity	M5, M6	It is the product of the descriptors g_m and *SD Similarity* where *SD Similarity* is the standard deviation of the similarity value of the selected close source compounds. This descriptor shows a positive contribution for models M5 and M6, where in the former one it has the highest contribution.

Standard Error (SE)	M6	It is the standard error of the observed response values of the selected close source compounds for each query compound which is calculated by using following equation: $$SE = \frac{SD\ Activity}{\sqrt{n}}$$ This descriptor shows a negative contribution to the PCE in the model M6 and also has least importance to the PCE.

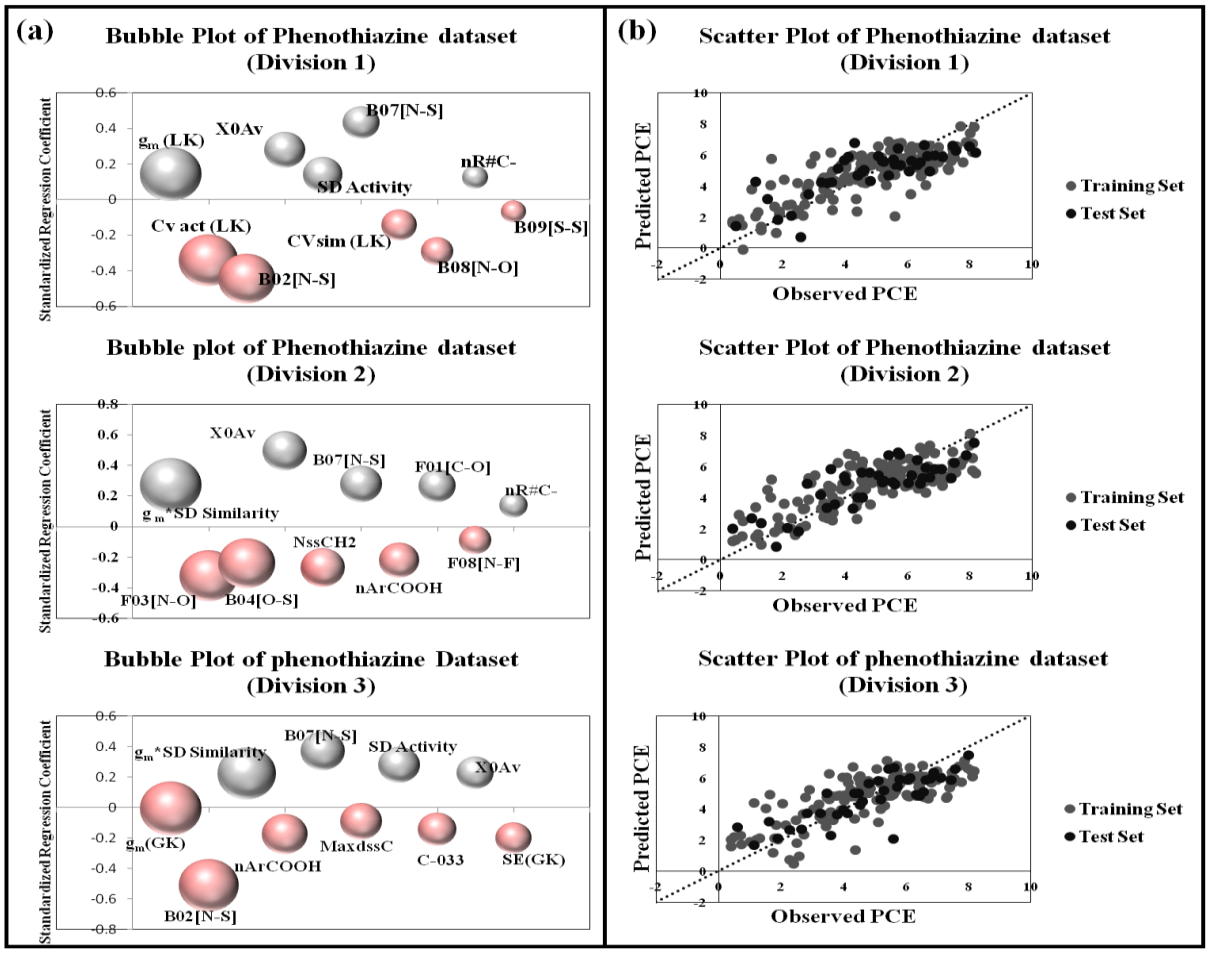


Figure 4.20: (a) Bubble plot of the phenothiazine dataset for three divisions to show the relative importance of the descriptors; (b) Scatter plot of the phenothiazine dataset for three divisions to show correlation between the observed and the predicted PCE.

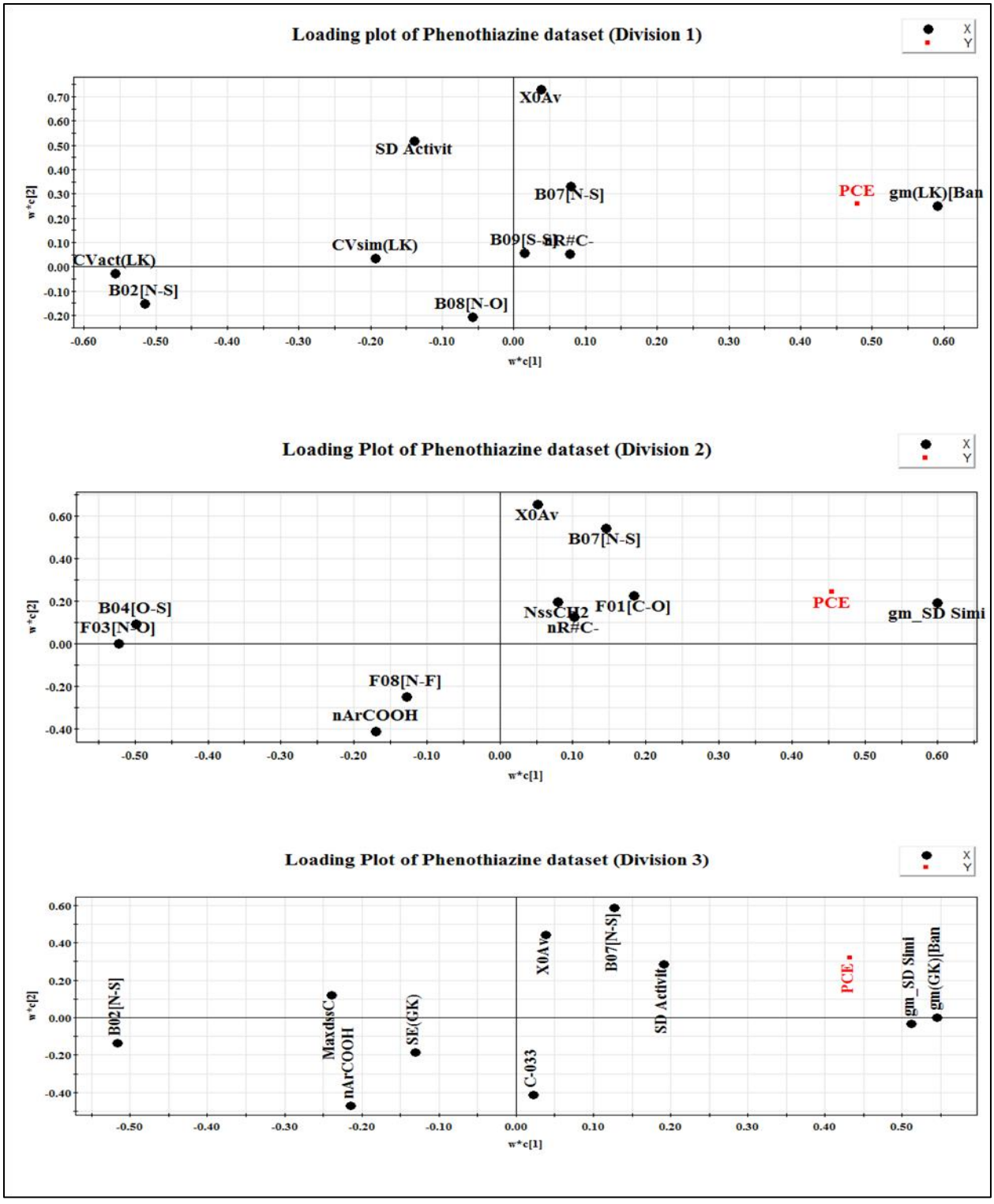


Figure 4.21: Loading plot of Phenothiazine dataset for all the respective divisions

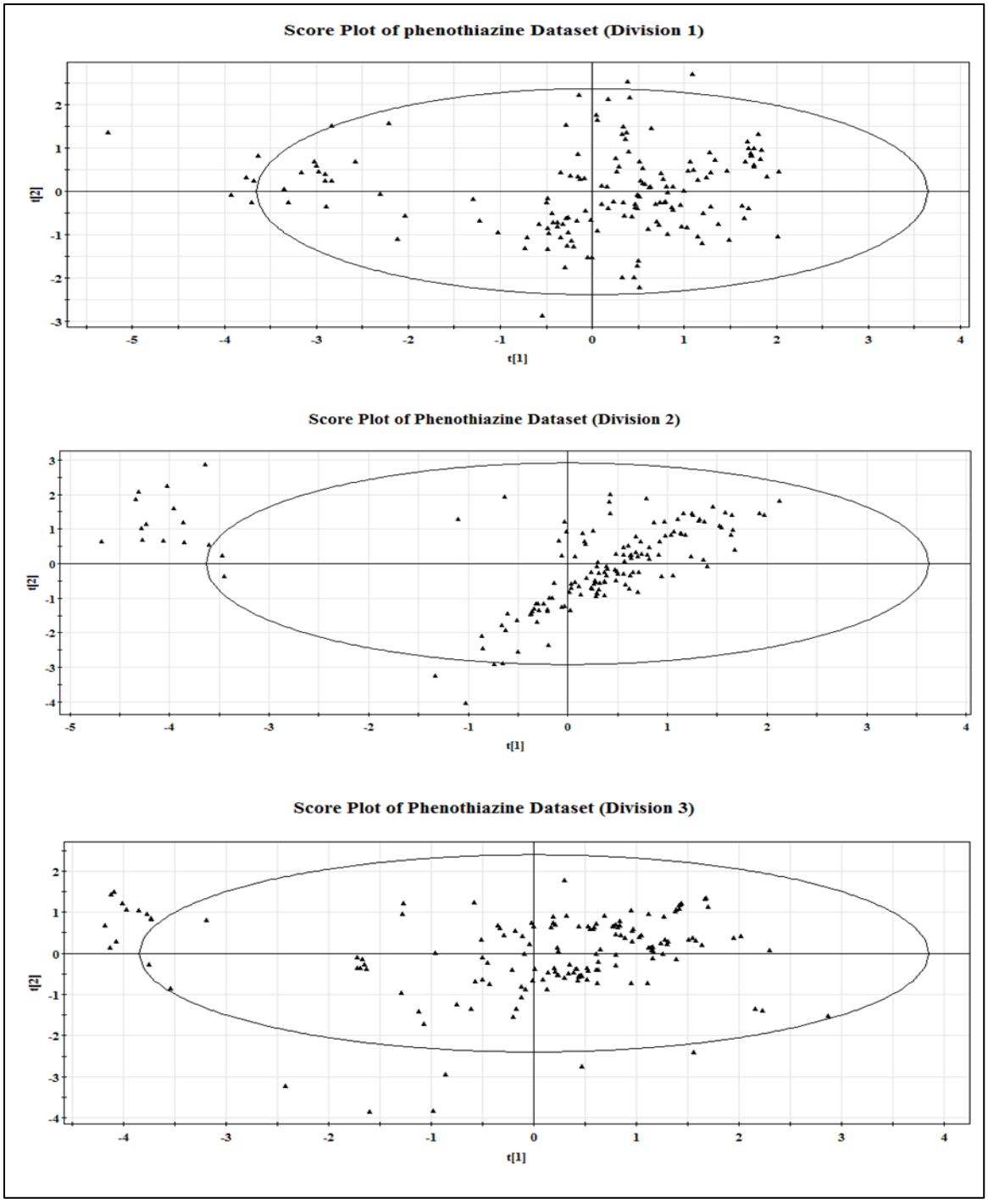


Figure 4.22: Score plot of Phenothiazine dataset for all the respective divisions

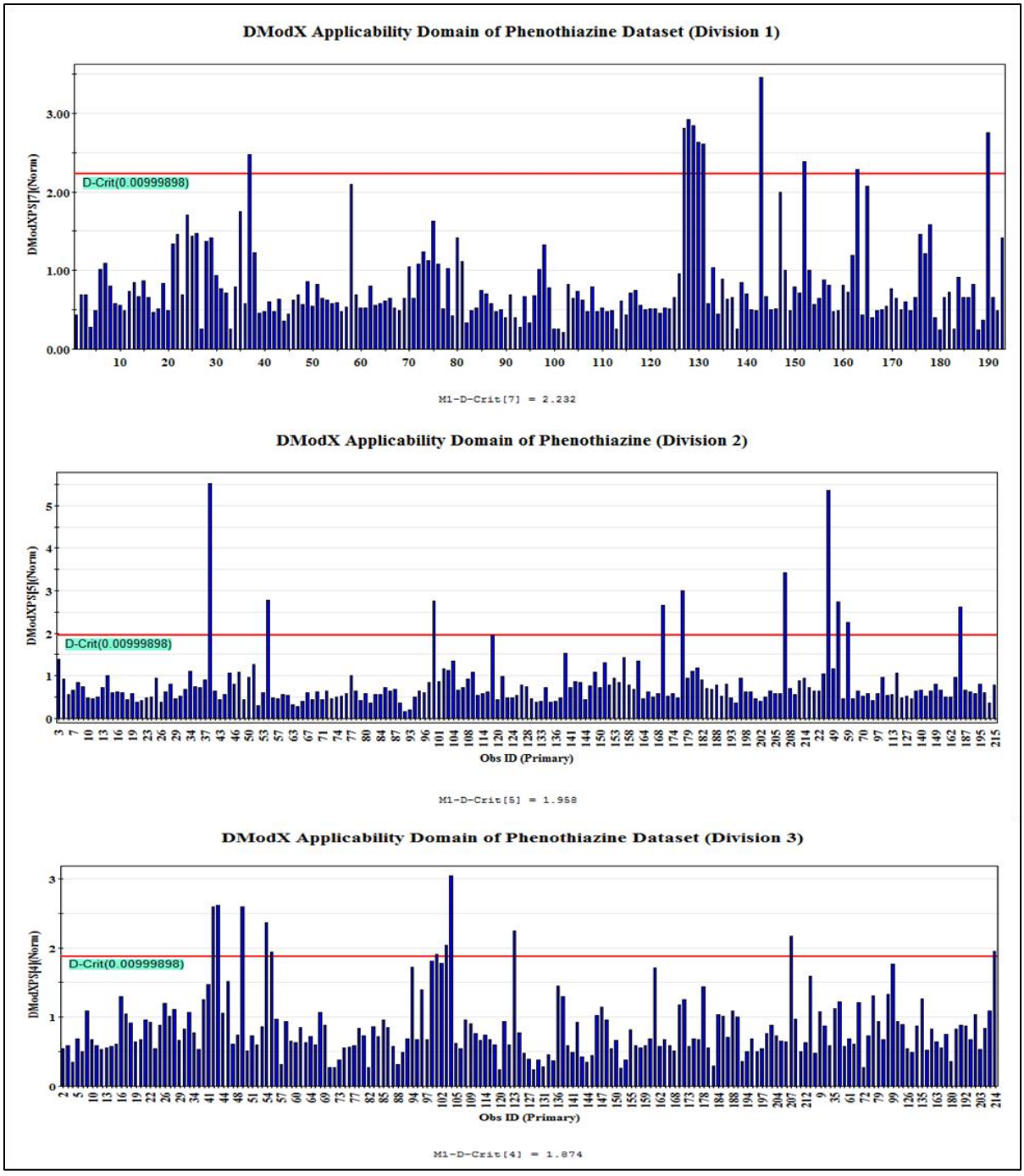


Figure 4.23: DModX Applicability domain plot of Phenothiazine dataset for all the respective divisions

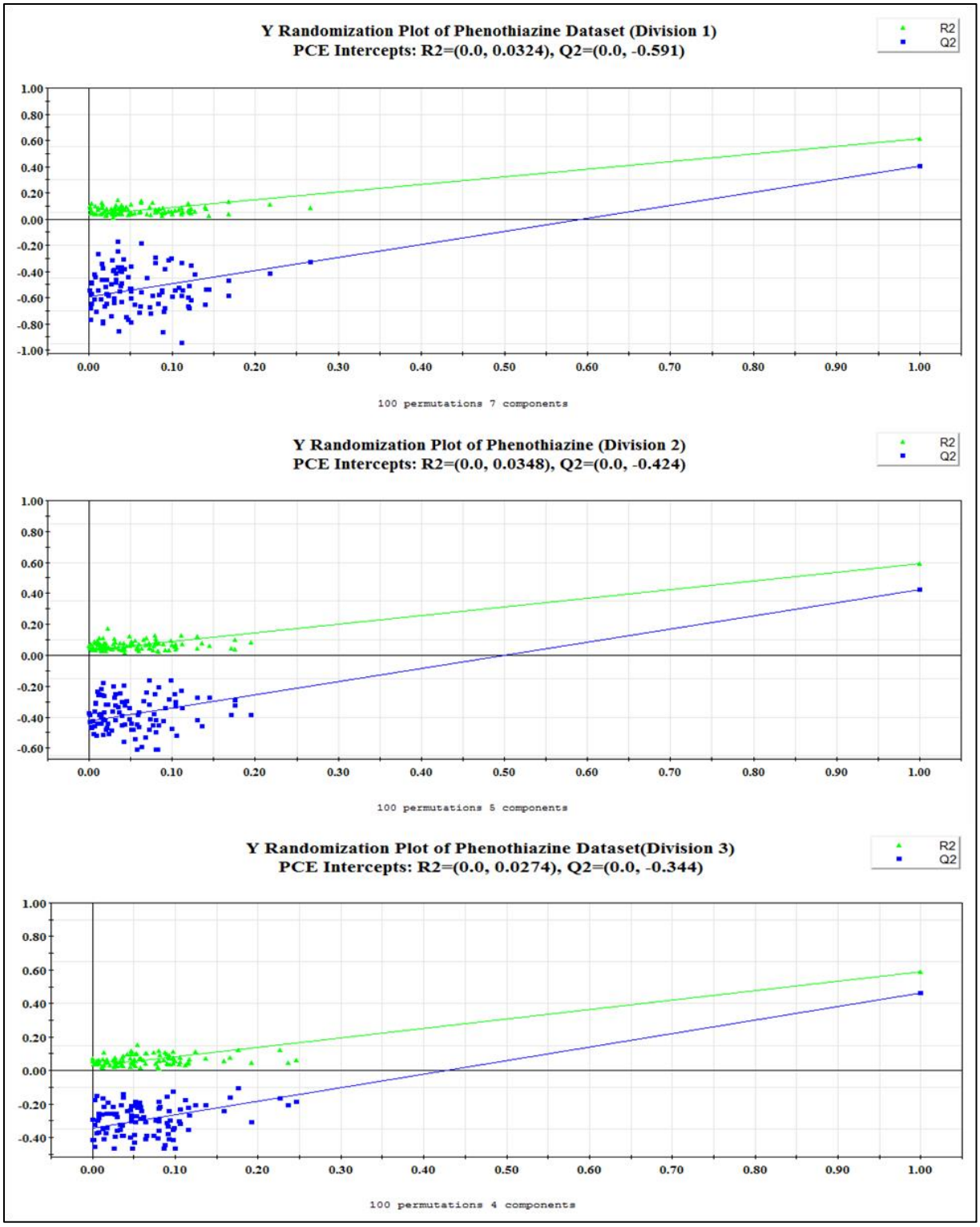


Figure 4.24: Y-Randomization plot of Phenothiazine dataset for all the respective divisions

4.2.1.2 Modeling of PCE of Porphyrin dyes

The most contributing descriptors to the PCE value of porphyrin dataset for all respective divisions are shown in the form of mathematical equations in **Box-2**. The description of the modelled structural and physicochemical descriptors of QSPR and q-RASPR PLS models are represented in the **Table 4.7**. The modelled RASPR descriptors are discussed below in **Table 4.8**, along with their influence on the PCE. Here, we have represented the importance of the descriptors of q-RASPR PLS model in the form of bubble plot, and correlation between observed and predicted PCE is represented in the form of a scatter plot, as shown in the **Figure 4.25**. The PLS plots for all the respective divisions are shown in the **Figure 4.26** to **4.29**.

Box-2

Porphyrin dataset

QSPR Models

M1 (Division 1)

$$PCE = 13.421 + 0.275 \times F07[C-X] + 0.529 \times MaxsssN - 0.025 \times TPSA(NO) + 0.820 \times F09[S-X] + 0.054 \times NssCH2 - 0.778 \times B09[O-X] + 2.028 \times F05[N-S] + 0.275 \times F10[N-O] - 4.419 \times MaxaaCH - 0.888 \times B08[O-X]$$

M2 (Division 2)

$$PCE = 2.723 - 1.847 \times nFuranes + 0.035 \times NssCH2 + 0.336 \times MaxsssN - 0.462 \times B09[O-X] + 0.124 \times F05[N-O] + 2.039 \times F05[N-S] - 0.151 \times F06[N-O] + 0.242 \times F07[C-X] + 0.244 \times F10[N-O] - 0.014 \times TPSA(NO)$$

M3 (Division 3)

$$PCE = 2.682 + 0.033 \times NssCH2 + 0.399 \times MaxsssN + 1.588 \times B03[N-O] - 1.112 \times B07[S-X] - 1.129 \times B08[O-X] + 2.274 \times F05[N-S] + 0.306 \times F06[O-X] + 0.217 \times F07[C-X] + 0.240 \times F10[N-O] - 0.014 \times TPSA(NO)$$

RASPR Models

M4 (Division 1)

$$PCE = 11.865 + 0.221 \times RA\ function(LK) - 0.284 \times CVact(LK) + 3.059 \times Pos.Avg.\ Sim. + 0.247 \times F07[C-X] - 0.021 \times TPSA(NO) + 1.074 \times F09[S-X] + 0.047 \times NssCH2 + 1.553 \times F05[N-S] + 0.109 \times F10[N-O] - 4.121 \times MaxaaCH$$

M5 (Division 2)	$PCE = 2.300 - 0.545 \times SE(LK) - 0.411 \times CVsim(LK) + 1.766 \times MaxPos(LK) + 0.032 \times NsssCH2 + 0.338 \times MaxsssN + 0.096 \times F05[N-O] + 1.962 \times F05[N-S] + 0.215 \times F07[C-X] + 0.233 \times F10[N-O] - 0.010 \times TPSA(NO)$
M6 (Division 3)	$PCE = 1.926 - 2.269 \times CVact(LK) + 2.597 \times Avg.Sim\,(LK) + 0.039 \times NssCH2 + 0.389 \times MaxsssN + 1.814 \times B03[N-O] - 0.824 \times B07[S-X] + 2.040 \times F05[N-S] + 0.224 \times F07[C-X] + 0.137 \times F10[N-O] - 0.003 \times TPSA(NO)$

Table 4.7: The description of the modelled structural and physicochemical descriptors of QSPR and q-RASPR PLS models of porphyrin dyes.

Feature	Meaning	Model
F07[C-X]	M1, M2, M3, M4, M5, M6	It is a 2D atom pair descriptor that represent the frequency of carbon and heavy metal (X) at a topological distance of 7.
MaxsssN	M1, M2, M3, M5, M6	It is an atom type E-state indices descriptor that represent Maximum single bond, single bond, and single bond nitrogen atom (>N−) E-state.
TPSA (NO)	M1, M2, M3, M4, M5, M6	The topological polar surface area (TPSA) of a molecule is determined by the summation of tabulated surface contributions of polar atom types. It is calculated by using following formula: $TPSA = \sum_i N_i . G_i$ N_i= frequency of the ith atom type in the molecule G_i = surface contribution of this polar atom Here, TPSA(NO) is calculated only from polar fragments with nitrogen and oxygen atoms.
F09[S-X]	M1, M4	It is a 2D atom pair descriptor that represent the frequency of sulfur and heavy metal (X) at a topological distance of 9.
NssCH2	M1, M2, M3, M4, M5, M6	It is an atom type E-state indices that represent the number of atoms of single bond, single bond CH2 (-CH2-).

B09[O-X]	M1, M2	It is a 2D atom pair descriptor that represent the presence or absence of oxygen and heavy metal (X) atoms at a topological distance of 9.
F05[N-S]	M1, M2, M3, M4, M5, M6	It is a 2D atom pair descriptor that represent the frequency of nitrogen and sulfur atoms at a topological distance of 5.
F10[N-O]	M1, M2, M3, M4, M5, M6	It is a 2D atom pair descriptor that represent the frequency of nitrogen and oxygen atoms at a topological distance of 10.
MaxaaCH	M1, M4	It is an atom type E-state indices descriptor that represent Maximum aromatic bond, aromatic bond, CH (aCHa) E-state.
B08[O-X]	M1, M3	It is a 2D atom pair descriptor that represent the presence or absence of oxygen and heavy metal (X) atoms at a topological distance of 8.
nFuranes	M2	It is a functional group count descriptor which indicate the number of furanes structure present in a molecule.
F05[N-O]	M2, M5	It is a 2D atom pair descriptor that represent the frequency of nitrogen and oxygen atoms at a topological distance of 5.
F06[N-O]	M2	It is a 2D atom pair descriptor that represent the frequency of nitrogen and oxygen atoms at a topological distance of 6.
B03[N-O]	M3, M6	It is a 2D atom pair descriptor that represent the presence or absence of nitrogen and oxygen atoms at a topological distance of 3.
B07[S-X]	M3, M6	It is a 2D atom pair descriptor that represent the presence or absence of sulfur and heavy metal (X) atoms at a topological distance of 7.
F06[O-X]	M3	It is a 2D atom pair descriptor that represent the frequency of oxygen and heavy metal (X) atoms at a topological distance of 6.

Table 4.8: Descriptors of the q-RASPR PLS models of the porphyrin dataset

Descriptors	Model	Description
RA function	M4	It is a RASPR descriptor which is like a latent variable containing information of all the selected structural and physicochemical descriptors and is obtained based on similarity with selected close

		source compounds. It is calculated by using following formula: $$RA\ function = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$ $$w_i = \frac{S_i}{\sum_{i=1}^{n} S_i}$$ w_i = weightage of the individual selected close source compounds S_i= Similarity between individual selected close source compounds and target compounds x_i=observed response values of the selected close source compounds This descriptor shows a positive contribution to the PCE for model M4 and has the highest relative importance for the PCE.
CV activity (CVact)	M4, M6	This descriptor indicates the coefficient of variation of the observed response values of selected close source compounds for each query compound. This descriptor shows a negative contribution to the PCE in both models M4 and M6.
Positive Average Similarity (*Pos.Avg.Sim.*)	M4	This RASPR descriptor represents the average similarity value of the selected positive close source compound where compounds having observed response values higher than training set mean observed response value are considered as positive close source compounds. This descriptor shows a positive contribution to the PCE value in the model M4.
Standard Error (SE)	M5	This descriptor represents the standard error of the observed response values of selected close source compounds for each query compounds, and this descriptor shows a negative contribution to the PCE in the model M5.
CV Similarity (CVsim)	M5	It is the coefficient of variation of the similarity values of the selected close source compounds for each query compound. This descriptor has a negative contribution to the PCE in model M5.
MaxPos	M5	This descriptor has a positive contribution to the PCE in model M6, and this descriptor represents the maximum similarity level to the positive close source compounds based on the mean training set observed response value.
Average Similarity (Avg.Sim.)	M6	It is the mean of similarity values of the selected close source compounds for each query compound. This descriptor has positive

		contributions to the PCE in the model M6.

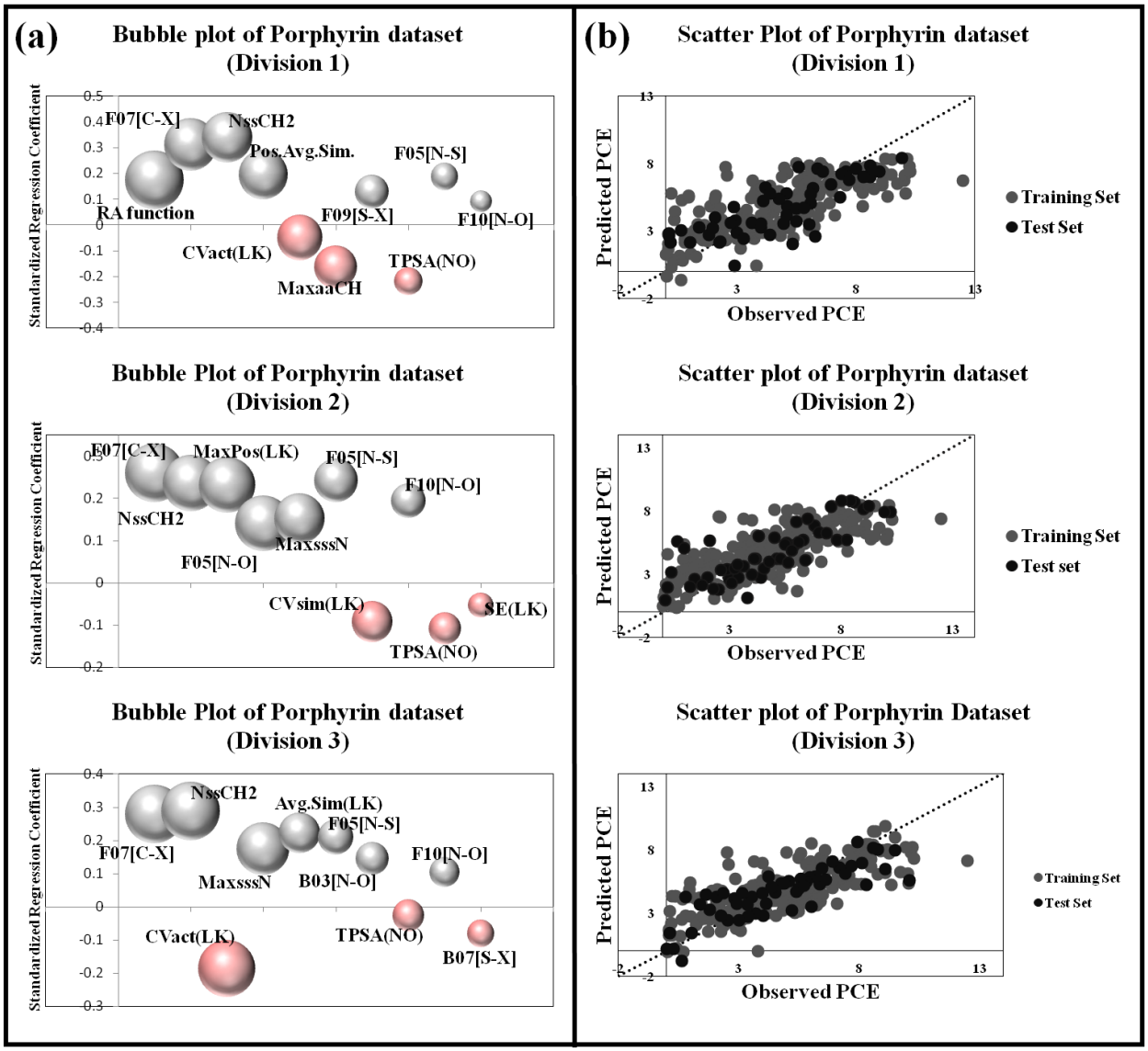


Figure 4.25: (a) Bubble plot of the porphyrin dataset for three divisions to show relative importance of the descriptors to the PCE; (b) Scatter plot of the porphyrin dataset for three divisions to show correlation between observed PCE and predicted PCE.

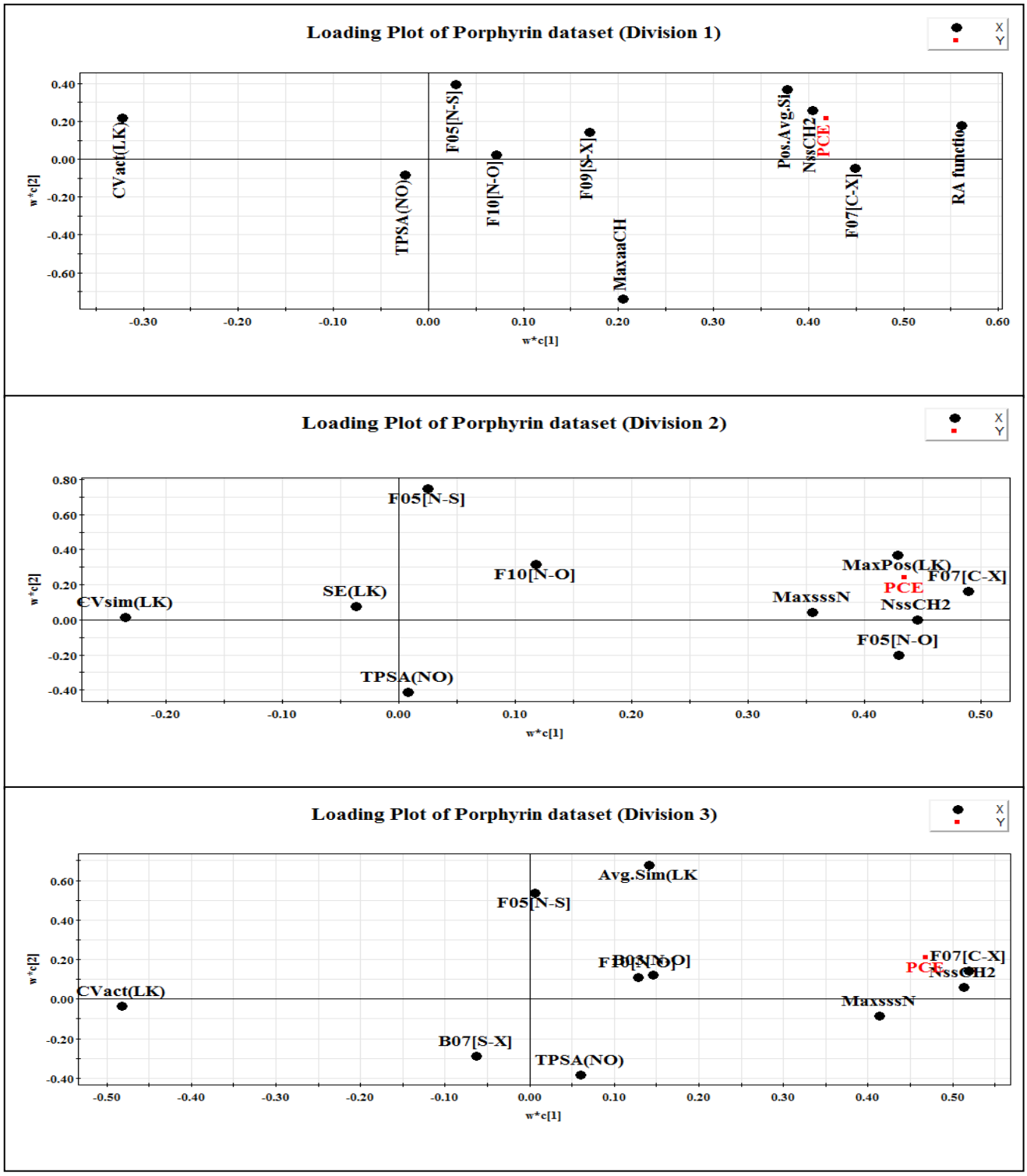


Figure 4.26: Loading plot of porphyrin dataset for all the respective divisions.

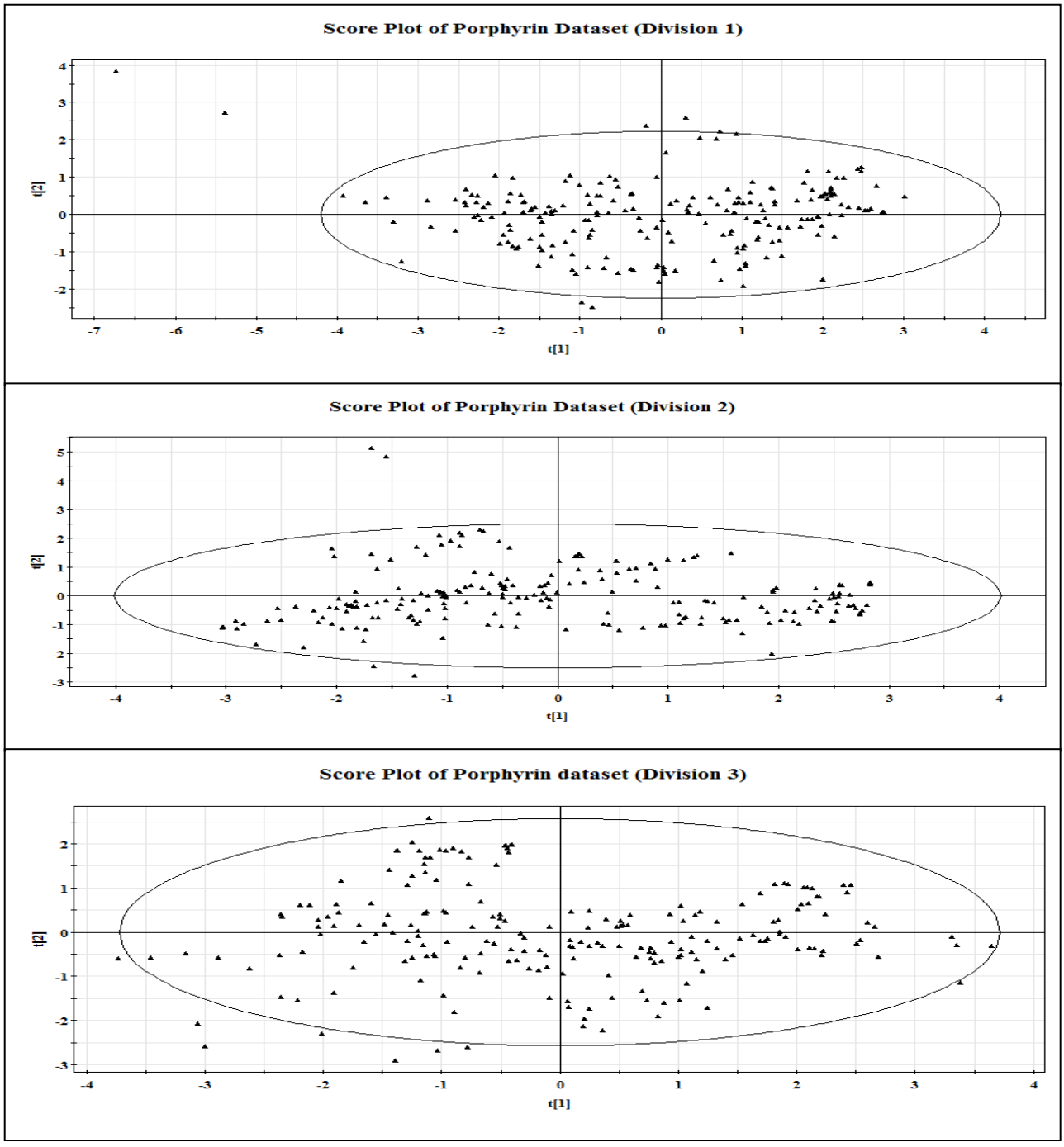


Figure 4.27: Score plot of porphyrin dataset for all the respective divisions

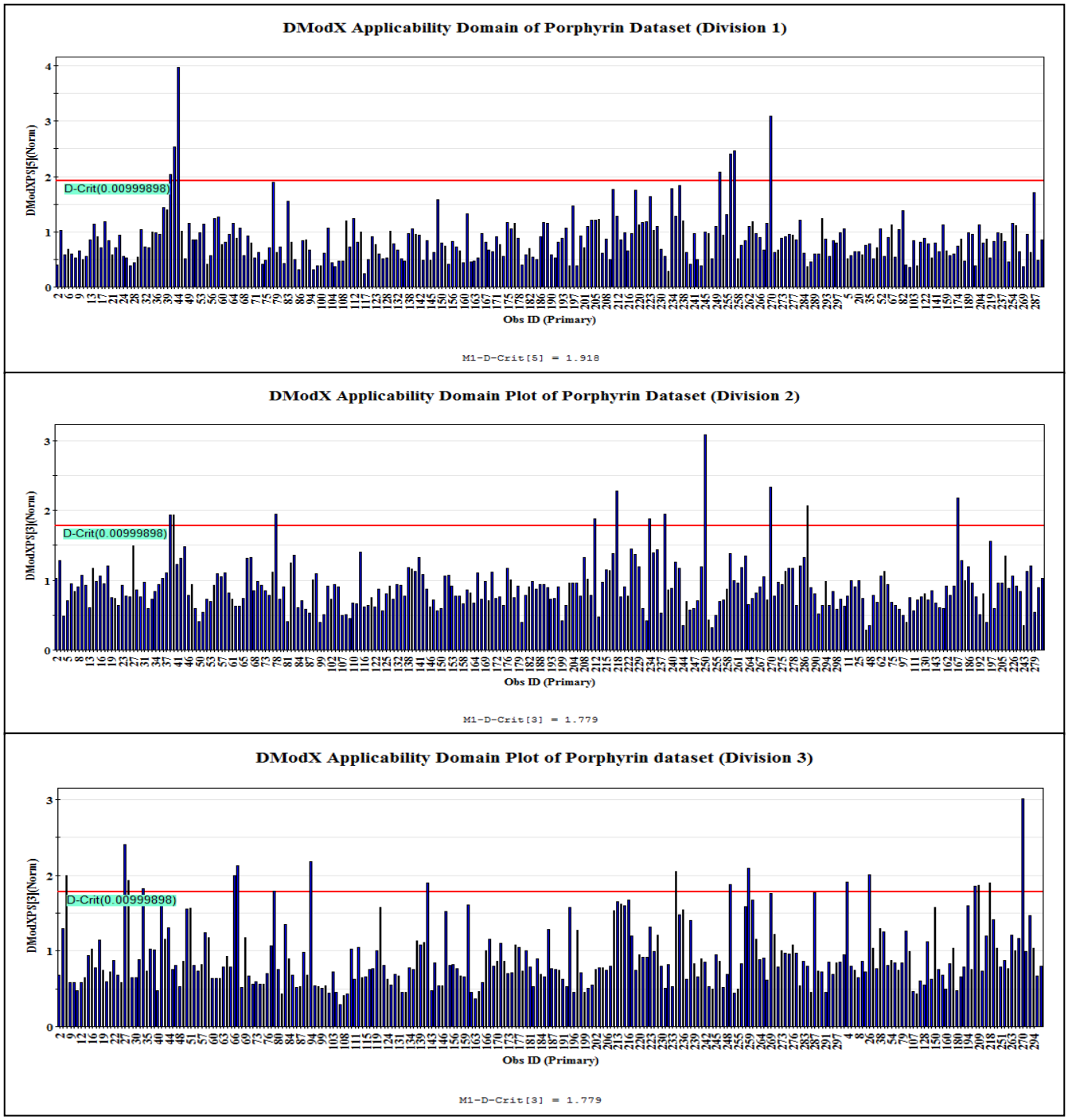


Figure 4.28: DModX Applicability domain plot of porphyrin dataset for all the respective divisions

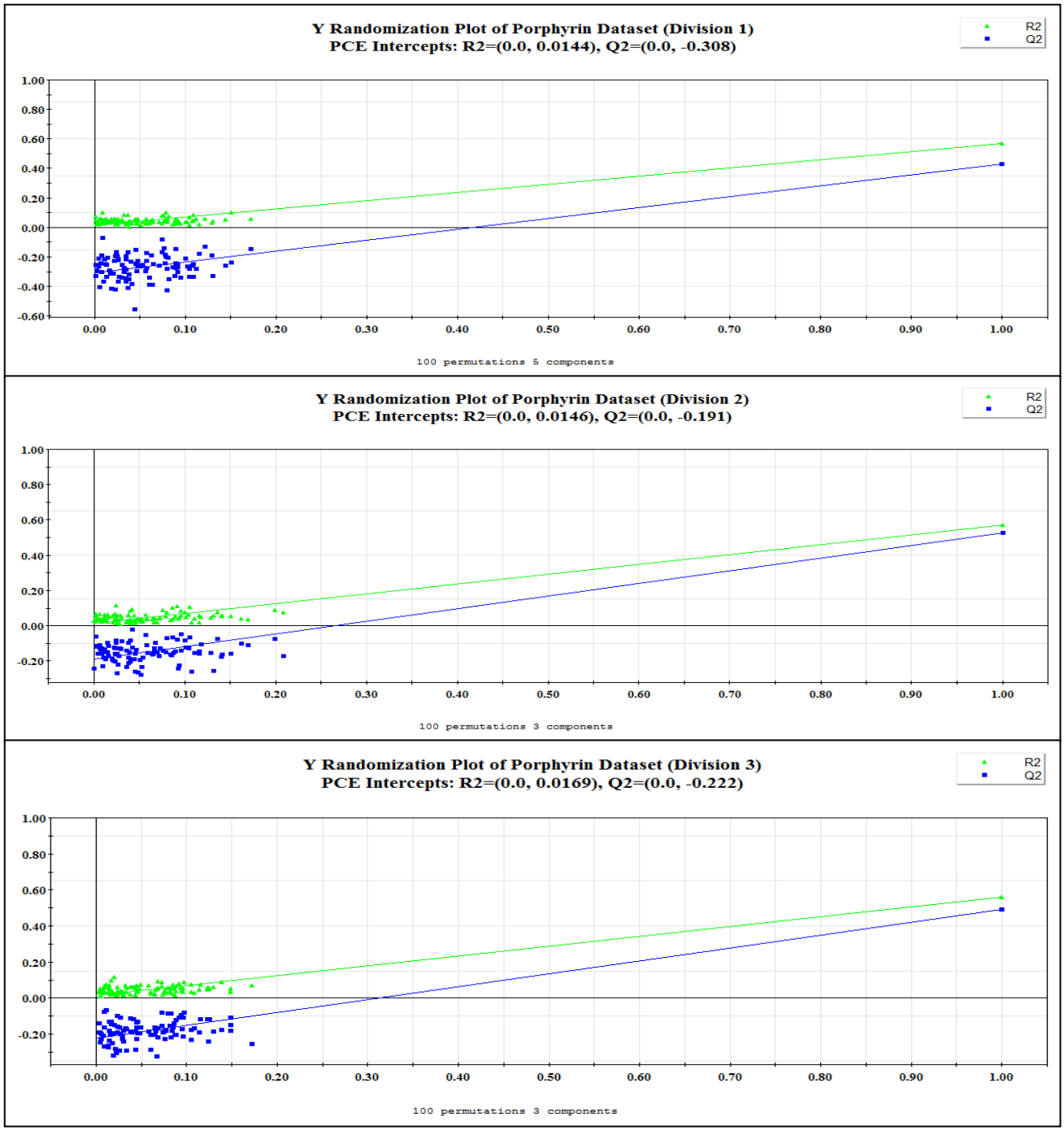


Figure 4.29: Y-Randomization plot of porphyrin dataset for all the respective divisions

4.2.1.3 Modeling of PCE of Triphenylamine dyes

The significant descriptors used for the modelling of PCE of the triphenylamine dyes are shown in the form of mathematical equations in the **Box-3**. The details of the physicochemical and

structural descriptors of the QSPR and q-RASPR PLS models are represented in **Table 4.9,** and the description of the RASPR descriptors are shown in the **Table 4.10**. The importance of the descriptors is represented in the form of bubble plots and is shown in **Figure 4.30**. The PLS plots of all the respective divisions are shown from **Figures 4.31** to **4.34**.

Box-3

Triphenylamine dataset

QSPR Models

M1 (Division 1) $PCE = 4.997 - 29.450 \times GD + 0.503 \times nRCOOH - 1.586 \times nArCOOH - 3.132 \times nImidazoles + 0.676 \times C - 019 + 0.543 \times SsssCH - 3.026 \times B02[N - S] + 1.355 \times B05[O - S] - 1.358 \times B07[O - S] + 2.100 \times F06[C - F]$

M2 (Division 2) $PCE = -0.034 + 1.064 \times nRCOOH - 1.315 \times nArCOOH + 0.234 \times C - 026 - 2.463 \times C - 043 + 0.161 \times SaaaC + 2.446 \times MaxsssCH + 0.438 \times MaxaaS + 0.972 \times B04[N - O] - 1.574 \times B07[O - S] + 1.339 \times B08[C - N]$

M3 (Division 3) $PCE = 10.021 + 3.218 \times minsssCH - 1.547 \times nArCOOH - 0.478 \times mindS + 0.748 \times NssS - 2.309 \times MaxaasC + 1.113 \times B05[O - S] - 60.619 \times GD - 1.621 \times B04[N - N] - 2.969 \times C - 043 - 1.753 \times B07[O - S]$

RASPR Models

M4 (Division 1) $PCE = 1.517 + 0.451 \times RA\ function(GK) + 3.559 \times g_m * SD\ Similarity + 0.686 \times Neg.\ Avg.\ Sim + 0.287 \times nRCOOH - 0.500 \times nArCOOH - 2.083 \times nImidazoles + 0.623 \times SsssCH + 0.602 \times B05[O - S] - 0.668 \times B07[O - S] + 1.778 \times F06[C - F]$

M5 (Division 2) $PCE = 3.434 - 2.480 \times CVact(LK) + 0.295 \times MaxPos(LK) - 1.120 \times Avg.Sim.\ (LK) + 0.717 \times g_m * Avg.Sim. + 0.496 \times g_m * SD\ Similarity - 1.446 \times C - 043 + 0.784 \times MaxaaS + 0.776 \times B04[N - O] - 1.642 \times B07[O - S] + 1.276 \times B08[C - N]$

M6 (Division 3) $PCE = 6.618 + 1.601 \times MaxPos(GK) + 4.017 \times g_m * SD\ Similarity +$

$$2.736 \times minsssCH - 0.982 \times nArCOOH - 0.300 \times mindS - 1.635 \times MaxaasC + 0.975 \times B05[O-S] - 36.879 \times GD - 0.628 \times B04[N-N] - 1.439 \times C - 043$$

Table 4.9: The description of the modelled structural and physicochemical descriptors of QSPR and q-RASPR PLS models of triphenylamine dyes.

Feature	Meaning	Model
GD	Graph density (GD) is a constitutional indices descriptor which is derived from H-depleted molecular graph and calculated by following equation: $$GD = \frac{2.nBo}{nSK.(nSK-1)}$$ nBo = number of non-hydrogen bonds (edges) nSK = number of vertices in the graph	M1, M3, M6
nRCOOH	It is functional group count descriptor that represent the number of aliphatic carboxylic acid present in the molecular structure.	M1, M2, M4
nArCOOH	It is functional group count descriptor that represent the number of aromatic carboxylic acid present in the molecular structure.	M1, M2, M3, M4, M6
nImidazoles	It is functional group count descriptor that represent the number of imidazole group present in the molecular structure.	M1, M4
C-019	It is an atom centred fragment descriptor that represent the fragment **=CRX** (R: any group linked through carbon, X: any electronegative atom (O, N, S, P, Se, halogens), =: double bond).	M1
SsssCH	It is an atom type E-state indices descriptor that represent the sum of sssCH (single, single and single bond) E-states.	M1, M4
B02[N-S]	It is a 2D atom pair descriptor that represent the presence or absence of nitrogen and sulfur atoms at a topological distance of 2.	M1
B05[O-S]	It is a 2D atom pair descriptor that represent the presence or absence of oxygen and sulfur atoms at a topological distance of 5.	M1, M3, M4, M6
B07[O-S]	It is a 2D atom pair descriptor that represent the presence or absence of oxygen and sulfur atoms at a topological distance of 7.	M1, M2, M3, M4, M5

F06[C-F]	It is a 2D atom pair descriptor that represent the frequency of carbon and sulfur atoms at a topological distance of 6.	M1, M4
C-026	It is an atom centred fragment descriptor that represent the fragment **R--CX--R** (R: any group linked through carbon, X: any electronegative atom (O, N, S, P, Se, halogens), --: an aromatic bond or delocalized bonds).	M2
C-043	It is an atom centred fragment descriptor that represent the fragment **X--CR..X** (R: any group linked through carbon, X: any electronegative atom (O, N, S, P, Se, halogens), --: an aromatic bond or delocalized bonds, ..: aromatic single bonds).	M2, M3, M5, M6
SaaaC	It is an atom centred fragment descriptor that represents the sum of aromatic carbons (-C(-)-) (-: aromatic bonds)	M2
MaxsssCH	It is an atom type E-state indices descriptor that represent Maximum single bond, single bond, and single bond CH atom (>CH−) E-state.	M2
MaxaaS	It is an atom type E-state indices descriptor that represent Maximum aromatic bond, aromatic bond sulfur atom (aSa) E-state.	M2, M5
B04[N-O]	It is a 2D atom pair descriptor that represent the presence or absence of nitrogen and oxygen atoms at a topological distance of 4.	M2, M5
B08[C-N]	It is a 2D atom pair descriptor that represent the presence or absence of carbon and nitrogen atoms at a topological distance of 8.	M2, M5
minsssCH	It is an atom type E-state indices descriptor that represent the minimum single bond, single bond, and single bond CH atom (>CH−) E-state.	M3, M6
mindS	It is an atom type E-state indices descriptor that represent the minimum double bond sulfur atom (=S) E-state.	M3, M6
NssS	It is an atom type E-state indices descriptor that represent the number of single bond, single bond sulfur atom (-S-) E-state.	M3
MaxaasC	It is an atom type E-state indices descriptor that represent the Maximum aromatic bond, aromatic bond and single bond carbon atom (aCa-) E-state.	M3, M6
B04[N-N]	It is a 2D atom pair descriptor that represent the presence or absence of two nitrogen atoms at a topological distance of 4.	M3, M6

Table 4.10: Descriptors of the q-RASPR PLS models of the Triphenylamine dataset

Descriptors	Model	Description
RA function	M4	It is a RASPR descriptor which is like a latent variable containing information of all the selected structural and physicochemical descriptors This descriptor shows a positive contribution to the PCE in the model M4 and also has the highest importance to the PCE, as shown in the **Figure 4.30**.
g_m**SD Similarity*	M4, M5, M6	It is the product of g_m (a concordance measure) and *SD Similarity* (standard deviation of the similarity value of the selected close source compounds). This descriptor shows a positive contribution to the PCE in all the q-RASPR PLS models.
Negative Average Similarity *(Neg.Avg.Sim.)*	M4	This RASPR descriptor represents the average similarity value of the selected negative close source compound where negative compounds are selected based on the training set mean value. This descriptor shows a positive contribution to the PCE values in model M4.
CV activity *(CVact)*	M5	This descriptor shows the highest importance to the predictions in model M5 which basically represents the coefficient of variation of the observed response values of the selected close source compounds. This descriptor shows a negative contribution to the PCE in model M5.
MaxPos	M5, M6	This descriptor represents the maximum similarity level to the Positive closest source compound based on the mean value of training set observed response. This descriptor has the highest importance to the predictions in model M6 and shows a positive contribution in both the models M5 and M6.
Average Similarity *(Avg.Sim.)*	M5	It is the average of the similarity values of the selected close source compounds, and it shows a negative contribution to the PCE values in model M5.
g_m** Avg.Sim.*	M5	This RASPR descriptor shows a positive contribution to the PCE in model M5 which is basically a product of g_m and *average similarity*.

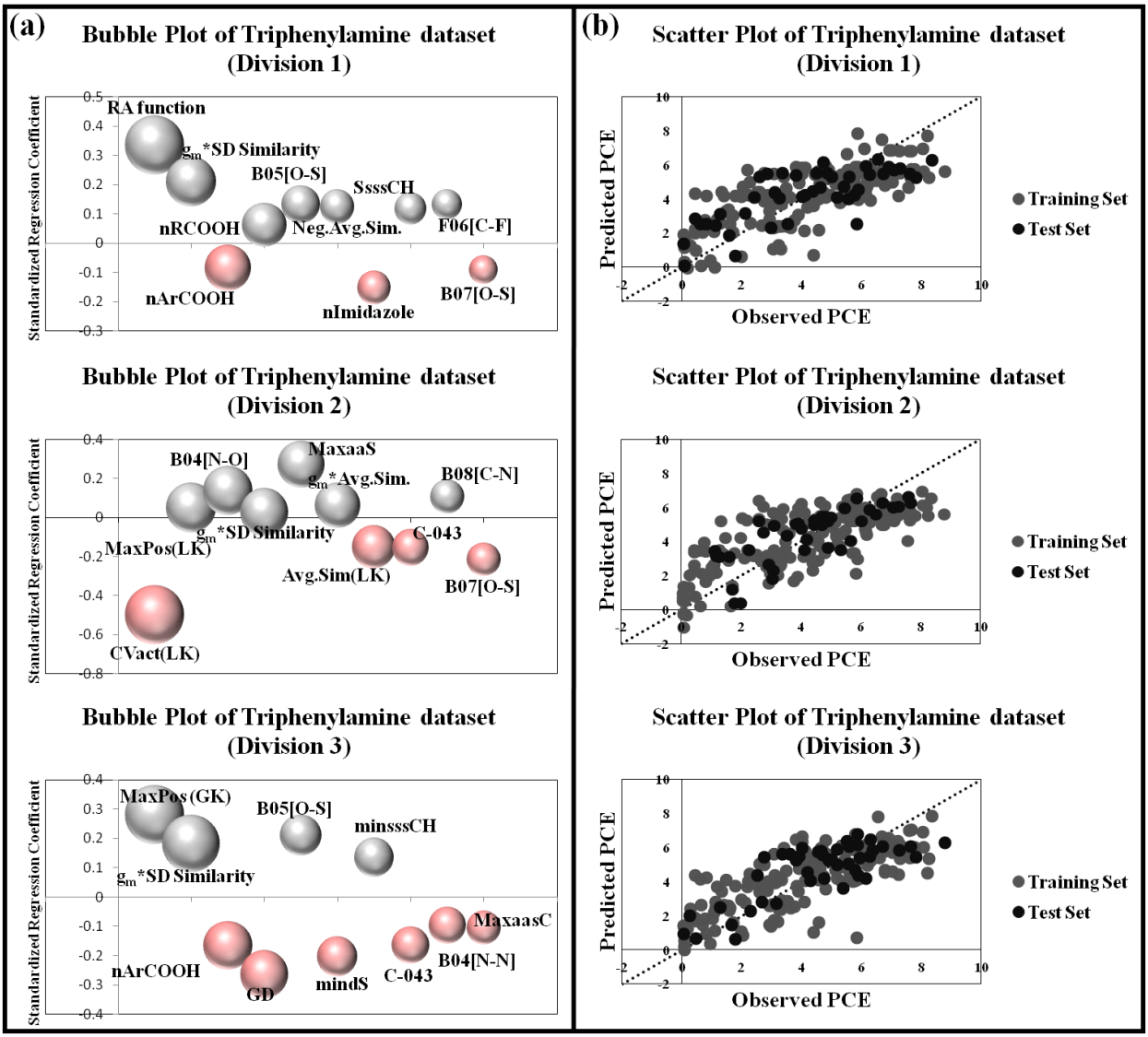


Figure 4.30: (a) Bubble plot of the triphenylamine dataset for three divisions to show relative importance of the descriptors to the PCE; (b) Scatter plot of triphenylamine dataset for the three divisions to show correlation between observed and predicted PCE.

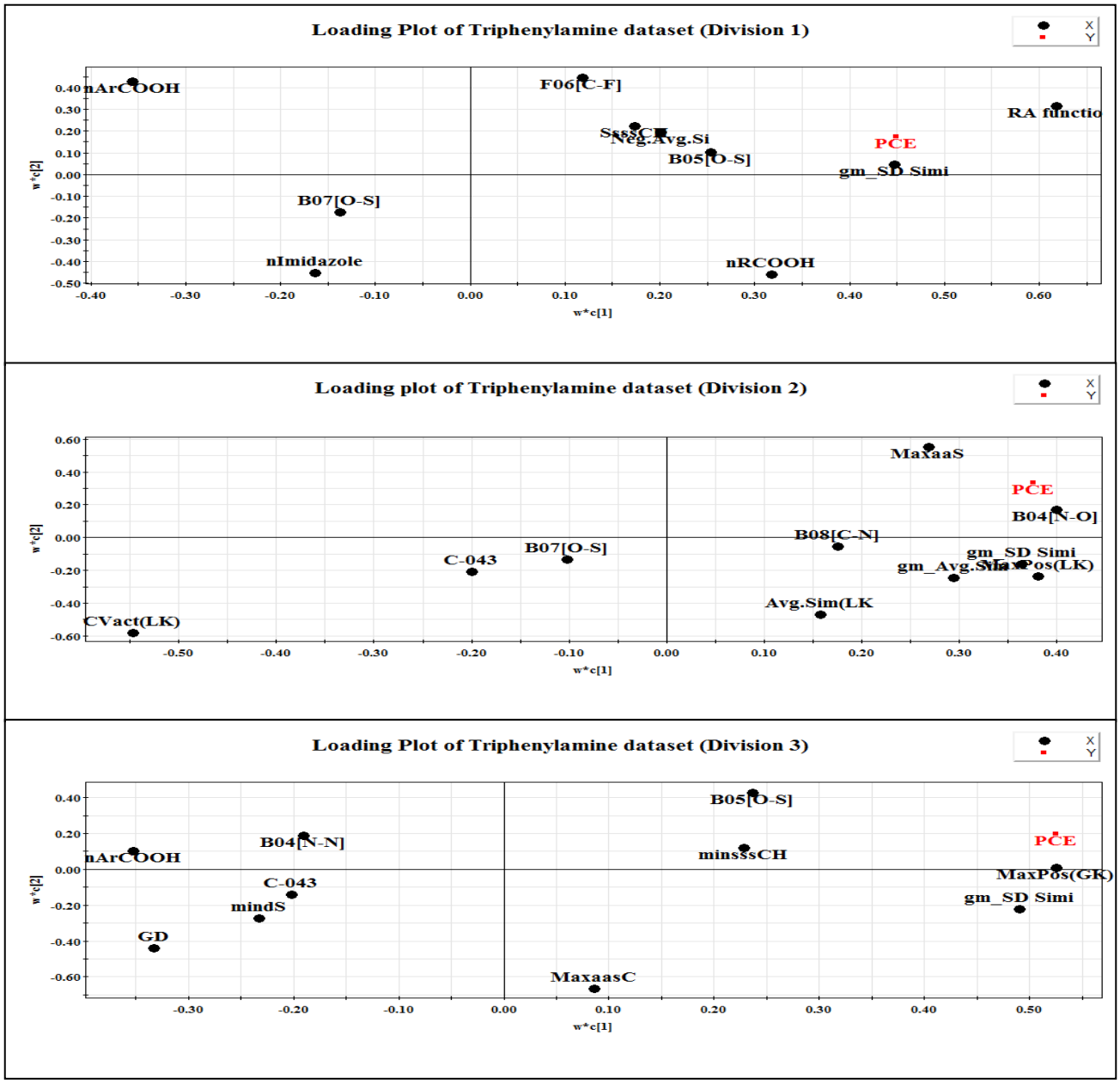


Figure 4.31: Loading plot of triphenylamine dataset for all the respective divisions.

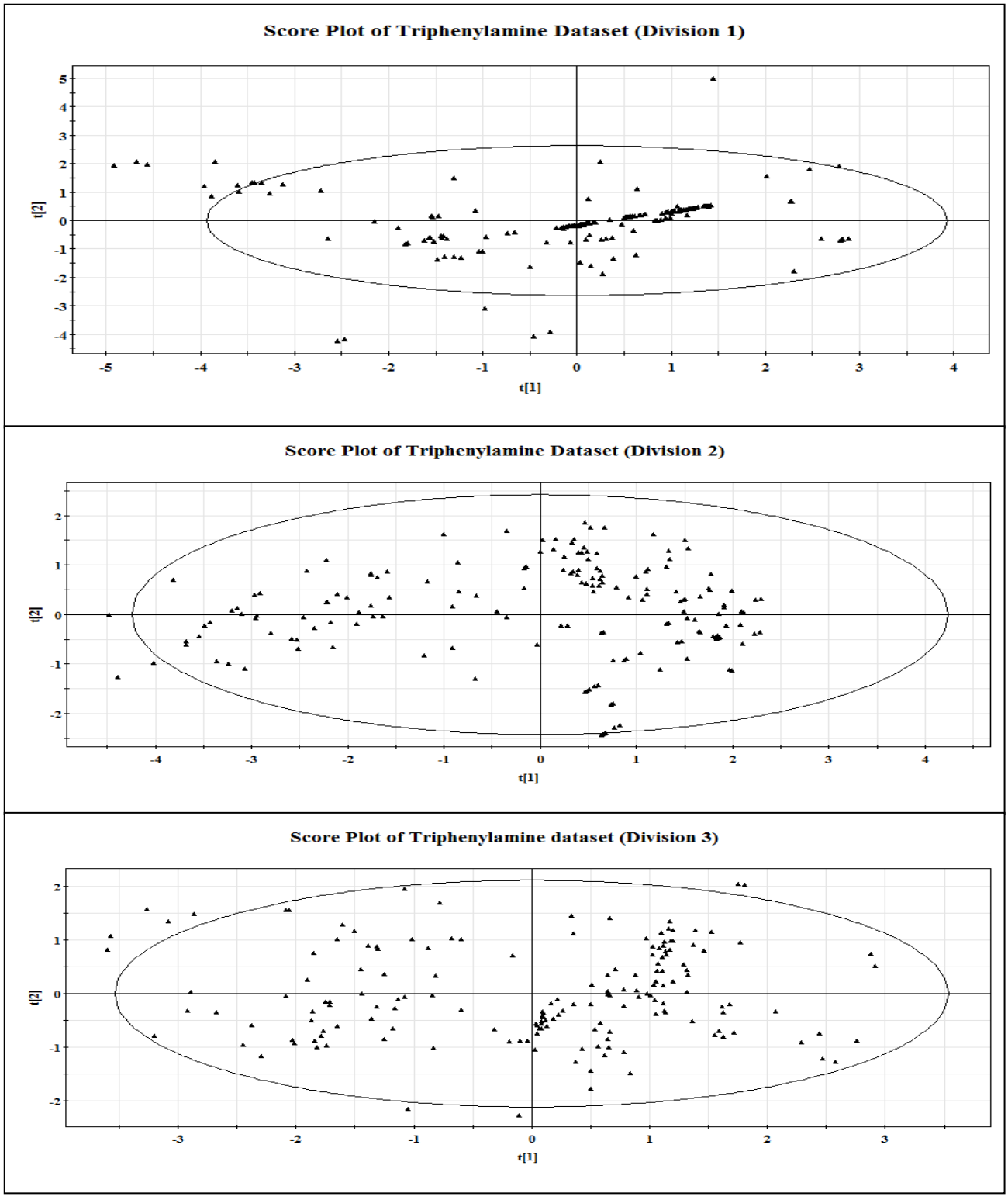


Figure 4.32: Score plot of triphenylamine dataset for all the respective divisions.

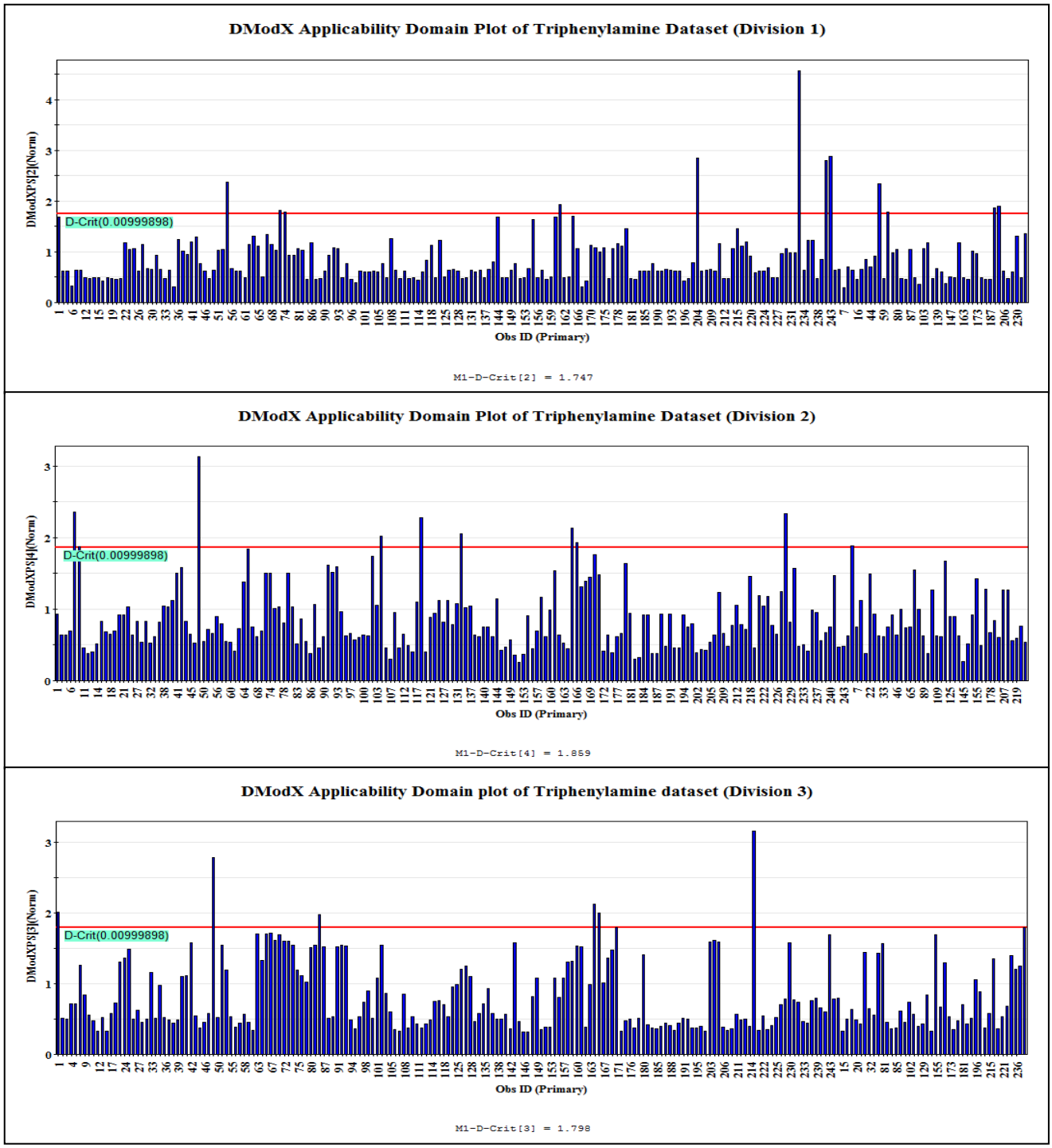


Figure 4.33: DModX Applicability domain plot of triphenylamine dataset for all the respective divisions.

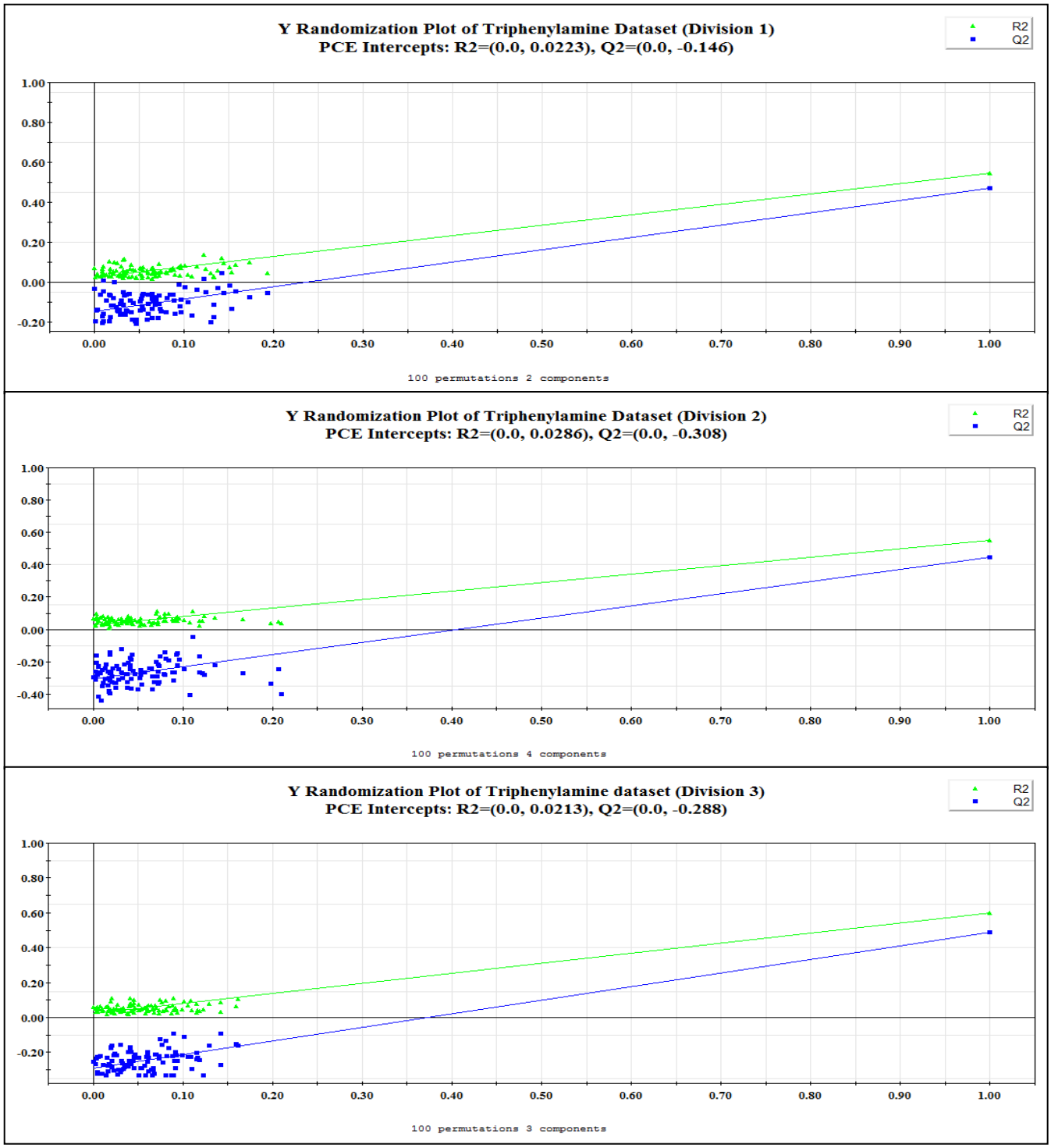


Figure 4.34: Y-Randomization plot of triphenylamine dataset for all the respective divisions.

4.2.2 Machine Learning model development

In this study, we have additionally developed machine learning models with QSPR and q-RASPR modelled descriptors at default hyper-parameter settings. The machine learning models are developed with standardized descriptors and response values where standardization is performed based on the mean and standard deviation of the training sets. The results of the machine learning models are represented in **Table 4.11**. For the purpose of comparison, we have also developed PLS models with standardized datasets using the same number of latent variables.

Table 4.11: Results of machine learning models

Dataset	Division	Method	Models	R^2	Q^2_{LOO}	MAE_{LOO}	MSE_{LOO}	MAE_{train}	RMSEC	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	MAE_{test}	RMSEP
Phenothiazines	1	RF	QSPR	0.92	0.46	0.59	0.54	0.23	0.29	0.46	0.46	0.45	0.58	0.74
			RASPR	0.93	0.50	0.56	0.50	0.20	0.26	0.43	0.43	0.42	0.62	0.76
		ADB	QSPR	0.63	0.39	0.64	0.61	0.52	0.61	0.46	0.46	0.45	0.59	0.74
			RASPR	0.74	0.49	0.58	0.51	0.45	0.51	0.42	0.42	0.41	0.63	0.77
		GB	QSPR	0.90	0.50	0.57	0.49	0.25	0.32	0.49	0.48	0.48	0.56	0.72
			RASPR	0.94	0.47	0.58	0.53	0.18	0.24	0.40	0.40	0.39	0.65	0.78
		XGB	QSPR	0.94	0.37	0.66	0.63	0.19	0.24	0.42	0.42	0.41	0.61	0.77
			RASPR	0.98	0.46	0.59	0.53	0.11	0.14	0.41	0.41	0.40	0.64	0.77
		SVM	QSPR	0.68	0.52	0.54	0.48	0.41	0.56	0.53	0.53	0.52	0.57	0.69
			RASPR	0.66	0.49	0.55	0.51	0.42	0.58	0.59	0.59	0.58	0.52	0.65
		LSVM	QSPR	0.60	0.55	0.55	0.45	0.49	0.63	0.57	0.57	0.56	0.52	0.66
			RASPR	0.60	0.55	0.53	0.44	0.48	0.63	0.62	0.62	0.62	0.51	0.62
		RR	QSPR	0.60	0.56	0.53	0.44	0.50	0.63	0.57	0.57	0.56	0.51	0.66

			RASPR	0.61	0.56	0.52	0.43	0.49	0.62	0.62	0.62	0.61	0.51	0.62
		PLS	QSPR	0.60	0.55	0.54	0.44	0.50	0.63	0.57	0.57	0.57	0.51	0.66
			RASPR	0.61	0.56	0.52	0.44	0.49	0.62	0.62	0.62	0.61	0.50	0.62
	2	RF	QSPR	0.91	0.46	0.57	0.54	0.22	0.30	0.39	0.39	0.36	0.65	0.80
			RASPR	0.92	0.40	0.60	0.59	0.22	0.29	0.38	0.38	0.35	0.67	0.80
		ADB	QSPR	0.65	0.40	0.61	0.59	0.50	0.59	0.45	0.45	0.42	0.60	0.76
			RASPR	0.65	0.40	0.61	0.60	0.50	0.59	0.46	0.46	0.43	0.61	0.75
		GB	QSPR	0.87	0.46	0.58	0.54	0.26	0.36	0.46	0.46	0.44	0.63	0.75
			RASPR	0.88	0.42	0.59	0.58	0.26	0.34	0.45	0.45	0.42	0.66	0.76
		XGB	QSPR	0.92	0.36	0.62	0.63	0.20	0.28	0.25	0.25	0.22	0.73	0.88
			RASPR	0.92	0.39	0.60	0.60	0.21	0.28	0.30	0.30	0.27	0.73	0.85
		SVM	QSPR	0.68	0.54	0.53	0.45	0.41	0.56	0.54	0.54	0.52	0.56	0.69
			RASPR	0.68	0.54	0.54	0.45	0.42	0.57	0.62	0.62	0.60	0.52	0.63
		LSVM	QSPR	0.57	0.50	0.56	0.49	0.51	0.65	0.61	0.61	0.59	0.50	0.64
			RASPR	0.58	0.52	0.55	0.47	0.50	0.64	0.65	0.65	0.64	0.53	0.60
		RR	QSPR	0.59	0.53	0.56	0.46	0.52	0.64	0.64	0.64	0.63	0.50	0.61
			RASPR	0.59	0.53	0.55	0.46	0.51	0.64	0.67	0.67	0.65	0.50	0.59
		PLS	QSPR	0.59	0.54	0.56	0.46	0.52	0.64	0.64	0.64	0.63	0.50	0.61
			RASPR	0.59	0.53	0.55	0.46	0.51	0.64	0.67	0.67	0.65	0.51	0.59
	3	RF	QSPR	0.89	0.39	0.62	0.60	0.25	0.33	0.58	0.58	0.61	0.47	0.62
			RASPR	0.92	0.44	0.55	0.56	0.21	0.29	0.44	0.44	0.48	0.54	0.72
		ADB	QSPR	0.63	0.40	0.62	0.59	0.52	0.61	0.52	0.52	0.55	0.54	0.67
			RASPR	0.71	0.42	0.59	0.58	0.46	0.54	0.53	0.53	0.56	0.54	0.66
		GB	QSPR	0.86	0.41	0.62	0.59	0.29	0.37	0.63	0.63	0.66	0.42	0.58
			RASPR	0.94	0.35	0.61	0.65	0.18	0.24	0.50	0.50	0.53	0.54	0.68
		XGB	QSPR	0.92	0.25	0.70	0.74	0.19	0.27	0.53	0.53	0.56	0.52	0.66
			RASPR	0.98	0.35	0.62	0.65	0.12	0.15	0.37	0.37	0.41	0.58	0.77
		SVM	QSPR	0.64	0.52	0.54	0.48	0.44	0.60	0.57	0.57	0.60	0.46	0.63
			RASPR	0.68	0.52	0.54	0.48	0.40	0.56	0.48	0.48	0.51	0.51	0.70
		V	QSPR	0.56	0.51	0.56	0.49	0.50	0.66	0.55	0.55	0.57	0.48	0.65

			RASPR	0.56	0.48	0.59	0.51	0.51	0.66	0.63	0.63	0.65	0.43	0.59
		RR	QSPR	0.59	0.53	0.56	0.47	0.52	0.64	0.59	0.59	0.62	0.47	0.61
			RASPR	0.60	0.54	0.56	0.46	0.52	0.63	0.67	0.67	0.69	0.42	0.55
		PLS	QSPR	0.59	0.53	0.55	0.47	0.51	0.64	0.60	0.60	0.62	0.47	0.61
			RASPR	0.59	0.54	0.56	0.46	0.52	0.64	0.67	0.66	0.69	0.42	0.56
Porphyrins	1	RF	QSPR	0.92	0.49	0.54	0.51	0.21	0.28	0.69	0.69	0.72	0.43	0.53
			RASPR	0.93	0.46	0.56	0.54	0.20	0.27	0.61	0.61	0.64	0.46	0.59
		ADB	QSPR	0.65	0.40	0.61	0.59	0.50	0.59	0.55	0.55	0.60	0.52	0.63
			RASPR	0.70	0.43	0.61	0.57	0.48	0.55	0.61	0.61	0.64	0.49	0.59
		GB	QSPR	0.89	0.41	0.58	0.58	0.26	0.33	0.65	0.65	0.68	0.45	0.56
			RASPR	0.92	0.36	0.61	0.64	0.23	0.29	0.58	0.58	0.62	0.49	0.61
		XGB	QSPR	0.93	0.42	0.57	0.57	0.20	0.27	0.65	0.65	0.68	0.46	0.56
			RASPR	0.98	0.41	0.58	0.58	0.12	0.16	0.60	0.60	0.64	0.48	0.60
		SVM	QSPR	0.67	0.49	0.54	0.51	0.40	0.57	0.62	0.62	0.66	0.46	0.58
			RASPR	0.67	0.48	0.55	0.52	0.40	0.57	0.70	0.70	0.73	0.41	0.52
		LSVM	QSPR	0.55	0.52	0.54	0.48	0.51	0.67	0.54	0.54	0.58	0.52	0.64
			RASPR	0.56	0.53	0.52	0.46	0.49	0.66	0.66	0.66	0.69	0.45	0.55
		RR	QSPR	0.56	0.52	0.54	0.48	0.51	0.66	0.51	0.51	0.56	0.55	0.66
			RASPR	0.57	0.52	0.52	0.47	0.50	0.66	0.65	0.65	0.68	0.47	0.56
		PLS	QSPR	0.56	0.52	0.54	0.48	0.51	0.66	0.50	0.50	0.55	0.56	0.67
			RASPR	0.57	0.52	0.52	0.48	0.50	0.66	0.64	0.64	0.68	0.47	0.57
	2	RF	QSPR	0.91	0.46	0.56	0.53	0.22	0.30	0.57	0.57	0.57	0.49	0.65
			RASPR	0.93	0.46	0.59	0.54	0.21	0.27	0.57	0.57	0.58	0.50	0.65
		ADB	QSPR	0.61	0.41	0.60	0.58	0.53	0.63	0.48	0.48	0.49	0.57	0.72
			RASPR	0.70	0.47	0.59	0.53	0.48	0.54	0.58	0.58	0.59	0.46	0.64
		GB	QSPR	0.86	0.50	0.55	0.50	0.28	0.37	0.55	0.55	0.56	0.52	0.67
			RASPR	0.91	0.46	0.59	0.53	0.24	0.30	0.44	0.44	0.45	0.55	0.74
		XGB	QSPR	0.88	0.49	0.55	0.51	0.25	0.35	0.60	0.60	0.60	0.50	0.63
			RASPR	0.97	0.44	0.59	0.55	0.14	0.19	0.48	0.48	0.49	0.55	0.71
		SVM	QSPR	0.65	0.51	0.53	0.49	0.41	0.59	0.60	0.60	0.61	0.45	0.62
			RASPR	0.70	0.51	0.54	0.49	0.39	0.55	0.65	0.65	0.65	0.42	0.59
		S V	QSPR	0.52	0.49	0.54	0.51	0.51	0.69	0.60	0.60	0.60	0.47	0.63

			RASPR	0.56	0.51	0.54	0.49	0.50	0.66	0.64	0.64	0.65	0.44	0.59
		RR	QSPR	0.55	0.50	0.55	0.50	0.53	0.67	0.62	0.62	0.62	0.46	0.61
			RASPR	0.57	0.53	0.53	0.47	0.51	0.65	0.66	0.66	0.66	0.43	0.58
		PLS	QSPR	0.54	0.50	0.55	0.49	0.53	0.67	0.61	0.61	0.62	0.46	0.62
			RASPR	0.57	0.53	0.53	0.47	0.51	0.65	0.66	0.66	0.67	0.43	0.58
	3	RF	QSPR	0.91	0.50	0.53	0.50	0.22	0.29	0.57	0.57	0.59	0.49	0.64
			RASPR	0.93	0.46	0.55	0.54	0.20	0.26	0.59	0.59	0.61	0.49	0.62
		ADB	QSPR	0.63	0.41	0.61	0.59	0.51	0.60	0.50	0.50	0.52	0.58	0.69
			RASPR	0.68	0.45	0.59	0.54	0.48	0.56	0.63	0.63	0.64	0.48	0.60
		GB	QSPR	0.83	0.45	0.56	0.55	0.30	0.41	0.58	0.58	0.60	0.47	0.63
			RASPR	0.90	0.46	0.57	0.54	0.25	0.32	0.56	0.56	0.58	0.49	0.65
		XGB	QSPR	0.86	0.47	0.56	0.53	0.28	0.38	0.58	0.58	0.59	0.47	0.64
			RASPR	0.95	0.46	0.57	0.53	0.17	0.23	0.47	0.46	0.48	0.57	0.72
		SVM	QSPR	0.66	0.52	0.53	0.48	0.41	0.58	0.59	0.59	0.60	0.46	0.63
			RASPR	0.68	0.49	0.54	0.50	0.40	0.57	0.64	0.64	0.65	0.43	0.59
		LSVM	QSPR	0.54	0.51	0.54	0.49	0.51	0.68	0.64	0.64	0.65	0.46	0.59
			RASPR	0.55	0.52	0.54	0.48	0.51	0.67	0.70	0.70	0.71	0.41	0.54
		RR	QSPR	0.55	0.51	0.55	0.49	0.52	0.67	0.63	0.63	0.64	0.48	0.60
			RASPR	0.56	0.52	0.54	0.48	0.51	0.66	0.70	0.70	0.71	0.41	0.54
		PLS	QSPR	0.55	0.51	0.55	0.49	0.52	0.67	0.62	0.62	0.64	0.48	0.60
			RASPR	0.56	0.51	0.54	0.48	0.52	0.66	0.69	0.69	0.70	0.41	0.55
Triphenylamines	1	RF	QSPR	0.88	0.33	0.66	0.66	0.28	0.34	0.50	0.50	0.51	0.56	0.70
			RASPR	0.93	0.50	0.55	0.49	0.21	0.26	0.47	0.47	0.48	0.61	0.72
		ADB	QSPR	0.54	0.39	0.64	0.60	0.57	0.68	0.49	0.49	0.50	0.62	0.71
			RASPR	0.68	0.45	0.60	0.55	0.48	0.56	0.45	0.45	0.46	0.64	0.73
		GB	QSPR	0.76	0.41	0.62	0.59	0.37	0.49	0.58	0.58	0.58	0.54	0.65
			RASPR	0.89	0.47	0.58	0.52	0.26	0.34	0.45	0.45	0.46	0.62	0.73
		XGB	QSPR	0.76	0.39	0.63	0.61	0.39	0.49	0.51	0.51	0.52	0.55	0.69
			RASPR	0.93	0.42	0.60	0.57	0.21	0.26	0.39	0.39	0.40	0.66	0.77
		SVM	QSPR	0.62	0.44	0.61	0.56	0.47	0.62	0.53	0.53	0.54	0.56	0.68
			RASPR	0.62	0.47	0.60	0.53	0.47	0.62	0.60	0.60	0.61	0.51	0.62

		LSVM	QSPR	0.53	0.47	0.59	0.52	0.54	0.68	0.53	0.53	0.53	0.55	0.68
			RASPR	0.52	0.47	0.59	0.53	0.53	0.69	0.59	0.59	0.60	0.52	0.63
		RR	QSPR	0.55	0.50	0.58	0.50	0.55	0.67	0.54	0.54	0.54	0.54	0.67
			RASPR	0.56	0.50	0.58	0.49	0.55	0.67	0.58	0.58	0.59	0.54	0.64
		PLS	QSPR	0.55	0.50	0.58	0.50	0.55	0.67	0.53	0.53	0.54	0.54	0.68
			RASPR	0.55	0.50	0.58	0.49	0.56	0.67	0.57	0.57	0.58	0.54	0.65
	2	RF	QSPR	0.86	0.45	0.58	0.55	0.27	0.37	0.34	0.34	0.61	0.50	0.63
			RASPR	0.92	0.45	0.57	0.54	0.21	0.28	0.34	0.34	0.60	0.51	0.63
		ADB	QSPR	0.54	0.39	0.63	0.61	0.56	0.67	0.40	0.40	0.64	0.49	0.60
			RASPR	0.67	0.46	0.59	0.54	0.49	0.57	0.43	0.43	0.66	0.48	0.58
		GB	QSPR	0.80	0.37	0.64	0.62	0.35	0.45	0.46	0.45	0.67	0.48	0.57
			RASPR	0.90	0.44	0.57	0.55	0.25	0.31	0.24	0.24	0.54	0.53	0.67
		XGB	QSPR	0.81	0.38	0.63	0.62	0.33	0.43	0.40	0.40	0.64	0.50	0.60
			RASPR	0.96	0.39	0.60	0.60	0.15	0.20	0.19	0.19	0.51	0.56	0.70
		SVM	QSPR	0.63	0.48	0.58	0.51	0.44	0.60	0.37	0.37	0.62	0.52	0.62
			RASPR	0.61	0.46	0.59	0.54	0.47	0.62	0.46	0.45	0.67	0.50	0.57
		LSVM	QSPR	0.52	0.46	0.60	0.54	0.54	0.69	0.42	0.42	0.65	0.47	0.59
			RASPR	0.55	0.49	0.57	0.50	0.52	0.67	0.60	0.60	0.76	0.42	0.49
		RR	QSPR	0.54	0.48	0.58	0.51	0.55	0.68	0.48	0.48	0.69	0.46	0.56
			RASPR	0.56	0.50	0.57	0.50	0.54	0.67	0.61	0.61	0.77	0.41	0.48
		PLS	QSPR	0.53	0.48	0.58	0.52	0.55	0.68	0.52	0.52	0.71	0.44	0.54
			RASPR	0.55	0.50	0.57	0.50	0.54	0.67	0.57	0.57	0.74	0.44	0.51
	3	RF	QSPR	0.91	0.40	0.62	0.59	0.24	0.29	0.50	0.48	0.60	0.50	0.63
			RASPR	0.91	0.41	0.60	0.59	0.23	0.29	0.53	0.51	0.63	0.46	0.61
		ADB	QSPR	0.47	0.33	0.66	0.67	0.60	0.72	0.43	0.41	0.55	0.53	0.67
			RASPR	0.71	0.41	0.61	0.58	0.47	0.54	0.47	0.45	0.58	0.52	0.64
		GB	QSPR	0.83	0.43	0.60	0.57	0.32	0.41	0.53	0.51	0.63	0.49	0.61
			RASPR	0.91	0.37	0.61	0.62	0.24	0.30	0.49	0.47	0.60	0.53	0.63
		XGB	QSPR	0.85	0.32	0.66	0.68	0.30	0.39	0.50	0.48	0.60	0.50	0.63
			RASPR	0.96	0.48	0.60	0.52	0.16	0.21	0.52	0.50	0.62	0.47	0.62
		V	QSPR	0.66	0.49	0.59	0.50	0.44	0.58	0.47	0.45	0.58	0.52	0.65

			RASPR	0.69	0.50	0.58	0.50	0.41	0.55	0.54	0.52	0.64	0.49	0.60
		LSVM	QSPR	0.56	0.48	0.59	0.51	0.52	0.66	0.45	0.43	0.56	0.54	0.66
			RASPR	0.57	0.50	0.56	0.50	0.50	0.65	0.60	0.58	0.68	0.47	0.56
		RR	QSPR	0.57	0.52	0.56	0.48	0.52	0.65	0.48	0.46	0.59	0.54	0.64
			RASPR	0.60	0.55	0.54	0.45	0.51	0.63	0.63	0.61	0.71	0.44	0.54
		PLS	QSPR	0.57	0.51	0.56	0.48	0.53	0.65	0.49	0.47	0.59	0.54	0.64
			RASPR	0.60	0.55	0.54	0.45	0.51	0.63	0.62	0.61	0.70	0.45	0.55

From the results of machine learning models, we can see that in all the q-RASPR models, external validation metrics are improved significantly compared to the QSPR models. Here, we have also performed SHAP analysis on the training set compounds for all the best q-RASPR machine learning models to identify the relative importance of a feature present in the models. The relative importance of a feature in a machine learning model which is obtained from SHAP analysis is represented in the form of summary plots, as shown in **Figure 4.35**. From the summary plots, we can see for phenothiazines, the B07[N-S] descriptor has highest contribution to the predictions made by the models in division 1 and division 3, and for division 2, the X0Av descriptor shows the highest importance. In case of the porphyrin dataset, the F07[C-X] descriptor shows the highest contribution to the predictions for all the three divisions, and for the triphenylamine dataset, the descriptor *RA function (GK)*, *CVact (GK)* and *GD* show the highest contributions for all the three divisions respectively.

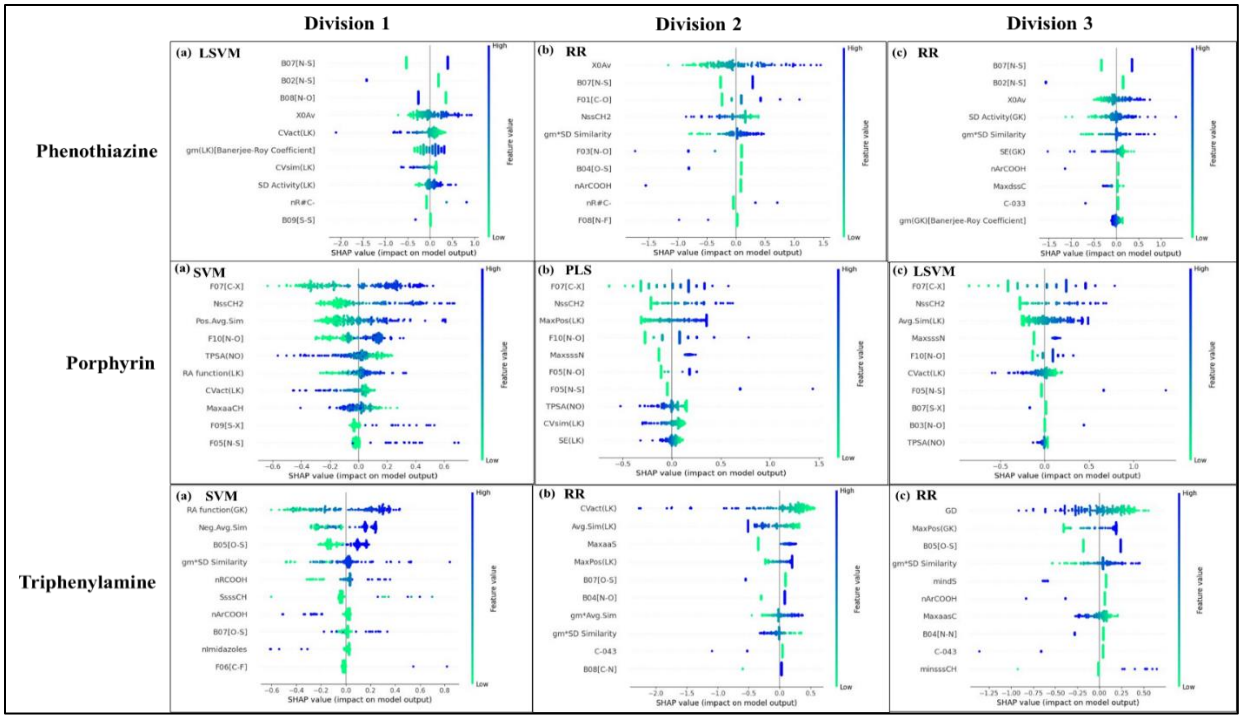


Figure 4.35: Summary plot of q-RASPR machine learning models of phenothiazine, porphyrin and triphenylamine dataset for all the respective divisions.

For all the selected q-RASPR machine learning models, we have developed 3D surface plots to identify under-predictions and over-predictions made by machine learning models, which is shown in **Figure 4.36**. Here, the 3D surface plots are generated by plotting the residual values along the z axis, response value along the y axis and the most important descriptor (determined by the SHAP values) value along the x axis, using the NCSS software (https://www.ncss.com/). The depth of the plot is represented by the z axis (residual value) which is colored based on the different range of values. From this color difference, we can find out under-prediction and over-prediction regions and also find out the reason of such predictions, i.e., whether it is influenced by the response value or the value of a descriptors. In this plot, the red colour indicates the under-predictions made by the models, while the over-predictions are indicated by the violet and blue colours. From this plot, we can find out that for all the datasets. The under-predictions are made by a model when the response values are higher side, while the models give over-predictions when the response values are lower side. For divisions 1 and 3 of the phenothiazine dataset, the under-predictions made by the model is mainly dependent on the response value irrespective of descriptor values, but over-predictions are made by a model when both the

descriptor values and the response values are in the lower side. For division 2 of the phenothiazine dataset, under-predictions are made by a model when both the descriptor values and the response values are in the higher side, and over-predictions are made by a model when both values are in the lower side. For all the divisions of the porphyrin dataset, over-predictions are made by a model with lower values of the descriptor and the response, while under-predictions are made by a model when both of these values are in the higher side. For division 1 of the triphenylamine dataset, under-predictions are made by a model when both the descriptor and response values are in the higher side, while over-predictions are made by the model with lower values of the response and the descriptor. For divisions 2 and 3 of the triphenylamine dataset, under-predictions are made by the models with lower values of the descriptor along with higher values of the response, and over-predictions are made when both descriptor and response values are in the higher side.

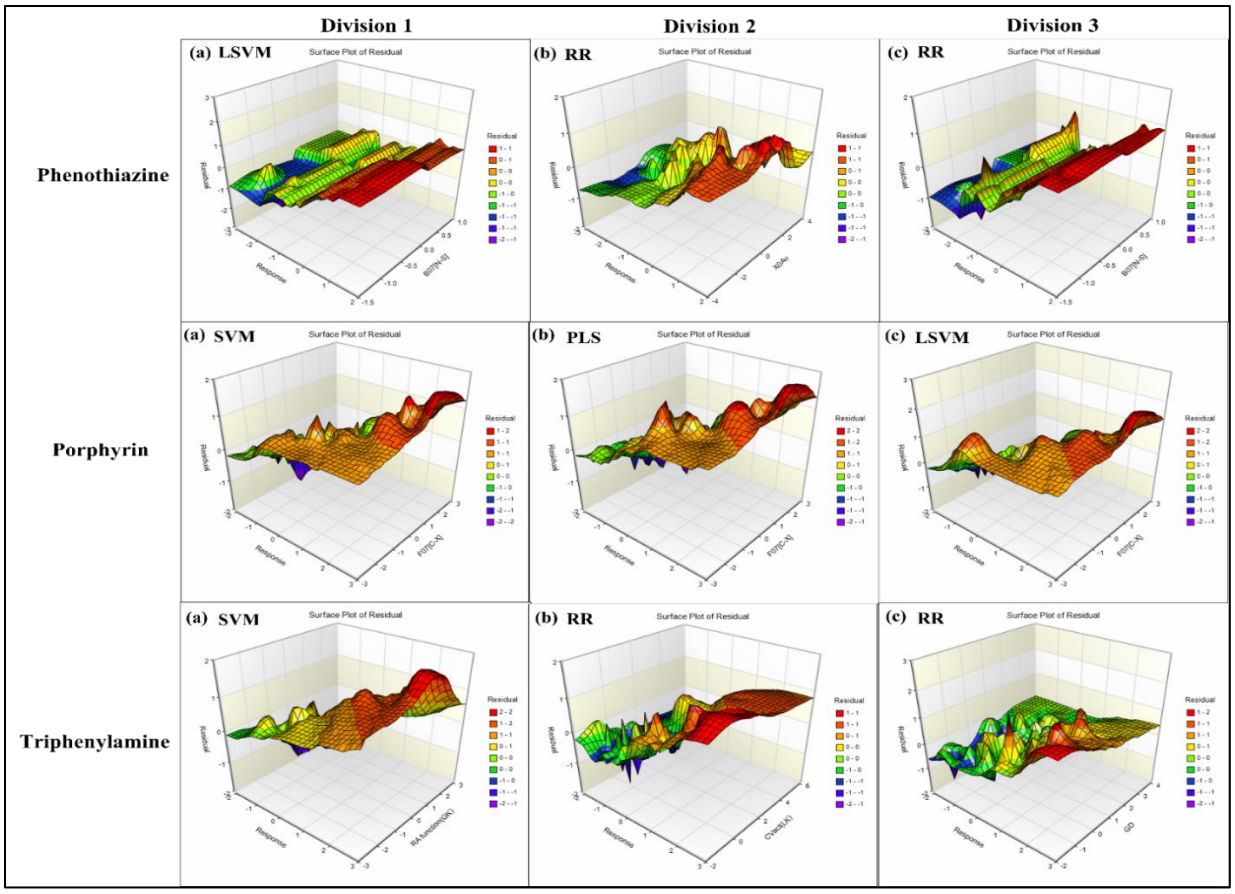


Figure 4.36: 3D surface plots of the q-RASPR machine learning models of the phenothiazine, porphyrin and triphenylamine datasets for all the respective divisions.

4.2.3 Predictions of the removed compounds (True external set)

All the developed machine learning models were used for the prediction of the compounds which were initially removed from the dataset before modeling based on their outlier behavior. The result of the predictions made by the selected machine learning model is represented in **Table 4.12**, in terms of MAE and RMSEP. From the results, we can find out that, in most of the cases MAE and RMSEP are reduced significantly for RASPR method compared to QSPR method, which signifies that the RASPR method helps to improve the external predictivity.

Table 4.12: Results of the predictions made by machine learning models for the true external set.

Dataset	Division	Method	Models	MAE	RMSEP
Phenothiazine	1	RF	QSPR	1.03	1.33
			RASPR	1.02	1.21
		ADB	QSPR	1.03	1.26
			RASPR	1.05	1.27
		GB	QSPR	1.02	1.33
			RASPR	1.02	1.32
		XGB	QSPR	1.04	1.43
			RASPR	1.03	1.24
		SVM	QSPR	1.06	1.34
			RASPR	0.99	1.19
		LSVM	QSPR	1.28	1.58
			RASPR	1.27	1.57
		RR	QSPR	1.24	1.56

			RASPR	1.20	1.51
		PLS	QSPR	1.21	1.54
			RASPR	1.21	1.52
	2	RF	QSPR	0.97	1.26
			RASPR	1.08	1.38
		ADB	QSPR	1.05	1.26
			RASPR	1.06	1.31
		GB	QSPR	0.88	1.19
			RASPR	1.05	1.45
		XGB	QSPR	1.08	1.39
			RASPR	1.18	1.48
		SVM	QSPR	0.99	1.27
			RASPR	1.04	1.34
		LSVM	QSPR	1.14	1.40
			RASPR	1.28	1.49
		RR	QSPR	1.11	1.39
			RASPR	1.22	1.48
		PLS	QSPR	1.14	1.42
			RASPR	1.25	1.50
	3	RF	QSPR	1.01	1.23
			RASPR	1.08	1.27
		ADB	QSPR	1.06	1.28
			RASPR	1.05	1.27
		GB	QSPR	0.93	1.20
			RASPR	1.14	1.36
		XGB	QSPR	1.02	1.32
			RASPR	1.02	1.34
		SVM	QSPR	1.01	1.28
			RASPR	0.97	1.18

		LSVM	QSPR	1.28	1.48
			RASPR	1.25	1.48
		RR	QSPR	1.19	1.49
			RASPR	1.15	1.39
		PLS	QSPR	1.20	1.51
			RASPR	1.17	1.41
Porphyrin	1	RF	QSPR	0.60	0.75
			RASPR	0.49	0.61
		ADB	QSPR	0.58	0.72
			RASPR	0.52	0.61
		GB	QSPR	0.68	0.86
			RASPR	0.46	0.57
		XGB	QSPR	0.57	0.74
			RASPR	0.52	0.65
		SVM	QSPR	0.63	0.73
			RASPR	0.59	0.68
		LSVM	QSPR	0.97	1.43
			RASPR	0.86	1.27
		RR	QSPR	0.95	1.37
			RASPR	0.83	1.23
		PLS	QSPR	0.94	1.32
			RASPR	0.83	1.20
	2	RF	QSPR	0.63	0.78
			RASPR	0.62	0.78
		ADB	QSPR	0.69	0.84
			RASPR	0.71	0.87
		GB	QSPR	0.62	0.80
			RASPR	0.64	0.87
		XGB	QSPR	0.65	0.86

			RASPR	0.65	0.83
		SVM	QSPR	0.58	0.70
			RASPR	0.64	0.76
		LSVM	QSPR	0.89	1.09
			RASPR	0.82	1.01
		RR	QSPR	0.78	0.98
			RASPR	0.75	0.93
		PLS	QSPR	0.81	1.00
			RASPR	0.75	0.92
	3	RF	QSPR	0.66	0.82
			RASPR	0.59	0.74
		ADB	QSPR	0.63	0.75
			RASPR	0.56	0.69
		GB	QSPR	0.62	0.77
			RASPR	0.64	0.78
		XGB	QSPR	0.64	0.79
			RASPR	0.60	0.73
		SVM	QSPR	0.57	0.68
			RASPR	0.52	0.65
		LSVM	QSPR	0.86	1.07
			RASPR	0.71	0.90
		RR	QSPR	0.77	0.98
			RASPR	0.68	0.88
		PLS	QSPR	0.76	0.96
			RASPR	0.67	0.87
Triphenylamine	1	RF	QSPR	0.57	0.72
			RASPR	0.62	0.80
		ADB	QSPR	0.49	0.67
			RASPR	0.60	0.77
		GB	QSPR	0.74	1.06

			RASPR	0.73	0.96
		XGB	QSPR	0.66	0.80
			RASPR	0.76	0.99
		SVM	QSPR	0.63	0.81
			RASPR	0.58	0.75
		LSVM	QSPR	1.00	1.53
			RASPR	1.18	1.81
		RR	QSPR	1.00	1.48
			RASPR	1.06	1.71
		PLS	QSPR	1.01	1.51
			RASPR	1.00	1.51
	2	RF	QSPR	0.64	0.83
			RASPR	0.60	0.77
		ADB	QSPR	0.71	0.84
			RASPR	0.54	0.69
		GB	QSPR	0.66	0.89
			RASPR	0.70	0.96
		XGB	QSPR	0.63	0.81
			RASPR	0.69	0.86
		SVM	QSPR	0.58	0.77
			RASPR	0.59	0.75
		LSVM	QSPR	0.67	0.88
			RASPR	0.59	0.72
		RR	QSPR	0.64	0.82
			RASPR	0.61	0.76
		PLS	QSPR	0.62	0.78
			RASPR	0.59	0.76
	3	RF	QSPR	0.62	0.76
			RASPR	0.49	0.64
		ADB	QSPR	0.60	0.78

			RASPR	0.52	0.68
		GB	QSPR	0.51	0.70
			RASPR	0.42	0.57
		XGB	QSPR	0.58	0.75
			RASPR	0.58	0.69
		SVM	QSPR	0.50	0.70
			RASPR	0.48	0.66
		LSVM	QSPR	0.95	1.29
			RASPR	0.53	0.86
		RR	QSPR	0.91	1.25
			RASPR	0.57	0.87
		PLS	QSPR	0.86	1.16
			RASPR	0.56	0.88

Chapter – 5

CONCLUSION

5. CONCLUSION

Conclusions convey both the success and failure of an experiment's findings. The success of any type of study is determined by the findings and conclusions reached, which may reveal previously unknown or unexplored scientific explanations. These findings may lead to a better understanding and deeper knowledge of the specific area where the studies were conducted. The results and explanations reflect the various concepts included in the work as well as the various methodologies used.

In the present study, we have developed different predictive models using RASPR descriptors, derived from the similarity-based read-across method. The RASPR descriptors are 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 put through a randomization test for validation.

5.1 Machine learning - based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells

Solar energy is one of the most important forms of renewable energy that meets the increasing demand for electrical energy. Various in-silico approaches have been used to predict the power conversion efficiency (PCE) of DSSCs using structural and physicochemical features of dyes. In the present study, we have used RASPR descriptors along with 2D structural descriptors for the development of Partial Least Squares (PLS) and Machine Learning models of coumarin, carbazole, indoline and diphenylamine dyes. The developed models are statistically robust, and all the models show acceptable results for the internal validation and enhanced values of external validation metrics which indicate the importance of RASPR descriptors. We have performed different cross-validation statistics to determine the quality of the developed model. We also performed SHAP analysis on the training sets for all four datasets to determine the feature

importance toward the endpoints. Using the developed and validated models, the PCE of a newly designed molecule can be determined before its synthesis and it can save a lot of time, money, and resources. The software tools used for our work are easy to operate and most of them are freely accessible which makes the modeling exercise simple and inexpensive. Hence, our developed models can give a direction for scientists and researchers present in different research areas or industrial organizations to design and synthesize new dyes. In this way, time, resources, and cost involved in the synthesis and experimentation can be reduced which helps to develop more efficient molecules.

5.2 Application of machine learning-based read-across structure-property relationship (RASPR) as a new tool for predictive modelling: Prediction of power conversion efficiency (PCE) for selected classes of organic dyes in dye-sensitized solar cells (DSSCs)

In this study, we have presented the q-RASPR approach for the prediction of power conversion efficiency of dye-sensitized solar cells. This similarity-based method uses non-linear functions to calculate the similarity between a target compound and selected close source compounds using different structural and physicochemical descriptors. In the q-RASPR method, different similarity and error measures obtained from read-across predictions, along with initial structural and physicochemical descriptors, are used for the development of predictive models. In the present study, three different classes of dyes were used for modeling, and from each dataset, 10 percent of compounds were removed based on high leverage values before modeling due to modelability issues. The reduced datasets were then divided into training and test sets with three different combinations, where the training set is used for statistical model development, and the test is used for model validation. The selection of features was performed using genetic algorithm (GA) and best subset feature selection method for the development of QSPR models. Initially we have developed QSPR partial least squares (PLS) models with these selected features, and then these modelled features were used for the calculation of RASPR descriptors. The q-RASPR PLS models were developed after data fusion between the calculated RASPR descriptors and the initial physicochemical and structural descriptors. All the developed q-RASPR models were statistically robust, and all the models show acceptable values of the external validation metrics to become statistically reliable predictive models. For all the developed q-RASPR PLS models, external validation metrics have improved significantly

compared to the corresponding QSPR PLS method, which signifies the importance of the RASPR descriptors. We have further developed machine learning (ML) models with the modelled descriptors of QSPR and RASPR PLS models, and here also RASPR ML models have higher values of the external validation metrics as compared to their corresponding QSPR ML models. The developed models can be used for the efficient predictions of PCE for newly designed molecules even before their synthesis, which will help to reduce the wastage of time, money, and resources. Hence, the developed models can help different researchers across different fields or different organizations to design and synthesize new dyes. In this way, the cost, time, and resources involved in the experimentation and synthesis of a molecule can be reduced which will help to design a molecule that can show improved performance compared to the previously used molecules.

Chapter – 6

REFERENCES

6. References

1. Makieła, K., Mazur, B. and Głowacki, J., 2022. The impact of renewable energy supply on economic growth and productivity. Energies, 15(13), p.4808.
2. Khezri, M., Karimi, M.S., Mamkhezri, J., Ghazal, R. and Blank, L., 2022. Assessing the impact of selected determinants on renewable energy sources in the electricity mix: The case of ASEAN countries. Energies, 15(13), p.4604.
3. Roosjen, S., Glushenkov, M., Kronberg, A. and Kersten, S., 2022. Waste Heat Recovery Systems with Isobaric Expansion Technology Using Pure and Mixed Working Fluids. Energies, 15(14), p.5265.
4. Sousa, C.C., Martins, J., Carvalho, Ó., Coelho, M., Moita, A.S. and Brito, F.P., 2022. Assessment of an exhaust thermoelectric generator incorporating thermal control applied to a heavy duty vehicle. Energies, 15(13), p.4787.
5. Wilberforce, T., Ijaodola, O., Baroutaji, A., Ogungbemi, E. and Olabi, A.G., 2022. Effect of bipolar plate material on proton exchange membrane fuel cell performance. Energies, 15(5), p.1886.
6. Alaswad, A., Omran, A., Sodre, J.R., Wilberforce, T., Pignatelli, G., Dassisti, M., Baroutaji, A. and Olabi, A.G., 2020. Technical and commercial challenges of proton-exchange membrane (PEM) fuel cells. Energies, 14(1), p.144.
7. Said, R., Bhatti, M.I. and Hunjra, A.I., 2022. Toward Understanding Renewable Energy and Sustainable Development in Developing and Developed Economies: A Review. Energies, 15(15), p.5349.
8. Meirinhos, G., Malebo, M., Cardoso, A., Silva, R. and Rêgo, R., 2022. Information and Public Knowledge of the Potential of Alternative Energies. Energies, 15(13), p.4928.
9. Ahmed, N., Sheikh, A.A., Hamid, Z., Senkus, P., Borda, R.C., Wysokińska-Senkus, A. and Glabiszewski, W., 2022. Exploring the causal relationship among green taxes, energy intensity, and energy consumption in nordic countries: Dumitrescu and Hurlin causality approach. Energies, 15(14), p.5199.
10. Kannan, N. and Vakeesan, D., 2016. Solar energy for future world:-A review. Renewable and sustainable energy reviews, 62, pp.1092-1105.

11. Granqvist, C.G., 2003. Solar energy materials. Advanced Materials, 15(21), pp.1789-1803.
12. Mekhilef, S., Saidur, R. and Safari, A., 2011. A review on solar energy use in industries. Renewable and sustainable energy reviews, 15(4), pp.1777-1790.
13. Şen, Z., 2004. Solar energy in progress and future research trends. Progress in energy and combustion science, 30(4), pp.367-416.
14. Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G. and Shelby, J., 2018. The national solar radiation data base (NSRDB). Renewable and sustainable energy reviews, 89, pp.51-60.
15. Seidlitz, H.K., Thiel, S., Krins, A. and Mayer, H., 2001. Solar radiation at the Earth's surface. Comprehensive series in photosciences, 3, pp.705-738.
16. Hayat, M.B., Ali, D., Monyake, K.C., Alagha, L. and Ahmed, N., 2019. Solar energy—A look into power generation, challenges, and a solar-powered future. International Journal of Energy Research, 43(3), pp.1049-1067.
17. Nayak, P.K., Mahesh, S., Snaith, H.J. and Cahen, D., 2019. Photovoltaic solar cell technologies: analysing the state of the art. Nature Reviews Materials, 4(4), pp.269-285.
18. Lovegrove, K. and Stein, W. eds., 2012. Concentrating solar power technology: principles, developments and applications.
19. Lakatos, L., Hevessy, G. and Kovács, J.J.W.F., 2011. Advantages and disadvantages of solar energy and wind-power utilization. World Futures, 67(6), pp.395-408.
20. Terzioglu, H., Kazan, F.A. and Arslan, M., 2015, April. A new approach to the installation of solar panels. In 2015 2nd International Conference on Information Science and Control Engineering (pp. 573-577). IEEE.
21. Chu, Y. and Meisen, P. (2011) Review and Comparison of Different Solar Energy Technologies. Report of Global
22. Choubey, P.C., Oudhia, A. and Dewangan, R. (2012) A Review: Solar Cell Current Scenario and Future Trends. Recent Research in Science and Technology, 4, 99-101.
23. McEvoy, A., Castaner, L. and Markvart, T. (2012) Solar Cells: Materials, Manufacture and Operation. 2nd Edition, Elsevier Ltd., Oxford, 3-25.
24. Fahrenbruch, A.L. and Bube, R.H. (1983) Fundamentals of Solar Cells. Academic Press Inc., New York.

25. Rappaport, P., 1959. The photovoltaic effect and its utilization. Solar Energy, 3(4), pp.8-18.
26. Palz, W., 2010. Power for the world: the emergence of electricity from the Sun. Pan Stanford Publishing.
27. Guarnieri, M., 2015. More light on information [historical]. IEEE Industrial Electronics Magazine, 9(4), pp.58-61.
28. Castellano, R.N., 2010. Solar panel processing. Archives contemporaines.
29. Kim, S., Hoang, V.Q. and Bark, C.W., 2021. Silicon-based technologies for flexible photovoltaic (Pv) devices: From basic mechanism to manufacturing technologies. Nanomaterials, 11(11), p.2944.
30. Kosyachenko, L.A. ed., 2011. Solar Cells: Silicon Wafer-Based Technologies. BoD–Books on Demand.
31. Khatibi, A., Razi Astaraei, F. and Ahmadi, M.H., 2019. Generation and combination of the solar cells: A current model review. Energy Science & Engineering, 7(2), pp.305-322.
32. Endrös, A.L., 2002. Mono-and tri-crystalline Si for PV application. Solar energy materials and solar cells, 72(1-4), pp.109-124.
33. Bertolli, M., 2008. Solar cell materials. Course: Solid State II. Department of Physics, University of Tennessee, Knoxville.
34. Lan, C.W., Lan, A., Yang, C.F., Hsu, H.P., Yang, M., Yu, A., Hsu, B., Hsu, W.C. and Yang, A., 2017. The emergence of high-performance multi-crystalline silicon in photovoltaics. Journal of crystal growth, 468, pp.17-23.
35. Chopra, K.L., Paulson, P.D. and Dutta, V., 2004. Thin-film solar cells: an overview. Progress in Photovoltaics: Research and applications, 12(2-3), pp.69-92.
36. Imamzai, M., Aghaei, M., Thayoob, Y.H.M. and Forouzanfar, M., 2012, November. A review on comparison between traditional silicon solar cells and thin-film CdTe solar cells. In Proceedings of National Graduate Conference (Nat-Grad (pp. 1-5).
37. Sreejith, S., Ajayan, J., Kollem, S. and Sivasankari, B., 2022. A comprehensive review on thin film amorphous silicon solar cells. Silicon, pp.1-17.
38. Kabir, M.I., Shahahmadi, S.A., Lim, V., Zaidi, S., Sopian, K. and Amin, N., 2012. Amorphous silicon single-junction thin-film solar cell exceeding 10% efficiency by design optimization. International Journal of Photoenergy, 2012.

39. Bonnet, D. and Meyers, P., 1998. Cadmium-telluride—Material for thin film solar cells. Journal of Materials Research, 13, pp.2740-2753.
40. Chu, T.L. and Chu, S.S., 1993. Recent progress in thin-film cadmium telluride solar cells. Progress in Photovoltaics: Research and Applications, 1(1), pp.31-42.
41. Fthenakis, V.M., Morris, S.C., Moskowitz, P.D. and Morgan, D.L., 1999. Toxicity of cadmium telluride, copper indium diselenide, and copper gallium diselenide. Progress in Photovoltaics: Research and Applications, 7(6), pp.489-497.
42. Akinoglu, B.G., Tuncel, B. and Badescu, V., 2021. Beyond 3rd generation solar cells and the full spectrum project. Recent advances and new emerging solar cells. Sustainable Energy Technologies and Assessments, 46, p.101287.
43. Sharma, S., Jain, K.K. and Sharma, A., 2015. Solar cells: in research and applications—a review. Materials Sciences and Applications, 6(12), p.1145.
44. Rühle, S., Shalom, M. and Zaban, A., 2010. Quantum-dot-sensitized solar cells. ChemPhysChem, 11(11), pp.2290-2304.
45. Pan, Z., Rao, H., Mora-Seró, I., Bisquert, J. and Zhong, X., 2018. Quantum dot-sensitized solar cells. Chemical Society Reviews, 47(20), pp.7659-7702.
46. Ludin, N.A., Mahmoud, A.A.A., Mohamad, A.B., Kadhum, A.A.H., Sopian, K. and Karim, N.S.A., 2014. Review on the development of natural dye photosensitizer for dye-sensitized solar cells. Renewable and Sustainable Energy Reviews, 31, pp.386-396.
47. Tian, J. and Cao, G., 2013. Semiconductor quantum dot-sensitized solar cells. Nano reviews, 4(1), p.22578.
48. Sharma, D., Jha, R. and Kumar, S., 2016. Quantum dot sensitized solar cell: recent advances and future perspectives in photoanode. Solar Energy Materials and Solar Cells, 155, pp.294-322.
49. McConnell, R.D., 2002. Assessment of the dye-sensitized solar cell. Renewable and Sustainable Energy Reviews, 6(3), pp.271-293.
50. Nazeeruddin, M.K., Baranoff, E. and Grätzel, M., 2011. Dye-sensitized solar cells: A brief overview. Solar energy, 85(6), pp.1172-1178.
51. Ferber, J., Stangl, R. and Luther, J., 1998. An electrical model of the dye-sensitized solar cell. Solar Energy Materials and Solar Cells, 53(1-2), pp.29-54.

52. Arka, G.N., Prasad, S.B. and Singh, S., 2021. Comprehensive study on dye sensitized solar cell in subsystem level to excel performance potential: A review. Solar Energy, 226, pp.192-213.
53. Hoppe, H. and Sariciftci, N.S., 2004. Organic solar cells: An overview. Journal of materials research, 19(7), pp.1924-1945.
54. Günes, S., Neugebauer, H. and Sariciftci, N.S., 2007. Conjugated polymer-based organic solar cells. Chemical reviews, 107(4), pp.1324-1338.
55. Abdulrazzaq, O.A., Saini, V., Bourdo, S., Dervishi, E. and Biris, A.S., 2013. Organic solar cells: a review of materials, limitations, and possibilities for improvement. Particulate science and technology, 31(5), pp.427-442.
56. Wei, Y., Yu, J., Qin, L., Chen, H., Wu, X., Wei, Z., Zhang, X., Xiao, Z., Ding, L., Gao, F. and Huang, H., 2021. A universal method for constructing high efficiency organic solar cells with stacked structures. Energy & Environmental Science, 14(4), pp.2314-2321.
57. Benanti, T.L. and Venkataraman, D., 2006. Organic solar cells: An overview focusing on active layer morphology. Photosynthesis research, 87, pp.73-81.
58. Weng, K., Ye, L., Zhu, L., Xu, J., Zhou, J., Feng, X., Lu, G., Tan, S., Liu, F. and Sun, Y., 2020. Optimized active layer morphology toward efficient and polymer batch insensitive organic solar cells. Nature communications, 11(1), p.2855.
59. Rivaton, A., Chambon, S., Manceau, M., Gardette, J.L., Lemaître, N. and Guillerez, S., 2010. Light-induced degradation of the active layer of polymer-based solar cells. Polymer Degradation and Stability, 95(3), pp.278-284.
60. Wang, M., Xie, F., Du, J., Tang, Q., Zheng, S., Miao, Q., Chen, J., Zhao, N. and Xu, J.B., 2011. Degradation mechanism of organic solar cells with aluminum cathode. Solar Energy Materials and Solar Cells, 95(12), pp.3303-3310.
61. Wang, K., Li, Y. and Li, Y., 2020. Challenges to the stability of active layer materials in organic solar cells. Macromolecular Rapid Communications, 41(4), p.1900437.
62. Park, H., Shi, Y. and Kong, J., 2013. Application of solvent modified PEDOT: PSS to graphene electrodes in organic solar cells. Nanoscale, 5(19), pp.8934-8939.
63. Hoppe, H. and Sariciftci, N.S., 2004. Organic solar cells: An overview. Journal of materials research, 19(7), pp.1924-1945.

64. Yuan, Y., Huang, J. and Li, G., 2011. Intermediate layers in tandem organic solar cells.
65. Ho, C.H.Y., Kothari, J., Fu, X. and So, F., 2021. Interconnecting layers for tandem organic solar cells. Materials Today Energy, 21, p.100707.
66. Xu, X., Li, Y. and Peng, Q., 2020. Recent advances toward highly efficient tandem organic solar cells. Small Structures, 1(1), p.2000016.
67. Jung, H.S. and Park, N.G., 2015. Perovskite solar cells: from materials to devices. small, 11(1), pp.10-25.
68. Green, M.A., Ho-Baillie, A. and Snaith, H.J., 2014. The emergence of perovskite solar cells. Nature photonics, 8(7), pp.506-514.
69. Raut, K.H., Chopde, H.N. and Deshmukh, D.W., 2018. A review on comparative studies of diverse generation in solar cell. Int. J. Electr. Eng. Ethics, 1, pp.1-9.
70. Zhao, X. and Park, N.G., 2015, November. Stability issues on perovskite solar cells. In Photonics (Vol. 2, No. 4, pp. 1139-1151). MDPI.
71. Lu, Q., Yang, Z., Meng, X., Yue, Y., Ahmad, M.A., Zhang, W., Zhang, S., Zhang, Y., Liu, Z. and Chen, W., 2021. A review on encapsulation technology from organic light emitting diodes to organic and perovskite solar cells. Advanced Functional Materials, 31(23), p.2100151.
72. Zhou, X., Qi, W., Li, J., Cheng, J., Li, Y., Luo, J., Ko, M.J., Li, Y., Zhao, Y. and Zhang, X., 2020. Toward efficient and stable perovskite solar cells: Choosing appropriate passivator to specific defects. Solar RRL, 4(10), p.2000308.
73. Schileo, G. and Grancini, G., 2021. Lead or no lead? Availability, toxicity, sustainability and environmental impact of lead-free perovskite solar cells. Journal of materials chemistry C, 9(1), pp.67-76.
74. Yang, S., Fu, W., Zhang, Z., Chen, H. and Li, C.Z., 2017. Recent advances in perovskite solar cells: efficiency, stability and lead-free perovskite. Journal of Materials Chemistry A, 5(23), pp.11462-11482.
75. Miyasaka, T., Kulkarni, A., Kim, G.M., Öz, S. and Jena, A.K., 2020. Perovskite solar cells: can we go organic-free, lead-free, and dopant-free?. Advanced Energy Materials, 10(13), p.1902500.
76. Sharma, S., Jain, K.K. and Sharma, A., 2015. Solar cells: in research and applications—a review. Materials Sciences and Applications, 6(12), p.1145.

77. Arka, G.N., Prasad, S.B. and Singh, S., 2021. Comprehensive study on dye sensitized solar cell in subsystem level to excel performance potential: A review. Solar Energy, 226, pp.192-213.
78. Sharma, K., Sharma, V. and Sharma, S.S., 2018. Dye-sensitized solar cells: fundamentals and current status. Nanoscale research letters, 13, pp.1-46.
79. Hagfeldt, A., Boschloo, G., Sun, L., Kloo, L. and Pettersson, H., 2010. Dye-sensitized solar cells. Chemical reviews, 110(11), pp.6595-6663.
80. Chen, H.W., Hsu, C.Y., Chen, J.G., Lee, K.M., Wang, C.C., Huang, K.C. and Ho, K.C., 2010. Plastic dye-sensitized photo-supercapacitor using electrophoretic deposition and compression methods. Journal of Power Sources, 195(18), pp.6225-6231.
81. Ye, S., Rathmell, A.R., Chen, Z., Stewart, I.E. and Wiley, B.J., 2014. Metal nanowire networks: the next generation of transparent conductors. Advanced materials, 26(39), pp.6670-6687.
82. Yu, H., Zhang, S., Zhao, H., Will, G. and Liu, P., 2009. An efficient and low-cost TiO2 compact layer for performance improvement of dye-sensitized solar cells. Electrochimica Acta, 54(4), pp.1319-1324.
83. Karim, N.A., Mehmood, U., Zahid, H.F. and Asif, T., 2019. Nanostructured photoanode and counter electrode materials for efficient Dye-Sensitized Solar Cells (DSSCs). Solar Energy, 185, pp.165-188.
84. Gurulakshmi, M., Meenakshamma, A., Susmitha, K., Charanadhar, N., Srikanth, V.V.S.S., Babu, S.N., Subbaiah, Y.V., Venkateswarlu, K. and Raghavender, M.J.S.E., 2019. A transparent and Pt-free all-carbon nanocomposite counter electrode catalyst for efficient dye sensitized solar cells. Solar Energy, 193, pp.568-575.
85. Baheti, A., Justin Thomas, K.R., Li, C.T., Lee, C.P. and Ho, K.C., 2015. Fluorene-based sensitizers with a phenothiazine donor: effect of mode of donor tethering on the performance of dye-sensitized solar cells. ACS applied materials & interfaces, 7(4), pp.2249-2262.
86. Zhang, W., Wu, Y., Zhu, H., Chai, Q., Liu, J., Li, H., Song, X. and Zhu, W.H., 2015. Rational molecular engineering of indoline-based DA-π-A organic sensitizers for long-wavelength-responsive dye-sensitized solar cells. ACS Applied Materials & Interfaces, 7(48), pp.26802-26810.

87. 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.
88. Takahashi, K. and Tanaka, Y., 2016. Materials informatics: a journey towards material design and synthesis. Dalton Transactions, 45(26), pp.10497-10499.
89. Agrawal, A. and Choudhary, A., 2019. Deep materials informatics: Applications of deep learning in materials science. Mrs Communications, 9(3), pp.779-792.
90. Lopez-Bezanilla, A. and Littlewood, P.B., 2020. Growing field of materials informatics: databases and artificial intelligence. MRS Communications, 10(1), pp.1-10.
91. Zhang, L., Miao, J., Li, J. and Li, Q., 2020. Halide perovskite materials for energy storage applications. Advanced Functional Materials, 30(40), p.2003653.
92. Zhang, L.S., Gao, X.L., Liu, X.H., Zhang, Z.J., Cao, R., Cheng, H.C., Wang, M.Y., Yan, X.Y. and Yang, S.C., 2022. CHAIN: unlocking informatics-aided design of Li metal anode from materials to applications. Rare Metals, 41(5), pp.1477-1489.
93. Kireeva, N. and Pervov, V.S., 2020. Materials Informatics Screening of Li-Rich Layered Oxide Cathode Materials with Enhanced Characteristics Using Synthesis Data. Batteries & Supercaps, 3(5), pp.427-438.
94. Cheng, L., Assary, R.S., Qu, X., Jain, A., Ong, S.P., Rajput, N.N., Persson, K. and Curtiss, L.A., 2015. Accelerating electrolyte discovery for energy storage with high-throughput screening. The journal of physical chemistry letters, 6(2), pp.283-291.
95. Takahashi, K. and Takahashi, L., 2022. Data in Materials and Catalysts Informatics. In Machine Learning in Materials Informatics: Methods and Applications (pp. 239-246). American Chemical Society.
96. Takahashi, K., Takahashi, L., Miyazato, I., Fujima, J., Tanaka, Y., Uno, T., Satoh, H., Ohno, K., Nishida, M., Hirai, K. and Ohyama, J., 2019. The rise of catalyst informatics: towards catalyst genomics. ChemCatChem, 11(4), pp.1146-1152.
97. Saeki, A. and Kranthiraja, K., 2020. A high throughput molecular screening for organic electronics via machine learning: present status and perspective. Japanese Journal of Applied Physics, 59(SD), p.SD0801.
98. Vivanco-Benavides, L.E., Martínez-González, C.L., Mercado-Zúñiga, C. and Torres-Torres, C., 2022. Machine learning and materials informatics approaches in the analysis

of physical properties of carbon nanotubes: A review. Computational Materials Science, 201, p.110939.

99. Kalidindi, S.R., Brough, D.B., Li, S., Cecen, A., Blekh, A.L., Congo, F.Y.P. and Campbell, C., 2016. Role of materials data science and informatics in accelerated materials innovation. Mrs Bulletin, 41(8), pp.596-602.
100. Roy, K., Kar, S. and Das, R.N., 2015. A primer on QSAR/QSPR modeling: fundamental concepts. Springer.
101. Ferreira, M.M., 2001. Polycyclic aromatic hydrocarbons: a QSPR study. Chemosphere, 44(2), pp.125-146.
102. 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.
103. 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.
104. 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.
105. 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.
106. Heritage, T.W. and Lowis, D.R., 1999. Molecular hologram QSAR.
107. 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.
108. 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.
109. 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.

110. 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.
111. 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.
112. 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.
113. Le, T.C. and Winkler, D.A., 2018. Applications in Materials Science. Applied Chemoinformatics: Achievements and Future Opportunities, pp.547-569.
114. H. Van de Waterbeemd, R.E. Carter, G. Grassy, H. Kubinyi, Y.C. Martin, M.S. Tute, and P. Willett, Glossary of terms used in computational drug design (IUPAC Recommendations 1997). Pure and Applied Chemistry, 69 (1997), pp. 1137-1152.
115. M. Randić, On characterization of chemical structure. Journal of Chemical Information and Computer Sciences, 37 (1997), pp. 672-687.
116. M. Segall, E. Champness, O. Obrezanova, and C. Leeding, Beyond profiling: using ADMET models to guide decisions. Chemistry & biodiversity, 6 (2009), pp. 2144-2151.
117. 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.
118. 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.

119. 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.
120. 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.
121. 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.
122. 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.
123. 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.
124. 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.
125. 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.
126. Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.
127. Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".
128. 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.
129. Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2020. Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), p.18.

130. Yosipof, A., Shimanovich, K. and Senderowitz, H., 2016. Materials informatics: statistical modeling in material science. Molecular Informatics, 35(11-12), pp.568-579.
131. Krishna, J.G., Ojha, P.K., Kar, S., Roy, K. and Leszczynski, J., 2020. Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy. Nano Energy, 70, p.104537.
132. Manganelli, S. and Benfenati, E., 2016. Use of read-across tools. In Silico Methods for Predicting Drug Toxicity, pp.305-322.
133. Roy, K., Kar, S. and Das, R.N., 2015. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press.
134. Todeschini, R. and Consonni, V., 2008. Handbook of molecular descriptors. John Wiley & Sons.
135. Goodarzi, M., Dejaegher, B. and Heyden, Y.V., 2012. Feature selection methods in QSAR studies. Journal of AOAC International, 95(3), pp.636-651.
136. Shahlaei, M., 2013. Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chemical reviews, 113(10), pp.8093-8103.
137. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. and Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6), pp.463-477.
138. Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.
139. Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), p.160.
140. Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".
141. McDonald, G.C., 2009. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), pp.93-100.
142. Burbidge, R., Trotter, M., Buxton, B. and Holden, S., 2001. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers & chemistry, 26(1), pp.5-14.

143. Chase, R.J., Harrison, D.R., Burke, A., Lackmann, G.M. and McGovern, A., 2022. A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. Weather and Forecasting, 37(8), pp.1509-1529.
144. 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).
145. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S. and Ivanov, P., 2016. Jupyter Notebooks-a publishing format for reproducible computational workflows. Elpub, 2016, pp.87-90.
146. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
147. 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, pp.1013-1026.
148. Moncada-Torres, A., van Maaren, M.C., Hendriks, M.P., Siesling, S. and Geleijnse, G., 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific reports, 11(1), p.6968.
149. Moriwaki, H., Tian, Y.S., Kawashita, N. and Takagi, T., 2018. Mordred: a molecular descriptor calculator. Journal of cheminformatics, 10(1), pp.1-14.
150. Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs, pp.801-820.
151. Bro, R. and Smilde, A.K., 2014. Principal component analysis. Analytical methods, 6(9), pp.2812-2831.
152. Gramatica, P., 2013. On the development and validation of QSAR models. Computational Toxicology: Volume II, pp.499-526.
153. Cai, J., Luo, J., Wang, S. and Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing, 300, pp.70-79.
154. Putavy, C., Devillers, J. and Domine, D., 1996. Genetic selection of aromatic substituents for designing test series. In Genetic Algorithms in Molecular Modeling (pp. 243-269). Academic Press.

155. 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.
156. Vinzi, V.E., Chin, W.W., Henseler, J. and Wang, H., 2010. Handbook of partial least squares (Vol. 201, No. 0). Berlin: Springer.
157. L. Shapley, In: 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317, Classics in Game Theory, (Ed.: H.W. Kuhn), Princeton: Princeton University Press, 1997, 69-79.
158. Holzinger, A., Saranti, A., Molnar, C., Biecek, P. and Samek, W., 2020, July. Explainable AI methods-a brief overview. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 13-38). Cham: Springer International Publishing.
159. Roy, K. and Mitra, I., 2011. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Combinatorial chemistry & high throughput screening, 14(6), pp.450-474.
160. Gangadhar, P.S., Jagadeesh, A., Rajesh, M.N., George, A.S., Prasanthkumar, S., Soman, S. and Giribabu, L., 2022. Role of π-spacer in regulating the photovoltaic performance of copper electrolyte dye-sensitized solar cells using triphenylimidazole dyes. Materials Advances, 3(2), pp.1231-1239.
161. Kalsi, P.S., 2007. Spectroscopy of organic compounds. New age international.
162. Chen, W.C., Nachimuthu, S. and Jiang, J.C., 2017. Revealing the influence of cyano in anchoring groups of organic dyes on adsorption stability and photovoltaic properties for dye-sensitized solar cells. Scientific Reports, 7(1), p.4979.
163. Zhang, L. and Cole, J.M., 2015. Anchoring groups for dye-sensitized solar cells. ACS applied materials & interfaces, 7(6), pp.3427-3455.
164. Li, Y., Liu, J., Liu, D., Li, X. and Xu, Y., 2019. DA-π-A based organic dyes for efficient DSSCs: A theoretical study on the role of π-spacer. Computational Materials Science, 161, pp.163-176.
165. Luo, J.S., Wan, Z.Q. and Jia, C.Y., 2016. Recent advances in phenothiazine-based dyes for dye-sensitized solar cells. Chinese Chemical Letters, 27(8), pp.1304-1318.

166. Yao, Z., Zhang, M., Wu, H., Yang, L., Li, R. and Wang, P., 2015. Donor/acceptor indenoperylene dye for highly efficient organic dye-sensitized solar cells. Journal of the American Chemical Society, 137(11), pp.3799-3802.
167. Banerjee, A., Gajewicz-Skretna, A. and Roy, K., 2023. A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites. Molecular Informatics, 42(4), p.2200261.
168. Yang, Z., Shao, C. and Cao, D., 2015. Screening donor groups of organic dyes for dye-sensitized solar cells. RSC advances, 5(29), pp.22892-22898.
169. Mahmood, A., 2016. Triphenylamine based dyes for dye sensitized solar cells: A review. Solar energy, 123, pp.127-144.
170. Hossain, M.K., Pervez, M.F., Mia, M.N.H., Mortuza, A.A., Rahaman, M.S., Karim, M.R., Islam, J.M., Ahmed, F. and Khan, M.A., 2017. Effect of dye extracting solvents and sensitization time on photovoltaic performance of natural dye sensitized solar cells. Results in physics, 7, pp.1516-1523.

Appendix

REPRINTS

Sustainable Energy & Fuels

ROYAL SOCIETY OF CHEMISTRY

PAPER

Cite this: *Sustainable Energy Fuels*, 2023, 7, 3412

Machine learning-based q-RASPR modeling of power conversion efficiency of organic dyes in dye-sensitized solar cells†

Souvik Pore, Arkaprava Banerjee and Kunal Roy *

Different computational tools are now popularly used as an alternative to experiments for predicting several property endpoints of industrial importance. Recently, read-across and quantitative structure–property relationship (QSPR) have been merged to develop a new modeling technique read-across structure–property relationship (RASPR) which appears to have much potential in predictive modeling. This approach is also promising for modeling relatively smaller data sets as the similarity-based RASPR descriptors are computed from multiple structural and physicochemical features. To understand the potential of RASPR in data gap filling, we have undertaken a case study of modeling Power Conversion Efficiency (PCE) of different classes of organic dyes used in Dye-Sensitized Solar Cells (DSSCs) for renewable energy generation. We have used a large dataset of 429 compounds covering 4 classes of organic dyes. We initially performed read-across analysis using different similarity measures with structural analogues for query compounds and calculated the weighted average predictions. Based on the read-across optimized settings, RASPR descriptors were calculated, and these were then merged with the chemical descriptors, and finally, a single partial least squares (PLS) model was developed for each of the dye classes after feature selection, followed by additional Machine Learning (ML) models. The external prediction quality of the final RASPR models superseded those of the previously developed QSPR models using the same level of chemical information. The important structural features and similarity measures contributing to the PCE have been extracted using the RASPR method which can be used to enhance the PCE values in the newly designed dyes. The RASPR method may also be efficiently applied in modeling other properties of interest in a similar manner.

Received 7th April 2023
Accepted 20th June 2023

DOI: 10.1039/d3se00457k

rsc.li/sustainable-energy

Introduction

Population growth has resulted in a significant rise in energy demand. To address this issue, alternative sources of energy in the form of renewable resources like solar energy, wind energy, geothermal energy, *etc.* should be increasingly used.[1] Solar energy has been one of the most significant forms of renewable energy where the energy coming from the sun in the form of heat, and the radiation is converted into electrical energy by photovoltaic (PV) solar cells.[2] The benefits of assimilating solar energy lie in its free availability, environmental friendliness, and sustainability.[3] PV solar cells have undergone significant changes in their structures and composition since their development. PV solar cells can be classified into four generations based on their structural composition; 1st generation (Silicon Wafer based), 2nd generation (thin film based), 3rd generation (organic material-based), and 4th generation (perovskite-based solar cells).[4] Dye-Sensitized Solar Cells (DSSCs) represent one of the important types of 3rd generation PV solar cells in which different types of organic dyes are used as photosensitizers.[5]

The basic architecture of DSSCs is shown in Fig. 1a. It consists of 7 layers, namely a transparent substance (mainly glass or polymer), a transparent conductive oxide (TCO) layer (mainly Fluorine doped tin oxide (FTO), and Indium doped tin oxide (ITO) are used), a blocking layer (ZnO, In_2O_3, MgO, *etc.*), a semi-conductive oxide (SCO) layer (mainly TiO_2) coated with photoactive dye, electrolyte solution and a counter electrode (Pt). In DSSCs, transparent substance – TCO layer – blocking layer – SCO layer – dye together form a photoanode (PA), and counter electrode – TCO layer – transparent substance is united to form the cathode. Electrolytes like iodide/triiodide (I^-/I_3^-) solution is used for the preparation of DSSCs, where the electrolytes play an important role in the regeneration of dye by redox reaction.[6] In DSSCs, the electrons are generated when the

Drug Theoretics and Chemoinformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, 188 Raja S C Mullick Road, 700032, Kolkata, India. E-mail: kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in; Fax: +91-33-2837-1078; Tel: +91 9831594140

† Electronic supplementary information (ESI) available: SI-1 contains some details of different ML methods, different PLS model plots, cross-validation plots and SHAP plots. SI-2 contains raw data files in the Excel format. SI-3 contains SHAP partial dependence plots and a list of new designed dyes. See DOI: https://doi.org/10.1039/d3se00457k

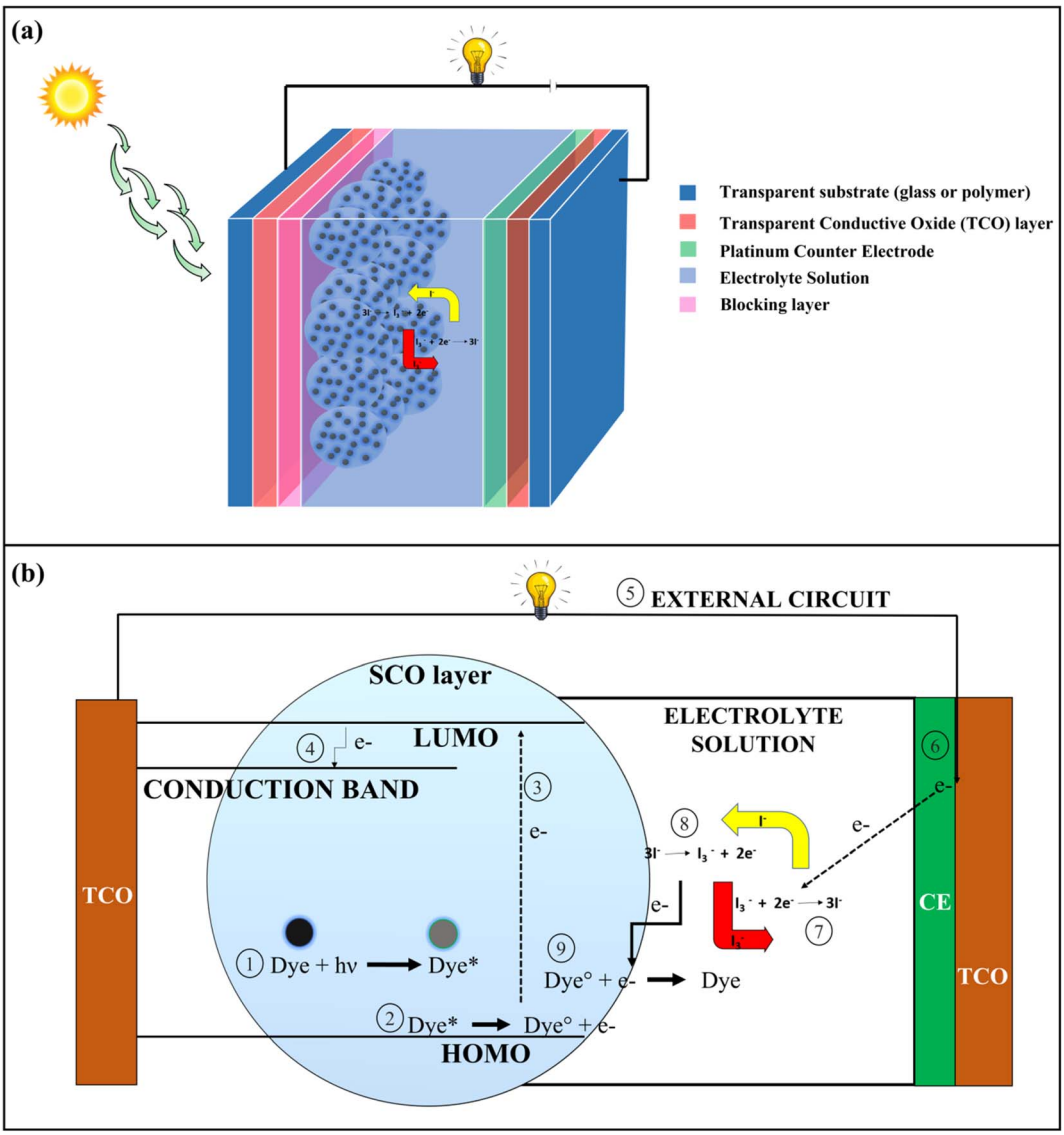


Fig. 1 (a) Basic structure of a Dye-Sensitized Solar Cells (DSSCs) (b) Mechanism of electricity generation in DSSCs.

dye undergoes photoexcitation by absorbing radiation coming from the sun. The electrons are excited to the lowest unoccupied molecular orbital (LUMO) from the highest occupied molecular orbital (HOMO), and subsequently, electrons are transported to the TCO layer by the conduction band of the nanostructured SCO layer. From the TCO layer, the electrons flow through the external circuit and get collected at the platinum counter electrode site. The electrons are then transferred to the HOMO of dyes for their regeneration by the redox reaction of electrolytes which is catalyzed by a platinum counter electrode.[6,7] The whole process for the generation of electrons is shown in Fig. 1b.

In DSSCs, the dye is the key element for the generation of solar power, because it controls photon harvesting and electron generation.[7] The dyes used in DSSCs can be classified into two groups namely metal-based inorganic dyes and metal-free organic dyes. The latter types are preferred due to having a low production cost, synthetically feasible, environment friendly, and easy to modify structure.[8] Most of the metal-free organic dyes have donor–π–acceptor (D–π–A) type structural configuration in which conjugated π-systems like polyenes and oligothiophenes act as π spacers and have a rod-like configuration for the effective intramolecular charge transfer (ICT) by photoexcitation. The donor units are composed of different aromatic moieties like coumarins, triphenylamines, and porphyrins while the acceptor end contains structures like carboxylic acids and cyanoacrylic acids.[7,8] The organic dyes have lower solar power conversion efficiency (PCE) as compared to the metal-based inorganic dyes due to the poor absorption at red and near-infrared spectrum of solar radiation, charge recombination at semi-conductive oxide layer surface and aggregation of dyes.[8] In the recent past, different types of structural modifications have been performed to increase the absorption of solar radiation and PCE, like increasing the

electron-donating ability of the donor and π-spacer by introducing an electron-donating group or increasing the electron-accepting ability of the acceptor by introducing the electron withdrawing group or increasing the length of π-spacer.[9] Therefore, by altering the structures, it is possible to generate new dyes with higher PCE values while maintaining the same properties for all other performance-controlling factors. For designing a new dye molecule, a well-known scheme should be developed and checked before the synthesis of the molecule.

In the last few years, due to the low cost in computational methods and faster generation of results, *in silico* approaches have been extensively used to explore molecules to determine their properties. *In silico* approaches help to identify the active structural moieties responsible for the desired property and thus reduce synthetic complexities.[10–12] Different types of *in silico* approaches like Quantitative Structure–Property Relationship (QSPR),[13,14] Read-Across (RA)[15,16] and various Machine learning (ML)[17–20] methods are being used in the field of materials science. QSPR is a method that represents a mathematical relationship between the chemical structure and the property, and are developed based on the Organization for Economic Co-operation and Development (OECD) principles.[13,14] Read-across (RA) is a similarity-based algorithm that predicts the response value of the query compounds by utilizing the similarity values of its close congeners, and this method is a potential alternative to the QSPR approach where lower number of data points are available.[15,16] ML is a subset of Artificial Intelligence (AI) that enables machines to learn from previous data and improve their performance.[17–20]

In the recent past, various *in silico* studies have been conducted to explore the different classes of organic dyes.[21–31] A cascaded QSPR model was developed by Li *et al.*[29] using quantum chemical molecular descriptors in which combined quantum chemical calculation and machine learning methods were used to establish a relationship between PCE and molecular structures of different organic dyes. The PCE of phenothiazine-containing DSSCs was modeled by Kumar and Kumar[24] using the CORAL software employing hybrid descriptors resulting from the combination of SMILES and hydrogen-suppressed graph (HSG). Combined QSPR modeling and quantum chemical analysis were performed by Kar *et al.*[21] for 273 arylamine organic dyes to understand the electron transfer mechanism and photo-physical properties of dye. Venkatraman *et al.*[31] developed a QSPR model for different phenothiazine derivatives using different structural descriptors and eigenvalue (EVA) descriptors obtained from vibrational frequencies. Krishna *et al.*[26] developed multiple Partial Least Squares (PLS) QSPR models for 1200 organic dyes of 7 classes, in order to know the important structural features contributing to higher PCE values. In the study, they have also designed 10 coumarin dyes using important structural feature obtained from the coumarin model with % PCE ranging from 8.93 to 10.62. Venkatraman *et al.*[30] designed 5 novel phenothiazine dyes by the *de novo* design method using QSPR analysis and all new dyes show PCE over 9%. Kar *et al.*[28] developed a QSPR model to establish the relationship between PCE and quantum chemical descriptors calculated from density functional theory (DFT) and time-dependent DFT (TD-DFT) methods to understand the basic electron transfer mechanism for arylamine-organic dye sensitizers. Seven indoline-based dyes with D–A–π–A molecular configuration designed using QSPR analysis were explored by Roy *et al.*[27] using density functional theory (DFT) and time-dependent DFT (TD-DFT) methods to understand the different optoelectrical properties of dyes used in DSSCs. A QSPR model was proposed by Wen *et al.*[22] which was obtained by combining the machine-learning approaches and computational quantum chemistry method and was used for virtual screening and to check the synthetic accessibility of the different organic dyes. *In silico* methods are thus important not only for the prediction of PCE values but also to explore the important structural and physicochemical properties of dyes that control the performance of DSSCs before synthesis of the dyes to save time, money, and resources.

In the present work, we have adopted a novel Quantitative Read-Across Structure–Property Relationship (q-RASPR) approach, which is analogous to the Quantitative Read-Across Structure Activity Relationship (q-RASAR) first reported by Banerjee and Roy,[32,33] to generate different predictive models for the PCE using a wide array of compounds from 4 different classes. The q-RASPR is a supervised machine learning (ML) approach and is a combination of Read-Across and QSPR. Compound specific similarity and error-based measures were used as RASPR descriptors and combined with the initial descriptors to generate different predictive models.[15,16,32] Different ML approaches in the form of Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (or XGBoosting), Support Vector Machine (SVM), Linear Support Vector Machine (Linear SVM), Ridge Regression (RR) and Partial Least Squares (PLS) models were adopted to predict the PCEs of organic dye-based DSSCs.

Materials and methods

Data collection

The current work deals with q-RASPR modeling of different organic dyes used in DSSCs for which experimental PCE and descriptor values are collected from the previous literature.[26] To demonstrate the performance of the q-RASPR approach, in comparison to the previous QSPR models using the same combination of original descriptors and similarity descriptors computed from them (same level of chemical information), we have collected a total of 429 compounds distributed across 4 representative datasets, *i.e.* coumarins, carbazoles, indolines, and diphenylamines. One of the primary objectives of the current work is to evaluate the novel q-RASPR approach for the enhancement of the quality of predictions of the power conversion efficiency of DSSCs. As shown by Banerjee *et al.* (2023),[34] q-RASPR models might be very useful even when the data set size is small as the similarity-based descriptors derived from the original structural and physicochemical variables act like latent variables thus allowing the usage of a greater amount of chemical information while using a lower number of regressing variables thus maintaining a more favorable degree of freedom. This is why we have taken here into consideration

specifically the dye families with a relatively smaller number of data points. The original training and test sets were retained for the q-RASPR modeling analysis aiming at comparing with the previously developed QSPR model by Krishna *et al.*[26] We used the same combination of structural and physicochemical descriptors as used in the original analysis. The dataset in each case was prepared by combining the descriptors of five individual models of the previous work and removing the common descriptors of these models. The logarithmic conversion of the endpoint parameter PCE was not required as it represents the performance of the solar cell. The datasets contain 56 coumarin dyes (42 training and 14 test compounds), 35 diphenylamine dyes (25 training and 10 test compounds), 179 carbazole dyes (125 training and 54 test compounds), and 159 indoline dyes (121 training and 38 test compounds). These datasets were used as input files for the calculation of the RASPR descriptors. The list of original structural and physicochemical descriptors used for read-across similarity computations (Table S1†) and the list of similarity descriptors used in q-RASPR model development (Table S2) are given in ESI SI-1.†

Read-across hyperparameter optimization

Read-across is a data-gap filling method in which information of one or more chemicals is/are used to predict the endpoint of a target chemical. In read-across, structural similarity between a source compound (or a training compound) and a target compound (or a test compound) is used for the calculation of the endpoint value.[15,16,35] A java-based tool Read-Across-v4.1 (available from **https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home**) was used for the computation of read-across-based predictions. This tool utilizes 3 different similarity-based methods, *i.e.*, Euclidean distance (ED)-based similarity, Gaussian kernel (GK)-based similarity and the Laplacian kernel (LK)-based similarity methods for the computation of the predictions for the query compound(s).[16] In compliance with the theory associated with Machine Learning, there is a need for the optimization of the hyperparameters associated with the three different similarity-based approaches. The Euclidean distance-based predictions require the optimum number of close source compounds, the Gaussian kernel-based predictions require the optimum value of σ and the number of close source compounds while the Laplacian kernel-based predictions require the optimum value of γ and the number of close source compounds.

To select the optimum values for σ, and γ, the training set is randomly divided into 5 sub-training and sub-test sets by the sorted response-based division algorithm.[36] Read-across-based predictions and validation metrics were calculated using these 5 sub-training and sub-test sets for each value of σ, γ, and CTC; and the average of external validation metrics for the subtest sets of 5 divisions was taken. The selection of the optimum σ and γ depends on whether the ${Q_{F1}}^2$ value (subtest set) is maximum for the GK and LK methods, respectively, and then the same values of σ and γ were applied for the original training and test sets at each value of close source compounds (CTC) between 2 to 10. The CTC value which corresponds to the maximum ${Q_{F1}}^2$ value (subtest set) in the ED approach is selected. These optimized settings of the σ, γ and CTC values were used for the computation of the RASPR descriptors. Note that the subtest sets are derived from the training set itself and different from the actual test set.

RASPR descriptor calculation

As defined by Todeschini and Consonni, a descriptor is "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiments".[37] In the present work, we have used different structural and physicochemical descriptors along with RASPR descriptors to develop predictive models. Different similarity and error-based measures obtained from the read-across-based predictions were used as RASPR descriptors.[38] We have used RASAR-Desc-Calc-v2.0 (available from **https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home**) for RASPR descriptor calculation for both training and test sets. The calculation of the RASPR descriptors for the training set involves the Leave-same-out (LSO) algorithm[38] which does not take the identical compound among the list of close source compounds. The optimized similarity-based method and the associated optimized hyperparameters were used for the computation of the RASPR descriptors. These RASPR descriptors were then combined with the initial structural and physicochemical descriptors and after subsequent feature selection based on cross-validation, q-RASPR models were generated. The optimized hyperparameter settings along with the similarity-based approach used for the calculation of the RASPR descriptors are shown in Table 1.

Feature selection and PLS model development

In the present work, we have used the Best Subset Selection (BSS) method to identify the significant descriptors or features that can influence the performance of DSSCs. Feature selection is important to reduce model noise and identify the significant descriptors for PCE in order to lower the risk of overtraining or overfitting.[39,40] Significant descriptors for the models were identified by using the tool Best Subset Selection v2.1 (available from **https://dtclab.webs.com/software-tools**) using the cross-validation statistics.

Best Subset Selection (BSS) is an algorithm that helps to identify the best descriptor combinations by developing models using a specific number of descriptor subset of input descriptors.

Table 1 Optimized hyperparameter settings and methods used for RASPR descriptor calculation[a]

Datasets	Method	σ	γ	CTC
Coumarins	GK	1.75	—	8
Carbazoles	GK	2	—	5
Indolines	LK	—	0.5	5
Diphenylamine	LK	—	0.5	2

[a] GK = Gaussian kernel, LK = Laplacian kernel, CTC = Close training compound.

This algorithm generates models using every possible combination of the descriptors, and the best combination is selected based on different internal validation metrics. Best Subset Selection is actually a grid search that identifies all possible combination of models from a given number of descriptors but the filters in the form of inter-correlation cut-off (<0.6) and R^2 cut-off (>0.5) makes it an "intelligent grid search" which shows only the significant models. The number of descriptors in the models was selected based on the cross-validation $Q_{LOO}{}^2$ score, and after that, we have developed several individual models for each data set, and the best models showing acceptable internal and external validation statistics are reported. For the present work, we have used the final models with 8 descriptors for coumarins, 5 descriptors for diphenylamines, 6 descriptors for indolines and 8 descriptors for carbazoles.

Partial Least Squares (PLS) is a generalized form of the multiple linear regression (MLR) that can be applied for collinear, correlated and noisy data containing multiple *X* variables (or descriptors) and one or more *Y* variable(s) (or endpoint(s)). The main idea behind PLS is to derive latent variables (LVs) *T* (or *X*-scores) and *U* (or *Y*-scores) from descriptors and response variables, respectively. These *X*-scores are then used to predict *Y*-scores which in turn are used to calculate the response.[36] Here, we have used PLS_SingleY_1.0_14May2020 tool (available from **https://dtclab.webs.com/software-tools**) for the development of PLS models of selected descriptors. We have also compared the derived PLS models with other machine-learning models obtained using the same feature combinations.

For the purpose of interpretation and explanation of individual descriptors, different PLS plots were generated using SIMCA-P v10.0 software (**https://www.sartorius.com/**). We have generated the score plots (individual compounds are defined in the LV space and show their distribution and similarity among compounds), the loading plots (loading of all descriptors among the plotted first two LVs, and distance from the origin denote the importance of these descriptors), the *Y*-randomization plot (plot developed by plotting R^2 and Q^2 value of random models (*Y* axis) *vs.* correlation coefficient between observed PCE and permuted PCE), scatter plots (plots of predicted PCE (*Y* axis) *vs.* observed PCE (*X* axis)) and the variable importance plots (in the form of bubble plots).

Machine-learning (ML) models

Machine learning (ML) is a part of artificial intelligence which enables machines to learn from its past data and improve performance based on past experiences for the future aspect.[41] In ML, machines are trained with a large amount of data and a suitable algorithm to accomplish a job. This trained algorithm is then applied to a query data set for reliable and accurate predictions. ML can be classified into 3 main groups: supervised (the labeled data is used to train the machine), unsupervised (the training data is not labeled) and reinforcement (feedback-based method, where the learning agents get a reward or penalty based on its action).[42,43] In the present study, we have performed regression analysis using different supervised learning methods namely ridge regression (RR),[44,45] linear support vector machine (LSVM),[46] support vector machine (SVM),[47] random forest (RF),[47] gradient boosting (GB),[44] and extreme gradient boost or XGBoost (XGB).[48] Some details of these methods are given in ESI SI-1.†

All the above-mentioned machine-learning models were developed using Anaconda Navigator software (version 2022.05) in Jupyter Notebook IDE (version 6.4.8)[49] with python 3.10.4 64-bit. Different python-based modules were used such as numpy (version 1.23.5), pandas (version 1.5.2), Scikit-learn (version 1.2.0), matplotlib (version 3.5.1) and xgboost (version 1.7.1) for model development. For all the machine-learning models, we have used the same inputs as used for the PLS model development and optimized all the hyper-parameters by the cross-validation method using GridSearchCV function of Scikit-learn. For ML modeling, we standardized the descriptors and endpoints values based on the training set mean and standard deviation which were then used as the input.

In this work, we developed machine learning models using moderate sized data sets and a sufficient number of compounds for the validation of the models. Here, we optimized the hyper-parameters using the GridSearchCV method which is basically a cross-validation method in which the training set is divided into five-folds, and four folds are used to build the models each time when the remaining fold is used to validate the model. After building machine learning models, we calculated various cross-validation metrics to check that the models are not over-fitted. The hyperparameter setting was chosen based on the best cross-validation statistics from the five-fold CV data. Again, a small difference between Mean Absolute Error (MAE) values for the training and test sets also indicates that the generated models are not overfitted.

Statistical quality and validation metrics

Validation of a model is important to justify the model quality and to determine its further application.[36] There are different statistical metrics available to check the model quality, goodness-of-fit, robustness, reliability and predictivity. The model quality is checked by the determination coefficient (R^2) and the explained variance ($R^2_{(adj)}$). The model quality is increased when its value become closer to 1. Model validation metrics can be classified into two groups – the internal validation metrics and the external validation metrics. Internal validation is performed only on the training set to check the goodness-of-fit and robustness of the model while the predictivity of a model is checked by an external validation performed on the test set. The robustness and goodness-of-fit of a model are checked by the internal validation metrics $Q_{LOO}{}^2$ (leave-one-out correlation coefficient) performed on the training set. The original dataset is divided into a training set and a test set; the test set is used to check the predictivity of a model by calculating external validation metrics like $Q_{F1}{}^2$, $Q_{F2}{}^2$ and MAE_{test}.[36] The validation metrics which demonstrated the quality of our PLS and ML models are shown in eqn (1)–(5).

$$R^2 = 1 - \frac{\sum \left(Y_{obs(train)} - Y_{cal(train)}\right)^2}{\sum \left(Y_{obs(train)} - \overline{Y}_{train}\right)^2} \quad (1)$$

$$Q_{LOO}^2 = 1 - \frac{\sum \left(Y_{obs(train)} - Y_{pred(train)}\right)^2}{\sum \left(Y_{obs(train)} - \overline{Y}_{train}\right)^2} \quad (2)$$

$$Q_{F1}^2 = 1 - \frac{\sum \left(Y_{obs(test)} - Y_{cal(test)}\right)^2}{\sum \left(Y_{obs(test)} - \overline{Y}_{train}\right)^2} \quad (3)$$

$$Q_{F2}^2 = 1 - \frac{\sum \left(Y_{obs(test)} - Y_{cal(test)}\right)^2}{\sum \left(Y_{obs(test)} - \overline{Y}_{test}\right)^2} \quad (4)$$

$$MAE_{test} = \frac{\sum \left|Y_{obs(test)} - Y_{cal(test)}\right|}{n_{test}} \quad (5)$$

Here, $Y_{obs(train)}$ = observed response values of the training set, $Y_{cal(train)}$ = calculated response values of the training set, $\bar{Y}_{train}$ = mean observed response value of the training set, $Y_{pred(train)}$ = LOO predicted response values of the training set, $Y_{obs(test)}$ = observed response values of the test set, $Y_{cal(test)}$ = calculated response values of the test set, $\bar{Y}_{test}$ = mean observed response value of the test set, n_{test} = the number of observations in the test set.

A model is considered to be well predictive if the values of Q_{F1}^2 and Q_{F2}^2 cross the threshold limit of 0.5 and MAE_{test} attains a minimum value.[36]

SHAP (SHapley additive exPlanation) analysis

The SHAP analysis is performed to identify the global and local contributions of each feature or descriptor for the predictions. By using SHAP, we can determine the feature's contribution in case of complex machine-learning models. SHAP uses the Shapley values to determine the feature contributions, the concept coming from the fair distribution in a cooperative game based on the player's importance.[50] The Shapley values determine the contributions of different features (by the magnitude of the Shapley values) and direction (sign). The positive sign indicates a positive contribution to the predictions and the negative sign indicates a negative contribution to the predictions. The Shapley value for each feature is calculated using the following formula:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S)\right] \quad (6)$$

where ϕ_i is the Shapley value for each feature, $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ is the model output for a subset of features including a particular feature, $f_S(x_S)$ is the model output for the subset of features without that feature, F is the number of input features and S is the number of features in a subset.[50–52]

The complete workflow for the current work is shown in the Fig. 2.

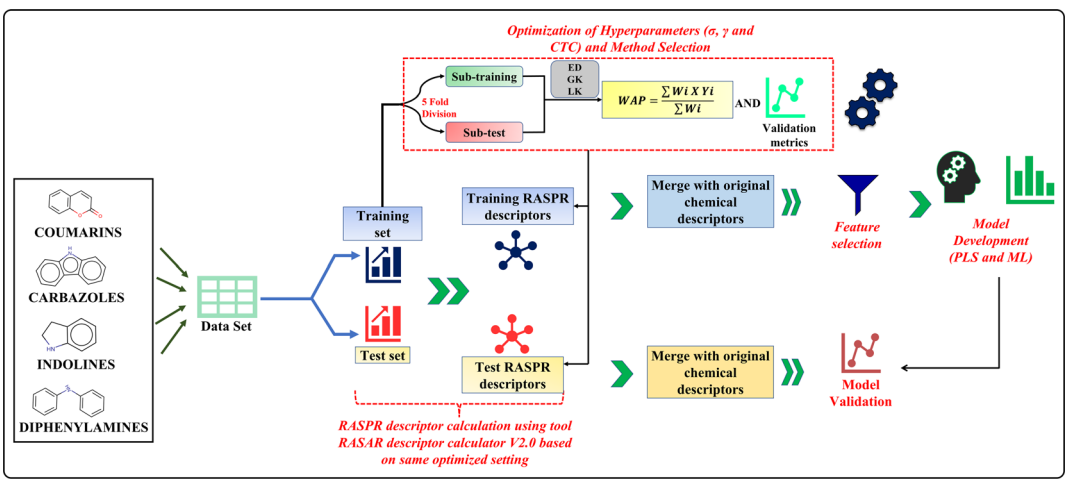
Result and discussion

We have initially developed PLS models for each of the considered data sets and then compared the quality of these models to ML-derived model predictions.

Partial least squares (PLS) models and interpretation of modeled descriptors

Most significant and statistically robust PLS models for different categories of dyes along with their quality in terms of different internal and external validation metrics are shown in Table 2. These models were developed with 8, 8, 6 and 5 descriptors for coumarins, carbazoles, indolines and diphenylamines, respectively, and all these models consist of both RASPR and 2D structural descriptors. The models were developed using 7, 3, 3 and 3 latent variables (LVs), respectively based on the leave-one-out Q^2 values. All the developed models satisfy the threshold limit required to become robust, reliable and good predictivity.[53]

The performance of the q-RASPR models toward the training set is in general inferior compared to the test set due to the



RASPR = Read-across structure-property relationships, PLS = Partial least squares, ML = Machine learning, CTC = Close training compound, ED = Euclidean distance, GK = Gaussian kernel, LK = Laplacian kernel, WAP = Weighted average predictions

Fig. 2 Schematic representation of complete work for q-RASPR and ML model development.

Table 2 Developed q-RASPR PLS models for different type of organic dye used in DSSCs[a]

Types of organic dyes	PLS models	Training set metrics		Test set metrics		
		R^2	$Q_{LOO}{}^2$	$Q_{F1}{}^2$	$Q_{F2}{}^2$	MAE_{test} (95%) (non-standardized)
Coumarins (LV = 7)	PCE = −1.71195 + 0.60957 × SD Activity(GK) + 0.94835 × MaxPos(GK) − 0.75671 × *n*RCN + 1.07215 × *n*Thiophenes − 1.01363 × *n*R#C + 1.18272 × *n*R = Ct − 0.12347 × T(S···S) + 0.76919 × *C* − 040	0.75	0.63	0.72	0.70	0.75
Carbazoles (LV = 3)	PCE = −0.23418 + 1.26064 × Avg.Sim(GK) − 1.51529 × F06[N–N] + 0.92434 × *n*R10 + 0.19841 × F04[C–N] + 2.12686 × B04[N–O] − 0.4133 × N% + 2.35133 × F06[N–O] + 1.42348 × B02[C–S]	0.71	0.66	0.77	0.76	0.61
Indolines (LV = 3)	PCE = 1.52408 + 0.88535 × RA function(LK) − 0.89273 × CV sim(LK) − 0.92139 × Neg. Avg. Sim + 0.01956 × F04 [C–N] + 0.70912 × B09[O–S] − 0.05307 × *n*Cconj	0.63	0.59	0.81	0.81	0.55
Diphenylamines (LV = 3)	PCE = 1.28039 + 0.8856 × RA function(LK) + 1.53133 × SD similarity(LK) − 0.14367 × F01[C–N] − 0.15417 × StsC + 0.35804 × F04[N–S]	0.83	0.73	0.90	0.90	0.62

[a] LV = Latent variables.

algorithm of the RASPR descriptor calculation (Tables 2 and 3). For the calculation of RASPR descriptors for the training set, the algorithm works based on the "Leave Same Out (LSO) method"[3,4] where identical compounds are not considered during the finding of close source compounds to avoid overfitting. In the case of any QSAR modeling study, chemical or physicochemical descriptors of a training compound are computed based on the structure or property of that particular compound. However, RASPR descriptors of a particular training compound are computed not from that particular compound, but from its close congeners based on the similarity features. Thus, the prediction aspect is in-built in the case of RASPR descriptor computation. A QSAR model is fitted based on the training set descriptor data while a RASPR model is fitted based on the leave-same-out "predicted" training set descriptor data. Again, during PLS model development, the number of components (LVs) of a PLS model is selected based on the cross-validation (Leave-One-Out (LOO) method). Due to the combined effect of leave-same-out descriptor computation followed by LOO cross-validation, q-RASPR models show inferior performance on the training data than on the test set data. Further details on this aspect are given while discussing other machine learning models (*vide infra*).

The importance of the features toward the PCE is represented in the form of the bubble plot (Fig. S1 in ESI SI-1†), in which variable importance scores and coefficient scores are calculated by using SIMCA-P v10.0 software (**https://www.sartorius.com/**). The importance of these descriptors is represented by the diameter of the bubbles and their relative position along the *y*-axis whereas color difference denotes positive and negative contribution. The information related to all the datasets are provided in the ESI SI-2.†

Modelling of PCE of coumarin dyes. The most significant descriptors in the form of a mathematical equation are shown in Table 2 and the contribution of these descriptors towards PCE in the form of a bubble plot in Fig. S1a in ESI SI-1.† The mechanistic interpretation of these descriptors is discussed below. It may be noted that the predictions made by the models do not depend on a single descriptor; instead, they are the resultant of many positively and negatively contributing features. Here, we give suitable examples to show how a particular descriptor can influence the performance of the model but there may be some other examples where the descriptor contribution is not so obvious, and some other important descriptors might contribute to the response for those data points.

MaxPos(GK) is a RASPR descriptor that represents the similarity value to the nearest positive close source compound based on training set mean, obtained by Gaussian kernel similarity-based method.[32] From the bubble plot, it was found that this descriptor has the highest contribution to the PCE, as shown in Fig. S1a.† MaxPos(GK) shows a positive contribution as reflected in following example: **19** (MaxPos(GK) = 1, PCE = 7.4), **20** (MaxPos(GK) = 1, PCE = 6.4) and *vice versa* for the dyes **56** (MaxPos(GK) = 0.014, PCE = 0.99), **22** (MaxPos(GK) = 0.004, PCE = 0.33). Any QSPR-derived predictions are based on the similarity assumptions; *i.e.*, structurally similar compounds will have similar property or activity values. Thus, it is obvious that

Table 3 Comparison of the quality of PLS and other machine learning models

		Training set metrics						Test set metrics		
		Model statistics	Cross-validation statistics					Prediction statistics		
Datasets	Methods	R^2	MAE_{LOO}	MAE ± SEM (20 times 5-fold CV)	r2 ± SEM (20 times 5-fold CV)	MAE ± SEM (1000 times ShuffleSplit CV)	r2 ± SEM (1000 times ShuffleSplit CV)	MAE_{test}	${Q_{F1}}^2$	Optimized hyperparameters
Coumarins	PLS	0.75	0.49	0.54 ± 0.015	0.41 ± 0.053	0.56 ± 0.004	0.44 ± 0.011	0.45	0.72	n_components:7
	RR	0.74	0.51	0.54 ± 0.011	0.46 ± 0.036	0.57 ± 0.003	0.48 ± 0.007	0.44	0.73	'Alpha': 1.0
	LSVM	0.72	0.52	0.60 ± 0.017	0.27 ± 0.077	0.62 ± 0.005	0.32 ± 0.015	0.49	0.67	'C': 1.0, 'max_iter': 1000
	SVM	0.74	0.66	0.67 ± 0.013	0.17 ± 0.053	0.67 ± 0.004	0.27 ± 0.009	0.5	0.68	'C': 1.0, 'degree': 2, 'gamma': 'auto'
	RF	0.7	0.58	0.61 ± 0.013	0.23 ± 0.062	0.61 ± 0.004	0.35 ± 0.009	0.47	0.68	'max_depth':2, 'min_samples_leaf':2, 'min_samples_split':2, 'n_estimators':200
	GB	0.85	0.57	0.62 ± 0.014	0.18 ± 0.065	0.62 ± 0.003	0.32 ± 0.009	0.55	0.58	'max_depth':2, 'min_samples_leaf':3, 'min_samples_split':3, 'n_estimators':50
	XGB	0.75	0.49	0.54 ± 0.014	0.42 ± 0.050	0.56 ± 0.004	0.44 ± 0.011	0.45	0.72	'booster':'gblinear', 'learning_rate':1.0, 'max_depth': none, 'n_estimators':90
Carbazoles	PLS	0.71	0.47	0.47 ± 0.006	0.62 ± 0.012	0.48 ± 0.002	0.62 ± 0.004	0.31	0.77	n_components:3
	RR	0.71	0.47	0.47 ± 0.006	0.62 ± 0.011	0.48 ± 0.002	0.62 ± 0.003	0.32	0.77	'Alpha': 0.5
	LSVM	0.6	0.53	0.51 ± 0.008	0.53 ± 0.016	0.51 ± 0.002	0.54 ± 0.005	0.43	0.64	'C': 5.0, 'max_iter': 100
	SVM	0.84	0.53	0.51 ± 0.009	0.50 ± 0.014	0.52 ± 0.002	0.48 ± 0.004	0.41	0.63	'C': 25.0, 'degree': 2, 'gamma': 'auto'
	RF	0.81	0.56	0.55 ± 0.009	0.44 ± 0.015	0.56 ± 0.002	0.41 ± 0.005	0.45	0.49	'max_depth':6, 'min_samples_leaf':1, 'min_samples_split':4, 'n_estimators':70
	GB	0.82	0.51	0.55 ± 0.009	0.44 ± 0.017	0.56 ± 0.002	0.42 ± 0.005	0.41	0.53	'max_depth':2, 'min_samples_leaf':5, 'min_samples_split':2, 'n_estimators':90
	XGB	0.71	0.47	0.47 ± 0.006	0.62 ± 0.012	0.48 ± 0.002	0.62 ± 0.003	0.32	0.77	'Booster': 'Gblinear', 'learning_rate':0.1, 'max_depth': none, 'n_estimators':90
Indolines	PLS	0.63	0.48	0.49 ± 0.008	0.55 ± 0.017	0.49 ± 0.002	0.56 ± 0.004	0.3	0.81	n_components:3
	RR	0.63	0.49	0.49 ± 0.007	0.55 ± 0.016	0.49 ± 0.002	0.56 ± 0.004	0.3	0.82	'Alpha': 1.0
	LSVM	0.58	0.47	0.51 ± 0.008	0.51 ± 0.021	0.51 ± 0.002	0.53 ± 0.004	0.36	0.73	'C': 5.0, 'max_iter': 100
	SVM	0.71	0.5	0.51 ± 0.008	0.51 ± 0.018	0.52 ± 0.002	0.51 ± 0.004	0.34	0.76	'C': 1.0, 'degree': 2, 'gamma': 'Scale'
	RF	0.83	0.45	0.48 ± 0.007	0.54 ± 0.016	0.49 ± 0.002	0.55 ± 0.004	0.42	0.65	'max_depth':5, 'min_samples_leaf':3, 'min_samples_split':2, 'n_estimators':80
	GB	0.81	0.49	0.51 ± 0.007	0.48 ± 0.017	0.51 ± 0.002	0.5 ± 0.004	0.36	0.73	'max_depth':2, 'min_samples_leaf':1, 'min_samples_split':5, 'n_estimators':50
	XGB	0.63	0.48	0.49 ± 0.008	0.55 ± 0.017	0.49 ± 0.002	0.56 ± 0.004	0.3	0.81	'Booster': 'Gblinear', 'learning_rate':1.0, 'max_depth': none, 'n_estimators':120

Table 3 (*Contd.*)

Datasets	Methods	Training set metrics: Model statistics R^2	Training set metrics: Cross-validation statistics MAE_{LOO}	MAE ± SEM (20 times 5-fold CV)	r2 ± SEM (20 times 5-fold CV)	MAE ± SEM (1000 times ShuffleSplit CV)	r2 ± SEM (1000 times ShuffleSplit CV)	Test set metrics: Prediction statistics MAE_{test}	Q_{F1}^2	Optimized hyperparameters
Diphenylamines	PLS	0.83	0.44	0.47 ± 0.013	0.39 ± 0.072	0.48 ± 0.005	0.49 ± 0.031	0.31	0.9	n_components:3
	RR	0.83	0.44	0.49 ± 0.012	0.37 ± 0.071	0.5 ± 0.004	0.53 ± 0.013	0.31	0.91	'Alpha': 0.5
	LSVM	0.77	0.47	0.51 ± 0.016	0.18 ± 0.133	0.54 ± 0.005	0.38 ± 0.018	0.34	0.87	'C': 15.0, 'max_iter': 100
	SVM	0.87	0.48	0.55 ± 0.019	0.32 ± 0.066	0.57 ± 0.005	0.42 ± 0.014	0.64	0.61	'C': 1.0, 'degree': 2, 'Gamma': 'auto'
	RF	0.88	0.48	0.51 ± 0.01	0.41 ± 0.059	0.51 ± 0.004	0.55 ± 0.01	0.4	0.8	'max_depth':2, 'min_samples_leaf':1, 'min_samples_split':3, 'n_estimators':120
	GB	1	0.56	0.56 ± 0.014	0.13 ± 0.109	0.54 ± 0.004	0.43 ± 0.013	0.43	0.78	'max_depth':3, 'min_samples_leaf':1, 'min_samples_split':5, 'n_estimators':60
	XGB	0.84	0.44	0.47 ± 0.014	0.37 ± 0.08	0.49 ± 0.005	0.5 ± 0.02	0.3	0.91	'Booster': 'Gblinear', 'learning_rate':0.1, 'max_depth': none, 'n_estimators':90

a data point showing structural similarity (MaxPos) to compounds having high response values will also have high response value and *vice versa*.

The functional group count descriptor *n*Thiophene denoting the number of thiophene rings in the coumarin dyes contributes positively to the PCE. Therefore, presence of such functional group in the dye increases the performance of DSSCs as represented by the following examples: **19** (*n*Thiophene = 2, PCE = 7.4), **32** (*n*Thiophene = 2, PCE = 6.5). The PCE value may reduce for the compounds where no such functional group is present as shown in the following examples: **56** (*n*Thiophene = 0, PCE = 0.99), **17** (*n*Thiophene = 0, PCE = 0.9). Thiophene groups are the part of the π-spacer which not only improves light absorption and dipole moment but also decreases the dihedral angle between donor/acceptor and π-spacer plane for better orbital overlap which in turn improve electron injection to TiO_2.[54]

Two other functional group count descriptors *n*R = Ct (number of an aliphatic tertiary carbon atom with the 'sp^2' hybridization) and *n*R#C– (number of a non-terminal carbon atom with the 'sp' hybridization) have positive and negative contributions to the PCE, respectively. The presence of aliphatic tertiary 'sp^2' hybridized C atom and the absence of non-terminal 'sp' hybridized C atom frequency of 's' is responsible for the enhancement of absorption.[55] The contribution of the descriptor *n*R = Ct is represented by the following examples: **35** (*n*R = Ct = 4, PCE = 6.2), **32** (*n*R = Ct = 3, PCE = 6.5), **17** (*n*R = Ct = 0, PCE = 0.9), **22** (*n*R = Ct = 0, PCE = 0.33); and the following examples represent the contribution of *n*R#C– **44** (*n*R#C– = 2, PCE = 1.35), **56** (*n*R#C– = 2, PCE = 0.99), **19** (*n*R#C– = 0, PCE = 7.4), **32** (*n*R#C– = 0, PCE = 6.5).

*n*RCN is a functional group count descriptor denoting the number of aliphatic nitriles in the dye which contributes negatively to the PCE of coumarin dyes. Therefore, with the increasing number of nitrile groups, the performance of DSSCs is reduced as indicated by the following examples: **44** (*n*RCN = 1, PCE = 1.35), **56** (*n*RCN = 1, PCE = 0.99) and *vice versa* for the dyes **10** (*n*RCN = 0, PCE = 3.7), **54** (*n*RCN = 0, PCE = 3.5) where no nitrile group is present. Anchoring groups are a part of the dye which involves adsorption on TiO_2 surface that determines electron injection ability and optoelectrical property of the dye. Nitrile groups are generally a part of this anchoring group which may increase adsorption stability when CN group itself is involved in the binding. On the other hand, nitrile groups may reduce the photovoltaic property when it is not involved in binding.[56]

T(S···S) is a 2D atom pair descriptor that indicates the sum of the topological distance between two sulfur atoms where they are part of two thiophene rings. The negative contribution of this descriptor signifies that with the increasing distance between sulfur atoms, the performance of the DSSCs may decrease as represented by the following examples: **22** (T(S···S) = 31, PCE = 0.33), **3** (T(S···S) = 28, PCE = 1.77), **44** (T(S···S) = 21, PCE = 1.35) and *vice versa* for the dyes **19** (T(S···S) = 3, PCE = 7.4), **32** (T(S···S) = 3, PCE = 6.5), **29** (T(S···S) = 3, PCE = 6.07). The possible reason for this may be due to the disruption of the planar structure of the π-spacer and increase of dihedral angle

between adjacent donor/acceptor and π-spacer. Another reason is that a hole is created in the π-spacer after injection of an electron to the TiO_2; this hole is transferred to the donor part and prevents charge recombination. Therefore, with the increasing length of π-spacer, the possibility of this hole transfer is reduced, and this may cause back transfer of electron and reduce DSSC performances.[54]

C-040 is an atom-centered fragment descriptor that represents fragments like R–C(=X)–X/R–C#X/X=C=X (R: any group linked through carbon; X: any electronegative atom like N, S, P, O, halogen; #: triple bond) in the dye which contributes positively to the PCE. The positive contribution of this descriptor signifies that the presence of such fragments in the dye may increase the performance of the dye as shown in the following examples: **19** (C-040 = 3, PCE = 7.4), **20** (C-040 = 3, PCE = 6.4), **35** (C-040 = 3, PCE = 6.2) and *vice versa* for the dyes **24** (C-040 = 2, PCE = 1.04), **17** (C-040 = 2, PCE = 0.9). These fragments are generally parts of the anchoring group containing carboxylic acid or cyanoacrylic acid as a binder which increases the stability of adsorption on TiO_2 surface and helps in efficient electron transfer.[56]

SD Activity(GK) is a RASPR descriptor that denotes the weighted standard deviation of the response value of the selected close source compound for each query compound. The positive contribution of this descriptor[32] is represented by the following examples: **29** (SD Activity(GK) = 1.53777, PCE = 6.07), **36** (SD Activity(GK) = 1.40648, PCE = 5.5), **7** (SD Activity(GK) = 1.04559, PCE = 1.1), **17** (SD Activity(GK) = 0.9868, PCE = 0.9).

The mechanistic interpretation of the 2D structural descriptor of the q-RASPR PLS model for the coumarin dyes is schematically represented in Fig. 3.

Modeling of PCE of carbazoles. The PLS q-RASPR model related to the carbazole dyes consisting of 8 descriptors has been shown in Table 2, and the contribution of the descriptors in the form of a bubble plot has been shown in Fig. S1b of ESI SI-1.† The mechanistic interpretation of these descriptors and their influence on PCE is discussed below:

The 2D atom pair descriptor B04[N–O] indicates the presence or absence of nitrogen and oxygen atoms at the topological distance 4, and this descriptor contributes positively to the PCE. This fragment is part of an anchoring group cyanoacrylic acid. Cyanoacrylic acid is one of the most common anchoring groups for metal oxide (TiO_2) surfaces because of its dual character of a strong adsorber and a good acceptor. Its strong binding with TiO_2 provides stability to the adsorbed dyes which in turn helps in the efficient transfer of an electron to TiO_2. This cyanoacrylic acid also has a strong electron-withdrawing ability which helps in intramolecular charge transfer from donor to metal oxide.[55] Therefore, when such fragments are present in the dye, the performance of DSSCs will increase as represented by the following examples: **132** (B04[N–O] = 1, PCE = 12.5), **133** (B04[N–O] = 1, PCE = 9.32), **101** (B04[N–O] = 1, PCE = 8.09) and *vice versa* for the dyes **160** (B04[N–O] = 0, PCE = 0.34), **157** B04[N–O] = 0, (PCE = 0.31), **159** (B04[N–O] = 0, PCE = 0.21).

B02[C–S] is a 2D atom pair descriptor that indicates the presence or absence of carbon and sulfur at the topological distance 2. The positive contribution of this descriptor indicates that the presence of such fragment increases the performance of DSSCs. This fragment is a part of the thiophene group that acts as a π-spacer present between donor and acceptor moieties. This electron rich π-spacer is responsible for the enhancement absorption of photon which in turn increases PCE of carbazole dye.[52] The positive contribution of this descriptor is represented

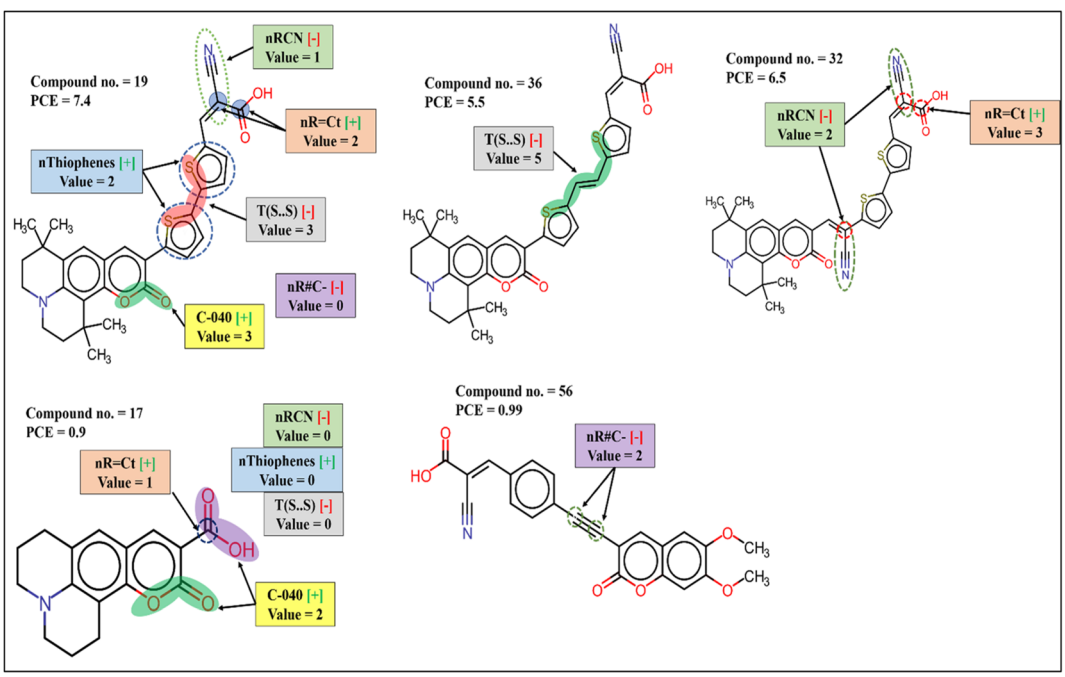


Fig. 3 Mechanistic interpretation of 2D structural descriptors of q-RASPR PLS model for the coumarin dataset.

by the following examples: **132** (B02[C–S] = 1, PCE = 12.5), **133** (B02[C–S] = 1, PCE = 9.32), **101** (B02[C–S] = 1, PCE = 8.09) and *vice versa* for the dyes **160** (B02[C–S] = 0, PCE = 0.34), **157** (B02[C–S] = 0, PCE = 0.31).

F06[N–O] is another 2D atom pair descriptor that indicates the frequency of nitrogen and oxygen atoms at the topological distance 6, and it contributes positively toward the PCE. This fragment is present in the dye either as a part of the phenyl moiety between acceptor (cyanoacrylic acid) and adsorber or as a part of the linker between the donor and π-spacer (furan or enedioxythiophene). A dye containing this fragment between acceptor and adsorber will have an improved performance by its diode like effect (which prevents the back transfer of electrons from TiO_2 to the dye).[57] It helps in an efficient intramolecular charge transfer for the dyes containing this fragment as a linker between the donor moiety and π-spacer.[58] The positive contribution of this descriptor is represented by the following examples: **132** (F06[N–O] = 3, PCE = 12.5), **133** (F06[N–O] = 2, PCE = 9.32) and *vice versa* for the dyes where no such fragment is present, **160** (F06[N–O] = 0, PCE = 0.34), **157** (F06[N–O] = 0, PCE = 0.31).

F04[C–N] is a 2D atom pair descriptor that denotes the frequency of carbon and nitrogen atoms at the topological distance 4, and this descriptor contributes positively to the PCE. This fragment is present mainly as a part of the main scaffold (carbazole moiety) of the dye, and also in some dyes it is present adjacent to the carbazole moiety as a part of π-spacer. This fragment helps in the generation of electrons by a donor group and helps in the efficient transfer of electrons toward the acceptor part which in turn increases the performance of the DSSCs.[59–61] The PCE value increases in the presence of such fragments in the dyes as indicated by the following examples: **50** (F04[C–N] = 22, PCE = 7.52), **101** (F04[C–N] = 17, PCE = 8.09), **130** (F04[C–N] = 15, PCE = 9.8) and *vice versa* for the dyes **97** (F04[C–N] = 0, PCE = 0.0538), **98** (F04[C–N] = 0, PCE = 0.0387) where no such fragment is present.

*n*R10 is a ring descriptor that indicates the number of 10 membered rings in a dye which contributes positively to the PCE. In this case, 6-membered or 5-membered aromatic rings are fused with the main carbazole scaffold of the dye and form a planar structure. These electron-rich centers help in the generation of electrons and due to their planar structure, the molar absorption coefficient and photon harvesting ability of the dye is increased which improve the performance of DSSCs.[62,63] In some dyes, this fragment is also present as a part of the π-spacer which helps in the efficient transfer of electrons from a donor part to the acceptor part. Therefore, performances of DSSCs should increase when such ring system is present in the structures, which is indicated by the following examples: **130** (*n*R10 = 6, PCE = 9.8), **131** (*n*R10 = 6, PCE = 7.6) and *vice versa* for the dyes **159** (*n*R10 = 0, PCE = 0.21), **91** (*n*R10 = 0, PCE = 0.19), **154** (*n*R10 = 0, PCE = 0.07) where no 10 membered rings are present.

The negative contribution of the constitutional descriptor N% (percentage of the nitrogen atoms in the dye) and 2D atom pair descriptor F06[N–N] (frequency of two nitrogen atoms at the topological distance 6) indicates that the presence of such fragments hinders the performance of DSSCs. Higher numerical values of these descriptors of a dye may decrease the PCE value which is represented by the following examples: **91** (N% = 6.25, PCE = 0.19), **118** (N% = 5.6338, PCE = 0.89), **53** (N% = 4.83871, PCE = 0.99), **112** (N% = 4.83871, PCE = 0.96) for the descriptor N%; **141** (F06[N–N] = 2, PCE = 2.58), **138** (F06[N–N] = 2, PCE = 2.17) for F06[N–N]. On the other hand, dyes with lower numerical value of this descriptor may have higher PCE values as shown in following examples: **99** (N% = 1.81818, PCE = 7.58), **103** (N% = 1.50376, PCE = 7.54), **130** (N% = 1.34529, PCE = 9.8) for N%, **132** (F06[N–N] = 0, PCE = 12.5), **130** (F06[N–N] = 0, PCE = 9.8), **133** (F06[N–N] = 0, PCE = 8.09) for the F06[N–N].

Avg. Sim(GK) is a RASPR descriptor that denotes the mean similarity value of the selected close source compounds for each query compound based on the Gaussian kernel similarity-based method. The positive contribution of this descriptor indicates a molecule having a higher Avg. Sim value may have a higher PCE value as represented by the following examples: **94** (Avg. Sim(GK) = 0.92848, PCE = 7.33), **56** (Avg. Sim(GK) = 0.89553, PCE = 6.04), **99** (Avg. Sim(GK) = 0.75399, PCE = 7.58) and *vice versa* for the dyes **156** (Avg. Sim(GK) = 0.24768, PCE = 0.06), **154** (Avg. Sim(GK) = 0.04227, PCE = 0.07).

The interpretation of 2D structural descriptors for the carbazole dyes is represented schematically in Fig. 4.

Modeling of PCE of indolines. The PLS q-RASPR model of indoline dyes consisting of 6 descriptors has been presented in Table 2, and the contribution of the descriptors in the form of a bubble plot is shown in Fig. S1c of the ESI SI-1.† The meaning of these descriptors and their influence on PCE are discussed below:

RA function is a Read-Across-derived RASPR descriptor which encodes information of all the selected structural and physicochemical descriptors.[33] It contributes positively to the PCE as indicated by the following examples: **141** (RA function = 8.1741, PCE = 8.38), **8** (RA function = 7.9697, PCE = 7.12), **24** (RA function = 7.88, PCE = 9.2) and *vice versa* for the dye **129** (RA function = 1.8131, PCE = 1.48), **32** (RA function = 1.5372, PCE = 0.63), **30** (RA function = 1.4248, PCE = 0.77).

Both RASPR descriptors Neg.Avg.Sim (denoting the mean of the similarity values of the negative close source compounds for a particular query compound) and CVsim(LK) (coefficient of variation of the similarity values of the selected close source compound for each query compound) contribute negatively to the PCE. This is represented by the following examples: **93** (Neg.Avg.Similarity = 0.2883, PCE = 0.35), **108** (Neg.Avg.Similarity = 0.2883, PCE = 0.046) for Neg.Avg.Similarity; **93** (CVsim(LK) = 1.404, PCE = 0.35), **108** (CVsim(LK) = 1.404, PCE = 0.046) for CVsim(LK); and *vice versa* for the dye **144** (Neg.Avg.Similarity = 0, PCE = 8.78), **24** (Neg.Avg.Similarity = 0, PCE = 9.2), **135** (Neg.Avg.Similarity = 0, PCE = 8.61) for Neg.Avg.Similarity; **135** (CVsim(LK) = 0.4427, PCE = 8.61), **78** (CVsim(LK) = 0.3868, PCE = 7.99) for CVsim(LK).

The functional group count descriptor *n*Cconj denotes the number of non-aromatic conjugated sp^2 hybridized carbon atoms that contributes negatively to the PCE. The negative contribution of this descriptor signifies that the PCE value may decrease when the number of non-aromatic conjugated sp^2

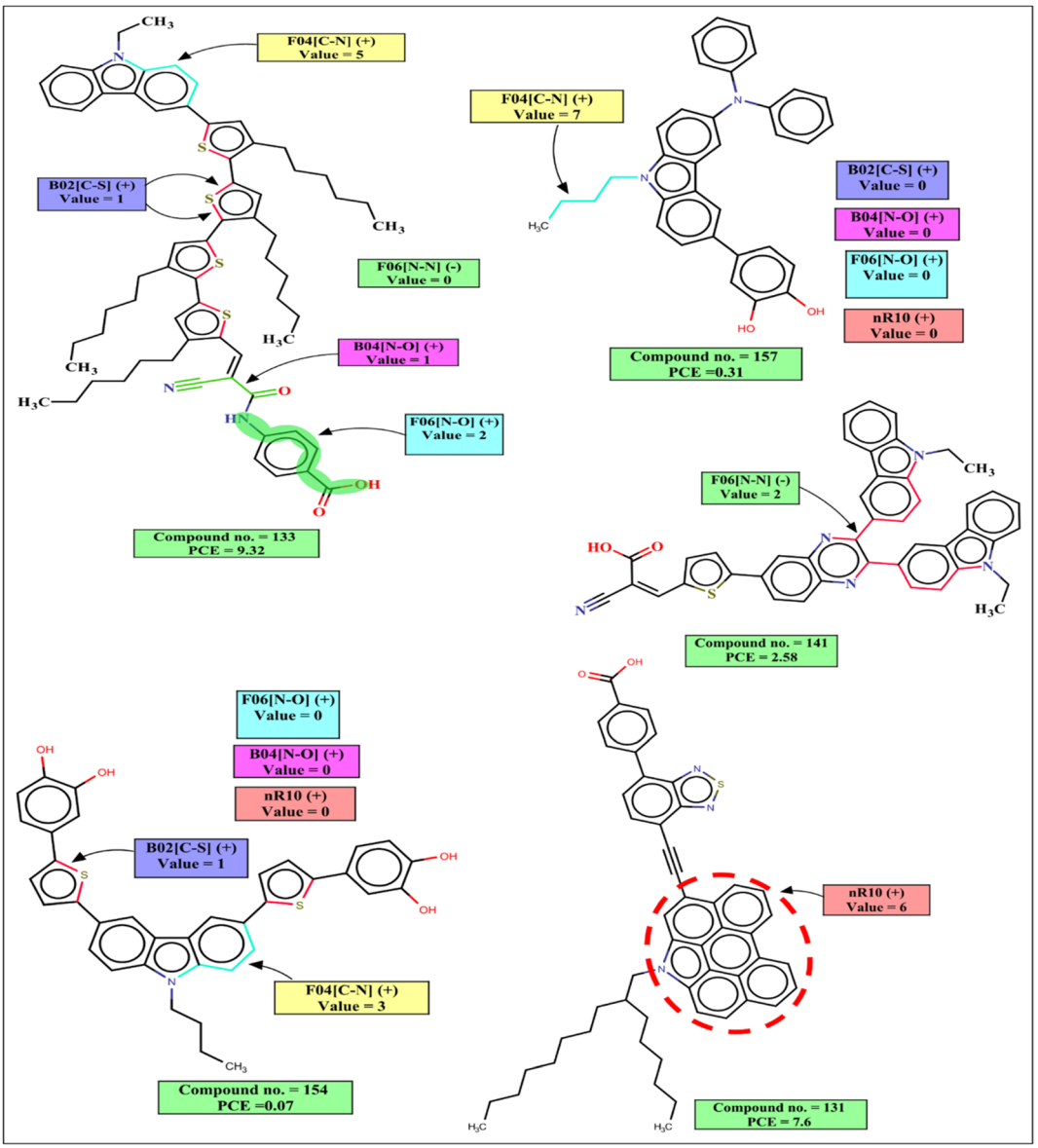


Fig. 4 Mechanistic interpretation of the 2D structural descriptors of the q-RASPR PLS model for the carbazole dataset.

carbon increases as represented by the following examples: **11** (*n*Cconj = 11, PCE = 2.65), **10** (*n*Cconj = 10, PCE = 2.7) and *vice versa* for the dyes with a low numerical value of *n*Cconj like **155** (*n*Cconj = 1, PCE = 5.61) and **152** (*n*Cconj = 1, PCE = 5.5).

F04[C–N] is a 2D atom pair descriptor that indicates the frequency of carbon and nitrogen atoms at the topological distance 4 in the dye, and this descriptor contributes positively to the PCE. It was found that if the donor group is present with a non-planar orientation with other groups, it may increase the PCE value. Although this fragment is present as a part of the dye in a non-planar structure, it may increase the performance of the DSSCs as indicated by its positive contribution to the PCE.[64,65] Therefore, the presence of such fragment increases the performance of DSSCs as shown by the following examples: **21** (F04[C–N] = 21, PCE = 8.43), **8** (F04[C–N] = 20, PCE = 7.12), **24** (F04[C–N] = 18, PCE = 9.2) and *vice versa* for the dyes **105** (F04[C–N] = 3, PCE = 2.53), **164** (F04[C–N] = 3, PCE = 2.08).

Another 2D atom pair descriptor B09[O–S] indicates the presence or absence of oxygen and sulfur atoms at the topological distance 9, and this descriptor contributes positively to the PCE. This is a part of the anchoring group for the dye which contains this fragment. It helps to transfer electrons from the dye to the TiO_2 surface through π-bond conjugation. Oxygen and sulfur atoms control electron density delocalization which helps in π bond conjugation. As a result, the molar extinction coefficient of the dye increases which may lead to shifting of the absorption maxima.[65] If the topological distance between O and S is reduced or increased, the conformation of the dye will change which may decrease the anchoring stability of the dye and the performance of the DSSCs will be reduced.[66,67] Dyes containing this type of fragment may increase PCE values as

represented by the following examples: **78** (B09[O–S] = 1, PCE = 7.99), **21** (B09[O–S] = 1, PCE = 6.12), **131** (B09[O–S] = 1, PCE = 6.11) and *vice versa* for the dyes **30** (B09[O–S] = 0, PCE = 0.77), **32** (B09[O–S] = 0, PCE = 0.63), **93** (B09[O–S] = 0, PCE = 0.35).The mechanistic interpretation of the relevant descriptors for the indoline dataset is schematically represented in Fig. 5.

Modelling of PCE of diphenylamines. The PLS q-RASPR model consisting of 5 descriptors has been shown in Table 2, and the contribution of the descriptors in the form of a bubble plot is shown in Fig. S1d of ESI SI-1.† The mechanistic interpretation of these descriptors is represented below:

F01[C–N] is a 2D atom pair descriptor that indicates the frequency of carbon and nitrogen atoms at the topological distance of 1, and this descriptor contributes negatively to the PCE. In the presence of these fragments, the overall polarity of the dye will change which may lead to an increased intermolecular interaction in terms of different weak forces like hydrogen bonding, aromatic ring stacking, van der Waals force, *etc.* These weak forces may cause aggregation of dyes on the surface of the TiO_2, and the performance of the DSSCs is reduced.[68] Therefore, the presence of such fragment reduces the performance of the DSSCs as represented by the following examples: **35** (F01[C–N] = 11, PCE = 0.4), **34** (F01[C–N] = 10, PCE = 1) and *vice versa* for the dyes **3** (F01[C–N] = 4, PCE = 5.4), **22** (F01[C–N] = 4, PCE = 5.22), where no such fragment is present.

Another 2D atom pair descriptor F04[N–S] denotes the frequency of nitrogen and sulfur atoms at the topological distance 4, and this descriptor contributes positively to the PCE. This can be represented by the following examples: **27** (F04[N–S] = 2, PCE = 8), **26** (F04[N–S] = 2, PCE = 7.1), **17** (F04[N–S] = 2, PCE = 6.19) and *vice versa* for the dyes **34** (F04[N–S] = 0, PCE = 1), **33** (F04[N–S] = 0, PCE = 0.44), **35** (F04[N–S] = 0, PCE = 0.4) where no such fragment is present.

SD_similarity is a RASPR descriptor that denotes the standard deviation of the similarity values of close source compounds for each query compound. A high numerical value of the descriptor may increase PCE value as shown in the following examples: **7** (SD similarity = 0.33695, PCE = 7.05), **8** (SD similarity = 0.33274, PCE = 7.64) and *vice versa* for the dyes **35** (SD similarity = 0.177502, PCE = 0.4), **33** (SD similarity = 0.016381, PCE = 0.44).

StsC is an atom type E-state descriptor that indicates the sum of tsC E-states (≡C–), which contributes negatively to the PCE property of the DSSCs, as observed for the dyes **10** (StsC = 8.292574, PCE = 1.99), **13** (StsC = 7.76829, PCE = 3.16) and *vice versa* for the dyes **8** (StsC = 1.671126, PCE = 7.64), **7** (StsC = 1.644351, PCE = 7.05).

The mechanistic interpretation of the significant 2D-structural descriptors for the diphenylamine dataset is schematically represented in Fig. 6.

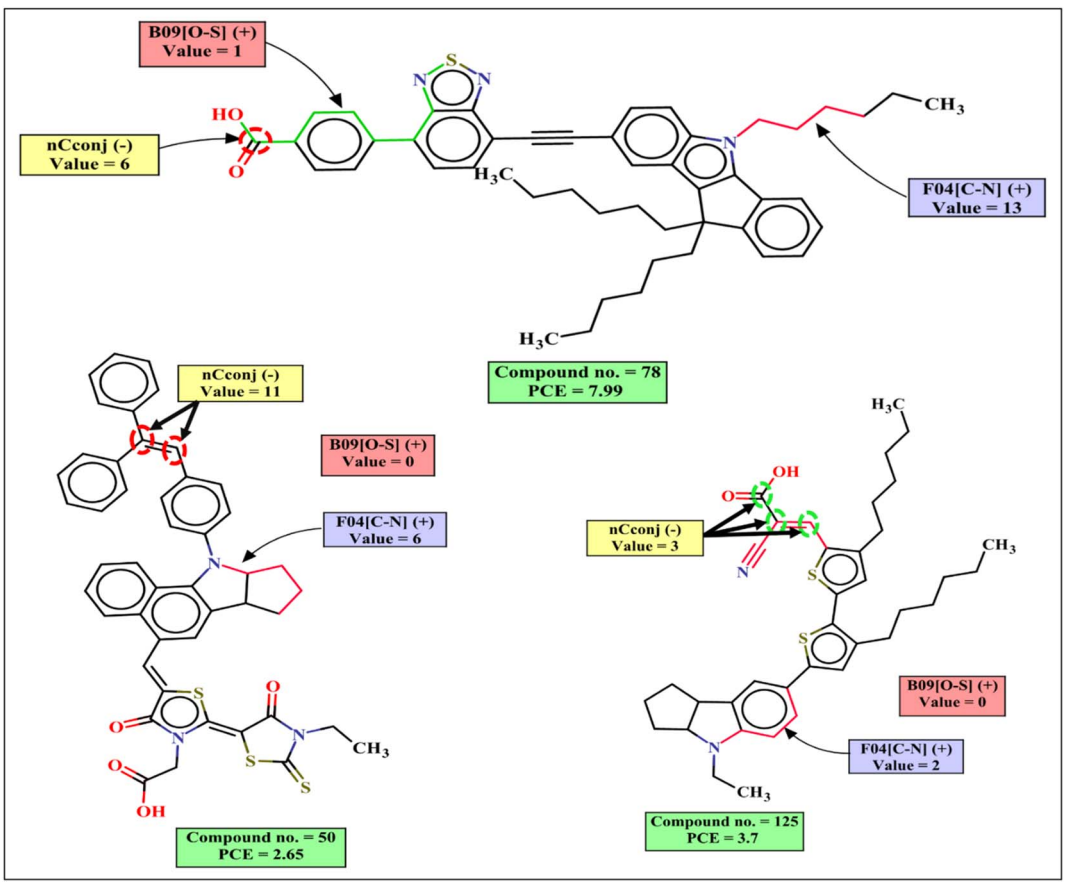


Fig. 5 Mechanistic interpretation of the 2D structural descriptors of q-RASPR PLS model for the indoline dataset.

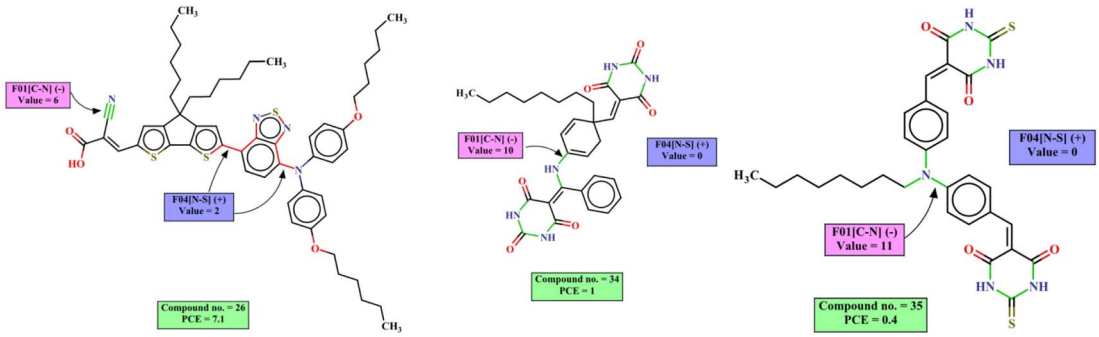


Fig. 6 Mechanistic interpretation of the 2D structural descriptors of q-RASPR PLS model for the diphenylamine dataset.

For all 4 datasets, different PLS plots like randomization plots, loading plots, and score plots were developed which are shown in the ESI SI-1.† For all the datasets, the PLS Scatter plots (Fig. S2†) show that there is not so much difference between observed and predicted PCE indicating the good quality of the test set predictions. The *Y*-randomization plots (Fig. S3†) show that all the models have R^2 and Q^2 intercept values within their threshold limits (0.3 for R^2 and 0.05 for Q^2), indicating that our models are not obtained by chance. The loading plots (Fig. S4†) show that the descriptors MaxPos(GK) (for coumarins), BO4[N–O] (for carbazoles) and RA function (for both indoline and diphenylamines) have the highest contributions to the PCE because they are present closest to the response variable (PCE). The score plots (Fig. S5†) show that there are 2 coumarin (**3**, **54**), 5 carbazole (**132**, **138**, **139**, **140**, **141**) and 1 indoline (**18**) molecules which are present outside the applicability domain of the corresponding models (located outside the ellipse drawn on based on Hotelling t^2 test).[69]

Machine-learning (ML) models

The qualities of the Machine-Learning (ML) models and the different validation metrics for all 4 datasets are shown in Table 3 along with their different optimized hyperparameter settings. To determine the models' quality and predictability, we calculated various quality and validation metrics on the training and test sets. For the purpose of comparison, we have also included PLS q-RASAR models for which the same method was used for calculation of the internal, external and cross-validation metrics. For all 4 datasets, ridge regression, XGBoost and PLS models show almost similar ${Q_{F1}}^2$, MAE_{test} and R^2 score (of training set) values. There are two different aspects here regarding the quality of predictions from the q-RASPR models for the test sets. The first is the enhancement of prediction quality for the test set in case of a q-RASPR model in comparison to the corresponding QSPR model (*vide infra* the comparison section) while the second is the comparison of the prediction quality of the q-RASPR models for the test set in comparison to its fitting quality in case of the corresponding training set. The test set prediction quality may better be compared with the training set fitting ability by considering MAE_{Test} for the test set and MAE_{LOO} for the training set. In our opinion, ${Q_{F1}}^2$ for the test set should not be directly compared with R^2 or Q^2 values of the training set as these metrics depend on the distribution of the observed response values of the training and test set compounds around the training set mean, and usually these patterns may be different in the training and test sets.[70] In our examples, MAE_{Test} values for the test sets in different models are lower than the corresponding MAE_{LOO} values for the training sets (Table 3). This also happened in the case of machine learning models like the random forest, Gradient Boost, and Extreme Gradient Boost for different data sets in Table 3. Thus, in terms of MAE as a metric (MAE_{LOO} for the training sets and MAE_{Test} for the test sets), the quality of the test set predictions is comparatively better in these examples.

Now, as per the q-RASPR algorithm, the RASPR descriptors of both the training and test sets are computed from the structural congeners in the training set. It is natural that a data set may contain a few activity cliffs, which are similar to other compounds in structural features but have quite different response values from their structural congeners. The fitting ability of such compounds in the training set and the prediction ability of such compounds in the test set will naturally be poor, especially when we use similarity-based descriptors like RASPR descriptors. In our present examples, the training set size is much bigger than the corresponding test set size in order to maximize the learning ability of the models (as usual in conventional QSPR studies). Thus, the probability of the occurrence of such activity cliffs in the training sets is more than that in the corresponding test sets, which may explain (at least partially) the lower MAE_{Test} values in comparison to the corresponding MAE_{LOO} values of the training sets. The activity cliff aspect in q-RASPR modeling has been extensively discussed in our recent work.[71]

We have checked the number of activity cliffs in the training and test sets of the four different data sets based on novel Banerjee–Roy similarity coefficients as per ref. 71. A compound is considered an activity cliff when both of the two similarity coefficients do not show values as per the expected category (positive/negative, considering the training set response mean as the threshold). From Table 4, it is evident that in the case of each data set, the number of activity cliffs in a training set is much higher than the number of activity cliffs in the corresponding test sets. In the case of QSAR analysis, descriptors are

Table 4 Number of activity cliffs in the training and test sets based on the analysis of similarity coefficients

Dataset	Number of training set compounds	Number of test set compounds	Number of activity cliffs in the training set[a]	activity cliff in the training set (%)	Number of activity cliffs in the test set[a]	activity cliff in the test set (%)
Carbazoles	124	54	**37**	29.84	**14**	25.93
Coumarins	42	14	**7**	16.66	**2**	14.29
Diphenylamines	25	10	**6**	24	**1**	10
Indolines	121	38	**20**	16.53	**7**	18.42

[a] Computed based on similarity coefficients described in ref. 71.

computed directly from the structures of the compounds in question; however, RASAR descriptors are computed from close congeners of the compounds under consideration. In the case of activity cliffs, the similarity principle is not obeyed and thus the similarity descriptors computed from the close congeners cannot capture the structure–response relationship properly. In the case of QSAR analysis, the model fitting is done based on the whole training set in which activity cliffs may penalize a model but not to the extent to a RASAR model as in the latter case the similarity descriptors of the activity cliffs (not obeying the similarity principle) heavily penalize the model. This is more evident in the case of regression-based predictions, as precise quantitative predictions are considered here as also seen in ref. 38. Due to the lower number of activity cliffs in the test sets, the quality of predictions is less impacted. Such observations are not common in case of QSAR analysis including ML methods as in the latter case descriptors are not computed from close congeners of the compounds under consideration, rather computed from the same compounds. In fact, one of the objectives of RASAR modeling is to enhance the quality of predictions for the test set which may be at the expense of lowering the prediction quality for the training set. Further, the novel similarity coefficients[71] may be used to identify activity cliffs and enhance the modelability of a data set.

To further evaluate the quality of developed models, we have also performed 20 times 5-fold repetitive cross-validation, and 1000 times shuffle-split cross-validation with 30% data holding in the validation set. The result of cross-validation for the coumarin dataset is shown in the Fig. 7 and that for the carbazole, indoline and diphenylamine datasets are shown in Fig. S6–S8 in ESI SI-1.† For the coumarin dataset, the mean R^2 value for both the repetitive CV and Shuffle-split CV indicates

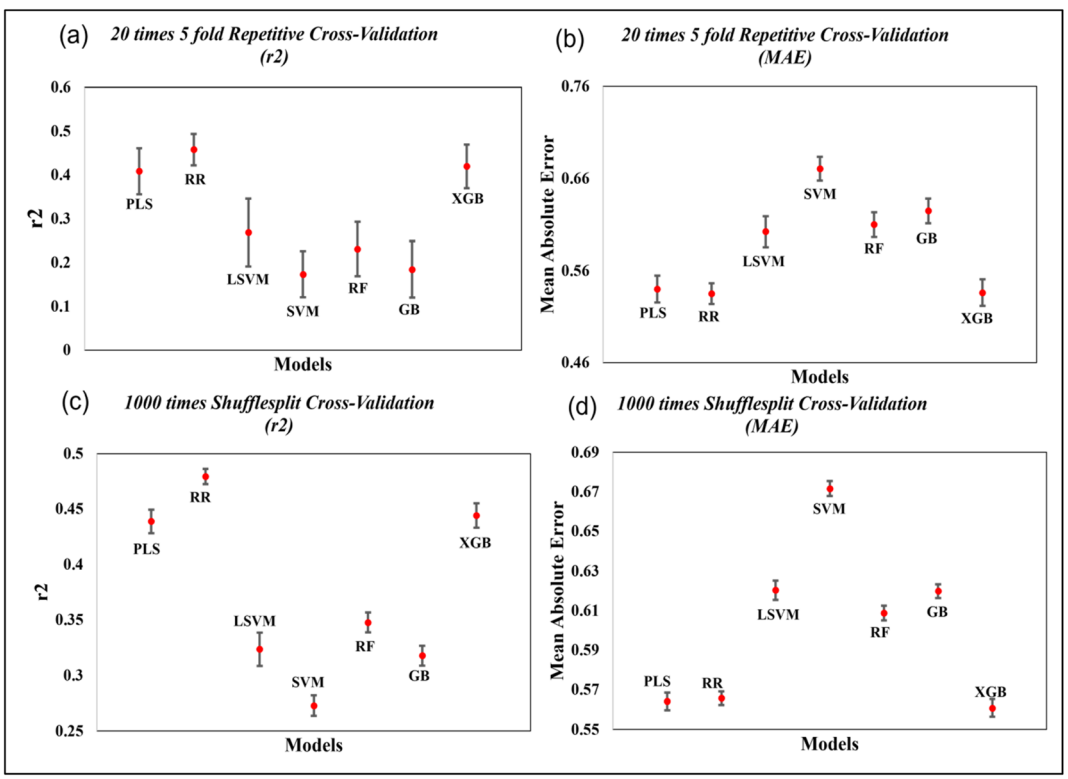


Fig. 7 Cross-validation statistics based on 20 times 5-fold repetitive CV and 1000 shuffle split CV method (mean ± SEM) for the coumarin dataset.

that the Ridge regression method is the best model among all models while the PLS and XGBoost models show comparable results. For the carbazole and indoline datasets, the ridge regression, PLS and XGBoost models show comparable results, as shown in Fig. S6 and S7.† For the diphenylamine dataset, the random forest model shows the highest mean R^2 value in both repetitive cross-validation and shuffle-split cross-validation method but ridge regression, XGBoost and PLS models show comparable results, as shown in Fig. S8.†

To evaluate the importance of descriptors in the machine-learning models, we have performed SHAP analysis on the training set data. We have represented the importance of descriptors in the form of heatmap plots of SHAP as shown in Fig. 8. The PLS, ridge regression and XGBoost models are considered here, and the plots of the remaining models are shown in the Fig. S9–S12 of ESI SI-1.† On the *Y*-axis of the heatmap plot, the features are arranged based on their mean absolute SHAP values, which in turn denotes their importance to the predictions. From the heatmap plot, we can also obtain how the model's prediction changes over every instance which is denoted by the wavy line above the plot. The colour difference in the plot indicates how the SHAP value of the features changes over every instance and how it affects the model's output.

A partial dependence plot shows the marginal effect of a feature (or two features) on the predicted outcome of a machine learning model. This plot can suggest the dependence interaction between two features. In case of an interaction with the other feature, a distinct vertical pattern of coloring will be seen. The partial dependence plots of selected ML models are shown in ESI SI-3.†

Design of new dyes with improved PCEs

We can design new dyes with improved power conversion efficiency (PCE) by incorporating structural fragments that have positive contributions to the PCE of solar cells or by removing fragments that contribute negatively to the PCE of solar cells. The favorable fragments improve PCE either by increasing intramolecular charge transfer or by increasing the anchoring stability of the dye on the TiO_2 semiconductor oxide surface. The descriptors which encode the structural information help to identify the necessary modifications that must be made to improve the PCE. In this work, we use the PLS variable importance plot to identify the relative importance of the positive and negatively contributing descriptors. Based on this, we have designed new dyes with improved predicted PCE values as shown in ESI SI-3.† We have also checked the synthetic accessibility of the designed dyes.[72]

For carbazole dyes, the incorporation of a cyano acrylic acid group increases the value of B04[N–O] (presence or absence of nitrogen and oxygen atoms at the topological distance 4) while a *para*-aminobenzoic acid group increases the value of F06[N–O] (count of nitrogen and oxygen at the topological distance 6). For example, the F06[N–O] value increases when one incorporates 4-(2-cyanoprop-2-enamido)benzoic acid (as shown in NCA1, NCA2 and NCA3) and 2-cyano-*N*-[4-(trimethoxysilyl)phenyl]

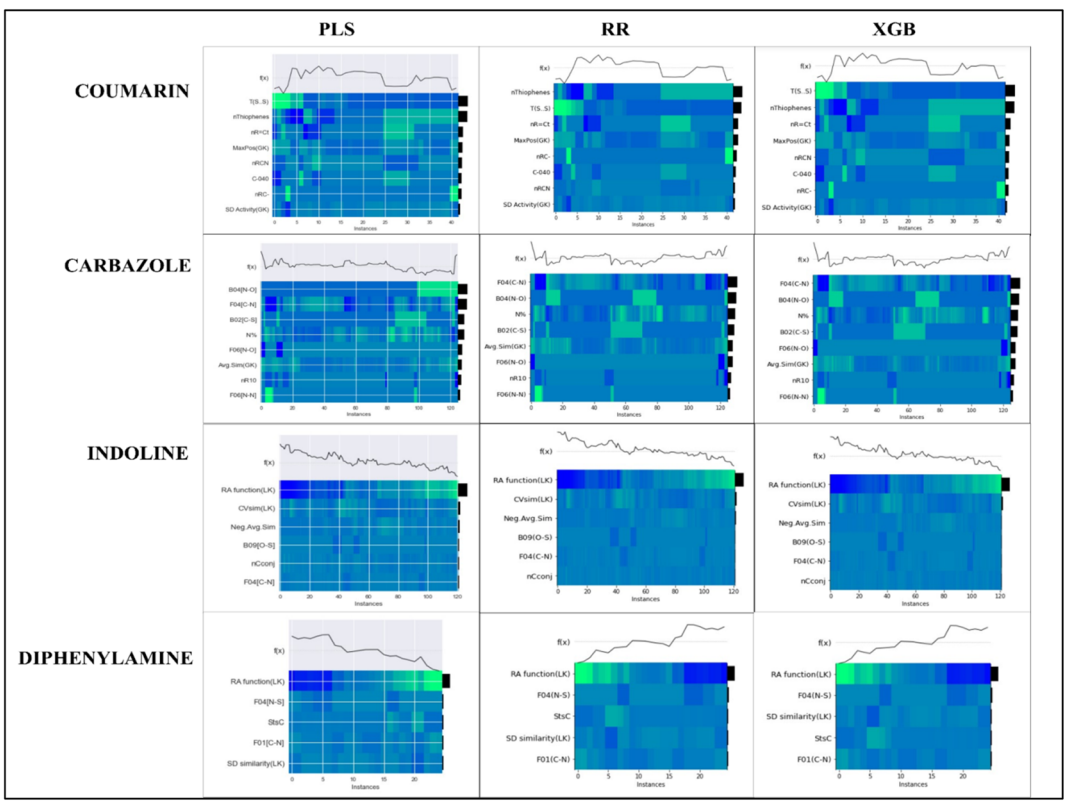


Fig. 8 Heatmap plots for the PLS, Ridge regression and XGBoost models for all datasets, indicating relative importance of descriptors.

prop-2-enamide (as shown in NCA4 and NCA5) moieties in the carbazole structure. These fragments are generally a part of the anchoring group which increases the stability of the binding of the dye with the TiO_2 surface. We can increase the value of B02 [C–S] (presence or absence of carbon and sulfur atoms at the topological distance of 2) by incorporating a thiophene group, increase the value of F04[C–N] (frequency of carbon and nitrogen atoms at the topological distance of 4) by attaching a long aliphatic chain to the nitrogen atoms and increase the value of *n*R10 (the number of 10-member rings) by incorporating 10 membered rings in the structure. All these fragments are responsible for the generation of electrons which are transferred to the TiO_2 surface.

For coumarin dyes, one can increase the value of positively contributing descriptors like *n*Thiophene (number of thiophene group), *n*R = Ct (number of aliphatic tertiary C atom with sp^2 hybridization) and C-040 (R–C(=X)–X/R–C#X/X=C=X). These descriptors are generally responsible for electron generation and intramolecular charge transfer. One can also try removing the negatively contributing descriptor *n*RCN (number of aliphatic nitrile groups). The nitrile group is a part of the anchoring group cyanoacrylic acid; therefore, one can try using other anchoring groups like carboxylic acid, pyridine, *etc.*

For diphenylamine dyes, one can increase the value of positively contributing descriptor F04[N–S] (frequency of nitrogen and sulfur atoms at the topological distance 4) by incorporating groups like a pyrimidine ring adjacent to a thiophene ring (as shown below in NDI1), 2,1,3-benzothiadiazole and 1,2,3-benzodithiazole groups adjacent to the thiophene and pyrimidine rings respectively (as shown below in NDI5 and NDI3). These fragments are generally a part of the linker between the donor part and the acceptor part which helps to improve performance by increasing intramolecular charge transfer.

For indoline dyes, one can increase the value of positively contributing descriptors like F04[C–N] (frequency of carbon and nitrogen atoms at the topological distance of 4) by attaching different aliphatic and aromatic groups to the nitrogen atoms, and B09[O–S] (presence of oxygen and sulfur atoms at the topological distance 9) by increasing the length of the π-spacer (for example, compound NIN1 is formed by incorporating a butylene group between the thiophene ring and the cyanoacrylic acid). These fragments help in the generation of electrons and improve intramolecular charge transfer.

Comparison with the previous work

In 2020, Krishna *et al.* worked on the prediction of PCE value of metal free organic DSSCs by PLS regression using 2D structural descriptors.[26] The models reported by them were statistically sound with good quality validation metrics. However, our q-RASPR PLS models outperformed them with better prediction quality with the same level of chemical information used, also explaining the importance of RASPR descriptor. The advantage of RASPR descriptors is that they are simple, easy to calculate and transferable. In comparison to the previous models, the present models have higher Q_{F1}^2 scores and lower MAE_{test} values. The comparison between the previous PLS QSPR models

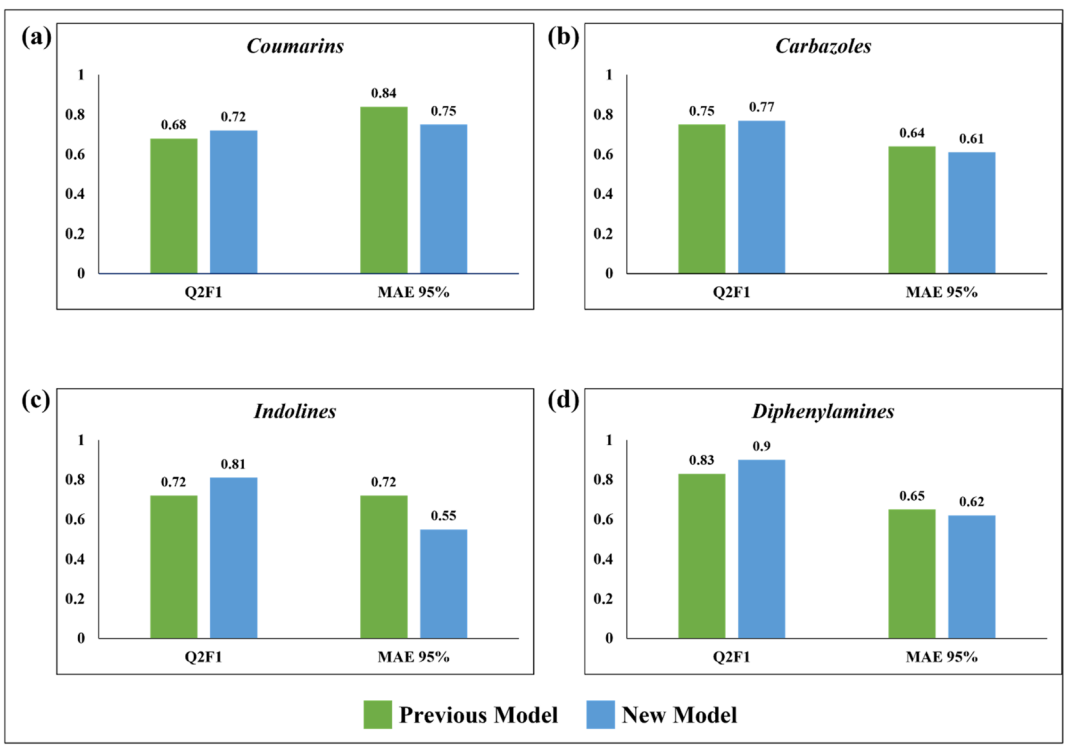


Fig. 9 Comparison between the previous PLS QSPR models and the newly developed q-RASPR PLS models.

and the present q-RASPR PLS models is shown in the form of column plots in Fig. 9. The previous study developed five individual models and consensus models of datasets, but here we have considered only the best individual models for comparison. It is clear from the column plots that our newly developed models are much more predictive. This study shows that the inclusion of RASPR descriptors along with structural and physicochemical descriptors enhances the predictivity of the models.

Conclusion

Solar energy is one of most important forms of renewable energy that meets the increasing demand of electrical energy. Various *in silico* approaches have been used to predict the power conversion efficiency (PCE) of DSSCs using structural and physicochemical features of dyes. In the present study, we have used RASPR descriptors along with 2D structural descriptors for the development of Partial Least Squares (PLS) and machine learning models of coumarin, carbazole, indoline and diphenylamine dyes. The developed q-RASPR models are statistically robust, and all the models show acceptable results for the internal validation and enhanced values of external validation metrics which indicate the importance of RASPR descriptors. The analysis of internal and external validation metrics of the developed q-RASPR models shows a surprising trend of better quality of predictions (in terms of MAE) for the test sets in comparison to the corresponding training sets which is contrary to the usual observations of better training set statistics than the corresponding test set statistics in case of QSAR models. This may happen due to a strange distribution of activity cliffs that for some reason affects the training set predictions more than the test set ones. This is indeed one of those strange (and rare) cases where the performances on the test set are better than on the training set. We have performed different cross-validation statistics to determine the quality of the developed model. We have also performed SHAP analysis on the training sets for all four datasets to determine the feature importance toward the endpoints. Using the developed and validated models, the PCE of a newly designed molecule can be determined before its synthesis, and it can save a lot of time, money, and resources. The software tools used for our work are easy to operate and most of them are freely accessible which make the modeling exercise simple and inexpensive. Hence, our developed models can give a direction for scientists and researchers present in different research areas or industrial organizations to design and synthesize new dyes. In this way, time, resources, and cost involved in the synthesis and experimentation can be reduced which may help to develop more efficient molecules.

Data availability

The DTC Lab software tools are available from **http://teqip.jdvu.ac.in/QSAR_Tools/**(MLR BestSubsetSelection and MLR plus Validation), **https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home** (Quantitative Read-Across v4.1 and RASAR Descriptor Calculator v2.0), and **https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis** (Machine Learning Model Development GUIs). The raw data files used to develop the models are provided in ESI.†

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. SP: computation, validation, software tool development, initial draft; AB: validation, editing, software tool development; KR: conceptualization, supervision, and editing.

Conflicts of interest

There are no conflicts to delare.

Acknowledgements

SP thanks the All India Council for Technical Education (AICTE), New Delhi for financial assistance. AB thanks Life Science Research Board, DRDO, New Delhi for a fellowship.

References

1 W. H. Chen and F. You, Sustainable building climate control with renewable energy sources using nonlinear model predictive control, *Renewable Sustainable Energy Rev.*, 2022, **168**, 112830.
2 T. Z. Ang, M. Salem, M. Kamarol, H. S. Das, M. A. Nazari and N. Prabaharan, A comprehensive study of renewable energy sources: Classifications, challenges and suggestions, *Energy Strategy Rev.*, 2022, **43**, 100939.
3 R. Venkateswari and S. Sreejith, Factors influencing the efficiency of photovoltaic system, *Renewable Sustainable Energy Rev.*, 2019, **101**, 376–394.
4 P. Tonui, S. O. Oseni, G. Sharma, Q. Yan and G. Tessema Mola, Perovskites photovoltaic solar cells: An overview of current status, *Renewable Sustainable Energy Rev.*, 2018, **91**, 1025–1044.
5 S. Sharma, K. K. Jain and A. Sharma, Solar cells: In research and applications—A review, *Mater. Sci. Appl.*, 2015, **06**, 1145–1155.
6 G. Nandan Arka, S. Bhushan Prasad and S. Singh, Comprehensive study on dye sensitized solar cell in subsystem level to excel performance potential: A review, *Sol. Energy.*, 2021, **226**, 192–213.
7 K. Sharma, V. Sharma and S. S. Sharma, Dye-sensitized solar cells: Fundamentals and current status, *Nanoscale Res. Lett.*, 2018, **131**, 1–46.
8 A. Baheti, K. R. Justin Thomas, C. T. Li, C. P. Lee and K. C. Ho, Fluorene-based sensitizers with a phenothiazine donor: Effect of mode of donor tethering on the performance of dye-sensitized solar cells, *ACS Appl. Mater. Interfaces*, 2015, **7**, 2249–2262.
9 W. Zhang, Y. Wu, H. Zhu, Q. Chai, J. Liu, H. Li, X. Song and W. H. Zhu, Rational molecular engineering of indoline-based d–a–π–a organic sensitizers for long-wavelength-

responsive dye-sensitized solar cells, *ACS Appl. Mater. Interfaces*, 2015, **7**, 26802–26810.
10 S. Ekins, J. Mestres and B. Testa, In silico pharmacology for drug discovery: applications to targets and beyond, *Br. J. Pharmacol.*, 2007, **152**, 21–37.
11 S. Brogi, T. C. Ramalho, K. Kuca, J. L. Medina-Franco and M. Valko, Editorial: *In silico* methods for drug design and discovery, *Front. Chem.*, 2020, **8**, 612.
12 B. Shaker, S. Ahmad, J. Lee, C. Jung and D. Na, In silico methods and tools for drug discovery, *Comput. Biol. Med.*, 2021, **137**, 104851.
13 R. A. Lewis and D. Wood, Modern 2D QSAR for drug discovery, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 505–522.
14 M. A. Lill, Multi-dimensional QSAR in drug discovery, *Drug Discovery Today*, 2007, **12**, 1013–1017.
15 A. Banerjee, M. Chatterjee, P. De and K. Roy, Quantitative predictions from chemical read-across and their confidence measures, *Chemom. Intell. Lab. Syst.*, 2022, **227**, 104613.
16 M. Chatterjee, A. Banerjee, P. De, A. Gajewicz-Skretna and K. Roy, A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data, *Environ. Sci.: Nano*, 2022, **9**, 189–203.
17 S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu and M. J. Ahsan, Machine learning in drug discovery: A review, *Artif. Intell. Rev.*, 2022, **55**, 1947–1999.
18 B. B. Goldman and W. P. Walters, Chapter 8. Machine learning in computational chemistry, *Annu. Rep. Comput. Chem.*, 2006, **2**, 127–140.
19 L. Patel, T. Shukla, X. Huang, D. W. Ussery and S. Wang, Machine learning methods in drug discovery, *Molecules*, 2020, **25**, 5277.
20 A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, *Drug Discovery Today*, 2015, **20**, 318–331.
21 S. Kar, J. K. Roy and J. Leszczynski, In silico designing of power conversion efficient organic lead dyes for solar cells using todays innovative approaches to assure renewable energy for future, *npj Comput. Mater.*, 2017, **31**, 1–12.
22 Y. Wen, L. Fu, G. Li, J. Ma and H. Ma, Accelerated discovery of potential organic dyes for dye-sensitized solar cells by interpretable machine learning models and virtual screening, *Sol. RRL*, 2020, **4**, 2000110.
23 J. K. Roy, S. Kar and J. Leszczynski, Optoelectronic properties of c60 and c70 fullerene derivatives: designing and evaluating novel candidates for efficient P3HT polymer solar cells, *Mater*, 2019, **12**, 2282.
24 A. Kumar and P. Kumar, Prediction of power conversion efficiency of phenothiazine-based dye-sensitized solar cells using Monte Carlo method with index of ideality of correlation, *SAR QSAR Environ. Res.*, 2021, **32**, 817–834.
25 H. Li, Y. Cui, Y. Liu, W. Li, Y. Shi, C. Fang, H. Li, T. Gao, L. Hu and Y. Lu, Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells, *IEEE Access*, 2018, **6**, 34118–34126.
26 J. G. Krishna, P. K. Ojha, S. Kar, K. Roy and J. Leszczynski, Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy, *Nano Energy*, 2020, **70**, 104537.
27 J. K. Roy, S. Kar and J. Leszczynski, Electronic structure and optical properties of designed photo-efficient indoline-based dye-sensitizers with D–A–π–A framework, *J. Phys. Chem. C*, 2019, **123**, 3309–3320.
28 S. Kar, J. K. Roy, D. Leszczynska and J. Leszczynski, Power conversion efficiency of arylamine organic dyes for dye-sensitized solar cells (DSSCs) explicit to cobalt electrolyte: understanding the structural attributes using a direct qspr approach, *Computation*, 2017, **5**, 2.
29 H. Li, Z. Zhong, L. Li, R. Gao, J. Cui, T. Gao, L. H. Hu, Y. Lu, Z. M. Su and H. Li, A cascaded QSAR model for efficient prediction of overall power conversion efficiency of all-organic dye-sensitized solar cells, *J. Comput. Chem.*, 2015, **36**, 1036–1046.
30 V. Venkatraman, M. Foscato, V. R. Jensen and B. K. Alsberg, Evolutionary *de novo* design of phenothiazine derivatives for dye-sensitized solar cells, *J. Mater. Chem. A*, 2015, **3**, 9851–9860.
31 V. Venkatraman and B. K. Alsberg, A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes, *Dyes Pigm.*, 2015, **114**, 69–77.
32 A. Banerjee and K. Roy, First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability, *Mol. Diversity*, 2022, **26**, 2847–2862.
33 A. Banerjee, A. Gajewicz-Skretna and K. Roy, A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites, *Mol. Inf.*, 2023, **42**, 2200261.
34 A. Banerjee, S. Kar, S. Pore and K. Roy, Efficient predictions of cytotoxicity of TiO2-based multi-component nanoparticles using a machine learning-based q-RASAR approach, *Nanotoxicology*, 2023, **17**, 78–93.
35 S. Manganelli and E. Benfenati, Use of read-across tools, *In Silico Methods for Predicting Drug Toxicity*, ed. Benfenati E., Springer, Milan, Italy, 2016, pp. 305–322.
36 K. Roy, S. Kar and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, NY, 2015.
37 R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, John Wiley & Sons, Germany, 2008.
38 A. Banerjee and K. Roy, Machine-learning-based similarity meets traditional QSAR:"q-RASAR" for the enhancement of the external predictivity and detection of prediction confidence outliers in an hERG toxicity dataset, *Chemom. Intell. Lab. Syst.*, 2023, **237**, 104829.
39 M. Goodarzi, B. Dejaegher and Y. V. Heyden, Feature selection methods in QSAR studies, *J. AOAC Int.*, 2012, **95**, 636–651.
40 M. Shahlaei, Descriptor selection methods in quantitative structure–activity relationship studies: a review study, *Chem. Rev.*, 2013, **113**, 8093–8103.
41 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer

and S. Zhao, Applications of machine learning in drug discovery and development, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477.
42 M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science*, 2015, **349**, 255–260.
43 I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, *SN Comput. Sci.*, 2021, **2**, 160.
44 A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Inc., USA., 2022.
45 G. C. McDonald, Ridge regression, *Wiley Interdiscip. Rev. Comput. Stat.*, 2009, **1**, 93–100.
46 R. Burbidge, M. Trotter, B. Buxton and S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, *Comput. Chem.*, 2001, **26**, 5–14.
47 R. J. Chase, D. R. Harrison, A. Burke, G. M. Lackmann and A. McGovern, A machine learning tutorial for operational meteorology. Part I: Traditional machine learning, *Weather Forecast.*, 2022, **37**, 1509–1529.
48 T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
49 T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay and P. Ivanov, *Jupyter Notebooks-a Publishing Format for Reproducible Computational Workflows*, 2016.
50 S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 4765–4774.
51 R. Rodríguez-Pérez and J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions, *J. Comput.-Aided Mol. Des.*, 2020, **34**, 1013–1026.
52 A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling and G. Geleijnse, Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival, *Sci. Rep.*, 2021, **11**, 6968.
53 K. Roy and I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, *Comb. Chem. High Throughput Screening*, 2011, **14**, 450–474.
54 P. S. Gangadhar, A. Jagadeesh, M. N. Rajesh, A. S. George, S. Prasanthkumar, S. Soman and L. Giribabu, Role of π-spacer in regulating the photovoltaic performance of copper electrolyte dye-sensitized solar cells using triphenylimidazole dyes, *Mater. Adv.*, 2022, **3**, 1231–1239.
55 P. S. Kalsi, *Spectroscopy of Organic Compounds*, New age international, New York, 2007.
56 W. C. Chen, S. Nachimuthu and J. C. Jiang, Revealing the influence of cyano in anchoring groups of organic dyes on adsorption stability and photovoltaic properties for dye-sensitized solar cells, *Sci. Rep.*, 2017, **7**, 1–13.
57 L. Zhang and J. M. Cole, Anchoring groups for dye-sensitized solar cells, *ACS Appl. Mater. Interfaces*, 2015, **7**, 3427–3455.
58 Y. Li, J. Liu, D. Liu, X. Li and Y. Xu, D–A–π–A based organic dyes for efficient DSSCs: A theoretical study on the role of π-spacer, *Comput. Mater. Sci.*, 2019, **161**, 163–176.
59 V. V. Divya and C. H. Suresh, Density functional theory study on the donating strength of donor systems in dye-sensitized solar cells, *New J. Chem.*, 2020, **44**, 7200–7209.
60 B. G. Kim, K. Chung and J. Kim, Molecular design principle of all-organic dyes for dye-sensitized solar cells, *Chem. - Eur. J.*, 2013, **19**, 5220–5230.
61 O. Britel, A. Fitri, A. T. Benjelloun, A. Slimi, M. Benzakour and M. Mcharfi, Theoretical investigation of the influence of π-spacer on photovoltaic performances in carbazole-based dyes for dye-sensitized solar cells applications, *J. Photochem. Photobiol., A*, 2022, **428**, 113870.
62 Z. Yao, M. Zhang, H. Wu, L. Yang, R. Li and P. Wang, Donor/acceptor indenoperylene dye for highly efficient organic dye-sensitized solar cells, *J. Am. Chem. Soc.*, 2015, **137**, 3799–3802.
63 Z. Yang, C. Shao and D. Cao, Screening donor groups of organic dyes for dye-sensitized solar cells, *RSC Adv.*, 2015, **5**, 22892–22898.
64 A. Mahmood, Triphenylamine based dyes for dye sensitized solar cells: A review, *Sol. Energy.*, 2016, **123**, 127–144.
65 M. K. Hossain, M. F. Pervez, M. N. Mia, A. A. Mortuza, M. S. Rahaman, M. R. Karim, J. M. Islam, F. Ahmed and M. A. Khan, Effect of dye extracting solvents and sensitization time on photovoltaic performance of natural dye sensitized solar cells, *Results Phys.*, 2017, **7**, 1516–1523.
66 L. Zhang and J. M. Cole, Anchoring groups for dye-sensitized solar cells, *ACS Appl. Mater. Interfaces*, 2015, **7**, 3427–3455.
67 F. Ambrosio, N. Martsinovich and A. Troisi, What is the best anchoring group for a dye in a dye-sensitized solar cell?, *J. Phys. Chem. Lett.*, 2012, **3**, 1531–1535.
68 F. Xu, T. T. Testoff, L. Wang and X. Zhou, Cause, regulation and utilization of dye aggregation in dye-sensitized solar cells, *Molecules*, 2020, **25**, 4478.
69 S. Wold, M. Sjöström and L. Eriksson, PLS-Regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
70 K. Roy, R. N. Das, P. Ambure and R. B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, *Chemom. Intell. Lab. Syst.*, 2016, **152**, 18–33.
71 A. Banerjee and K. Roy, Prediction-inspired intelligent training for the development of c-RASAR models for organic skin sensitizers: Assessment of classification error rate from novel similarity coefficients, *ChemRxiv*, 2023, preprint, DOI: **10.26434/chemrxiv-2023-20v0k**.
72 P. Ertl and A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, *J. Cheminf.*, 2009, **1**, 8.

Reg id: 1465
Poster id ADD4

Application of Read-Across Structure-Property Relationship (RASPR) as a New Tool For Predictive Modelling: Prediction of Power Conversion Efficiency (PCE) For Different Classes of Organic Dyes in Dye-Sensitized Solar Cells (DSSCs)

A) Souvik Pore
B) Arkaprava Banerjee
C) Kunal Roy
A) Drug Theoretics & Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
B) Drug Theoretics & Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
C) Drug Theoretics & Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

As an alternative to experiments, different computational tools are now popularly used for the prediction of activity/property/toxicity assessment. Read-across is one of the most widely used similarity-based unsupervised approach used for the prediction and data gap filling of property or biological activity endpoints. Quantitative structure-activity/property relationship (QSAR/QSPR) is a statistical modeling technique widely used for prediction purpose in different areas. Recently, read-across and QSAR/QSPR have been merged to develop a new modeling technique read-across structure-activity/property relationship (RASAR/RASPR) which appears to have much potential in predictive modeling [1]. To understand the potential of RASAR/RASPR in data gap filling, we have undertaken a case study of modeling Power Conversion Efficiency (PCE) of different classes of dyes used in Dye-Sensitized Solar Cell (DSSCs). We have used a large dataset of 1147 compounds covering 7 classes of organic dyes. We have initially performed read-across analysis using different similarity measures like Euclidian distance-based similarity, Gaussian kernel function similarity, Laplacian kernel function similarity with structural analogues for query compounds and calculated the weighted average predictions using the tool Read-across v4.1 available from (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). Based on the read-across optimized setting, RASAR descriptors were calculated using RASAR Descriptor Calculator v2.0 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home), and these were then merged with the chemical descriptors and finally a single partial least squares (PLS) model was developed for each of the endpoints after feature selection using the best subset selection approach. The external prediction quality of the final RASPR models superseded those of the previously developed QSPR models. The important structural features and similarity measures contributing to the PCE have been extracted using the RASPR method which can be used to enhance the PCE values in the newly designed dyes. The RASAR/RASPR method may as well be efficiently applied in modelling other property and biological activity of interest in a similar manner.

1) Banerjee, A., Roy, K., 2022. First report of q-RASAR modeling towards an approach of easy interpretability and efficient transferability. Mol. Divers. https://doi.org/10.1007/s11030-022-10478-6
2) Krishna, J. G., Ojha, P. K., Kar, S., Roy, K., & Leszczynski, J. (2020). Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy. Nano Energy, 70, 104537

Kunal Roy
kunalroy_in@yahoo.com