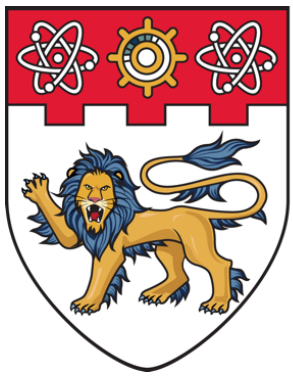


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MACHINE LEARNING APPROACHES FOR SCREENING OF MATERIALS IN FLEXIBLE ELECTRONIC DEVICES

DENG SIYAN

SCHOOL OF MATERIALS SCIENCE AND ENGINEERING

2024

MACHINE LEARNING APPROACHES FOR SCREENING OF MATERIALS IN FLEXIBLE ELECTRONIC DEVICES

DENG SIYAN

SCHOOL OF MATERIALS SCIENCE AND ENGINEERING

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

2024-01-03

. . . . . .
Date

. . . . . .
Deng Siyan

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

2024-01-03

. . . . . .

Date

. . . . . .

Assoc. Prof. Li Shuzhou

Authorship Attribution Statement

This thesis contains material from 1 paper published in the following peer-reviewed journal in which I am listed as the first author.

Chapter 4 are published as

Siyan Deng, Chao Chen, Ke Li, Xi Chen, Kelin Xia*, and Shuzhou Li*. Structure-Based Multilevel Descriptors for High-throughput Screening of Elastomers. *The Journal of Physical Chemistry B* (2023).

The contributions of the co-authors are listed as follows:

- Prof. Li Shuzhou proposed the initial project direction and conceptual framework.
- I was responsible for a multitude of tasks including the preparation and preprocessing of data, the development of predictive models, the establishment of the high-throughput screening pipeline, as well as drafting and revising the manuscript.
- Dr. Li Ke and Mr. Chen Xi performed the molecular dynamics simulations.
- Prof. Li Shuzhou, Prof. Xia Kelin, and Dr. Chen Chao provided revision suggestion.

2024-01-03

. . . . . .

Date

. . . . . .

Deng Siyan

Abstract

In the rapidly evolving field of flexible electronics, the development of advanced materials is a cornerstone for innovation. Both active and substrate materials are key to the functionality and performance of flexible devices, which are becoming increasingly integral across various applications. The integration of machine learning into the field of flexible electronics represents a groundbreaking shift in materials science, allowing for accelerated discovery and optimization of materials. This thesis presents two pivotal projects that leverage machine learning to promote advancements in flexible electronics.

The first project employs machine learning-assisted high-throughput screening (HTS) to identify elastomers with desired mechanical properties, utilizing innovative structure-based multilevel (SM) descriptors derived solely from the molecular structure of the materials. Existing elastomer descriptors necessitate both experimental and simulation data for precise prediction of elastomer properties, which may not be available for all candidates of interest. This impedes the discovery of new elastomers through HTS. Our SM descriptors are derived solely from the universally accessible molecular structure. These SM descriptors are hierarchically organized to capture both local and global structures of elastomers. With the SM-Morgan Fingerprint (SM-MF) descriptors, one of our SM descriptors, a machine learning model accurately predicts elastomer toughness with a remarkable accuracy of 0.91. Furthermore, an HTS pipeline is established to swiftly screen elastomers with targeted toughness. We also demonstrate the generality and applicability of SM descriptors by constructing HTS pipelines for screening elastomers with targeted critical strain or Young's modulus.

The second project applies deep learning-assisted HTS, enhanced by transfer learning, to identify conjugated oligomers suitable for photovoltaic materials from a vast pool of candidates. The study employs transfer learning techniques to overcome the challenge of limited data, particularly prevalent in the study of conjugated oligomers. By transferring knowledge from a extensive dataset, the models accurately predicted essential

optoelectronic properties of conjugated oligomers, including Highest Occupied Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO), and HOMO-LUMO gap, reducing mean absolute errors (MAE) significantly and outperforming direct learning models.

Throughout this thesis, the efficacy of machine learning technologies in addressing the challenges of data scarcity and complexity in material prediction is demonstrated. The SM descriptors and transfer learning models developed here not only streamline the process of material selection for flexible electronics but also set a precedent for the use of these advanced computational methods in broader materials science applications. The success of these projects underscores the potential of machine learning to revolutionize the discovery and design of new materials, promising a new era of innovation in flexible electronic devices.

Lay Summary

The world of flexible electronics is rapidly transforming with the need for innovative materials that bend and stretch while maintaining their electronic capabilities. This exciting area is crucial for the future of everything from foldable smartphones to medical sensors that can wrap around a finger. To speed up the discovery of such futuristic materials, this research taps into the power of machine learning – a form of artificial intelligence that allows computers to learn from existing data.

The research is split into two main projects. The first project creates a new way to screen through potential materials for flexible devices, focusing on elastomers – materials that can stretch and return to their original shape. By developing a smart computer model that looks at the very building blocks of these materials, we were able to predict how tough an elastomer would be without having to make and test it in the lab first. This model helps to quickly find the best candidates that could lead to flexible materials with the right kind of stretchiness and strength.

The second project dealt with a specific type of material needed for solar panels. The challenge here was that researchers didn't have enough data to train their computers to recognize which materials would work best. By using a technique called transfer learning, the computers could learn from a large set of existing data and then apply that knowledge to make accurate predictions about new materials.

In essence, this thesis showcases how machine learning can be a game-changer in finding and fine-tuning materials for the next generation of flexible electronic devices. The success of these projects not only makes material discovery faster and cheaper but also paves the way for smarter and more adaptable electronics in our daily lives.

Acknowledgements

I am deeply grateful to my mentors, Professor Li Shuzhou, for welcoming me into the world of doctoral research. His invaluable counsel and direction have been the compass during my four-year voyage towards a Ph.D. He have offered me insightful feedback and unwavering support, allowing me the liberty to delve into my research passions. His generosity in sharing his profound expertise has been crucial in the formulation of this thesis. I consider myself incredibly privileged to have been guided by such outstanding mentors at Nanyang Technological University.

Moreover, my heartfelt thanks go to all my co-researchers and every individual in our research team for their camaraderie, assistance, and motivation over these years. Their wisdom has been instrumental in shaping my research acumen. Beyond the lab, they have stood by me as cherished friends, sharing in every high and low along this journey.

I am also thankful to Nanyang Technological University for awarding me a research scholarship, which has been essential to my studies. The financial support from the National Research Foundation, Singapore, through the funding program for the Singapore Hybrid-Integrated Next-Generation μ-Electronics (SHINE) Centre, along with the computational resources from the National Supercomputing Centre Singapore, have been instrumental in facilitating the research that forms the core of this thesis.

Lastly, I owe a debt of gratitude to my family, whose love and encouragement have been the pillars of my strength throughout the duration of my Ph.D. studies. Their unwavering belief in my abilities has been a source of constant inspiration.

Table of Contents

Table Captions

Figure Captions

Figure 2.8 (a) Schematic depiction of the message passing operation for molecules and crystalline materials. (b) QM9 benchmark. Mean absolute error of the prediction of internal (red circles), highest occupied molecular orbital (HOMO, orange triangles), and lowest unoccupied molecular orbital (LUMO, inverted blue triangles) energies for different GNN models since 2017. Reproduced with permission from [108].

Figure 2.9 Machine learning model complexity and possible effects on interpretability, model performance, and the required amount of training data. Reproduced with permission from [114].

Figure 2.10 (a) Discharge capacity curves for 100th and 10th cycles for a representative cell. (b) Observed and predicted cycle lives. Reproduced with permission from [121].

Figure 2.11 Parity plots of prediction values (CNN) vs. actual values of different properties on the test set (the red line indicates exact prediction, i.e., the predicted value equals the actual value). Reproduced with permission from [122].

Figure 2.12 Predictions from the fourth generation ML model with evidential learning and molecular descriptors as the concatenated feature on the test dataset for properties vartical ionization energies (top) and hole reorganization energies (bottom). The histograms on the left plot represent the distribution of the corresponding DFT evaluated property in the test dataset. Scatter plots on the right represent the chemical space of the test dataset. The data points where the uncertainty is greater than 10% of the DFT values are in gray. Reproduced with permission from [123].

Figure 2.13 (a) The correlation matrix of elemental properties and alloy phase. (b) Gaussian process classification receiver operating characteristic (ROC) with confidence band for single-phase versus multi-phase (top) and face-centered cubic (FCC) versus body-centered cubic (BCC) versus hexagonal closest packed (HCP) single-phase (bottom) classification, respectively. (c) The area under curve (AUC) is calculated for each ROC

curve. (d) Gaussian process classification (GPC) probability as a single-phase alloy versus atomic size difference. Reproduced with permission from [124].

Figure 2.14 (a) Prediction accuracy for GPR and GPR-RFE models trained using different train set sizes, averaged over 100 runs. The corresponding test sets in (a) is the difference between total data and train sets. (b) illustrates example parity plot obtained from the GPR-RFE model (29 features) with train and test set of 76 and 19 points, respectively. Parity plots obtained from the GPR-RFE model with 95 train points and 5 unseen test points including, Sc_2O_3, Ga_2O_3, MnO, $AlCuO_2$, and $Ca_5Al_2Sb_6$, using (c) 29 features and (d) 28 features, eliminating the space group number feature from the 29 features. Reproduced with permission from [125].

Figure 2.15 Time spent for calculations(and similarly for experiments) as a function of technological developments. With the computer technological advances, the calculation step can be less time consuming than the setup construction and the results analysis. Reproduced with permission from [131].

Figure 2.16 Comparison of first- and second-generation predictions with HiTp experimental results for Co-Ti-Zr (first row), Co-Fe-Zr (second row), and Fe-Ti-Nb (third row) ternary. (A1 to A3) Prediction of GFL from the first-generation ML model. (B1 to B3) Revised predictions from the second-generation ML model. (C1 to C3) High-throughput (HiTp) experimental map of the full width at half maximum (FWHM) of the first sharp diffraction peak (FSDP) in x-ray diffraction (XRD) measurements. (D1 to D3) Experimental map of the glass-forming region (GFR) derived after application of the glass formation threshold based on amorphous silica applied to data in (C1) to (C3). Purple, glass; yellow, not glass. Reproduced with permission from [132].

Figure 2.17 (a) Schematic representation of the feedback mechanism in the dark reactions project. Machine-learning models generated from historical reaction data are used to recommend new reactions to perform, and to generate human-interpretable hypotheses about crystal formation. SVM, support vector machine. (b) Graphical representation of the

three hypotheses generated from the model, and representative structures for each hypothesis. Reproduced with permission from [133].

Figure 2.18 Summary of all the screened stable perovskites of different charge carriers. Different colored symbols indicate different classes of perovskites: red solid symbols are pure type I (AO + BO2) perovskites, blue are type II (A2O3 + B2O3) perovskites, yellow are the A-site doped type I perovskites, green are A-site doped type II, purple are the B-site doped type I, and gray are the B-site doped type II perovskites. Type I are doped with M2O3 and type II are doped with MO type oxides. The top candidate for each class has been labeled. The conductivity of BZY at 400 °C is shown by a black line for reference. Reproduced with permission from [134].

Figure 2.19 High-throughput screening of high Tg polymers with the DNN_Fingerprint model. The Tg distribution of the dataset-1, dataset-2, and dataset-3 are plotted in green, yellow, and red, respectively. The polymer samples on the right are following by their predicated Tg and true Tg values. For the sample in dataset-1 (green box), true Tg is the collected experimental value. For the samples in dataset-2 (yellow box) and dataset-3 (red box), true Tg is the MD-simulated value. More than 1,000 real polymers and 65,000 hypothetical polymers were discovered with Tg > 200℃. Reproduced with permission from [135].

Figure 2.20 Visualization of predicted permeabilities for hypothetical polymers in datasets. The data are visualized for (A) O2/N2, (B) CO2/CH4, (C) CO2/N2, and (D) H2/CO2 separations, with thousands of promising polymers lying at or above the Robeson upper bounds. Units of permeability are Barrers. Reproduced with permission from [136].

Figure 2.21 Verification of ML models with experiment. (a) Comparison of the results from four different models. (b) Schematic diagram of the cell architecture used in this study. (c) J-V curve of the solar cell with the active layer using the predicted donor material. (d) Prediction results versus experimental data for the predicted donor materials with the RF algorithm and Daylight fingerprints. Reproduced with permission from [137].

Figure 2.22 (a) Scheme of polymer design by combining the RF screening and manual screening/modification. The picked-up molecule or polymer in each stage is shown. (b) Synthesized random copolymer for OPV analysis. (c) Photoabsorption spectrum of P1 in the film state. (d) Current density–voltage curve of the best performing device. Reproduced with permission from [139].

Figure 2.23 The predicted $|V_{ij}|$ for (a) RF with 300 decision trees and (b) ANN with 6 hidden layers (the numbers of neurons in each layer are 100, and the unit of MAE value is in eV). Reproduced with permission from [140].

Figure 2.24 (a) Process chart of (I) soft material design system development and (II) soft material design process using the developed design system. (b) Comparison of train RMSE and test RMSE values obtained from LR, SVR, and NN using initial 25 data points as a training set. LR showed the smallest train RMSE and NN showed the smallest test RMSE. (c),(d) RMSE value changes when the number of data points in the training set (N) changes in both cases LR and NN were used. Reproduced with permission from [141].

Figure 2.25 (a) Nine models to predict the reorganization energy from an inexpensive (MMFF94) 3D geometry or a SMILES string. (b) Learning curves for the models obtained with 5-fold cross-validation on 80% of the data using the [1000, 1000, 1000] network with dropout values of 0.1 for the signature and molecular transform (MT) descriptors and 0.2 for the circular fingerprint (CF) descriptors. Reproduced with permission from [142].

Figure 2.26 (a) Diagram of the collaborative discovery approach: the search space decreases by over five orders of magnitude as the screening progresses. The cubes represent the size of the chemical space considered at any given stage of the process. The distinct screening stages, from left to right, involve different theoretical and computational approaches as well as experimental input and testing. (b) Number of screened molecules as a function of singlet–triplet splitting (ΔE_{ST}) and oscillator strength (*f*). Contour lines represent estimated k_{TADF} (μs^{-1}) assuming S_1 at 3.0 eV. (c) Number of screened molecules

while Group 2 incorporates two types of modified units into PDMS. The modified units serve as the hard segment (HS), crosslinking to form the network structure, while the PDMS serves as the soft segment (SS). (b) Schematic representation of the structure-based multilevel (SM) descriptors.

Figure 4.2 Model performance using different descriptors and algorithms. (a) Five-fold cross-validation accuracy scores for models trained using five different descriptors (SM-MF, SM-PF, SM-R, SM-P, and Control) and four different algorithms (GNB, SVC, AdaBoost, and RF). The grey dashed line represents the accuracy of 0.8. (b) ROC curves of models trained using the SVC algorithm and five different descriptors. (c) ROC curves of models trained using the SM-MF descriptor and four different algorithms.

Figure 4.3 Performance comparison for different level descriptor. (a) Evaluation metrics, including accuracy, precision, recall, and F1 score, for models trained using different level descriptor: SL, SP, and SM, with the SVC algorithm. The SM descriptor employed here is SM-MF descriptor. The grey dashed line represents the accuracy of 0.8. (b) ROC curves for the models trained using the different level descriptors.

Figure 4.4 Probability of candidate dataset with 460 entries being classified as "durable" elastomers. Each 2*2 grid represents a unique combination of modified unit, with each small square within a grid representing a distinct combination of SS mass and polymer mass. The color of each square indicates the probability of the corresponding combination of SS mass, polymer mass, and modified units being classified as "durable" elastomers. The grey square indicates that the corresponding modified unit combination was included in the training dataset. The embedded image displays simulated stress-strain curves for combinations m4-m9 and m3-m10.

Figure 4.5 Performance and feature importance of models predicting different mechanical properties of elastomers. (a) Evaluation metrics, including accuracy, precision, recall, and

F1 score, for models predicting toughness, critical strain, and Young's modulus. The grey dashed line represents the accuracy of 0.8. (b) ROC curves for the models predicting toughness, critical strain, and Young's modulus. (c) Feature importance of the model predicting toughness. (d) Feature importance of the model predicting critical strain. (e) Feature importance of the model predicting Young’s modulus.

Figure 4.6 Visualization of Morgan Fingerprints

Figure 4.7 Probability of candidate dataset with 460 entries being classified as "stretchable" elastomers in terms of critical strain (a) and “flexible” elastomers in terms of Young's modulus (b). Each 2*2 grid represents a unique combination of modified unit, with each small square within a grid representing a distinct combination of SS mass and polymer mass. The color of each square indicates the probability of the corresponding combination of SS mass, polymer mass, and modified units being classified as "stretchable" elastomers or “flexible” elastomers, respectively. The grey square indicates that the corresponding modified unit combination was included in the training dataset.

Figure 5.1 The comparison between PolyMaS software and our proprietary software

Figure 5.2 Schematic of the transfer learning protocol used in this work

Figure 5.3 Comparative analysis of the PubChemQC-100k and CO-610 datasets. (a) The size diversity within each dataset. (b) The chemical diversity present in the datasets.

Figure 5.4 (a) HOMO Deviation: Displays oligomers of different polymerization degrees (4-10) with the same monomer, benchmarked to degree 7. Color indicates actual HOMO values. (b) LUMO Deviation: Columns represent oligomers with degrees of polymerization (4-10), using degree 7 as reference. Color shows actual LUMO values. (c) Gap Deviation: Shows oligomers across polymerization degrees (4-10), referenced to degree 7. Color denotes actual energy gap values.

Figure 5.5 Comparative model accuracy and dataset distribution. Graphs (a), (c), and (e) depict the mean absolute error (MAE) for HOMO, LUMO, and HOMO-LUMO gap energy levels, respectively, comparing predictions by SchNet-D (black) with SchNet-T (red) models across varying polymerization degrees. Histograms (b), (d), and (f) show the energy level distribution for HOMO, LUMO, and HOMO-LUMO gap within the PubChemQC-100k (upper histograms) and CO-610 (lower histograms) datasets.

Figure 5.6 Predictive performance of SchNet-D and SchNet-T models. The correlation between the calculated values and predicted (a) HOMO, (b) LUMO, and (c) HOMO-LUMO gap using SchNet-D (top) and SchNet-T (bottom) models.

Figure 5.7 Scatter plot of predicted HOMO and LUMO levels. The plot visualizes the predicted HOMO and LUMO energy levels for a candidate dataset of 3,710 oligomers, with 85 oligomers highlighted (red stars) as promising candidates for photovoltaic applications.

Figure 6.1. Model performance using different descriptors. (a) Five-fold cross-validation scores for models trained using five different descriptors (SM-MF, SM-PF, SM-R, SM-P, and SM-HCM) and SVC algorithm. The grey dashed line represents the accuracy of 0.8. (b) ROC curves of models trained using the five different descriptors.

Figure 6.2 Probability of candidate dataset with 176 entries being classified as "high-efficient" surfactants. Each grid represents a candidate, and each column represents a distinct combination of hydrophilic and hydrophobic group. The color of each gird indicates the probability of the corresponding candidate being classified as "high-efficient" surfactant.

Figure 6.3 Overview of the SchNetE algorithm

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
DFT	Density Functional Theory
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
CNN	Convolutional Neural Networks
GNN	Graph Neural Networks
RNN	Recurrent Neural Networks
SVM	Support Vector Machines
PCA	Principal Component Analysis
TDDFT	Time-Dependent Density Functional Theory
FCC	Face-Centered Cubic
BCC	Body-Centered Cubic
HCP	Hexagonal Closest Packed
GPC	Gaussian process classification
DNN	Deep Neural Network
OPV	Organic Photovoltaic
PCE	Power Conversion Efficiency
ANN	Artificial Neural Network
QM	Quantum Mechanics
LR	Linear Regression
GBDT	Gradient Boosting Decision Tree
AFM	Atomic Force Microscope
PEDOT	Poly(3,4-ethylenedioxythiophene)
PU	Polyurethane
RE	Reorganization Energy
OSCs	Organic Semiconductors

OLED	Organic Light-Emitting Diode
EQE	External Quantum Efficiencies
RFE	Recursive Feature Elimination
AdaBoost	Adaptive Boosting
DTNNs	Deep Tensor Neural Networks
FEA	Finite Element Analysis
H-F	Hartree-Fock
B-O	Born-Oppenheimer
3D	Three-Dimensional
HTS	High-Throughput Screening
SM	Structure-based Multilevel
SDR	Simplified Dimer Representation
SM-MF	SM-Morgan Fingerprint Descriptors
SM-PF	SM-PubChem Fingerprint Descriptors
SM-R	SM-RDkit Descriptors
SM-P	SM-PaDel Descriptors
PDMS	Polydimethylsiloxane
HS	Hard Segment
SS	Soft Segment
GNB	Gaussian Naïve Bayes
SVC	Support Vector Machine Classifiers
RF	Random Forest
ROC	Receiver Operating Characteristic
CV	Cross-Validation
AUC	Area Under Curve
SL	Structure-Based Local-Level
SP	Structure-Based Polymer-Level
MD	Molecular Dynamic
V_{oc}	Open Circuit Potential
J_{SC}	Short Circuit Density
V_{ij}	Transfer Integral

Chapter 1

Introduction

This chapter provides a concise overview of the topic and the challenges that the thesis aims to tackle. It begins by exploring the imperative for developing innovative materials suitable for use in flexible electronics, highlighting the critical role of machine learning in facilitating new material discoveries. Following this, the objectives and scope are articulated, grounded in the issues identified. The structure and principal content of the dissertation are then systematically outlined. The chapter concludes the outcomes and findings of the dissertation.

1.1 Research Background

Flexible electronics, a rapidly evolving domain of modern electronics, present capabilities far beyond the scope of traditional, rigid electronics. Characterized by their malleability, these electronic devices can be bent, folded, stretched, and even rolled up without impacting their functionality.[1, 2] As shown in **Figure 1.1**, from wearable sensors and electronic skin to flexible displays and circuit, their potential applications are broad and continually expanding.[3]

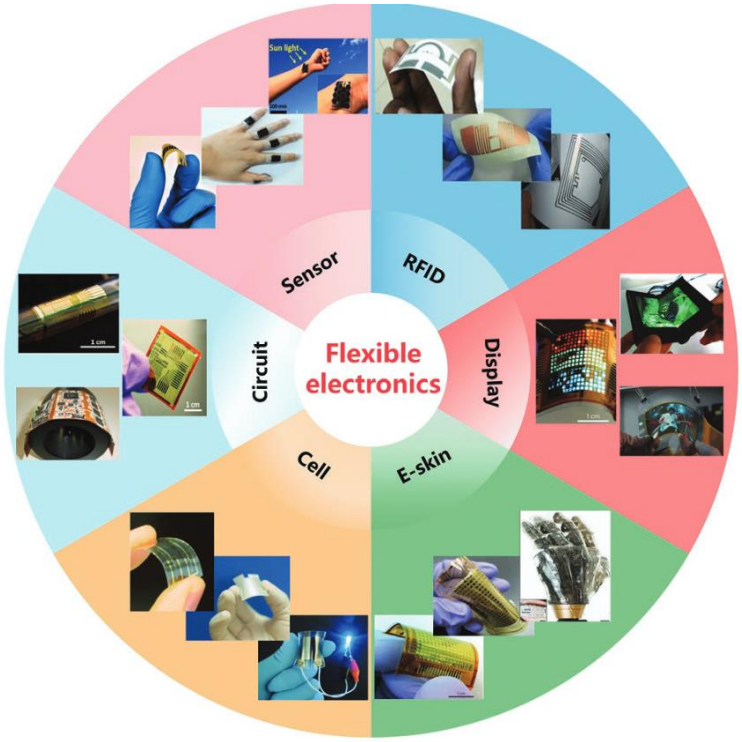


Figure 1.1 Typical flexible electronics and their applications. Reproduced with permission from [3].

At the heart of this technological revolution is the ingenious use of materials. The construction of flexible electronics typically involves a range of materials, including organic semiconductors (OSCs),[4-7] inorganic nanomaterials,[8, 9] and conductive polymers,[10-12] all of which are chosen for their ability to preserve electronic functionality

even when subjected to mechanical stress. Pioneering work by chemists and materials scientists has led to the discovery and refinement of materials and methods that support these cutting-edge devices. Over the past two decades, there has been a concerted effort to merge soft materials with electronically active ones, thereby achieving the desired flexibility and stretchability in electronic devices.[13-16] The contribution of materials chemistry to this field is indispensable, playing a crucial role in the development of new active compounds,[17-19] flexible substrates,[20-22] and innovative integration techniques.[23-25] The growing demand for such advanced materials places immense pressure on the pace of material discovery. Fortunately, the ascent of Artificial Intelligence (AI) offers potent tools to address this challenge.[26] AI and Machine Learning (ML) algorithms stand at the forefront of accelerating the discovery and optimization of novel materials,[27-29] enabling us to meet the growing needs of flexible electronic technology.

1.2 Problem Statement and Hypothesis

The development of flexible electronic devices requires materials that possess unique properties such as high electrical conductivity, mechanical flexibility, and robustness under stress. The traditional methods for screening suitable materials for such devices are largely empirical, time-consuming, and resource-intensive. They involve a trial-and-error approach that can lead to significant material and financial waste. Moreover, the rapidly advancing field of flexible electronics demands the discovery and implementation of novel materials at an accelerated pace to meet both market and technological trends. However, the vast combinatorial space of possible material compositions and configurations poses a significant challenge for conventional experimental and computational screening methods.[30]

ML approaches, particularly those utilizing advanced algorithms capable of handling large datasets and identifying complex patterns, can significantly streamline the material screening process for flexible electronic devices. By learning from existing data, ML models can predict the suitability of materials with high accuracy, thereby reducing the need for extensive physical experimentation. Our hypothesis is that by employing ML

techniques, we can accurately predict the electronic and mechanical properties of potential materials, thereby identifying candidates for flexible electronics more efficiently than traditional methods. This hypothesis extends to the assumption that ML can facilitate the discovery of materials with tailored properties that align with the specific requirements of flexible electronic applications, such as wearability, bendability, and durability. If proven correct, this hypothesis could not only expedite the material discovery process but also reduce costs and encourage innovation within the field of flexible electronics.

1.3 Objectives and Scope

The overall aim of this thesis is to use advanced ML techniques to identify and validate new materials that are applicable in the domain of flexible electronics. This work prioritizes the development of novel materials, with a special emphasis on elastomers and organic photovoltaics (OPV). The specific objectives and scope of this thesis are delineated as follows:

a) To address the constraints of machine learning-assisted high-throughput screening (HTS) in the discovery of new elastomers, this thesis proposes the development of innovative elastomer descriptors. These descriptors, predicated solely on molecular structure, aim to provide universally applicable predictive tools that can accurately forecast the properties of potential materials. The absence of experimental and/or simulation data for many elastomer candidates currently impedes HTS processes. By introducing descriptors grounded in molecular structure, this research endeavors to broaden the horizons for elastomer discovery.

b) To tackle the challenge posed by the limited data availability for conjugated oligomers, which hinders the construction of highly accurate models, this thesis employs transfer learning strategies. The process involves the meticulous selection and filtration of relevant source datasets, the acquisition of target datasets via Density Functional Theory (DFT) calculations, and the strategic choice of deep learning algorithms. The high-precision models derived from this approach are intended to establish a HTS pipeline. This pipeline

will be instrumental in identifying OPV materials that meet the stringent criteria set forth for flexible electronics.

These objectives underscore the thesis's commitment to pushing the frontiers of material discovery through sophisticated computational approaches, thereby contributing to the advancement of flexible electronic technologies.

1.4 Dissertation Overview

This thesis is dedicated to the screening of new materials used in flexible electronic applications, employing machine learning-assisted HTS methodologies. The structure of this thesis is organized as follows:

Chapter 1 introduces the field of flexible electronics, underscoring the significance of ML in the discovery of materials. It outlines the problem statement, defines the objectives of the research, and outlines the expected outcomes.

Chapter 2 provides a thorough review of the transformative impact of ML on materials science. This section presents the prevalent datasets employed in materials science, the progression of state-of-the-art descriptors, commonly used algorithms, and the broadening of application fields, with a particular focus on the integration of ML in the realm of flexible electronics.

Chapter 3 elaborates on the ML approaches and the simulation techniques employed in this thesis. The chapter covers the entire workflow, from the preparation of dataset to the development of ML models.

Chapter 4 introduces the novel set of structure-based multilevel (SM) descriptors for elastomer system. The models trained using diverse SM descriptors and algorithms were assessed in this chapter. Additionally, the screening results of elastomers exhibiting targeted toughness, achieved through HTS based on these models, are presented. Finally,

the chapter examines the generality and applicability of SM descriptors by applying the approach to screen elastomers with targeted critical strain or Young's modulus.

Chapter 5 focus on predicting electronic properties for conjugated oligomers using transfer learning techniques. This chapter provides a detailed account of the process involved in preparing the source dataset to train the base model and the target dataset to train the transfer learning model. The performance of both direct learning models and transfer learning models is thoroughly assessed. Additionally, the chapter showcases the OPV materials identified through HTS based on the transfer learning models.

Chapter 6 concludes the thesis by summarizing the key findings from the two major projects. It reflects on the contributions made to the field and proposes avenues for future research, suggesting directions that may yield further advancements in the development of flexible electronic materials.

1.5 Findings and Outcomes

The novel outcomes of this research are summarized as follows:

1. We introduced SM descriptors specifically tailored for elastomers, enabling the development of ML models capable of predicting a variety of mechanical properties with high accuracy. These properties include toughness, critical strain, and Young's modulus, for which the models achieved impressive accuracy scores of 0.91, 0.89, and 0.87, respectively.

2. Utilizing transfer learning technology, the prediction of electronic properties for conjugated oligomers, including the Highest Occupied Molecular Orbital (HOMO), the Lowest Unoccupied Molecular Orbital (LUMO), and the HOMO-LUMO energy gap, was significantly improved. The mean absolute error (MAE) for these predictions saw a notable reduction from 1.34, 0.68, and 0.71 to 0.74, 0.46, and 0.54, respectively, demonstrating enhanced accuracy in predicting these three properties.

3. HTS pipelines have been established to screen potential materials for use in flexible electronics. This includes the identification of elastomers with specified mechanical properties and the screening of conjugated oligomers suitable for use as OPV materials.

References

[1] Chen, X.; Rogers, J. A.; Lacour, S. P.; Hu, W.; Kim, D.-H. Materials chemistry in flexible electronics. *Chemical Society Reviews* **2019**, 48, 1431-1433.

[2] Huang, S.; Liu, Y.; Zhao, Y.; Ren, Z.; Guo, C. F. Flexible electronics: Stretchable electrodes and their future. *Advanced Functional Materials* **2019**, 29, 1805924.

[3] Jia, X.; Guo, R.; Tay, B. K.; Yan, X. Flexible ferroelectric devices: Status and applications. *Advanced Functional Materials* **2022**, 32, 2205933.

[4] Ling, H.; Liu, S.; Zheng, Z.; Yan, F. Organic flexible electronics. *Small Methods* **2018**, 2, 1800070.

[5] Mei, J.; Diao, Y.; Appleton, A. L.; Fang, L.; Bao, Z. Integrated materials design of organic semiconductors for field-effect transistors. *Journal of the American Chemical Society* **2013**, 135, 6724-6746.

[6] Root, S. E.; Savagatrup, S.; Printz, A. D.; Rodriquez, D.; Lipomi, D. J. Mechanical properties of organic semiconductors for stretchable, highly flexible, and mechanically robust electronics. *Chemical Reviews* **2017**, 117, 6467-6499.

[7] Wang, Y.; Sun, L.; Wang, C.; Yang, F.; Ren, X.; Zhang, X.; Dong, H.; Hu, W. Organic crystalline materials in flexible electronics. *Chemical Society Reviews* **2019**, 48, 1492-1530.

[8] Liu, Z.; Xu, J.; Chen, D.; Shen, G. Flexible electronics based on inorganic nanowires. *Chemical Society Reviews* **2015**, 44, 161-192.

[9] Sun, Y.; Rogers, J. A. Inorganic semiconductors for flexible electronics. *Advanced Materials* **2007**, 19, 1897-1916.

[10] Li, D.; Lai, W. Y.; Zhang, Y. Z.; Huang, W. Printable transparent conductive films for flexible electronics. *Advanced Materials* **2018**, 30, 1704738.

[11] Liu, H.; Li, Q.; Zhang, S.; Yin, R.; Liu, X.; He, Y.; Dai, K.; Shan, C.; Guo, J.; Liu, C. Electrically conductive polymer composites for smart flexible strain sensors: A critical review. *Journal of Materials Chemistry C* **2018**, 6, 12121-12141.

[12] Shown, I.; Ganguly, A.; Chen, L. C.; Chen, K. H. Conducting polymer-based flexible supercapacitor. *Energy Science & Engineering* **2015**, 3, 2-26.

[13] Stinner, F. S.; Lai, Y.; Straus, D. B.; Diroll, B. T.; Kim, D. K.; Murray, C. B.; Kagan, C. R. Flexible, high-speed CdSe nanocrystal integrated circuits. *Nano Letters* **2015**, 15, 7155-7160.

[14] Talapin, D. V.; Murray, C. B. PbSe nanocrystal solids for n- and p-channel thin film field-effect transistors. *Science* **2005**, 310, 86-89.

[15] Xu, S.; Yan, Z.; Jang, K.-I.; Huang, W.; Fu, H.; Kim, J.; Wei, Z.; Flavin, M.; McCracken, J.; Wang, R.; et al. Assembly of micro/nanomaterials into complex, three-dimensional architectures by compressive buckling. *Science* **2015**, 347, 154-159.

[16] Yang, J.; Choi, M. K.; Kim, D.-H.; Hyeon, T. Designed assembly and integration of colloidal nanocrystals for device applications. *Advanced Materials* **2016**, 28, 1176-1207.

[17] Wang, C.; Xia, K.; Wang, H.; Liang, X.; Yin, Z.; Zhang, Y. Advanced carbon for flexible and wearable electronics. *Advanced Materials* **2019**, 31, 1801072.

[18] Wu, Z.; Wang, Y.; Liu, X.; Lv, C.; Li, Y.; Wei, D.; Liu, Z. Carbon-nanomaterial-based flexible batteries for wearable electronics. *Advanced Materials* **2019**, 31, 1800716.

[19] Wang, C.; Xia, K.; Zhang, Y.; Kaplan, D. L. Silk-based advanced materials for soft electronics. *Accounts of Chemical Research* **2019**, 52, 2916-2927.

[20] MacDonald, W. A. Latest advances in substrates for flexible electronics. *Large Area and Flexible Electronics* **2015**, 291-314.

[21] Malik, A.; Kandasubramanian, B. Flexible polymeric substrates for electronic applications. *Polymer Reviews* **2018**, 58, 630-667.

[22] Zardetto, V.; Brown, T. M.; Reale, A.; Di Carlo, A. Substrates for flexible electronics: A practical investigation on the electrical, film flexibility, optical, temperature, and solvent resistance properties. *Journal of Polymer Science Part B: Polymer Physics* **2011**, 49, 638-648.

[23] Liu, X.; Long, Y.; Liao, L.; Duan, X.; Fan, Z. Large-scale integration of semiconductor nanowires for high-performance flexible electronics. *ACS Nano* **2012**, 6, 1888-1900.

[24] Vidor, F. F.; Meyers, T.; Hilleringmann, U. Flexible electronics: Integration processes for organic and inorganic semiconductor-based thin-film transistors. *Electronics* **2015**, 4, 480-506.

[25] Palavesam, N.; Marin, S.; Hemmetzberger, D.; Landesberger, C.; Bock, K.; Kutter, C. Roll-to-roll processing of film substrates for hybrid integrated flexible electronics. *Flexible and Printed Electronics* **2018**, 3, 014002.

[26] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, 559, 547-555.

[27] Cai, J.; Chu, X.; Xu, K.; Li, H.; Wei, J. Machine learning-driven new material discovery. *Nanoscale Advances* **2020**, 2, 3115-3130.

[28] Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *Journal of Materiomics* **2017**, 3, 159-177.

[29] Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials* **2022**, 8, 84.

[30] Corzo, D.; Tostado-Blázquez, G.; Baran, D. Flexible electronics: Status, challenges and opportunities. *Frontiers in Electronics* **2020**, 1, 594003.

Chapter 2

Literature Review

This chapter presents a thorough review of the transformative impact of machine learning on materials science. This section presents the prevalent datasets employed in materials science, the progression of state-of-the-art descriptors, commonly used algorithms, and the broadening of application fields, with a particular focus on the integration of machine learning in the realm of flexible electronics.

2.1 Overview

The development of scientific research has gone through four transformational stages, each with significant technological innovations, as shown in **Figure 2.1**.[1] The first stage is experimental science, in which empirical knowledge is accumulated through trial and error in experiments, with a long research cycle that consumes large amounts of resources. The second stage is theoretical science, which builds a theoretical framework based on empirical knowledge. However, owing to the complexity of material systems, theoretical models have also become intricate, and purely theoretical derivations gradually fall short of achieving the necessary accuracy. The third stage, computational science, involves leveraging computer technology, which synthesizes theoretical science with computational power to predict material performance. Density functional theory (DFT)[2-5] and molecular dynamics (MD)[6-10] simulation have been extensively applied in material computation simulations; however, the complexity of material systems often results in long computational times, impeding the swift exploration of material space that is required. The fourth stage is to utilize the excellent processing capability of machine learning (ML) on data to promote the exploration of material space and the development of material science.[11]

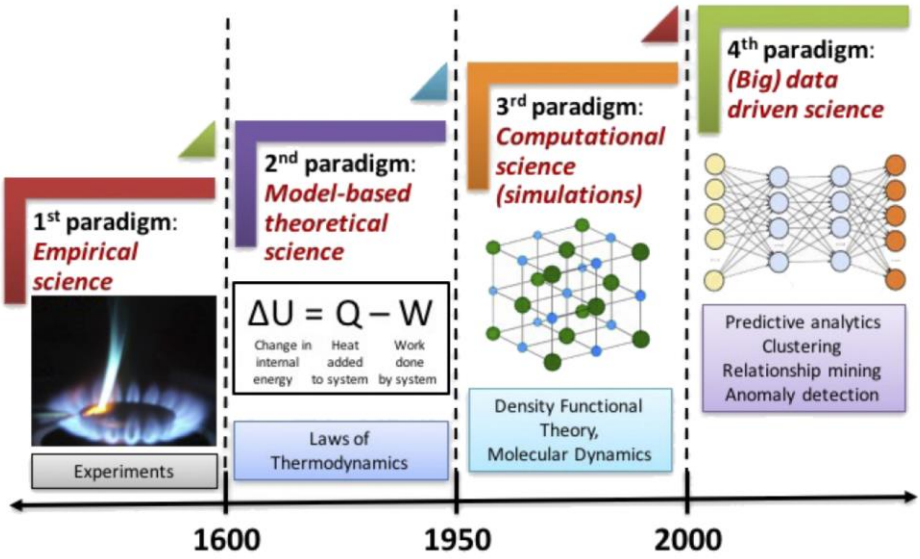


Figure 2.1 The four paradigms of science: empirical, theoretical, computational, and data-driven. Reproduced with permission from [1].

In the field of materials science, research on materials discovery through data-driven approaches has increased dramatically. Researchers use experimental[12-14] and simulation[15-17] data to provide knowledge for ML models, which are in turn used to predict the properties of materials, with the aim of screening out candidate materials with promising applications. The approach is scalable, automated, and cost-effective, offering many advantages over traditional experimental methods. However, machine learning-assisted material discovery faces many challenges, ranging from datasets, descriptor engineering to algorithmic developments.

This review aims to provide an overview of the progression of ML in the field of materials science and the challenges faced. Furthermore, the key contributions of ML in advancing materials science are reviewed. In the field of flexible electronics, materials are considered to be a key component to determine device performance. Therefore, discovering or designing suitable materials for flexible electronics has become a crucial area of endeavor. This review also reviews the progress of ML techniques in predicting and screening materials relevant to flexible electronics, a field where the convergence of materials innovation and computational power is crucial.

2.2 Machine Learning in Materials Science

When solving material problems with ML, a typical workflow is as shown in **Figure 2.2**.[18] This process begins with the careful curation and preparation of datasets, followed by the development of innovative descriptors to capture material characteristics. ML algorithms are then employed to model intricate relationships within the data, facilitating the prediction of material properties or behaviors.

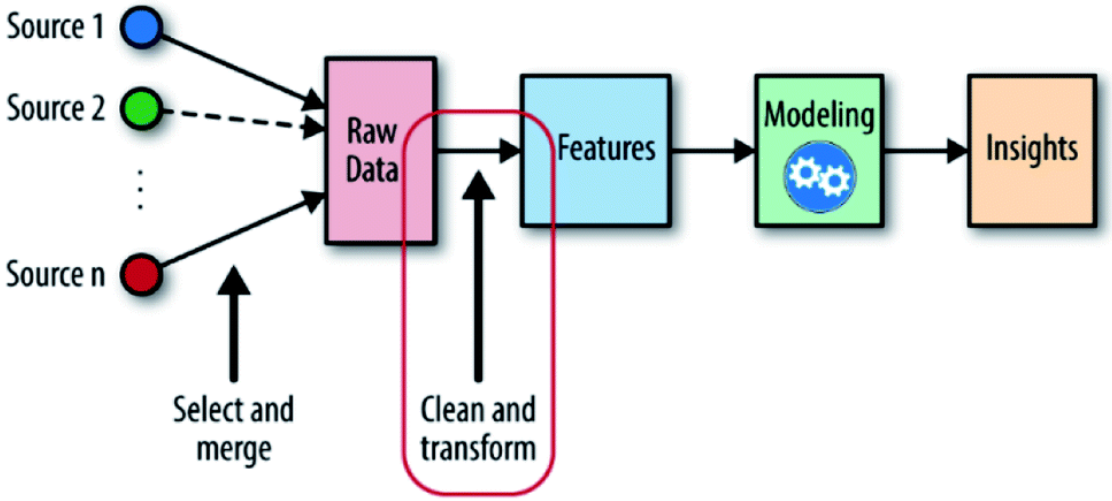


Figure 2.2 The machine learning workflow. Reproduced with permission from [18].

2.2.1 Dataset

The importance of datasets for ML in materials science is self-evident. The core of ML algorithms relies on large amounts of data to learn and discover underlying patterns in complex systems. In materials science, high-quality datasets are the cornerstone for advancing the discovery of new materials, property prediction, and scientific research.[19, 20] First, the diversity of datasets is crucial for the generalization ability of the model.[21-23] An extensive dataset covering different classes of materials can help the model better understand the differences and connections between various material properties, challenges of dataset in materials science and thus make accurate predictions even for new materials. A superior dataset not only needs to contain a sufficient number of samples, but also needs to have good data quality, such as accurate labeling, reasonable feature selection, and data consistency. These are prerequisites to ensure that ML algorithms can learn and predict effectively. Finally, with the rapid development of AI technology, the construction and maintenance of datasets have become more automated and intelligent. Through methods such as data mining,[24-27] automated experiments,[28-30] and high-throughput computing,[31-33] we can quickly generate new datasets to support the application of ML in materials science. Therefore, building comprehensive, accurate, and diverse datasets is the key to the success of ML in materials science. It not only accelerates the process of materials discovery, but

also helps scientists to deeply understand the nature of materials and promote the development and progress of materials science.

With the widespread application of ML in materials science, a large number of publicly accessible databases have become the cornerstone of research. They cover a wide range of areas in materials, from basic data on chemical structures and properties to highly specialized information on specific application areas. For example, comprehensive chemical databases such as PubChem[34] and ChemSpider,[35] which aggregate information on chemical substances worldwide, provide valuable data resources for chemistry and materials science research. Databases such as Materials Project[36] and AFLOWLIB[37] play an important role in high-throughput computation and materials property prediction, and they provide a large amount of material property data to support the design and optimization of new materials. In addition, biomacromolecular structure databases, such as Protein Data Bank (PDB),[38] provide fundamental data for organic materials research and promote the intersection of biochemistry and materials science. **Table 2.1** lists some of the crucial databases that have been widely used in materials science in recent years.[39-49] These databases not only cover various types of compounds and materials, but also provide detailed information on their properties, such as structure, thermodynamic properties, and electronic properties. This is crucial for the development and validation of new ML models.

Table 2.1 Publicly accessible databases in material science

Database Name	Description	Primary Focus
PubChem	An extensive database of chemical molecules and their activities against biological assays.	Chemical substances
ChemSpider	A free chemical structure database providing access to over 63 million structures from various sources.	Chemical structures and properties

Database Name	Description	Primary Focus
Materials Project	A database providing open access to computed information on known and predicted materials.	High-throughput computational materials
AFLOWLIB	A repository for high-throughput computational materials science.	Crystal structures and material properties
PDB (Protein Data Bank)	A database for the three-dimensional structural data of large biological molecules.	Biomacromolecular structures
ICSD (Inorganic Crystal Structure Database)	A comprehensive collection of crystal structures for inorganic compounds.	Inorganic crystal structures
Cambridge Structural Database (CSD)	A repository of small molecule crystal structures.	Organic and metal-organic crystal structures
NOMAD Repository	A database storing input and output files of density-functional theory calculations.	Computational materials science
The Crystallography Open Database (COD)	An open-access collection of crystal structures.	Crystal structures
JCAP Data Hub	A platform for sharing and analyzing data related to photoelectrochemical research.	Photoelectrochemical data
Reaxys	A database of chemical compounds, bibliographic data and chemical reactions.	Chemical reactions and properties

Database Name	Description	Primary Focus
SciFinder	A research discovery tool for accessing a wide variety of research from many scientific disciplines.	Chemical literature and patents
ZINC Database	A free database for virtual screening and purchasing of commercially-available compounds.	Drug-like molecules
CoRE MOF Database	A curated database of metal-organic frameworks.	Metal-organic frameworks
The Human Metabolome Database (HMDB)	Provides detailed information about small molecule metabolites found in the human body.	Human metabolites
DrugBank	A comprehensive database containing information on drugs and drug targets.	Pharmaceuticals and drug action
Tox21	A federal collaboration aimed at developing methods to rapidly and efficiently screen and test chemicals for potential toxicity.	Chemical safety and toxicity
The Computational Materials Repository (CMR)	A repository for materials science data, designed for the storage and sharing of data from computational materials science.	Computational materials science
The Thermodynamics Research Center (TRC)	Provides thermochemical, thermophysical, and ion energetics data compiled by NIST under the Standard Reference Data Program.	Thermodynamics data

Specifically, as polymers play a crucial role in flexible electronics, we have additionally summarized some polymer datasets as shown in **Table 2.2**.[50-53]

Table 2.2 Publicly accessible databases in polymer field

Database Name	Description	Primary Focus
Polymer Genome	A comprehensive dataset that includes various polymer properties such as glass transition temperature, dielectric constant, and elastic modulus.	Polymer property prediction and design
PolyInfo	A detailed polymer database containing information on polymer structures, properties, and applications.	Structural and property data for polymers
CAMPUS Plastics Database	Provides detailed information on properties of different plastic materials, including mechanical, thermal, and electrical characteristics.	Plastics material properties
Polymer Handbook	A collection of data on polymer properties, including mechanical, thermal, and physical properties.	Reference data for various polymer properties
PI1M Database	A database focused on high-throughput property prediction of polymers using machine learning methods, containing large datasets of polymer properties.	High-throughput property prediction of polymers
CHEMnetBASE-Polymers	Part of the CHEMnetBASE suite, this database includes detailed information on various polymers, including their structures, properties, and uses.	Comprehensive polymer data and information

In the application of ML techniques within the field of materials science, we are confronted with numerous challenges associated with datasets. These challenges not only impact the efficacy of model training but also constrain the further development of ML in this domain. The following are some of the challenges that need to be addressed:

1. Data quality and consistency: the quality of materials science datasets directly determines the accuracy of ML models. Noise, errors, or inconsistencies in the data can lead to inaccurate model predictions, affecting their performance and reliability.

Therefore, ensuring the high quality and consistency of the dataset is key to building effective ML models.[54]

2. Data size limit: the amount of data for certain material classes or attributes can be very limited, which restricts the model's ability to be trained and generalized. Insufficient data size becomes a significant challenge when dealing with complex or rare material systems.[54]

3. Data imbalance: certain types of data may be much more abundant than others in material datasets. This imbalance can potentially skew the model towards favoring data types that occur more frequently, while underrepresenting less common types. Such a bias can adversely affect the model's overall performance and its ability to generalize effectively across diverse data sets.[55]

4. Data access and sharing: data access and sharing remain significant challenges, as a substantial volume of research data is still in an unexplored state, dispersed across numerous publications. Extracting this data often entails considerable time and effort. Moreover, data sharing between various organizations encounters obstacles, further constraining data availability and hindering collaborative efforts in the field.

5. Data processing and integration: Data from different sources require proper pre-processing and integration to ensure data consistency and comparability. This process is often both complex and time-consuming.

Hence, for the effective implementation of ML in materials science, it is important to address these data-centric challenges. This includes enhancing the quality and consistency of datasets, expanding the data volume, addressing issues of data imbalance, augmenting the accessibility and sharing of data across platforms, and formulating more efficient methods for data processing and integration. Overcoming these hurdles is crucial for harnessing the full potential of ML in advancing materials science research.

2.2.2 Descriptors

Descriptors play a crucial role in ML applications to materials science. They not only affect the performance of models, but also determine the scope and efficiency of ML techniques in the discovery and performance optimization of new materials.

As shown in **Figure 2.3**, descriptors are the bridge between the atomic or molecular structure of a material and its properties, and are pivotal to the ability of ML models to understand and predict material properties. They serve as inputs to ML models that can provide detailed information about the structure and composition of a material. Importantly, the selection and design of descriptors directly affects the accuracy and generalization ability of the model. Appropriate descriptors can capture the key factors that determine material properties, while irrelevant or imprecise descriptors may lead to inaccurate or overfitting model predictions. Therefore, the development of effective descriptors is crucial to improve the effectiveness of ML applications in materials science. In order to achieve good performance, descriptors in materials science need to satisfy several key requirements:

1. Accuracy and relevance: descriptors should accurately capture the key factors that influence material properties. This means that the descriptor needs to be highly relevant to the physical, chemical or functional properties of the material. This information includes, but is not limited to, the geometry of the molecule, the electron distribution, and the type of chemical bonds. With this data, the ML model is able to recognize the correlation between the material properties and its structure for effective prediction.

2. Distinguishing ability: descriptors must have sufficient granularity to distinguish between different molecules or materials. This means that descriptors should be able to capture subtle differences that affect the properties of a material, even though these differences may be very subtle in chemical or physical structure. This ability of the descriptor to differentiate is critical to ensure that the model can accurately identify and distinguish between different compounds, alloys, or other composite materials. If descriptors are not able to effectively differentiate between similar structures or

compositions, this may result in a degradation of the predictive performance of the model, especially when dealing with material systems with high similarities or subtle differences.

3. Generalizability: descriptors should have good generalizability to many different types of materials and structures. This is to ensure that ML models can accurately predict the properties of new materials or unknown structures. With the development of the field of materials science, the emergence of novel materials raises new requirements for descriptors. These new materials may have complex structures or unknown properties, and traditional descriptors may no longer be applicable, so researchers need to continuously develop descriptors that can be adapted to new materials.

4. Computational efficiency: descriptors should be computationally efficient, especially when dealing with large-scale datasets. This is because in high-throughput screening (HTS) and big data analytics, computational efficiency is critical to the overall research schedule and cost. In high-throughput material screening, the ability to handle large amounts of data is an important indicator for evaluating the effectiveness of descriptors.

5. Interpretability: descriptors should provide intuitive physical or chemical meaning that enables researchers to understand the predictions of the model. Good interpretability helps to increase transparency and trust in the model while facilitating the development of new theories or material design principles.

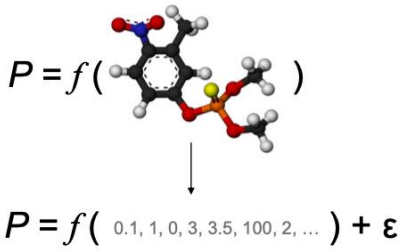


Figure 2.3 The relationship between descriptors and properties.

In the field of materials science, descriptors are categorized into two main types: experimental descriptors and theoretical descriptors.

Experimental descriptors are derived from empirical data that directly reflect the material's physical and chemical properties. They are typically utilized to characterize how a material behaves under practical conditions. For instance, physical properties such as hardness, density, and thermal conductivity describe how the material performs in physical applications. Chemical properties, like solubility, pKa value, and reactivity, delineate the material's behavior in chemical environments. Spectroscopic data, including infrared and UV-Vis spectra, provide insights into the molecular structure and functional groups present.

Theoretical descriptors, on the other hand, originate from symbolic representations of molecules, such as structural or empirical formulas. These descriptors facilitate an understanding of material properties at the atomic and molecular levels. As shown in **Figure 2.4**, theoretical descriptors are subdivided into five categories, each offering varying levels of detail and computational complexity.[56] 0D descriptors encapsulate descriptors that do not convey structural or atomic connectivity information. Examples include atom counts, bond counts, or molecular weights. The advantage of these descriptors lies in their simplicity and ease of acquisition. However, they tend to offer limited insights into molecular structure and are often used in conjunction with other descriptors to provide a more comprehensive picture. 1D descriptors include molecular descriptors calculable from a series of substructures, like functional groups. The most prevalent 1D descriptors are molecular fingerprints, which may be binary vectors indicating the presence or absence of structural features. Like 0D descriptors, 1D descriptors are straightforward to obtain and provide a basic level of structural information. 2D descriptors encompass descriptors that render information on molecular topology based on the graph representation of molecules. Classic examples of 2D descriptors include adjacency matrices,[57-59] Coulomb matrices,[60] or distance matrices.[61] These descriptors are particularly sensitive to the structural intricacies of molecules, such as size, shape, and symmetry, making them a favored choice in molecular characterization. 3D descriptors comprise geometric descriptors that detail the spatial coordinates of a molecule's atoms.[62-64] Renowned 3D descriptors include molecular

matrices and 3D-MoRSE descriptors.[65] These descriptors are invaluable for their detailed spatial information and ability to distinguish between isomers. However, the computation of 3D descriptors can be more time-intensive due to their complexity. 4D descriptors, also known as grid-based descriptors, add a fourth dimension to molecular geometry, often characterizing interactions with receptor active sites or the molecule's various conformational states.[66-69] Despite providing a wealth of information, the complexity of 4D descriptors makes them more challenging to compute.

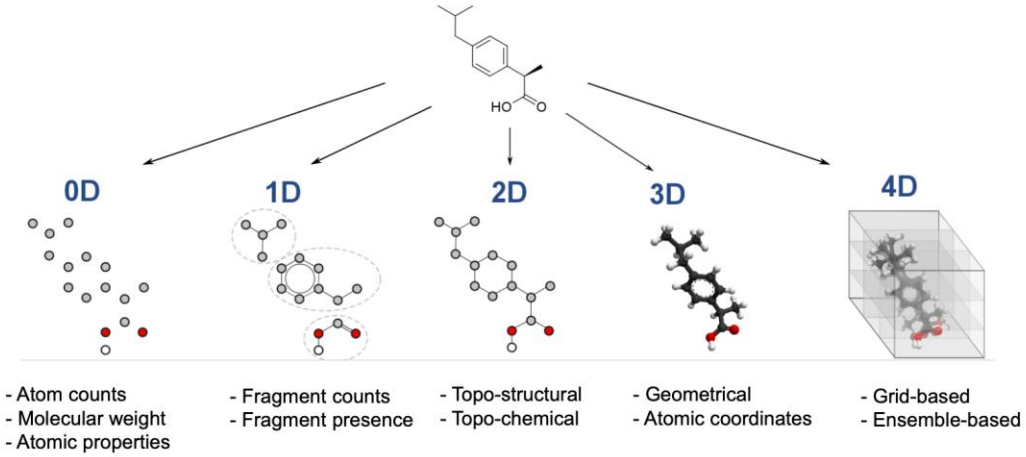


Figure 2.4 Graphical example of different molecular representations of the same structure (ibuprofen, here depicted as a 2D structure). Reproduced with permission from [56].

Theoretical descriptors have received increased attention over experimental descriptors due to their computationally accessible nature, allowing for direct calculation from molecular structures without reliance on complex experimental setups and conditions. **Figure 2.5** shows some representative theoretical descriptors.[70] A variety of software programs are available to facilitate the generation of such theoretical descriptors.[71-79] **Table 2.3** shows some examples of the software.

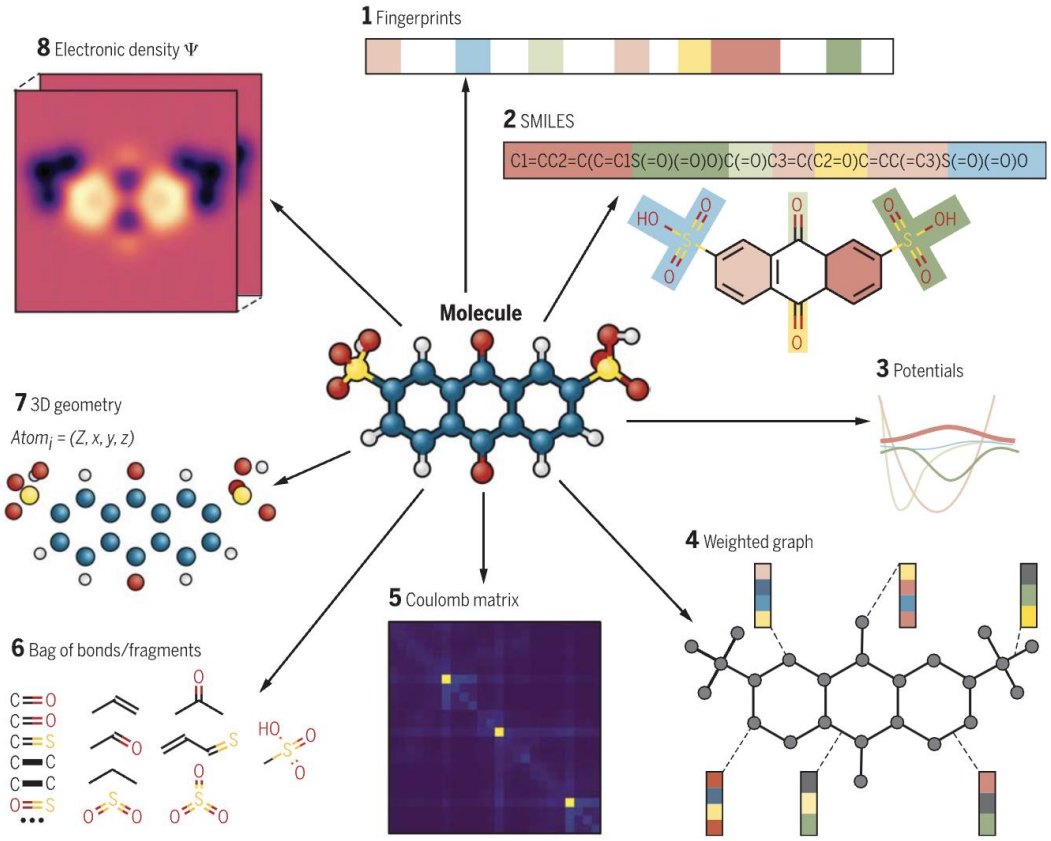


Figure 2.5 Different types of molecular representations applied to one molecule. Clockwise from top: (1) A fingerprint vector that quantifies presence or absence of molecular environments; (2) SMILES strings that use simplified text encodings to describe the structure of a chemical species; (3) potential energy functions that could model interactions or symmetries; (4) a graph with atom and bond weights; (5) Coulomb matrix; (6) bag of bonds and bag of fragments; (7) 3D geometry with associated atomic charges; and (8) the electronic density. Reproduced with permission from [66].

Table 2.3 Software programs used for generating descriptors

Software Tool	Description	Free/Proprietary
RDKit	Open-source cheminformatics software that provides tools for molecular descriptor generation, including chemical fingerprints and topological descriptors.	Free
Dragon	Provides a wide range of molecular descriptors, including topological, geometrical, 3D-stereo, and electronic descriptors.	Proprietary
Mold2	Developed by the FDA, capable of calculating a large number of 2D molecular descriptors for drug design and toxicity prediction.	Free
PaDEL	Offers a variety of 2D and 3D molecular descriptors commonly used in the prediction of molecular properties in drug discovery.	Free
CDK (Chemistry Development Kit)	An open-source Java library offering tools for generating molecular descriptors, chemical fingerprints, and other cheminformatics applications.	Free
Open Babel	A chemical file conversion tool that also provides molecular descriptor calculation functionalities.	Free
Molecular Operating Environment (MOE)	A comprehensive software platform offering a range of functionalities from molecular modeling to descriptor computation.	Proprietary
ADMET Predictor	Specialized software for predicting the ADMET properties of drugs, providing a suite of molecular descriptors.	Proprietary
QuantumATK	A software dedicated to materials modeling and computational properties, offering descriptors for material properties calculation.	Proprietary

Compared to small molecules, encoding polymer structures presents more challenges due to their complex structures and the variety of polymers. Consequently, there are fewer mature descriptors for polymers compared to those for small molecules. Polymer descriptors can be divided into monomer-level descriptors and polymer-level descriptors. For monomer-level descriptors, the descriptors suitable for small molecules can also be applied. For polymer-level descriptors, we typically use the features shown in **Table 2.4** to capture important aspects of polymer structure and properties. This table includes a mix of chemical, physical, mechanical, thermal, and optical descriptors, providing a comprehensive overview of the properties used to characterize polymers in informatics studies.

Table 2.4 Summary of common descriptors for polymers

Descriptor	Description
Molecular weight (Mw)	Average molecular weight of the polymer
Number-average molecular weight (Mn)	Average molecular weight based on the number of molecules
Polydispersity index (PDI)	Measure of the distribution of molecular weight in a polymer sample
Glass transition temperature (Tg)	Temperature at which the polymer transitions from a hard, glassy material to a soft, rubbery state
Melting temperature (Tm)	Temperature at which the polymer transitions from solid to liquid
Density	Mass per unit volume of the polymer
Crystallinity	Fraction of the polymer that is crystalline
Solubility parameter	Measure of the solubility of the polymer in different solvents
Dielectric constant	Measure of the polymer's ability to store electrical energy in an electric field
Thermal conductivity	Measure of the polymer's ability to conduct heat
Heat capacity (Cp)	Amount of heat required to raise the temperature of the polymer by one degree Celsius
Chain length	Length of the polymer chain
Permeability	Measure of the polymer's ability to allow gases or liquids to pass through

It is noteworthy that the advancement of ML techniques has endowed neural networks, especially deep learning models like Convolutional Neural Networks (CNN) and Graph Neural Networks (GNN), with the capability to autonomously extract features.[80-82] This stands in contrast to the manual selection of features traditionally required. The specifics of this process are further elaborated upon in the subsequent section on Algorithms.

2.2.3 Algorithms

Algorithms are critical to the successful application of ML in materials science. In ML, the role of algorithms is similar to that of a navigation system, guiding the model through the maze of data to find the optimal path to understand the nature of the material. The selection and optimization of algorithms directly affects the efficiency of model training and the accuracy of prediction results. An appropriate algorithm can ensure that the model learns effectively during the training process, avoid overfitting or underfitting, and ensure that the model has good generalization ability.

Among many ML algorithms, classical ML models stand out for their simplicity, interpretability and efficiency. Classical ML models can be categorized into four groups: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning, as shown in **Figure 2.6**.[83] Supervised learning algorithms, such as linear regression (LR),[84] logistic regression[85], support vector machines (SVM),[86] and random forests (RF)[87] have become important tools for researchers in material property prediction, classification, and regression analysis due to their powerful data fitting and prediction capabilities. These algorithms are not only highly accurate in predicting material properties such as mechanical strength,[88-90] thermal conductivity,[91, 92] and electronic band structure,[93-95] but they are also capable of handling problems with well-defined output objectives, such as classification,[96-98] enabling researchers to rapidly identify candidate materials with potential applications in massive amounts of material data. Unsupervised learning algorithms, including clustering and dimensionality reduction techniques such as K-mean clustering,[99] and Principal Component Analysis (PCA),[100] demonstrate their powerful data exploration capabilities on unlabeled datasets. By discovering the intrinsic structure and

patterns of data, these algorithms help researchers understand the complexity and diversity of materials and identify different classes or states of materials, thus providing a scientific basis for material design and functional optimization.[101-103] Reinforcement learning algorithms are applicable to decision-making problems in dynamic systems by modeling the decision-making process of an intelligent body in its environment and learning the optimal policy. In materials science, reinforcement learning is used to optimize material processing and synthesis processes,[104-106] as well as to develop design strategies for new materials,[107-109] which are complex problems requiring a series of decisions under changing conditions.

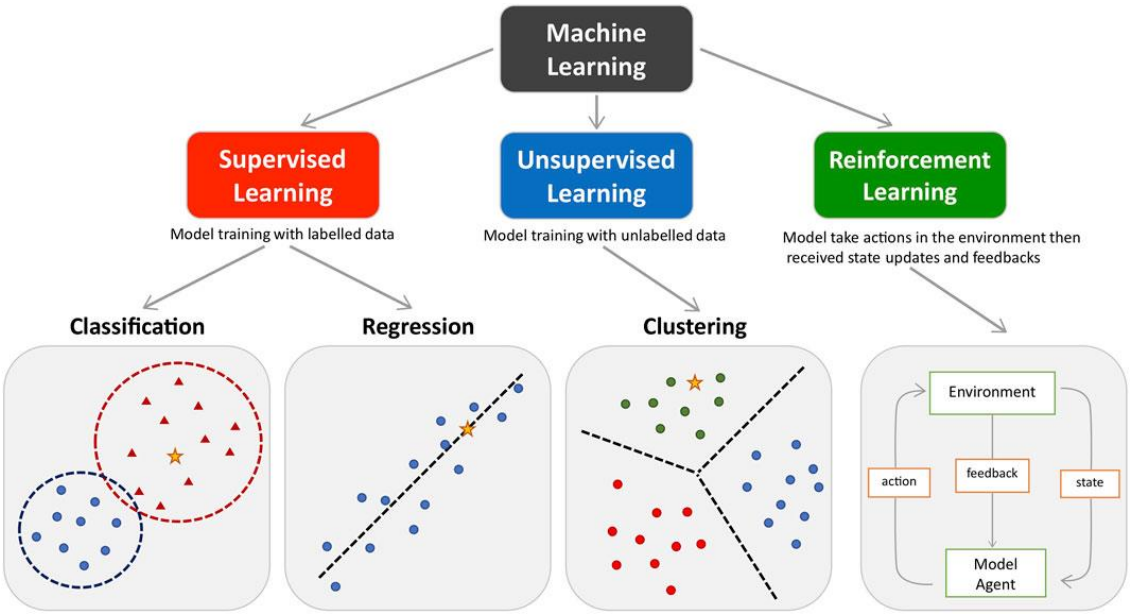


Figure 2.6 The main types of machine learning. Main approaches include classification and regression under the supervised learning and clustering under the unsupervised learning. Coloured dots and triangles represent the training data. Yellow stars represent the new data which can be predicted by the trained model. Reproduced with permission from [83].

In summary, classical ML algorithms play diverse roles in materials science, from simple supervised learning to complex reinforcement learning. They are each applicable to different problems and data types, and together they constitute a powerful toolset that continues to advance materials science and facilitate the discovery and application of new materials. However, despite the significant progress that has been made, the application of

ML in materials science still faces many challenges, such as interpretability of algorithms, and generalization ability of predictive models. Future research will continue to deepen the application of these algorithms, optimize their performance, and explore new algorithms to better meet the needs of materials science.

Deep learning, a jewel in the field of Artificial Intelligence (AI), is showing its strong potential for applications in materials science. Different types of deep learning models, such as CNN, GNN, Recurrent Neural Networks (RNN) as shown in **Figure 2.7**, each with its own strengths, provide new perspectives on structure identification, performance prediction, and the design of new materials.[110] CNN is a powerful tool for processing image data, which extracts localized features in an image by mimicking the workings of the human visual system. In materials science, CNN is used to analyze microstructural images of materials, such as scanning electron microscopy images,[111] to identify and classify different material phases or defects. CNN is able to learn patterns and textures in images, which is crucial for understanding the microstructure of materials and predicting their macroscopic properties. GNN is one of the most suitable deep learning models for processing molecular and crystal structure data as shown in **Figure 2.8**.[112] They are able to operate directly on graph data, recognizing the topology of molecules or lattices and automatically extracting features about the chemical and physical properties of materials. With GNN, researchers are able to predict the properties of materials,[113, 114] thereby accelerating the design and discovery process of new materials. RNN is particularly suited for processing sequential data, such as time-series material performance data. The ability of RNN to memorize information from the preceding sequential data through their internal recurrent structure allows RNN to take advantage of their unique strengths in analyzing and predicting time-dependent performance of materials under different conditions.[115, 116]

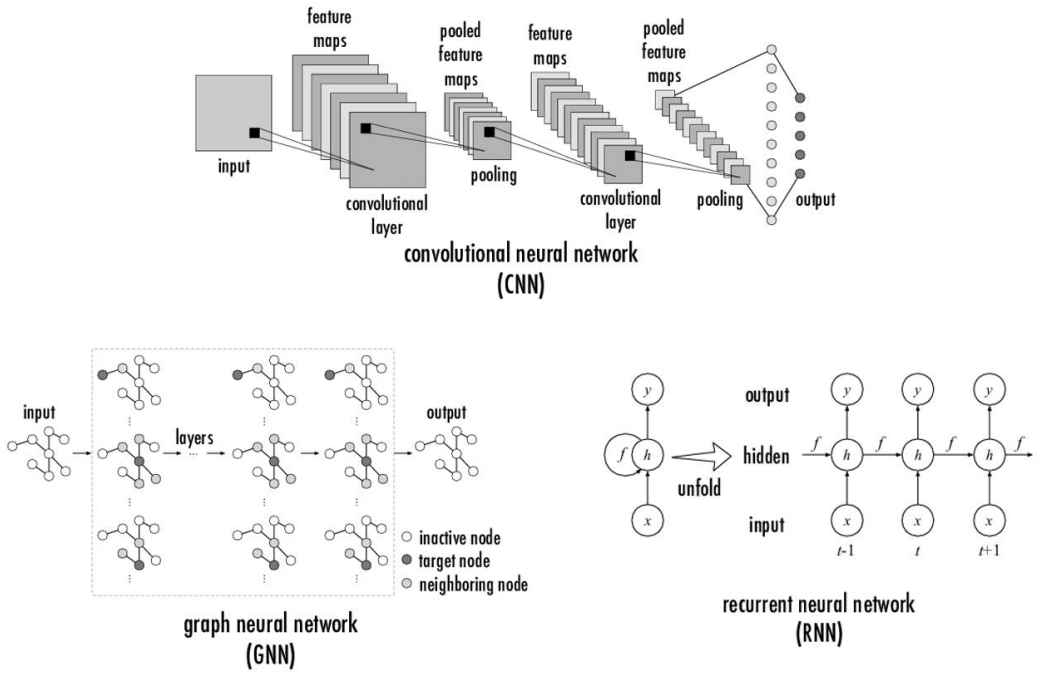


Figure 2.7 Illustrative examples of CNN, GNN, and RNN network architectures. Reproduced with permission from [110].

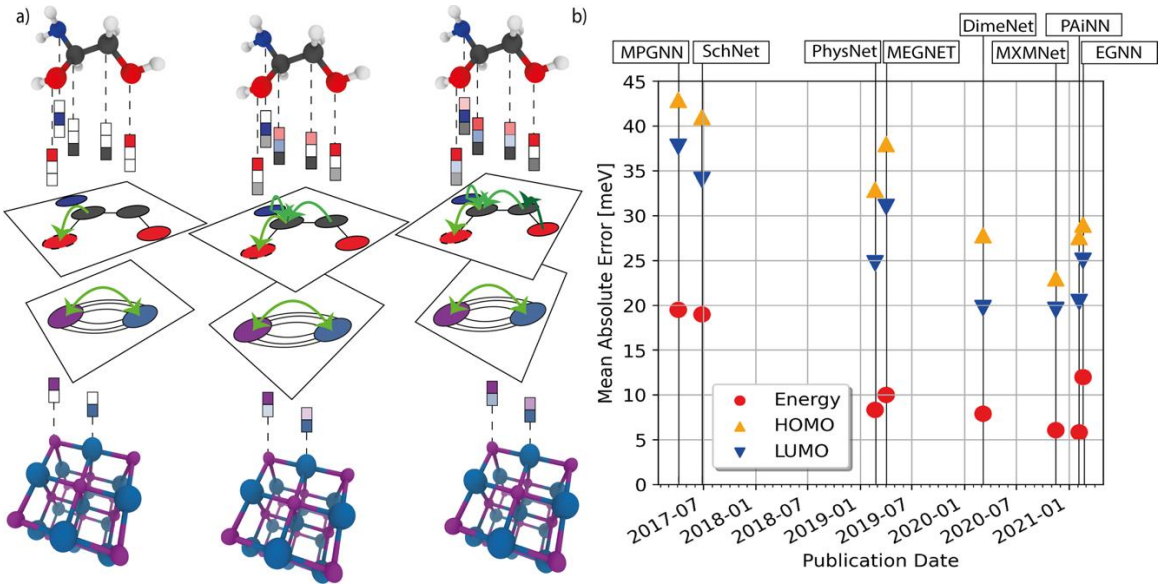


Figure 2.8 (a) Schematic depiction of the message passing operation for molecules and crystalline materials. (b) QM9 benchmark. Mean absolute error of the prediction of internal (red circles), highest occupied molecular orbital (HOMO, orange triangles), and lowest unoccupied molecular orbital (LUMO, inverted blue triangles) energies for different GNN models since 2017. Reproduced with permission from [112].

One of the most desirable features of these deep learning models is their ability to automatically extract and learn features from data without having to manually set specific descriptors.[117] This feature greatly reduces the effort of pre-processing the data, allowing the models to learn useful information directly from the raw data and automatically optimize the feature representation to improve the accuracy of the predictions. Such automatic feature extraction is extremely valuable for dealing with large-scale material databases, as it allows the model to continuously adapt and improve its understanding of material properties during training. However, there are challenges to the application of deep learning models in materials science. As shown in **Figure 2.9**, with the increase of model complexity, interpretability of models is a key issue, as the decision-making process of deep networks is often difficult to understand.[118] In addition, deep learning models usually require a large amount of labeled data for training, which is not always feasible in materials science. Therefore, how to design interpretable and data-efficient deep learning models and how to combine a small amount of labeled data with a large amount of unlabeled data are important directions for future research.

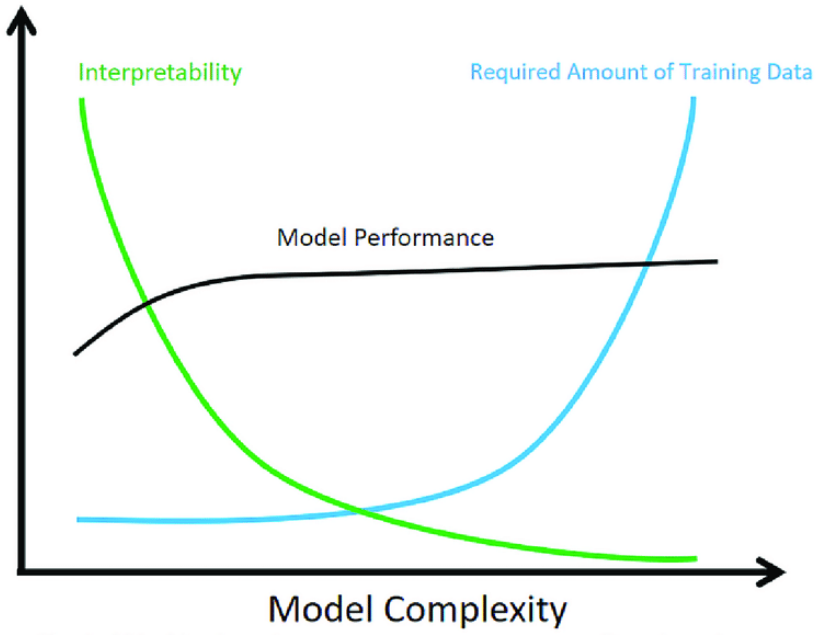


Figure 2.9 Machine learning model complexity and possible effects on interpretability, model performance, and the required amount of training data. Reproduced with permission from [118].

2.3 Applications

2.3.1 Property Prediction

An important role of ML in the field of materials is material property prediction.[119-124] ML has great potential for fast and efficient prediction of material properties compared to traditional experimental or simulation methods.

Severson et al. use ML models to accurately predict battery lifetime using early-cycle data.[125] This study involves generating a comprehensive dataset comprising 124 commercial lithium iron phosphate/graphite cells, subjected to fast-charging conditions and exhibiting diverse cycle lives ranging from 150 to 2300 cycles. The primary focus is on employing ML tools to predict and classify cells by cycle life, based on early discharge voltage curves that show no signs of capacity degradation as shown in **Figure 2.10a**. The research showcases the utilization of ML models to predict battery behavior accurately, even before significant degradation occurs. It could revolutionize the battery manufacturing process, enabling rapid validation of new production techniques and efficient sorting of cells based on expected lifetimes.

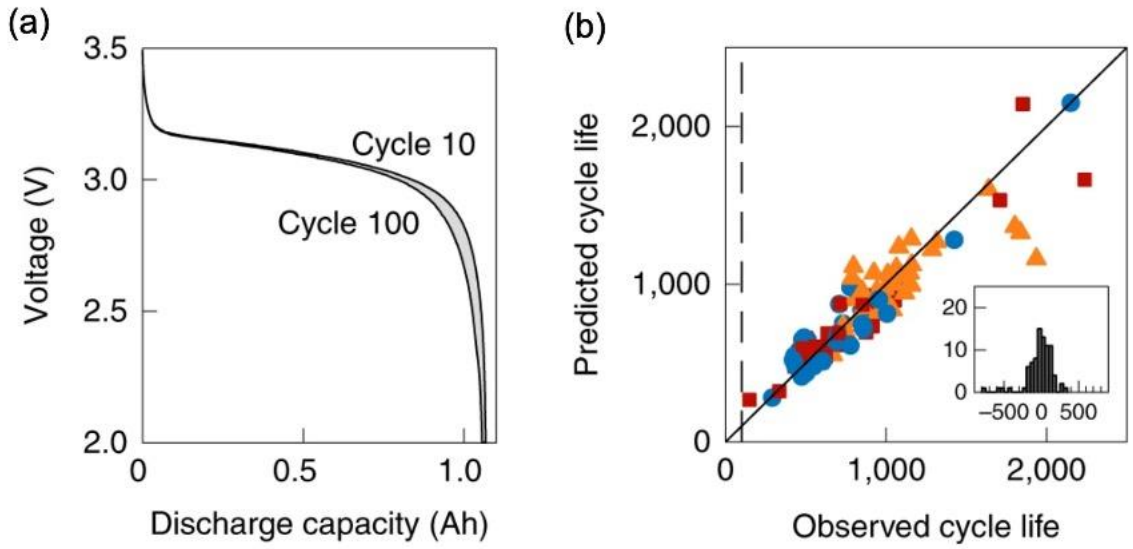


Figure 2.10 (a) Discharge capacity curves for 100th and 10th cycles for a representative cell. (b) Observed and predicted cycle lives. Reproduced with permission from [125].

Casey et al. use the neural network model trained on over 20,000 molecules to predict various properties of energetic materials including dipole moment, total electronic energy, detonation velocity, pressure, temperature, crystal density, HOMO-LUMO gap, and solid phase heat of formation.[126] The authors develop a CNN that can parse the 3D electronic structure of molecules, represented as 4D tensors combining charge density and electrostatic potential. The results as shown in **Figure 2.11** demonstrate that this method can predict multiple properties of energetic materials with high accuracy, highlighting the potential of 3D CNN in material science. This approach can revolutionize the way properties of materials are predicted, offering a more efficient, cost-effective, and less labor-intensive alternative to traditional methods.

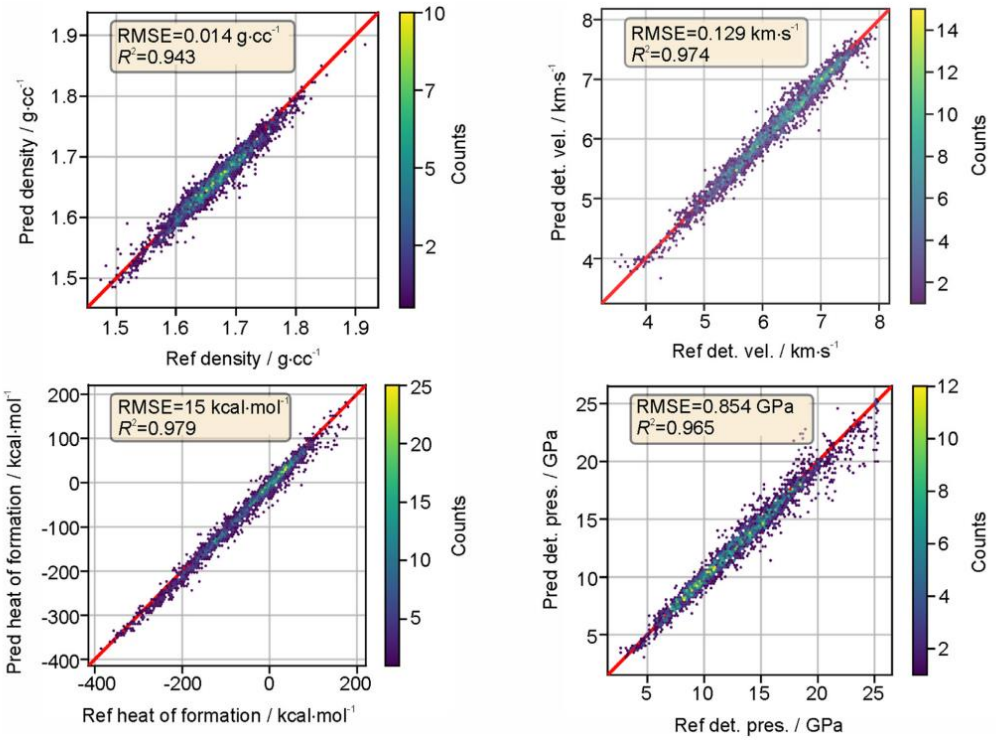


Figure 2.11 Parity plots of prediction values (CNN) vs. actual values of different properties on the test set (the red line indicates exact prediction, i.e., the predicted value equals the actual value). Reproduced with permission from [126].

Bhat et al. present an extensive dataset comprising 25,000 molecular entities characterized by a spectrum of properties ascertained through the methodologies of DFT and its time-dependent variant (TDDFT).[127] This rich dataset has been utilized to train a suite of ML

models, ranging from traditional approaches such as ridge regression to sophisticated GNN that interpret molecular SMILES strings. The investigation delineates the superiority of GNN, particularly those augmented with context-rich information, in delivering enhanced predictive accuracy for a myriad of molecular properties. As shown in **Figure 2.12,** an innovative aspect of their study is the integration of uncertainty quantification within their top models, infusing a measure of confidence in the predictive outcomes. In a laudable move towards open science, Bhat et al. have launched an interactive online portal, ensuring broad accessibility to this dataset and the corresponding ML models, thus making a substantial contribution to the advancement of materials science and computational chemistry.

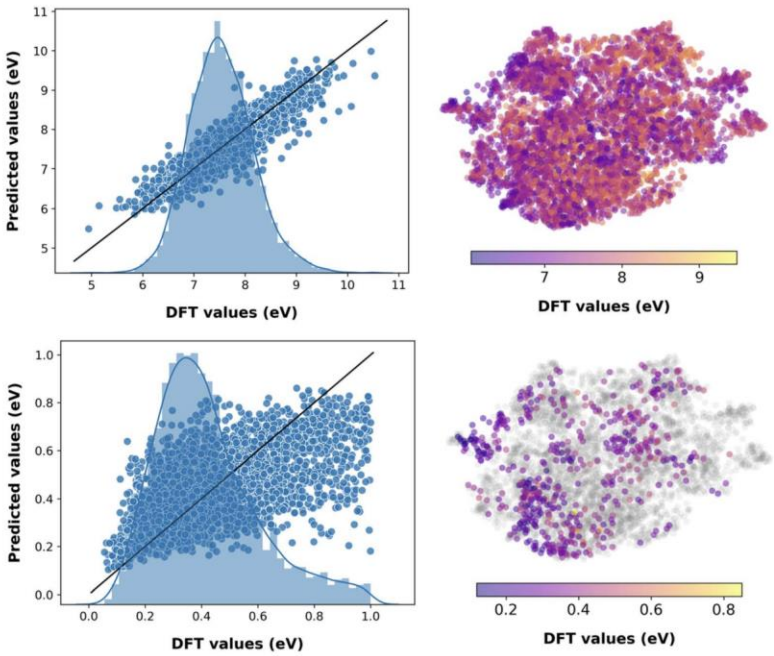


Figure 2.12 Predictions from the fourth generation ML model with evidential learning and molecular descriptors as the concatenated feature on the test dataset for properties vartical ionization energies (top) and hole reorganization energies (bottom). The histograms on the left plot represent the distribution of the corresponding DFT evaluated property in the test dataset. Scatter plots on the right represent the chemical space of the test dataset. The data points where the uncertainty is greater than 10% of the DFT values are in gray. Reproduced with permission from [127].

In a pivotal study conducted by Pei, the limitations of empirical rules for predicting the formation of solid solutions were addressed through the utilization of a substantial dataset comprising 1252 multicomponent alloys.[128] The study further elucidated key features such as molar volume, bulk modulus, and melting temperature, instrumental in the formation of solid solutions. The performance of the ML models as shown in **Figure 2.13b, c** demonstrated the method's predictive prowess. Building on these insights, Pei introduced a new thermodynamics-based rule that, while slightly less precise at 73% accuracy, is grounded in the physical aspects of the problem. This rule was adeptly applied to predict solid solutions across various elemental blocks, covering FCC, BCC, and HCP structures, with a high-throughput approach. The simplicity of the new rule, relying solely on elemental properties, marks a significant stride in the efficient screening of high-entropy alloys for solid solution formation, enhancing its practical utility in material science.

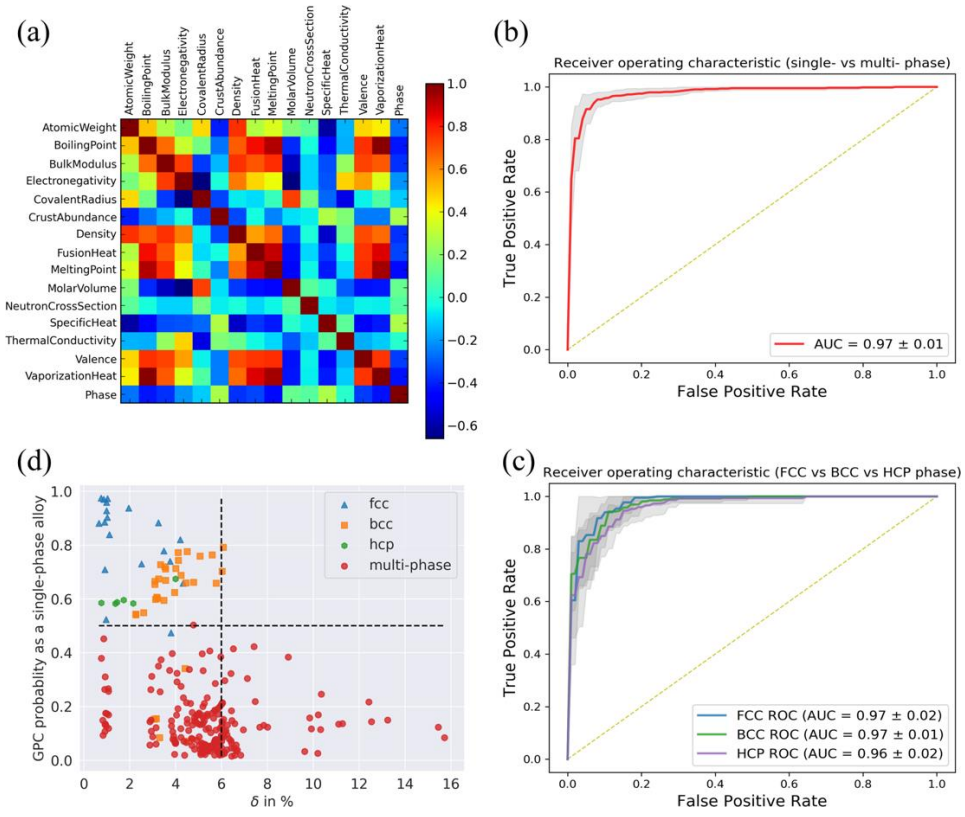


Figure 2.13 (a) The correlation matrix of elemental properties and alloy phase. (b) Gaussian process classification receiver operating characteristic (ROC) with confidence band for single-phase versus multi-phase (top) and face-centered cubic (FCC) versus body-centered cubic (BCC) versus hexagonal closest packed (HCP) single-phase (bottom) classification, respectively. (c) The area under curve (AUC) is calculated for each ROC curve. (d) Gaussian process classification (GPC)

probability as a single-phase alloy versus atomic size difference. Reproduced with permission from [128].

Chen's research addresses the critical challenge of measuring lattice thermal conductivity in materials such as thermoelectrics and semiconductors, where empirical methods are notably difficult.[129] To overcome the limitations of accuracy and computational expense in existing theoretical methods, Chen leverages ML to predict thermal conductivity for inorganic materials. Utilizing a dataset of experimental thermal conductivity measurements from roughly 100 inorganic materials, coupled with sophisticated feature engineering and the Gaussian process regression algorithm, Chen's ML model showcases remarkable predictive accuracy and speed. As shown in **Figure 2.14**, they compared the performance of models trained with training sets of different sizes and with different numbers of feature to optimize the training dataset and features selected. The optimized model not only outperforms or matches the precision of traditional computational methods but also provides insights into the critical factors influencing thermal transport in non-metals. Chen's work culminates in a ML framework that significantly enhances the ability to design and screen new materials with optimized thermal conductivity, marking a substantial contribution to the understanding and application of heat transport in solid-state physics.

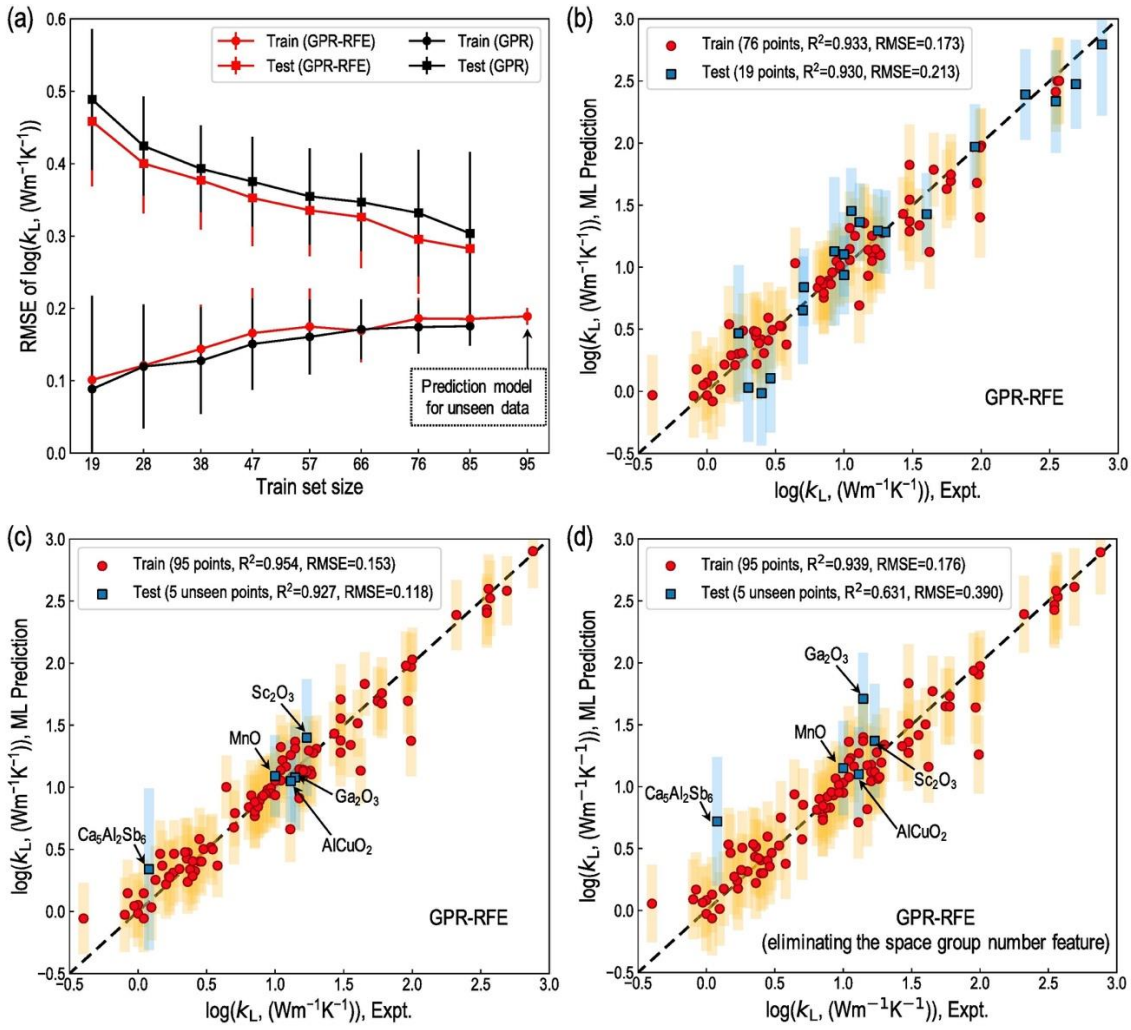


Figure 2.14 (a) Prediction accuracy for GPR and GPR-RFE models trained using different train set sizes, averaged over 100 runs. The corresponding test sets in (a) is the difference between total data and train sets. (b) illustrates example parity plot obtained from the GPR-RFE model (29 features) with train and test set of 76 and 19 points, respectively. Parity plots obtained from the GPR-RFE model with 95 train points and 5 unseen test points including, Sc_2O_3, Ga_2O_3, MnO, $AlCuO_2$, and $Ca_5Al_2Sb_6$, using (c) 29 features and (d) 28 features, eliminating the space group number feature from the 29 features. Reproduced with permission from [129].

Significant progress has been made in the application of ML to property prediction. By constructing large-scale datasets and combining them with advanced algorithms, researchers have been able to accurately predict a variety of important properties of materials. These models not only improve the speed and accuracy of predictions, but also reveal key factors that affect material properties. In addition, these studies provide

materials science researchers with a large amount of data and tools through open sharing, providing a solid foundation for future materials innovation and applications. Therefore, the successful application of ML in materials property prediction will undoubtedly continue to advance materials science and bring more innovative materials solutions.

2.3.2 Materials Discovery

Researchers have long been dedicated to discovering or designing new materials that will improve the performance of material applications and reduce the cost of industrialized material preparation and applications.[130-132] The new scientific paradigm shift is being led by ML and data-driven approaches in today's materials science. Scientific progress has undergone several major shifts and is now in the fourth scientific paradigm - data-driven scientific research.[1] At the core of this paradigm is the use of advanced computational technologies, artificial intelligence, and massive data resources to enable the acceleration of scientific discovery and theoretical innovation.[133, 134] As time advances, the computational power is swiftly expanding. This leads to a significant decrease in the duration needed for calculations, shifting more time towards the setup and analysis of simulations. This evolution has altered the conventional workflow and paved the way for novel research methodologies. Rather than conducting numerous simulations prepared by hand, it's now feasible to automate the generation of inputs and execute numerous simulations, potentially in the millions, either in parallel or one after another. This progression is depicted in **Figure 2.15**, and this method is referred to as high-throughput. In materials science, this shift is reflected in the HTS and analysis of huge material databases through ML algorithms to rapidly discover new materials with potential applications.[135] This approach not only greatly improves research efficiency, but also pushes the boundaries of new material design, providing new directions and strategies for the future development of materials science.

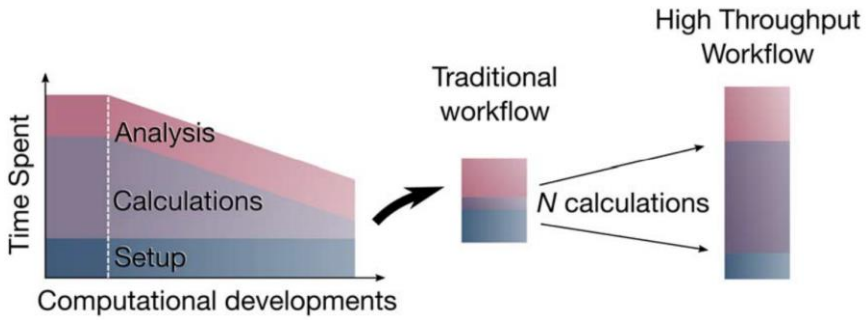


Figure 2.15 Time spent for calculations(and similarly for experiments) as a function of technological developments. With the computer technological advances, the calculation step can be less time consuming than the setup construction and the results analysis. Reproduced with permission from [135].

Ren et al. have made a compelling contribution to the field of new materials discovery with their use of ML to streamline high-throughput experiments.[136] Focusing on the Co-V-Zr ternary system, they trained an ML model incorporating physiochemical theories and synthesis method factors, which led to the successful prediction and discovery of new metallic glasses. As shown in **Figure 2.16**, despite initial discrepancies in the precise compositions predicted, the authors iteratively refined the ML model using experimental feedback, resulting in improved accuracy for not only the Co-V-Zr system but also other validation datasets. This enhanced ML model subsequently guided the discovery of metallic glasses in two unexplored ternaries. Ren et al.'s research illustrates the power of integrating ML with high-throughput experimentation, not only in the discovery of three new glass-forming systems but also in establishing a synthesis method–sensitive predictive tool that gains accuracy from continued use. This iterative, data-driven discovery approach holds great potential for accelerating the development of various technologically significant materials, especially those whose properties are synthesis path–dependent and challenging for current physiochemical theories to predict.

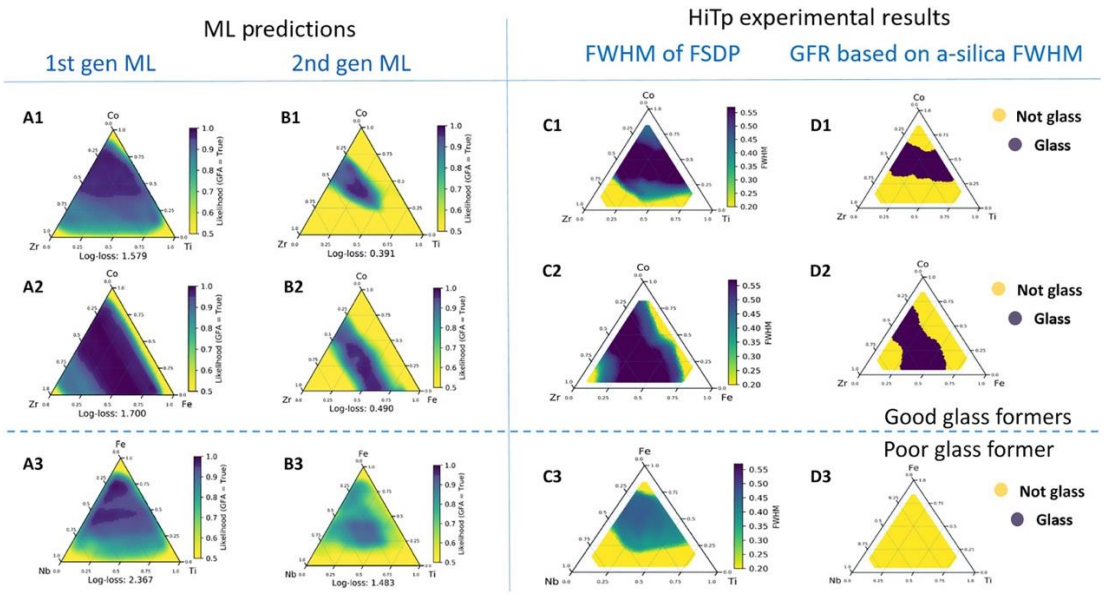


Figure 2.16 Comparison of first- and second-generation predictions with HiTp experimental results for Co-Ti-Zr (first row), Co-Fe-Zr (second row), and Fe-Ti-Nb (third row) ternary. (A1 to A3) Prediction of GFL from the first-generation ML model. (B1 to B3) Revised predictions from the second-generation ML model. (C1 to C3) High-throughput (HiTp) experimental map of the full width at half maximum (FWHM) of the first sharp diffraction peak (FSDP) in x-ray diffraction (XRD) measurements. (D1 to D3) Experimental map of the glass-forming region (GFR) derived after application of the glass formation threshold based on amorphous silica applied to data in (C1) to (C3). Purple, glass; yellow, not glass. Reproduced with permission from [136].

Raccugli and colleagues present an innovative ML approach to predict the outcomes of hydrothermal synthesis reactions, specifically for the crystallization of templated vanadium selenites.[137] As shown in **Figure 2.17,** utilizing archived data from unsuccessful reactions along with cheminformatics techniques, the researchers trained a machine-learning model that significantly outperformed traditional methods, achieving an 89% success rate in predicting new organically templated inorganic product formation. This approach not only provides a new predictive capability to guide synthetic efforts but also offers fresh insights into the underlying conditions for successful material synthesis. Their work marks a significant advance in the use of historical chemical reaction data, suggesting that such data-driven approaches can uncover new chemical principles and significantly improve the discovery and development of inorganic–organic hybrid materials.

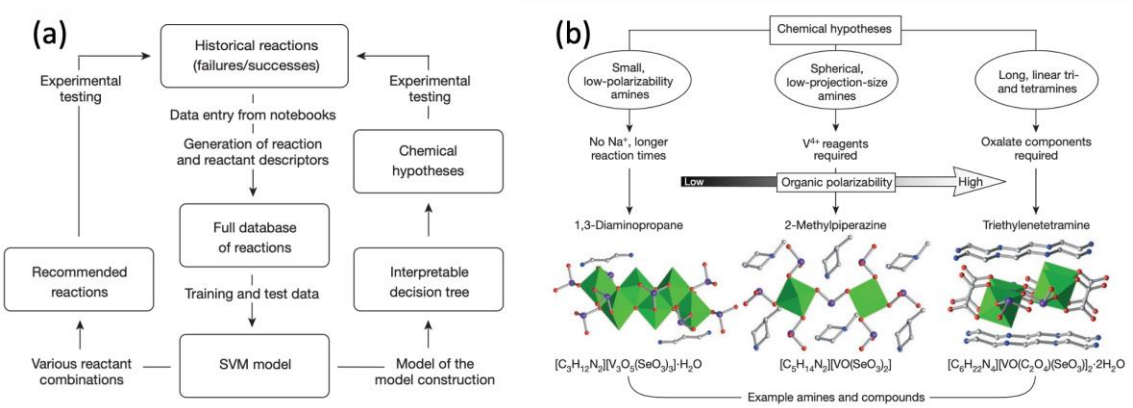


Figure 2.17 (a) Schematic representation of the feedback mechanism in the dark reactions project. Machine-learning models generated from historical reaction data are used to recommend new reactions to perform, and to generate human-interpretable hypotheses about crystal formation. SVM, support vector machine. (b) Graphical representation of the three hypotheses generated from the model, and representative structures for each hypothesis. Reproduced with permission from [137].

Priya et al. employ ML to sift through literature data on ABO3-type perovskite oxides, crucial for energy-related applications, aggregating over 7000 data points.[138] Their work focuses on predicting total conductivity and categorizing perovskites based on charge carriers under varying temperatures and environments. Key predictive features like average ionic radius and minimum electronegativity were identified to determine conductivity and charge carrier types. With an XGBoost algorithm, they screened thousands of doped and undoped perovskites as shown in **Figure 2.18**, pinpointing promising candidates for high conductivity. Some candidates, such as EuNbO3 and EuSnO3, demonstrated potential as low-temperature proton-conducting electrolytes.

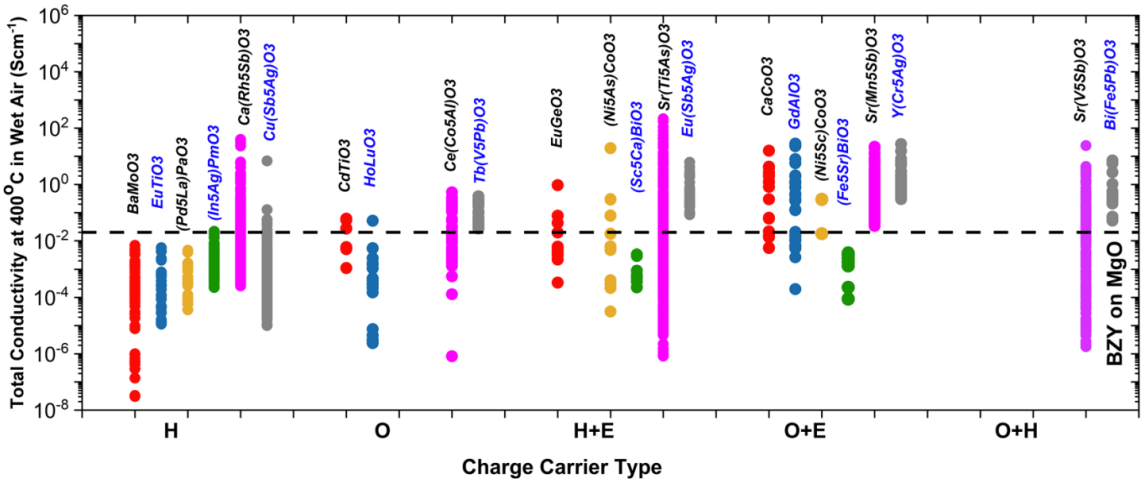


Figure 2.18 Summary of all the screened stable perovskites of different charge carriers. Different colored symbols indicate different classes of perovskites: red solid symbols are pure type I (AO + BO2) perovskites, blue are type II (A2O3 + B2O3) perovskites, yellow are the A-site doped type I perovskites, green are A-site doped type II, purple are the B-site doped type I, and gray are the B-site doped type II perovskites. Type I are doped with M2O3 and type II are doped with MO type oxides. The top candidate for each class has been labeled. The conductivity of BZY at 400 °C is shown by a black line for reference. Reproduced with permission from [138].

Tao et al. incorporate a ML approach with high-fidelity MD simulations to predict the glass transition temperature (Tg) of polymers based on their chemical structure.[139] Drawing from a vast dataset of nearly 13,000 homopolymers, they effectively trains a deep neural network (DNN) using experimental Tg values. The DNN's predictive capabilities are validated against MD simulations and experimental results, showcasing its potential for HTS. As shown in **Figure 2.19**, this innovative method led to the identification of over 65,000 polymers with Tg above 200°C, significantly expanding the scope of known high-temperature polymers. This advancement marks a substantial leap in the discovery and design of materials suitable for high-temperature applications.

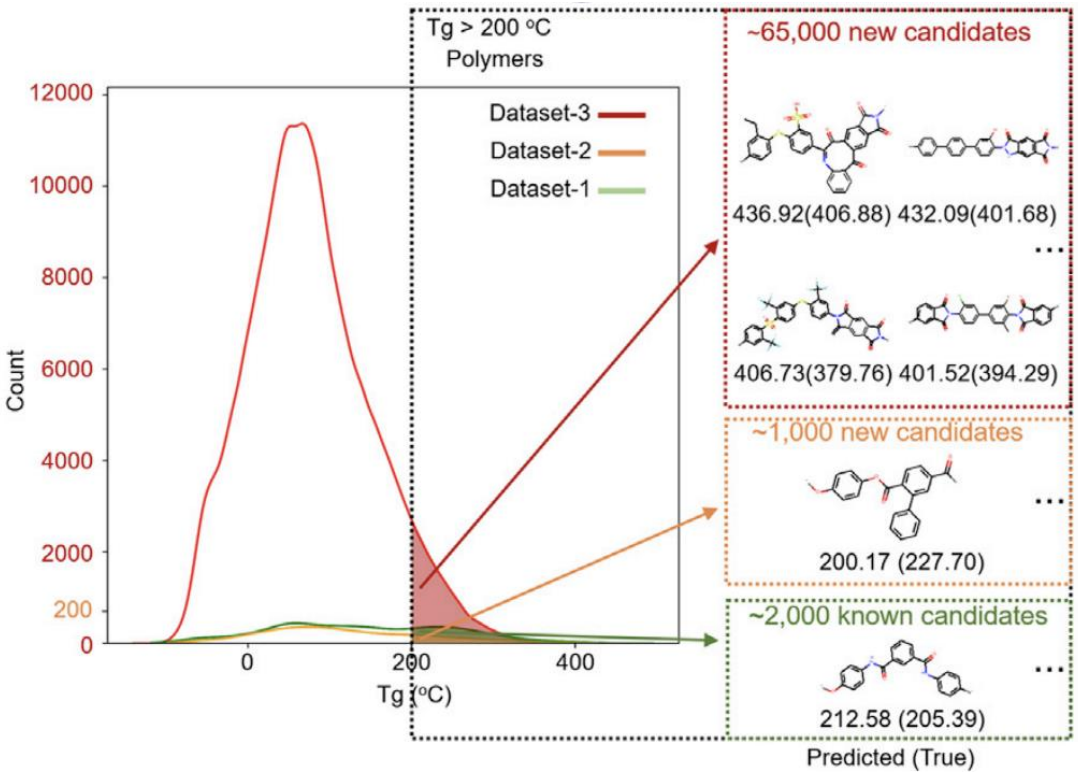


Figure 2.19 High-throughput screening of high Tg polymers with the DNN_Fingerprint model. The Tg distribution of the dataset-1, dataset-2, and dataset-3 are plotted in green, yellow, and red, respectively. The polymer samples on the right are following by their predicated Tg and true Tg values. For the sample in dataset-1 (green box), true Tg is the collected experimental value. For the samples in dataset-2 (yellow box) and dataset-3 (red box), true Tg is the MD-simulated value. More than 1,000 real polymers and 65,000 hypothetical polymers were discovered with Tg > 200℃. Reproduced with permission from [139].

Yang et al. employed a ML model to revolutionize the design of polymer membranes, traditionally a trial-and-error process, into a data-driven discovery method.[140] Their multitask ML models, trained on a vast array of experimental data, effectively link polymer chemistry with gas permeabilities of various gases. As shown in **Figure 2.20**, the models not only provide deep insights into the impact of different chemical groups on permeability and selectivity but also facilitate the screening of over 9 million hypothetical polymers. This extensive screening identified thousands of polymers exceeding current performance benchmarks, including hundreds of ultrapermeable membranes with exceptionally high O2 and CO2 permeabilities. The validity of these ML predictions is further reinforced through

high-fidelity MD simulations, underscoring the potential of these polymers to be actualized. The study concludes by offering the membrane design community numerous high-performance polymer candidates and key molecular insights, potentially serving as a blueprint for a variety of material discovery and design tasks across multiple domains.

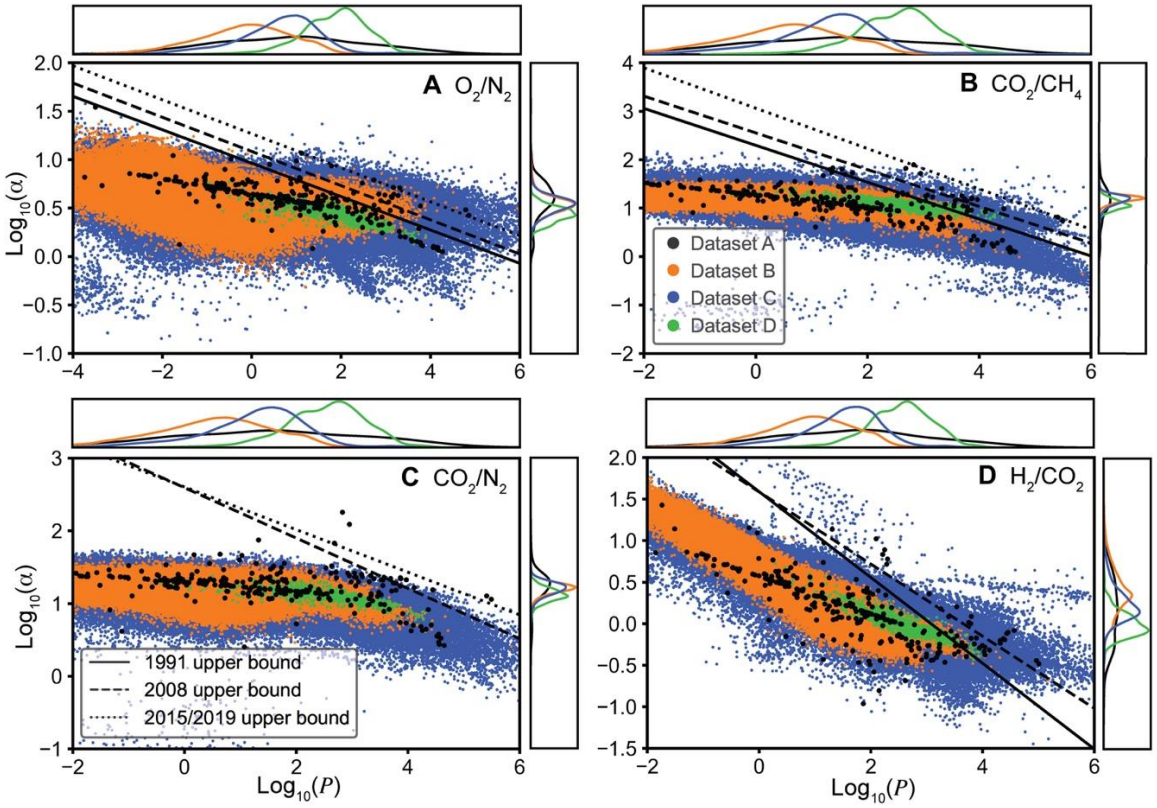


Figure 2.20 Visualization of predicted permeabilities for hypothetical polymers in datasets. The data are visualized for (A) O2/N2, (B) CO2/CH4, (C) CO2/N2, and (D) H2/CO2 separations, with thousands of promising polymers lying at or above the Robeson upper bounds. Units of permeability are Barrers. Reproduced with permission from [140].

In the realm of materials science, ML methodologies have emerged as transformative tools in the discovery and design of novel materials. Utilizing the capability to analyze vast datasets and identify intricate patterns, ML has not only expedited the process of material prediction and screening but also offered more precise and holistic understandings in materials research. Recent advancements have highlighted ML's potent capabilities in forecasting and identifying both inorganic and organic materials, thereby streamlining the material discovery process and reducing associated research expenditures. Furthermore,

these developments provide innovative perspectives and methodologies, enriching the traditional approaches to materials research and development. As computational power continues to grow and algorithmic refinements progress, ML's role in materials science is set to delve deeper and expand wider, paving the way for groundbreaking advancements in materials innovation and their practical applications.

2.3.3 Machine Learning in Flexible Electronics

In the burgeoning field of flexible electronics, the exploration and advancement of new materials are paramount for driving technological progress and fulfilling evolving market demands. As technological innovation accelerates, flexible electronic devices such as wearable sensors, flexible displays, and bendable solar panels are increasingly becoming an integral part of our lives. The performance and functionality of these devices are profoundly influenced by the materials used. To meet the demands of diverse bending and folding applications, not only must these materials exhibit excellent electronic and mechanical characteristics, but they must also maintain robust flexibility and durability. Consequently, the pursuit of novel materials for flexible electronics is not merely a quest for enhanced device performance; it also unlocks a realm of possibilities for pioneering applications and innovative designs. The advent of AI and ML heralds a new era in this domain, offering more rapid and efficient methodologies for the prediction and identification of promising materials.

In recent years, there has been a surge in research utilizing ML for the study of Organic Photovoltaics (OPVs). Sun et al. demonstrate the powerful capabilities of ML in advancing the field of OPVs.[141] By constructing a database of over 1700 donor materials from existing literature, the study successfully employs supervised learning to establish a correlation between chemical structures and photovoltaic properties, even prior to synthesis. The researchers explore various molecular structure expressions, such as images, ASCII strings, descriptors, and fingerprints, to feed different ML algorithms. As shown in **Figure 2.21a**, fingerprints with a length exceeding 1000 bits achieve high predictive accuracy. This approach's efficacy is further substantiated by accurately screening 10 novel donor

materials exhibited in **Figure 2.21c**, showcasing a strong alignment between ML predictions and experimental results. The findings underscore ML's potential as a powerful tool for pre-screening OPV materials, thereby catalyzing the evolution of OPV technology. Meftahi et al. conducted a comprehensive study on the application of machine-learning approaches to predict key properties of OPV materials.[142] Their research demonstrated the utility of chemically interpretable fragment-based descriptors in training ML models for the prediction of OPV device properties, including power conversion efficiency (PCE), open circuit potential (V_{OC}), short circuit density (J_{SC}), HOMO energy, LUMO energy, and the HOMO–LUMO gap. The most robust models exhibited a remarkable predictive accuracy for PCE, enabling accurate estimations with a standard error of ±0.5%. This approach facilitates the rapid screening of potential donor and acceptor materials for OPV applications, thereby expediting the design of green energy devices. Furthermore, the study emphasized the importance of using nonlinear ML methods and signature descriptors to create chemically interpretable and predictive models for OPV properties. Nagasawa et al. have addressed the challenge of data-driven molecular design in OPV applications with bulk heterojunction frameworks through the workflow as shown **Figure 2.22a**.[143] Their study employs supervised learning methods, specifically artificial neural network (ANN) and RF, to screen conjugated molecules for polymer-fullerene OPV applications. The authors manually collect approximately 1000 experimental parameters from the literature, encompassing factors like PCE, molecular weight, and electronic properties, and integrate them with digitized chemical structures for ML. While ANN exhibits a low correlation coefficient, RF demonstrates improved accuracy, particularly in PCE classification. Notably, the RF model guides the design, synthesis, and characterization of a conjugated polymer as depicted in **Figure 2.22b-d,** expediting the development of optoelectronic materials. This research showcases the potential of supervised learning in OPV molecular design.

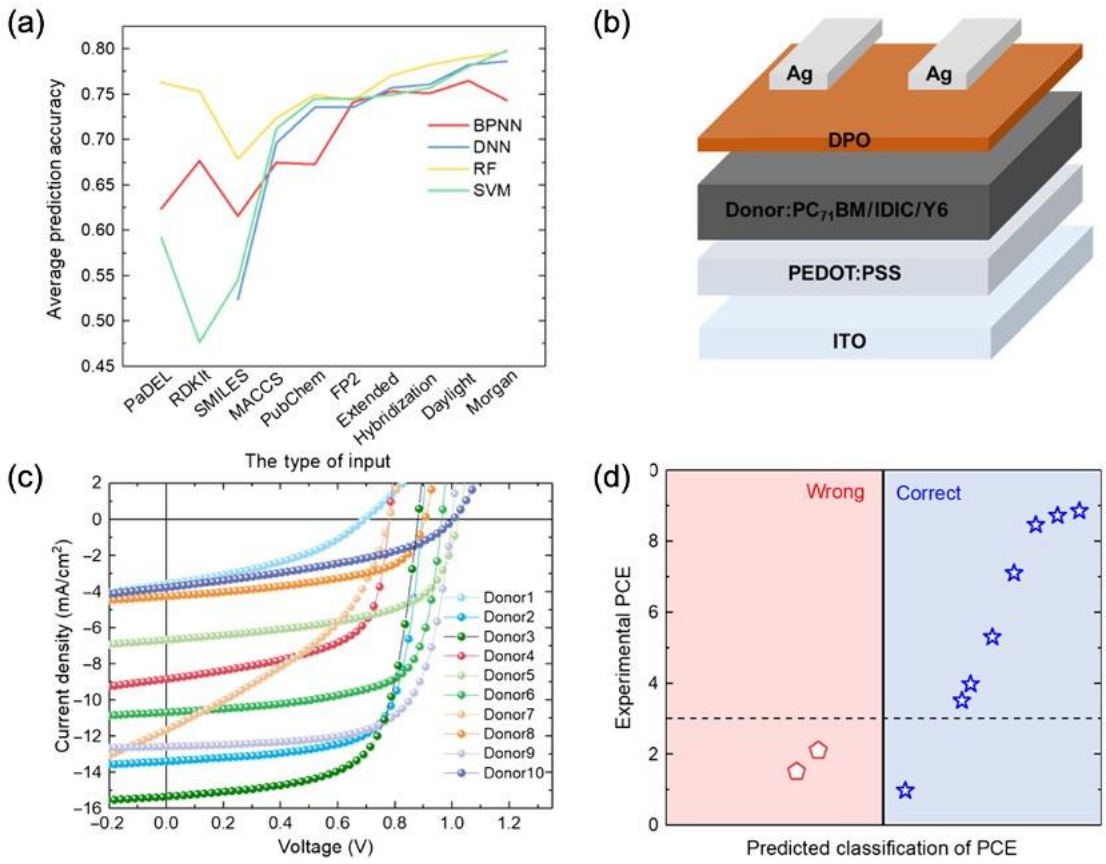


Figure 2.21 Verification of ML models with experiment. (a) Comparison of the results from four different models. (b) Schematic diagram of the cell architecture used in this study. (c) J-V curve of the solar cell with the active layer using the predicted donor material. (d) Prediction results versus experimental data for the predicted donor materials with the RF algorithm and Daylight fingerprints. Reproduced with permission from [141].

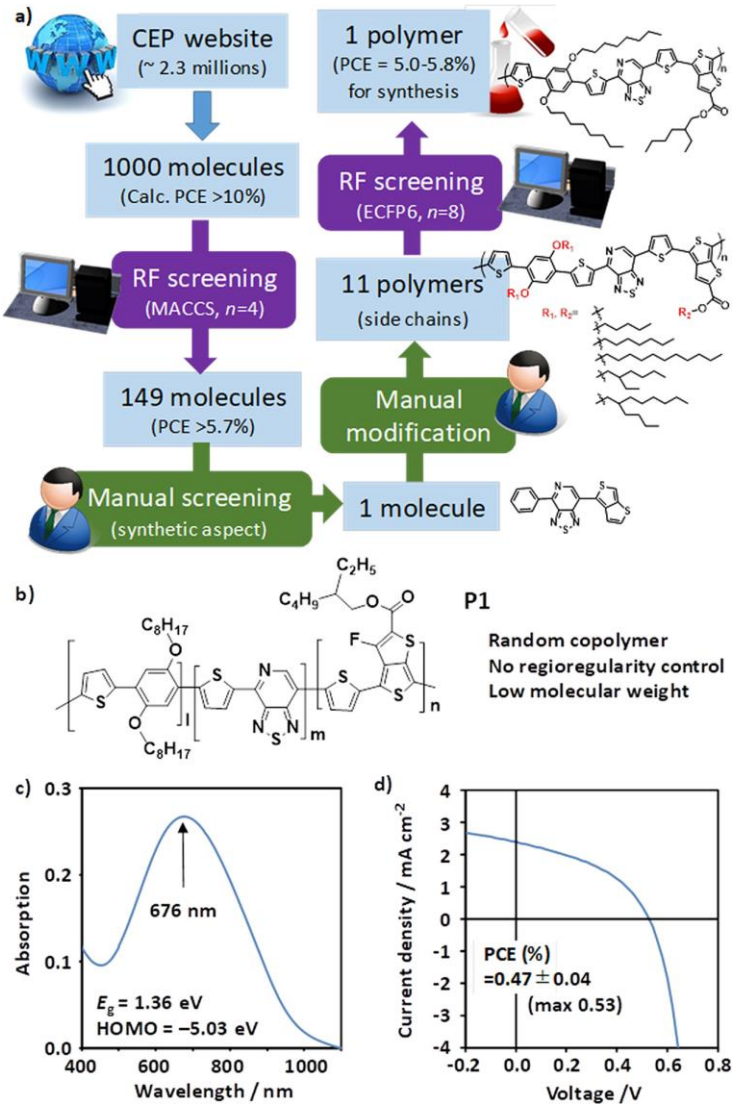


Figure 2.22 (a) Scheme of polymer design by combining the RF screening and manual screening/modification. The picked-up molecule or polymer in each stage is shown. (b) Synthesized random copolymer for OPV analysis. (c) Photoabsorption spectrum of P1 in the film state. (d) Current density–voltage curve of the best performing device. Reproduced with permission from [143].

Ding et al.'s research introduces ML technology for predicting the hole mobility of ionic liquid Poly(3,4-ethylenedioxythiophene) (PEDOT) systems, a critical parameter for their performance in advanced electronics.[144] Traditionally, assessing hole mobility involved MD simulations to obtain conformers, followed by quantum mechanics (QM) calculations or quantum Hall effect measurements. However, this study presents an alternative approach

utilizing supervised learning algorithms, including LR, ANN, RF, and gradient boosting decision tree (GBDT), to predict transfer integral (V_{ij}) values accurately. As shown in **Figure 2.23**, this ML model demonstrates exceptional predictive power (R2 > 0.9, MAE = 10^{-3} eV) and significantly reduces prediction time by six orders of magnitude when compared to the traditional MD and QM method. The model's predictive ability is validated across multiple IL-PEDOT systems. Additionally, experimental characterization with Atomic Force Microscope (AFM) and conductivity measurements aligns with the estimated mobility changes. This ML model not only elucidates key feature descriptors influencing V_{ij}|, ut also identifies the optimal feature quantity to maximize V_{ij}, dcom of 4 Å favoring the largest V_{ij}. The results highlight the potential of ML technology to expedite hole mobility prediction in IL-PEDOT systems across various morphologies for photovoltaic and thermoelectric applications, reducing development time and experimental costs.

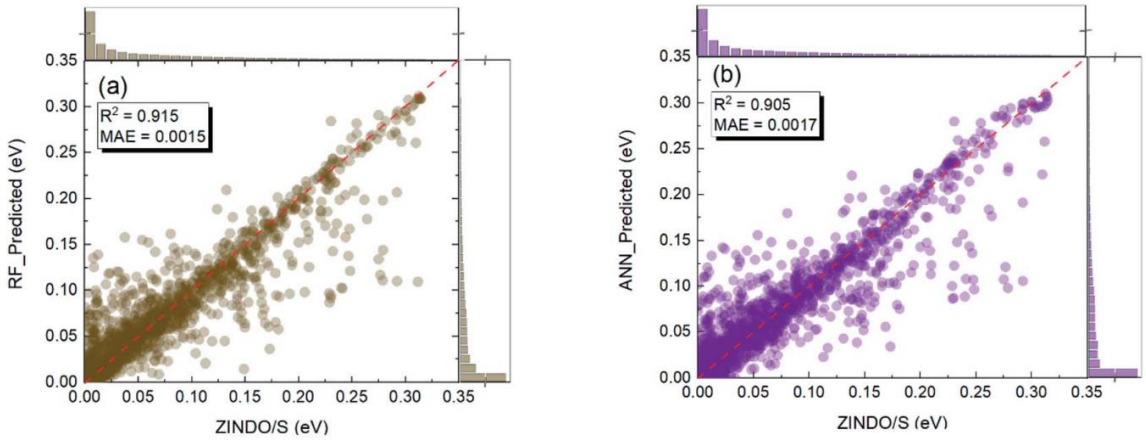


Figure 2.23 The predicted |V_{ij}| for (a) RF with 300 decision trees and (b) ANN with 6 hidden layers (the numbers of neurons in each layer are 100, and the unit of MAE value is in eV). Reproduced with permission from [144].

Leem et al. have presented an innovative data-driven approach to tailor the mechanical properties of soft materials by utilizing ML predictions based on experimental synthetic recipes.[145] This research addresses the challenge of directly modulating the properties of soft materials in a laboratory setting with specific ingredients. **Figure 2.24a** demonstrated

the process chart of the system. They employed polyurethane (PU) elastomer as a model soft material and manipulated the mechanical properties by adjusting the mixing ratio of its components. By utilizing a design of experiment, they collected data from 25 experimental conditions to train a LR model. This model takes desired mechanical properties as input and generates synthetic recipes for soft materials, which were subsequently validated through experiments. The study further conducted a comparative analysis of the predictive accuracies of various ML algorithms, as illustrated in **Figure 2.24b**. Additionally, it examined models trained with varying quantities of training data points, as depicted in **Figures 2.24c, d**. This data-driven approach has the potential to be applied across various material systems to establish experimental conditions and property relationships for soft materials, offering a valuable tool for experimental labs in the design of soft materials with desired mechanical characteristics.

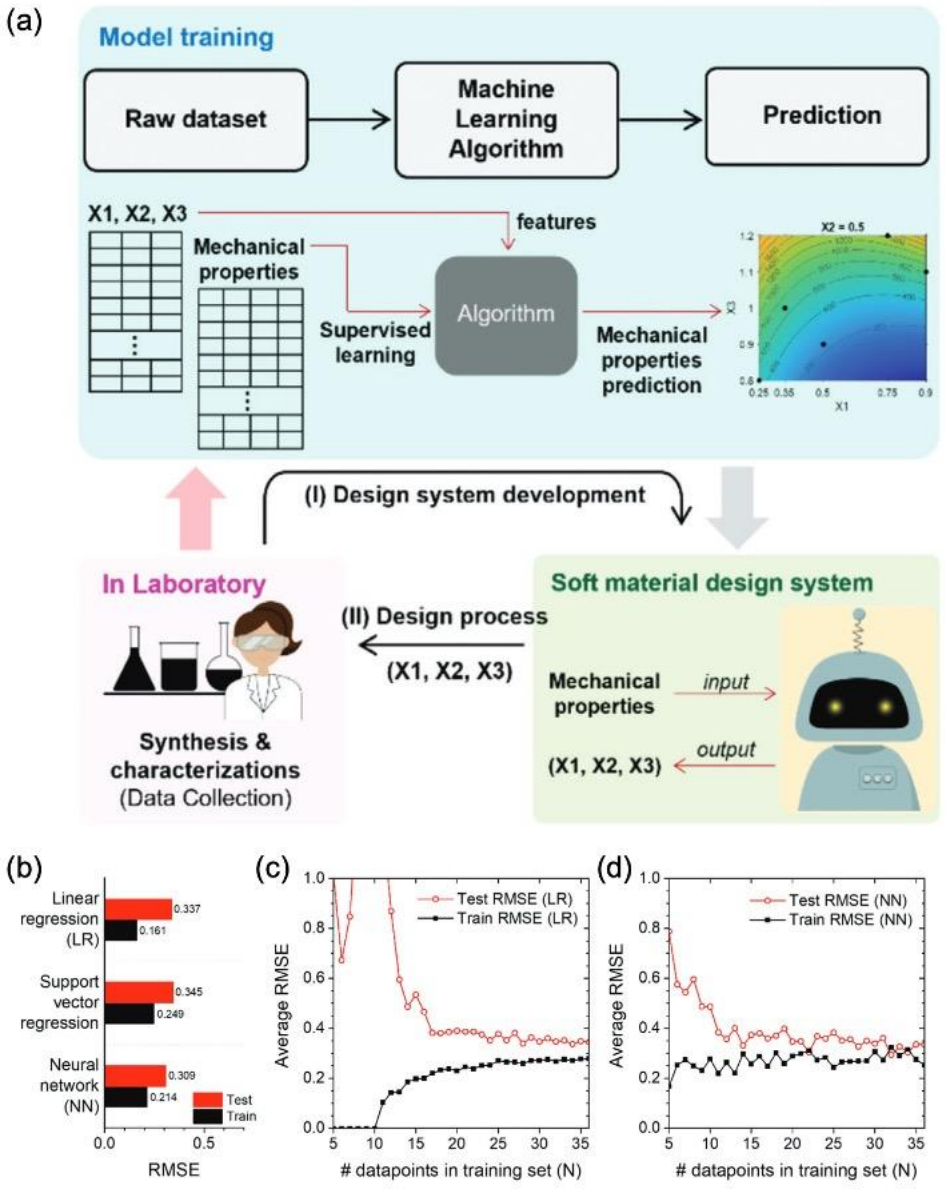


Figure 2.24 (a) Process chart of (I) soft material design system development and (II) soft material design process using the developed design system. (b) Comparison of train RMSE and test RMSE values obtained from LR, SVR, and NN using initial 25 data points as a training set. LR showed the smallest train RMSE and NN showed the smallest test RMSE. (c),(d) RMSE value changes when the number of data points in the training set (N) changes in both cases LR and NN were used. Reproduced with permission from [145].

Atahan-Evrenk et al. conducted a study focused on the prediction of intramolecular reorganization energy (RE) for hole transport in organic semiconductors (OSCs) using ML methods.[146] With the aim of enabling HTS for novel OSCs, they generated a molecular library comprising 5631 molecules with extended conjugated backbones. These molecules were characterized using various descriptors and the target electronic data were obtained through quantum-chemical calculations. As shown in **Figure 2.25**, the research compared ridge, kernel ridge, and deep neural net (DNN) regression models for predicting RE and found that DNNs outperformed the other methods, achieving a coefficient of determination of 0.92 and a RMSE of approximately 12 meV. This study demonstrates that the intramolecular REs of organic semiconductor molecules can be accurately predicted from their molecular structures, providing a valuable tool for the design and discovery of efficient OSCs.

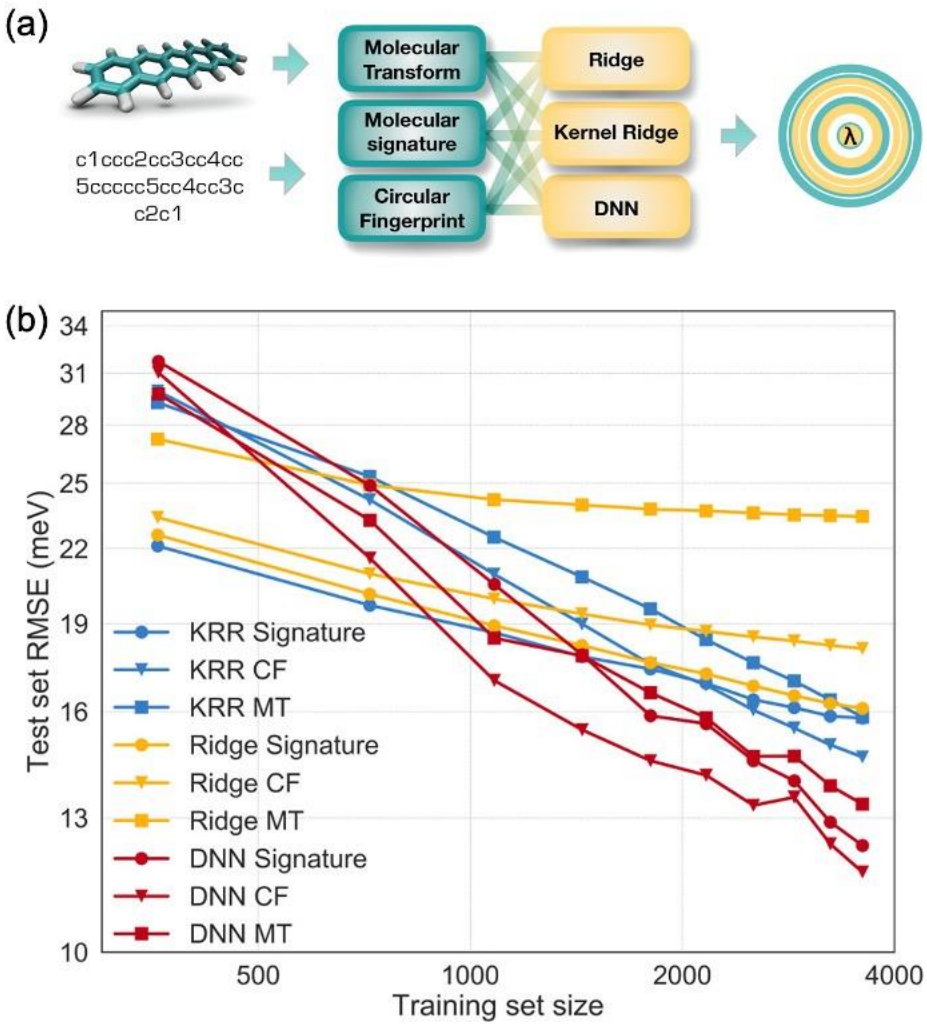


Figure 2.25 (a) Nine models to predict the reorganization energy from an inexpensive (MMFF94) 3D geometry or a SMILES string. (b) Learning curves for the models obtained with 5-fold cross-validation on 80% of the data using the [1000, 1000, 1000] network with dropout values of 0.1 for the signature and molecular transform (MT) descriptors and 0.2 for the circular fingerprint (CF) descriptors. Reproduced with permission from [146].

Gómez-Bombarelli et al. have presented an integrated approach to the design of organic functional materials, showcasing the power of virtual screening in molecular discovery.[147] As depicted in **Figure 2.26a**, their research combines theoretical insights, quantum chemistry, cheminformatics, ML, industrial expertise, organic synthesis, molecular characterization, device fabrication, and optoelectronic testing. Through the exploration of a vast search space of 1.6 million molecules and the screening of over 400,000 of them

using time-dependent density functional theory (TDDFT), the study identified thousands of promising novel organic light-emitting diode (OLED) molecules spanning the visible spectrum, as illustrated in **Figure 2.26b, c**. The collaboration between experts led to the selection of the most promising candidates, which were then experimentally synthesized and tested. Remarkably, these candidates achieved external quantum efficiencies (EQE) of up to 22%. The research highlights the potential of computational exploration of chemical space to not only identify new molecules but also provide fundamental chemical insights.

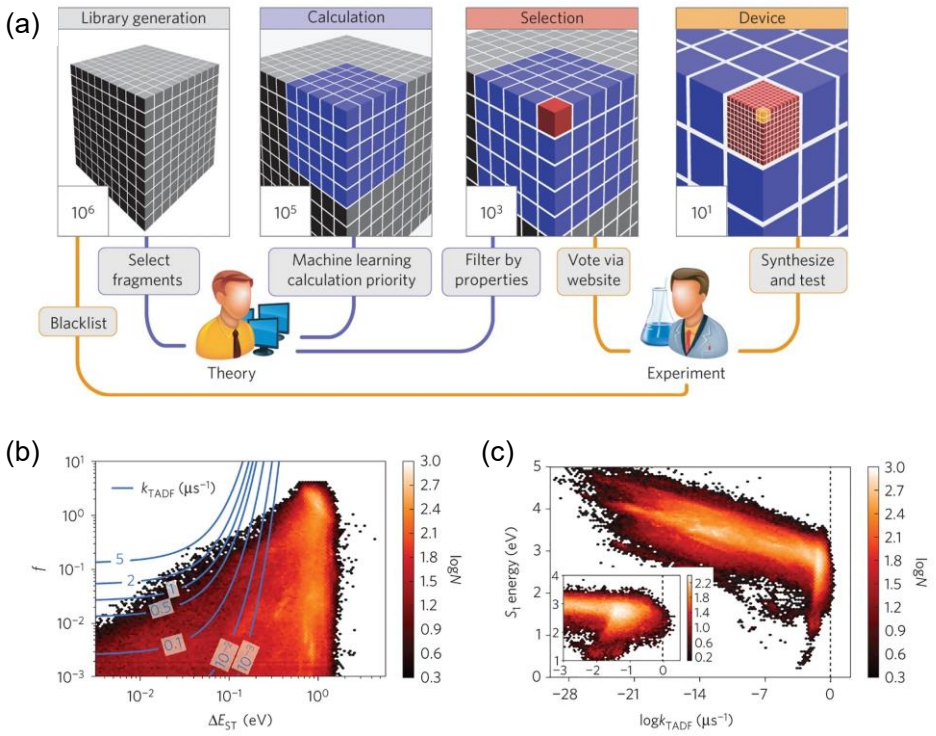


Figure 2.26 (a) Diagram of the collaborative discovery approach: the search space decreases by over five orders of magnitude as the screening progresses. The cubes represent the size of the chemical space considered at any given stage of the process. The distinct screening stages, from left to right, involve different theoretical and computational approaches as well as experimental input and testing. (b) Number of screened molecules as a function of singlet–triplet splitting (ΔE_{ST}) and oscillator strength (f). Contour lines represent estimated k_{TADF} (μs^{-1}) assuming S_1 at 3.0 eV. (c) Number of screened molecules as a function of k_{TADF} and S_1 energy. Vertical dashed line corresponds to $k_{TADF} = 1\ \mu s^{-1}$. Reproduced with permission from [147].

The application of ML in flexible electronics highlights a growing trend in research, where ML techniques are being leveraged to enhance the development and performance of flexible electronic devices. Numerous studies have investigated the integration of ML algorithms to optimize materials, device design, and manufacturing processes. These approaches have shown promising results in improving the efficiency, reliability, and flexibility of electronic components and systems.

References

[1] Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* **2016**, 4 (5).

[2] Baerends, E. J.; Gritsenko, O. V. A quantum chemical view of density functional theory. *The Journal of Physical Chemistry A* **1997**, 101 (30), 5383-5403.

[3] Becke, A. D. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of Chemical Physics* **2014**, 140 (18).

[4] Burke, K. Perspective on density functional theory. *The Journal of Chemical Physics* **2012**, 136 (15).

[5] Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Reviews of Modern Physics* **2015**, 87 (3), 897.

[6] Binder, K. *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*; Oxford University Press, **1995**.

[7] Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, **2006**; pp 84-es.

[8] Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, 99 (6), 1129-1143.

[9] Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology* **2015**, 31, 64-74.

[10] Rapaport, D. C. *The Art of Molecular Dynamics Simulation*; Cambridge University Press, **2004**.

[11] Tolle, K. M.; Tansley, D. S. W.; Hey, A. J. The fourth paradigm: Data-intensive scientific discovery [point of view]. *Proceedings of the IEEE* **2011**, 99 (8), 1334-1337.

[12] Käming, N.; Dawid, A.; Kottmann, K.; Lewenstein, M.; Sengstock, K.; Dauphin, A.; Weitenberg, C. Unsupervised machine learning of topological phase transitions from experimental data. *Machine Learning: Science and Technology* **2021**, 2 (3), 035037.

[13] Li, L.; Yang, Y.; Zhang, D.; Ye, Z.; Jesse, S.; Kalinin, S. V.; Vasudevan, R. K. Machine learning–enabled identification of material phase transitions based on experimental data: Exploring collective dynamics in ferroelectric relaxors. *Science Advances* **2018**, 4 (3), eaap8672.

[14] Smith, A.; Keane, A.; Dumesic, J. A.; Huber, G. W.; Zavala, V. M. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Applied Catalysis B: Environmental* **2020**, 263, 118257.

[15] Haffejee, R. A.; Laubscher, R. Application of machine learning to develop a real-time air-cooled condenser monitoring platform using thermofluid simulation data. *Energy and AI* **2021**, 3, 100048.

[16] Pattnaik, P.; Raghunathan, S.; Kalluri, T.; Bhimalapuram, P.; Jawahar, C.; Priyakumar, U. D. Machine learning for accurate force calculations in molecular dynamics simulations. *The Journal of Physical Chemistry A* **2020**, 124 (34), 6954-6967.

[17] Rahman, A.; Deshpande, P.; Radue, M. S.; Odegard, G. M.; Gowtham, S.; Ghosh, S.; Spear, A. D. A machine learning framework for predicting the shear strength of carbon nanotube-polymer interfaces based on molecular dynamics simulation data. *Composites Science and Technology* **2021**, 207, 108627.

[18] Cai, J.; Chu, X.; Xu, K.; Li, H.; Wei, J. Machine learning-driven new material discovery. *Nanoscale Advances* **2020**, 2 (8), 3115-3130.

[19] Jha, A.; Chandrasekaran, A.; Kim, C.; Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition temperatures. *Modelling and Simulation in Materials Science and Engineering* **2019**, 27 (2), 024002.

[20] Rodrigues, J. F.; Florea, L.; de Oliveira, M. C.; Diamond, D.; Oliveira, O. N. Big data and machine learning for materials science. *Discover Materials* **2021**, 1, 1-27.

[21] Jo, E. S.; Gebru, T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, **2020**; pp 306-316.

[22] Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and robust automated machine learning. *Advances in Neural Information Processing Systems* **2015**, 28.

[23] Kamulegeya, L.; Bwanika, J.; Okello, M.; Rusoke, D.; Nassiwa, F.; Lubega, W.; Musinguzi, D.; Börve, A. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *African Health Sciences* **2023**, 23 (2), 753-763.

[24] Jain, A.; Hautier, G.; Ong, S. P.; Persson, K. New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships. *Journal of Materials Research* **2016**, 31 (8), 977-994.

[25] Saad, Y.; Gao, D.; Ngo, T.; Bobbitt, S.; Chelikowsky, J. R.; Andreoni, W. Data mining for materials: Computational experiments with AB compounds. *Physical Review B* **2012**, 85 (10), 104104.

[26] Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M. Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **2018**, 152, 60-69.

[27] Yosipof, A.; Nahum, O. E.; Anderson, A. Y.; Barad, H. N.; Zaban, A.; Senderowitz, H. Data mining and machine learning tools for combinatorial material science of all-oxide photovoltaic cells. *Molecular Informatics* **2015**, 34 (6-7), 367-379.

[28] Kalinin, S. V.; Ziatdinov, M.; Hinkle, J.; Jesse, S.; Ghosh, A.; Kelley, K. P.; Lupini, A. R.; Sumpter, B. G.; Vasudevan, R. K. Automated and autonomous experiments in electron and scanning probe microscopy. *ACS Nano* **2021**, 15 (8), 12604-12627.

[29] Shi, Y.; Prieto, P. L.; Zepel, T.; Grunert, S.; Hein, J. E. Automated experimentation powers data science in chemistry. Accounts of Chemical Research 2021, 54 (3), 546-555.

[30] Ziatdinov, M. A.; Liu, Y.; Morozovska, A. N.; Eliseev, E. A.; Zhang, X.; Takeuchi, I.; Kalinin, S. V. Hypothesis learning in automated experiment: Application to combinatorial materials libraries. *Advanced Materials* **2022**, 34 (20), 2201345.

[31] Nakata, M.; Shimazaki, T. PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling* **2017**, 57 (6), 1300-1308.

[32] Nakata, M.; Shimazaki, T.; Hashimoto, M.; Maeda, T. PubChemQC PM6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *Journal of Chemical Information and Modeling* **2020**, 60 (12), 5891-5899.

[33] Yang, J.; Manganaris, P.; Mannodi-Kanakkithodi, A. A high-throughput computational dataset of halide perovskite alloys. *Digital Discovery* **2023**, 2(3), 856-870.

[34] Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A. PubChem substance and compound databases. *Nucleic Acids Research* **2016**, 44 (D1), D1202-D1213.

[35] Pence, H. E.; Williams, A. ChemSpider: an online chemical information resource. *ACS Publications* **2010**, 1123-1124.

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

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

[38] Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein Crystallography: Methods and Protocols* **2017**, 627-641.

[39] Frenkel, M. Global information systems in science: Application to the field of thermodynamics. *Journal of Chemical & Engineering Data* **2009**, 54 (9), 2411-2428.

[40] Landis, D. D.; Hummelshøj, J. S.; Nestorov, S.; Greeley, J.; Dułak, M.; Bligaard, T.; Nørskov, J. K.; Jacobsen, K. W. The computational materials repository. *Computing in Science & Engineering* **2012**, 14 (6), 51-57.

[41] Richard, A. M.; Huang, R.; Waidyanatha, S.; Shinn, P.; Collins, B. J.; Thillainadarajah, I.; Grulke, C. M.; Williams, A. J.; Lougee, R. R.; Judson, R. S. The Tox21 10K compound library: collaborative chemistry advancing toxicology. *Chemical Research in Toxicology* **2020**, 34 (2), 189-216.

[42] Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **2018**, 46 (D1), D1074-D1082.

[43] Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S. HMDB: the human metabolome database. *Nucleic Acids Research* **2007**, 35 (suppl_1), D521-D526.

[44] Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data* **2019**, 64 (12), 5985-5998.

[45] Irwin, J. J.; Shoichet, B. K. ZINC− a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* **2005**, 45 (1), 177-182.

[46] Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database–an open-access collection of crystal structures. *Journal of Applied Crystallography* **2009**, 42 (4), 726-729.

[47] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. Acta Crystallographica Section B: Structural Science, *Crystal Engineering and Materials* **2016**, 72 (2), 171-179.

[48] Hellenbrandt, M. The inorganic crystal structure database (ICSD)-present and future. *Crystallography Reviews* **2004**, 10 (1), 17-22.

[49] Draxl, C.; Scheffler, M. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin* **2018**, 43 (9), 676-682.

[50] Brandrup, J.; Immergut, E. H.; Grulke, E. A.; Abe, A.; Bloch, D. R. *Polymer Handbook*, **1999**; Wiley New York: Vol. 89.

[51] Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C* **2018**, 122 (31), 17575-17585.

[52] Ma, R.; Luo, T. PI1M: a benchmark database for polymer informatics. *Journal of Chemical Information and Modeling* **2020**, 60 (10), 4684-4690.

[53] Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. In PoLyInfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, **2011**; IEEE: 2011; pp 22-29.

[54] Nandy, A.; Duan, C.; Kulik, H. J. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Current Opinion in Chemical Engineering* **2022**, 36, 100778.

[55] Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine learning for materials scientists: an introductory guide toward best practices. *Chemistry of Materials* **2020**, 32 (12), 4954-4965.

[56] Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular descriptors for structure–activity applications: a hands-on approach. *Computational Toxicology: Methods and Protocols* **2018**, 3-53.

[57] Elspas, B.; Turner, J. Graphs with circulant adjacency matrices. *Journal of Combinatorial Theory* **1970**, 9 (3), 297-307.

[58] Estrada, E. Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles. *Journal of Chemical Information and Computer Sciences* **1998**, 38 (1), 23-27.

[59] Marrero-Ponce, Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *Journal of Chemical Information and Computer Sciences* **2004**, 44 (6), 2010-2026.

[60] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters* **2012**, 108 (5), 058301.

[61] Todeschini, R.; Consonni, V. Molecular descriptors. *Recent Advances in QSAR Studies* **2010**, 29-102.

[62] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *Journal of Computer-Aided Molecular Design* **1997**, 11, 79-92.

[63] Consonni, V.; Todeschini, R.; Pavan, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences* **2002**, 42 (3), 682-692.

[64] Todeschini, R.; Gramatica, P. The WHIM theory: New 3D molecular descriptors for QSAR in environmental modelling. *SAR and QSAR in Environmental Research* **1997**, 7 (1-4), 89-115.

[65] Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *Journal of Molecular Graphics and Modelling* **2014**, 54, 194-203.

[66] Kim, K. H.; Greco, G.; Novellino, E. A critical review of recent CoMFA applications. *Perspectives in Drug Discovery and Design* **1998**, 12 (14), 257-315.

[67] Ash, J.; Fourches, D. Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *Journal of Chemical Information and Modeling* **2017**, 57 (6), 1286-1299.

[68] Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, 363 (6424), eaau5631.

[69] Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Feng, J.; Zheng, W.; Tropsha, A. QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. *SAR and QSAR in Environmental Research* **2005**, 16 (1-2), 93-102.

[70] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, 361 (6400), 360-365.

[71] Smidstrup, S.; Markussen, T.; Vancraeyveld, P.; Wellendorff, J.; Schneider, J.; Gunst, T.; Verstichel, B.; Stradi, D.; Khomyakov, P. A.; Vej-Hansen, U. G. QuantumATK: An integrated platform of electronic and atomic-scale modelling tools. *Journal of Physics: Condensed Matter* **2019**, 32 (1), 015901.

[72] Sohlenius-Sternbeck, A.-K.; Terelius, Y. Evaluation of ADMET predictor in early discovery drug metabolism and pharmacokinetics project work. *Drug Metabolism and Disposition* **2022**, 50 (2), 95-104.

[73] Vilar, S.; Cozza, G.; Moro, S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Current Topics in Medicinal Chemistry* **2008**, 8 (18), 1555-1572.

[74] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, 3 (1), 1-14.

[75] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *Journal of Chemical Information and Computer Sciences* **2003**, 43 (2), 493-500.

[76] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 2011, 32 (7), 1466-1474.

[77] Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **2006**, 56 (2), 237-248.

[78] Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, 8, 31.

[79] Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of Chemical Information and Modeling* **2008**, 48 (7), 1337-1344.

[80] Bhuvaneswari, V.; Priyadharshini, M.; Deepa, C.; Balaji, D.; Rajeshkumar, L.; Ramesh, M. Deep learning for material synthesis and manufacturing systems: A review. *Materials Today: Proceedings* **2021**, 46, 3263-3269.

[81] Ge, M.; Su, F.; Zhao, Z.; Su, D. Deep learning analysis on microscopic imaging in materials science. *Materials Today Nano* **2020**, 11, 100087.

[82] Lin, Y. z.; Nie, Z. h.; Ma, H. w. Structural damage detection with automatic feature-extraction through deep learning. *Computer-Aided Civil and Infrastructure Engineering* **2017**, 32 (12), 1025-1046.

[83] Peng, J.; Jury, E. C.; Dönnes, P.; Ciurtin, C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. *Frontiers in Pharmacology* **2021**, 12, 720694.

[84] Montgomery, D. C.; Peck, E. A.; Vining, G. G. *Introduction to Linear Regression Analysis*; John Wiley & Sons, **2021**.

[85] Menard, S. *Applied Logistic Regression Analysis*; Sage, **2002**.

[86] Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2011**, 2 (3), 1-27.

[87] Breiman, L. Random forests. *Machine Learning* **2001**, 45, 5-32.

[88] Diao, Y.; Yan, L.; Gao, K. A strategy assisted machine learning to process multi-objective optimization for improving mechanical properties of carbon steels. *Journal of Materials Science & Technology* **2022**, 109, 86-93.

[89] Rupp, M.; Ramakrishnan, R.; Von Lilienfeld, O. A. Machine learning for quantum mechanical properties of atoms in molecules. *The Journal of Physical Chemistry Letters* **2015**, 6 (16), 3309-3313.

[90] Kwak, S.; Kim, J.; Ding, H.; Xu, X.; Chen, R.; Guo, J.; Fu, H. Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model. *Journal of Materials Research and Technology* **2022**, 18, 520-530.

[91] Sharma, P.; Ramesh, K.; Parameshwaran, R.; Deshmukh, S. S. Thermal conductivity prediction of titania-water nanofluid: A case study using different machine learning algorithms. *Case Studies in Thermal Engineering* **2022**, 30, 101658.

[92] Wang, G.; Cai, C.; Pei, J.; Zhu, X. Prediction of thermal conductivity of polymer-based composites by using support vector regression. *Science China Physics, Mechanics and Astronomy* **2011**, 54, 878-883.

[93] Huang, Y.; Yu, C.; Chen, W.; Liu, Y.; Li, C.; Niu, C.; Wang, F.; Jia, Y. Band gap and band alignment prediction of nitride-based semiconductors using machine learning. *Journal of Materials Chemistry C* **2019**, 7 (11), 3238-3245.

[94] Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. **Physical Review B** *2016*, 93 (11), 115104.

[95] Owolabi, T. O.; Abd Rahman, M. A. Prediction of band gap energy of doped graphitic carbon nitride using genetic algorithm-based support vector regression and extreme learning machine. *Symmetry* **2021**, 13 (3), 411.

[96] Pilania, G.; Balachandran, P. V.; Kim, C.; Lookman, T. Finding new perovskite halides via machine learning. *Frontiers in Materials* **2016**, 3, 19.

[97] Balachandran, P. V. Machine learning guided design of functional materials with targeted properties. *Computational Materials Science* **2019**, 164, 82-90.

[98] Butnariu, C.; Lisa, C.; Leon, F.; Curteanu, S. Prediction of liquid-crystalline property using support vector machine classification. *Journal of Chemometrics* **2013**, 27 (7-8), 179-188.

[99] Ahmad, A.; Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering* **2007**, 63 (2), 503-527.

[100] Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, 2 (1-3), 37-52.

[101] Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised learning methods for molecular simulation data. *Chemical Reviews* **2021**, 121 (16), 9722-9758.

[102] Jia, X.; Deng, Y.; Bao, X.; Yao, H.; Li, S.; Li, Z.; Chen, C.; Wang, X.; Mao, J.; Cao, F. Unsupervised machine learning for discovery of promising half-Heusler thermoelectric materials. *npj Computational Materials* **2022**, 8 (1), 34.

[103] Shen, S. C.-y.; Buehler, M. J. Nature-inspired architected materials using unsupervised deep learning. *Communications Engineering* **2022**, 1 (1), 37.

[104] Nguyen, P. C.; Vlassis, N. N.; Bahmani, B.; Sun, W.; Udaykumar, H.; Baek, S. S. Synthesizing controlled microstructures of porous media using generative adversarial networks and reinforcement learning. *Scientific Reports* **2022**, 12 (1), 9034.

[105] Rajak, P.; Krishnamoorthy, A.; Mishra, A.; Kalia, R.; Nakano, A.; Vashishta, P. Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials. *npj Computational Materials* **2021**, 7 (1), 108.

[106] Whitelam, S.; Tamblyn, I. Learning to grow: Control of material self-assembly using evolutionary reinforcement learning. *Physical Review E* **2020**, 101 (5), 052604.

[107] Ma, R.; Zhang, H.; Luo, T. Exploring high thermal conductivity amorphous polymers using reinforcement learning. *ACS Applied Materials & Interfaces* **2022**, 14 (13), 15587-15598.

[108] Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nature Communications* **2023**, 14 (1), 1403.

[109] Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of molecules via deep reinforcement learning. *Scientific Reports* **2019**, 9 (1), 10752.

[110] Wang, C.; Han, J. Dl4scivis: A state-of-the-art survey on deep learning for scientific visualization. *IEEE Transactions on Visualization and Computer Graphics* **2022**.

[111] Bangaru, S. S.; Wang, C.; Zhou, X.; Hassan, M. Scanning electron microscopy (SEM) image segmentation for microstructure analysis of concrete using U-net convolutional neural network. *Automation in Construction* **2022**, 144, 104602.

[112] Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T. Graph neural networks for materials science and chemistry. *Communications Materials* **2022**, 3 (1), 93.

[113] Schweidtmann, A. M.; Rittig, J. G.; Weber, J. M.; Grohe, M.; Dahmen, M.; Leonhard, K.; Mitsos, A. Physical pooling functions in graph neural networks for molecular property prediction. *Computers & Chemical Engineering* **2023**, 172, 108202.

[114] Ye, S.; Liang, J.; Liu, R.; Zhu, X. Symmetrical graph neural network for quantum chemistry with dual real and momenta space. *The Journal of Physical Chemistry A* **2020**, 124 (34), 6945-6953.

[115] Li, C.; Wang, C.; Sun, M.; Zeng, Y.; Yuan, Y.; Gou, Q.; Wang, G.; Guo, Y.; Pu, X. Correlated RNN Framework to Quickly Generate Molecules with Desired Properties for Energetic Materials in the Low Data Regime. *Journal of Chemical Information and Modeling* **2022**, 62 (20), 4873-4887.

[116] Yang, J.; Wang, X.; Wang, R.; Wang, H. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis–NIR spectroscopy. *Geoderma* **2020**, 380, 114616.

[117] Shaheen, F.; Verma, B.; Asafuddoula, M. Impact of automatic feature extraction in deep learning architecture. In *2016 International conference on digital image computing: techniques and applications (DICTA)*, **2016**; IEEE: pp 1-8.

[118] Varghese, J. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. *Visceral Medicine* **2020**, 36 (6), 443-449.

[119] Wei, J.; Chu, X.; Sun, X. Y.; Xu, K.; Deng, H. X.; Chen, J.; Wei, Z.; Lei, M. Machine learning in materials science. *InfoMat* **2019**, 1 (3), 338-358.

[120] Walters, W. P.; Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research* **2020**, 54 (2), 263-270.

[121] Ford, E.; Maneparambil, K.; Rajan, S.; Neithalath, N. Machine learning-based accelerated property prediction of two-phase materials using microstructural descriptors and finite element analysis. *Computational Materials Science* **2021**, 191, 110328.

[122] Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. Scientific Reports 2013, 3 (1), 2810.
[123] Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B* **2017**, 95 (14), 144110.
[124] Tian, X.; Song, S.; Chen, F.; Qi, X.; Wang, Y.; Zhang, Q.. Machine learning-guided property prediction of energetic materials: Recent advances, challenges, and perspectives. *Energetic Materials Frontiers* **2022**, 3(3), 177-186.
[125] Severson, K. A.; Attia, P. M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M. H.; Aykol, M.; Herring, P. K.; Fraggedakis, D. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy* **2019**, 4 (5), 383-391.
[126] Casey, A. D.; Son, S. F.; Bilionis, I.; Barnes, B. C. Prediction of energetic material properties from electronic structure using 3D convolutional neural networks. *Journal of Chemical Information and Modeling* **2020**, 60 (10), 4457-4473.
[127] Bhat, V.; Sornberger, P.; Pokuri, B. S. S.; Duke, R.; Ganapathysubramanian, B.; Risko, C. Electronic, redox, and optical property prediction of organic π-conjugated molecules through a hierarchy of machine learning approaches. *Chemical Science* **2023**, 14 (1), 203-213.
[128] Pei, Z.; Yin, J.; Hawk, J. A.; Alman, D. E.; Gao, M. C. Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules. *npj Computational Materials* **2020**, 6 (1), 50.
[129] Chen, L.; Tran, H.; Batra, R.; Kim, C.; Ramprasad, R. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Computational Materials Science* **2019**, 170, 109155.
[130] Xiang, X.-D.; Sun, X.; Briceno, G.; Lou, Y.; Wang, K.-A.; Chang, H.; Wallace-Freedman, W. G.; Chen, S.-W.; Schultz, P. G. A combinatorial approach to materials discovery. *Science* **1995**, 268 (5218), 1738-1740.
[131] Hautier, G.; Jain, A.; Ong, S. P. From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science* **2012**, 47, 7317-7340.

[132] Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure prediction drives materials discovery. *Nature Reviews Materials* **2019**, 4 (5), 331-348.

[133] Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *Journal of Materiomics* **2017**, 3 (3), 159-177.

[134] Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. Genetic algorithms for computational materials discovery accelerated by machine learning. *npj Computational Materials* **2019**, 5 (1), 46.

[135] Schleder, G. R.; Padilha, A. C.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials* **2019**, 2 (3), 032001.

[136] Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances* **2018**, 4 (4), eaaq1566.

[137] Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, 533 (7601), 73-76.

[138] Priya, P.; Aluru, N. Accelerated design and discovery of perovskites with high conductivity for energy applications through machine learning. *npj Computational Materials* **2021**, 7 (1), 90.

[139] Tao, L.; Chen, G.; Li, Y. Machine learning discovery of high-temperature polymers. *Patterns* **2021**, 2 (4).

[140] Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Science Advances* **2022**, 8 (29), eabn9545.

[141] Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances* **2019**, 5 (11), eaay4275.

[142] Meftahi, N.; Klymenko, M.; Christofferson, A. J.; Bach, U.; Winkler, D. A.; Russo, S. P. Machine learning property prediction for organic photovoltaic devices. *npj Computational Materials* **2020**, 6 (1), 166.

[143] Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *The Journal of Physical Chemistry Letters* **2018**, 9 (10), 2639-2646.

[144] Ding, W.-L.; Lu, Y.; Peng, X.-L.; Dong, H.; Chi, W.-J.; Yuan, X.; Sun, Z.-Z.; He, H. Accelerating evaluation of the mobility of ionic liquid-modulated PEDOT flexible electronics using machine learning. *Journal of Materials Chemistry A* **2021**, 9 (45), 25547-25557.

[145] Leem, J.; Jiang, Y.; Robinson, A.; Xia, Y.; Zheng, X. Data-Driven Approach to Tailoring Mechanical Properties of a Soft Material. *Advanced Functional Materials* **2023**, 2304451.

[146] Atahan-Evrenk, S.; Atalay, F. B. Prediction of intramolecular reorganization energy using machine learning. *The Journal of Physical Chemistry A* **2019**, 123 (36), 7855-7863.

[147] Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **2016**, 15 (10), 1120-1127.

Chapter 3

Experimental Methodology

This chapter initially introduces the preparation and preprocessing of the dataset. It then presents the selected algorithms, including both traditional machine learning and deep learning approaches. Furthermore, transfer learning techniques are discussed. Regarding simulation techniques, the chapter will cover the density functional theory method and the molecular dynamics simulation method.

3.1 Data Collection and Preparation

Our data collection methodology includes a multifaceted strategy that integrates information from a variety of sources: published literature, readily available databases, and computational simulations. We extracted raw data from the selected literature sources. Extensive data cleansing of the raw data is required to identify and resolve outliers, inconsistencies, and missing values in the collected dataset. The publicly accessible databases of materials are also an important source of data for us. Through specific search criteria, the databases can be filtered to obtain databases that satisfy specific requirements. In cases where experimental data were limited or unavailable, we use computational methods to obtain the dataset, and the specific computational methods used will be described in later sections.

Feature generation is a pivotal step in our methodology, wherein we transform the raw chemical structures and properties of materials into informative descriptors that are suitable for Machine Learning (ML). We employ the widely recognized software tools RDKit[1] and PaDel[2] for this purpose. RDKit is utilized to generate a diverse set of molecular descriptors from chemical structures. These descriptors encompass a broad range of chemical information, including molecular weight, atom counts, bond types, and topological properties. We employ molecular fingerprints derived from RDKit, such as Morgan Fingerprints,[3] to capture molecular structural patterns. PaDel, a comprehensive software tool, is employed to calculate property-based molecular descriptors. These descriptors are derived from molecular properties, electronic properties, and quantum chemical calculations. PubChem Fingerprint[4] can be also generated through the PaDel software to capture the chemistry information.

To avoid overfitting issue and enhance the efficiency of our ML models, we employ dimension reduction techniques. This involves the identification and removal of less informative features while retaining the most relevant ones. Our methodology incorporates several dimension reduction strategies, including low variance filters,[5] high correlation filters, and the recursive feature elimination (RFE) algorithm.[6] Low variance filters are

applied to eliminate features with minimal variation across the dataset. By setting a predefined variance threshold, features exhibiting low variability are removed. This step helps us focus on descriptors that capture significant changes in material properties, thus reducing noise and enhancing model interpretability. High correlation filters assess the pairwise correlation between features. Features that are highly correlated with others can be redundant and may not contribute substantially to the model's predictive power. We identify and remove one of the highly correlated features to reduce multicollinearity and improve model stability. The RFE algorithm systematically evaluates the importance of each feature by iteratively training the ML model and eliminating the least significant feature at each step. This iterative process continues until the desired number of features or a predefined performance criterion is met. RFE ensures that only the most influential descriptors are retained in the final feature set, optimizing model accuracy and generalization. By implementing these dimension reduction techniques, we ensure that the most informative features are retained for accurate predictions.

Data preprocessing is a crucial phase in our methodology, aimed at ensuring the quality, consistency, and compatibility of our dataset. Two fundamental techniques employed in this process are normalization and standardization.[7] To bring all features to a common scale and prevent any particular feature from dominating the modeling process due to its magnitude, we apply normalization. Each feature is transformed to have a range between 0 and 1. This technique is particularly useful when dealing with descriptors measured in different units or with varying ranges, ensuring that all features contribute equally to the ML models. Standardization is employed to center the data distribution around zero with a standard deviation of one. By subtracting the mean and dividing by the standard deviation for each feature, we achieve a standardized dataset. Standardization is valuable when dealing with ML algorithms that are sensitive to feature scales, ensuring that features with larger numerical values do not unduly influence model training.

3.2 Machine Learning Models

3.2.1 SVM

The Support Vector Machine (SVM) is a powerful ML algorithm that has gained widespread recognition for its versatility and effectiveness in both classification and regression tasks. Each SVM component is displayed in **Figure 3.1**.[8] At its core, the SVM algorithm is founded on the principle of finding an optimal hyperplane that maximally separates data points belonging to different classes in a high-dimensional feature space. This hyperplane is strategically positioned to have the maximum margin, defined as the distance between the hyperplane and the nearest data points (support vectors) from each class. SVM achieves this by transforming the original data into a higher-dimensional space using a kernel function, which implicitly maps data points into a space where they can be linearly separated. The choice of the appropriate kernel function plays a critical role in SVM's ability to handle complex, nonlinear relationships within the data. SVM excels not only in linearly separable scenarios but also in situations where data exhibits intricate boundaries, making it a versatile tool in various domains, including image classification, bioinformatics, and financial forecasting. Its robustness, capacity for handling high-dimensional data, and the ability to balance bias-variance trade-offs make SVM a valuable asset in the arsenal of ML techniques.

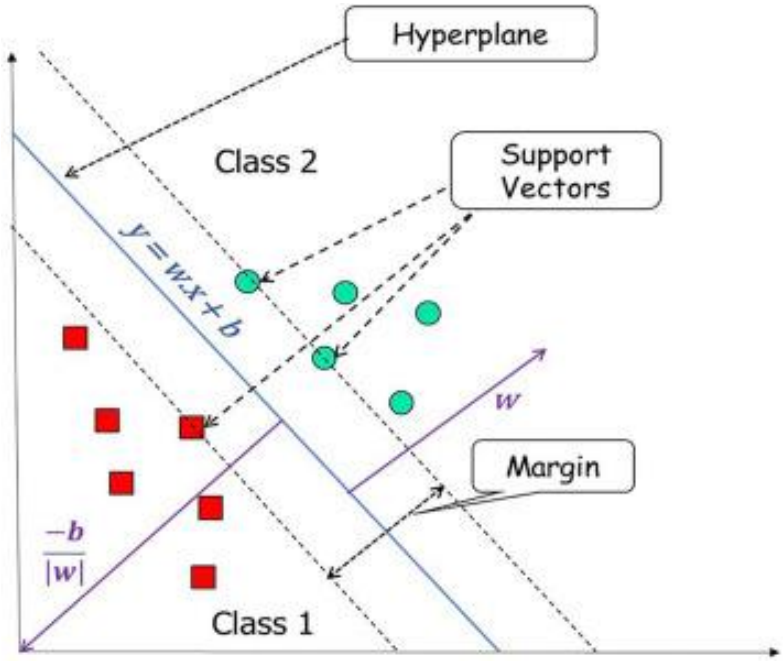


Figure 3.1 Components of SVM. Reproduced with permission from [8].

3.2.2 Random Forest

The Random Forest (RF) algorithm is a powerful ensemble learning technique that has gained prominence in the field of ML for its remarkable accuracy and versatility. At its core, the RF algorithm is built upon the principle of ensemble learning, where a multitude of decision trees are constructed and their outputs are aggregated to make predictions. **Figure 3.2** illustrates that each decision tree in the forest is trained on a random subset of the dataset and a random subset of the features, a technique known as bootstrapped aggregating or "bagging".[9] The brilliance of RF lies in its ability to mitigate overfitting, enhance generalization, and reduce variance by combining the predictions from multiple individual decision trees. During the training process, the algorithm assigns an importance score to each feature, indicating its contribution to the overall prediction. This property not only makes RF adept at handling high-dimensional data but also allows it to serve as a feature selection tool. Furthermore, the algorithm's ability to naturally handle missing values, outliers, and maintain robustness in noisy datasets has made it a go-to choice for a wide range of applications, including classification, regression, and feature selection, making it an invaluable asset in the realm of ML.

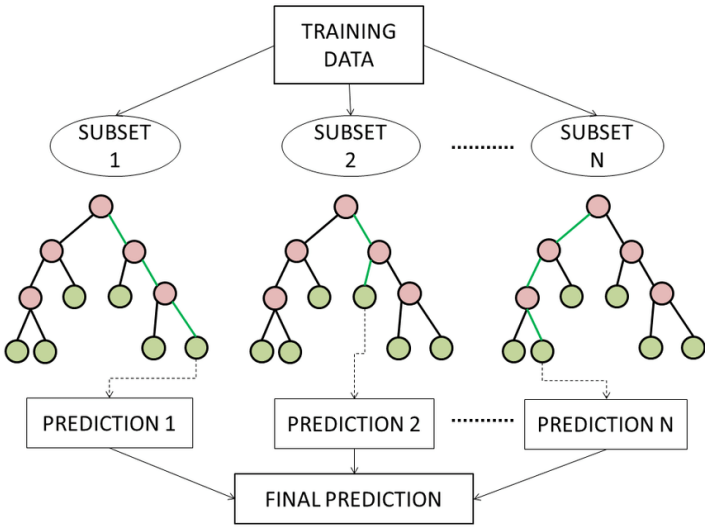


Figure 3.2 Example of a Random Forest workflow. Reproduced with permission from [9].

3.2.3 Gaussian Naive Bayes

The Gaussian Naive Bayes algorithm is a probabilistic ML technique based on the principles of Bayes' theorem and is particularly well-suited for classification tasks. At its core, the algorithm leverages the assumption that features are conditionally independent within each class, given the class label. This "naïve" assumption simplifies the calculation of probabilities and makes the algorithm computationally efficient. As shown in **Figure 3.3**,[10] Gaussian Naive Bayes is specifically designed for datasets where the features follow a Gaussian distribution, also known as a normal distribution. The z-score, or the difference between the distance from the mean and the standard deviation, is computed for each dimension (only one dimension is displayed in the graphic). The formula for the z-score distance from Class A is $(x - \mu_A)/\sigma_A$, while the formula for the z-score distance from Class B is similarly expressed. A p-value may be obtained immediately from each z-score distance. The algorithm calculates the likelihood of observing a particular set of feature values within each class, as well as the prior probability of each class occurring. By combining these probabilities, Gaussian Naive Bayes estimates the posterior probability of each class for a given input, enabling it to make probabilistic predictions. Despite its simplistic assumptions, Gaussian Naive Bayes often performs remarkably well in practice, particularly when dealing with high-dimensional datasets and cases where the independence assumption is not severely violated. It has found applications in various domains, including text classification, spam detection, and medical diagnosis, showcasing its effectiveness as a versatile classification tool.

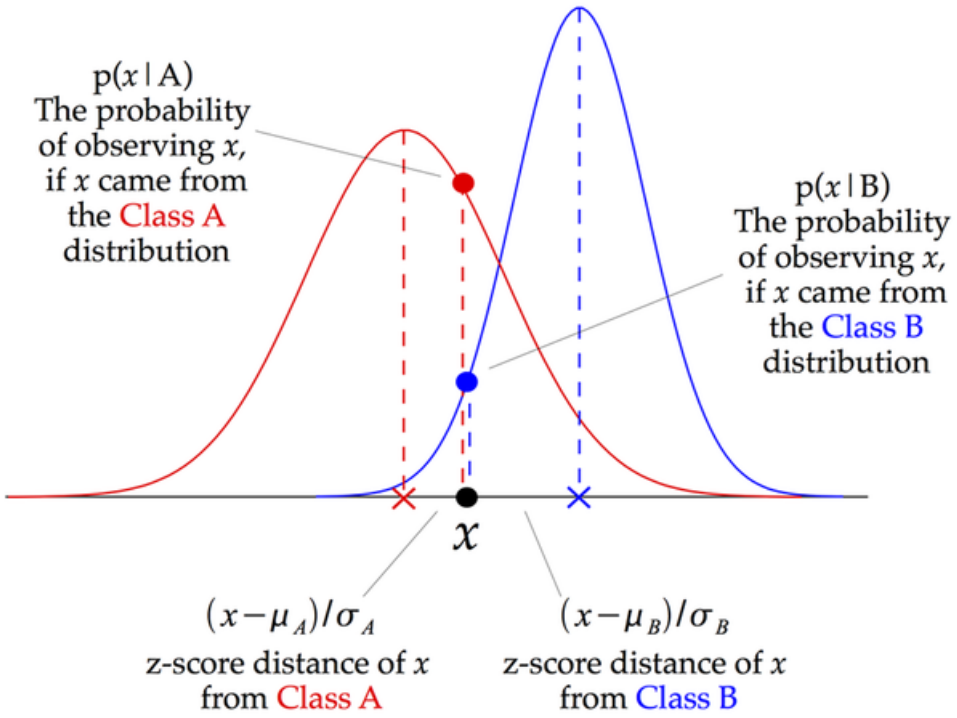


Figure 3.3 Illustration of how a Gaussian Naive Bayes (GNB) classifier works. Reproduced with permission from [10].

3.2.4 AdaBoost Algorithm

The AdaBoost (Adaptive Boosting) algorithm stands as a prominent ensemble learning method recognized for its capacity to improve the accuracy of weak learners and create a strong predictive model. The core principle underlying AdaBoost is to sequentially train multiple weak learners, typically decision trees with limited depth, and assign weights to each data point. **Figure 3.4** demonstrates that initially, all data points are given equal weight, and a weak learner is trained to minimize classification errors.[11] In subsequent iterations, the algorithm focuses more on the data points that were misclassified in the previous step by assigning them higher weights. This iterative process continues, with each new weak learner addressing the misclassified data points from the previous ones. Finally, AdaBoost combines the predictions of these weak learners by giving each of them a weight, resulting in a strong ensemble model. AdaBoost excels in identifying complex decision boundaries and adapting to noisy data, making it a versatile tool for a wide range of

classification tasks. Its ability to prioritize challenging instances and weigh their importance in the final prediction process has earned AdaBoost a reputation as a powerful algorithm for boosting the performance of ML models.

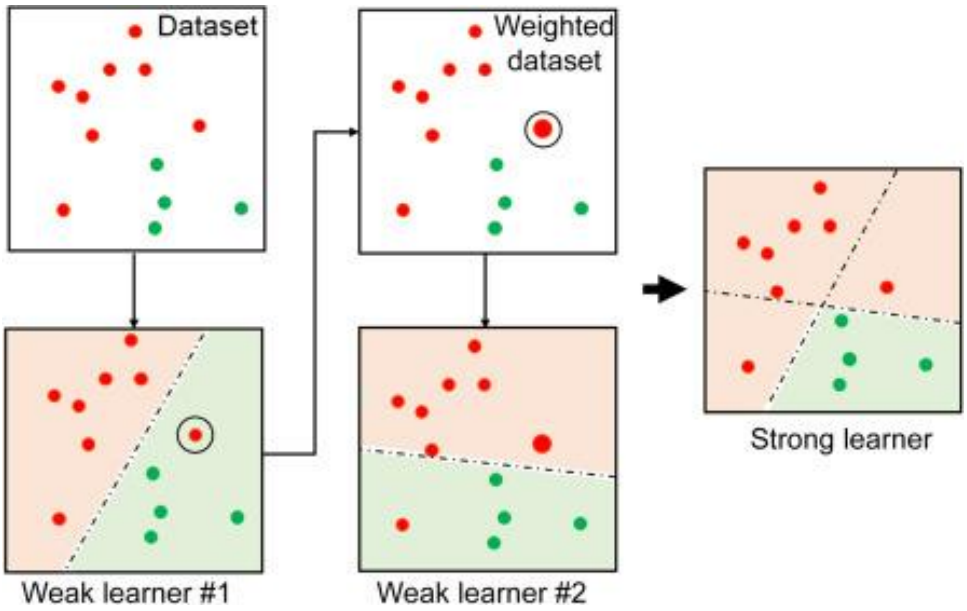


Figure 3.4 Implementation of AdaBoost classifier on a dataset that has two features and two classes. Weak learner #2 improves on the mistake made by weak learner #1, such that the decision boundaries learnt by the two weak learners can be combined to form a strong learner. In this case, each weak learner is a decision tree, and AdaBoost classifier (i.e., strong learner) combines the weak learner in series. Reproduced with permission from [11].

3.2.5 SchNet

The SchNet algorithm represents a significant advancement in deep learning architecture tailored to modeling intricate atomic interactions within molecular and material systems.[12] Building upon the foundation laid by Deep Tensor Neural Networks（DTNNs),[13] SchNet is engineered to capture the complex interplay of atoms, enabling the prediction of potential-energy surfaces and accelerating the exploration of chemical space.

Within the SchNet framework, a unique description of an atomistic system can be generated through a set of n atom sites with nuclear charges $Z = (Z_1, \dots, Z_n)$, and positions $R = (r_1, \dots, r_n)$. A graphical overview of the SchNet algorithm is depicted in **Figure 3.5**.

The atoms are represented by a tuple of features $X^l = (x_1^l, \dots, x_n^l)$, where n denotes the number of atoms, and l is the current layer. Each atom, i, is initially represented using an atom type-dependent embedding, Z_i, and a linear dependency on the single-atom contribution, E_i.

$$x_i^o = m_{Z_i} + n_{E_i}$$

These embeddings m_Z and linears n_E are randomly initialized and optimized during the training phase. Both the atom-wise layers and interaction blocks of SchNetE borrow from the original SchNet building blocks. Atom-wise layers are dense layers applied independently to each atom's representations x_i^l.

$$x_i^{l+1} = W^l x_i^l + b^l$$

where the weights W^l and biases b^l are optimized during training. The interaction blocks model the interaction between atoms and their surroundings, using continuous-filter convolutional layers that are represented as

$$x_i^{l+1} = (X^l * W^l)_i = \sum_{j=0}^{n} x_j^l \circ W^l(r_j - r_i)$$

where W^l are generated from the filter-generating networks. Shifted softplus will be used as activation functions throughout the network. To ensure indexing invariance, we will sum up the atom-wise contributions. Following these layers, a property P of a a molecule or material is predicted from the final layer.

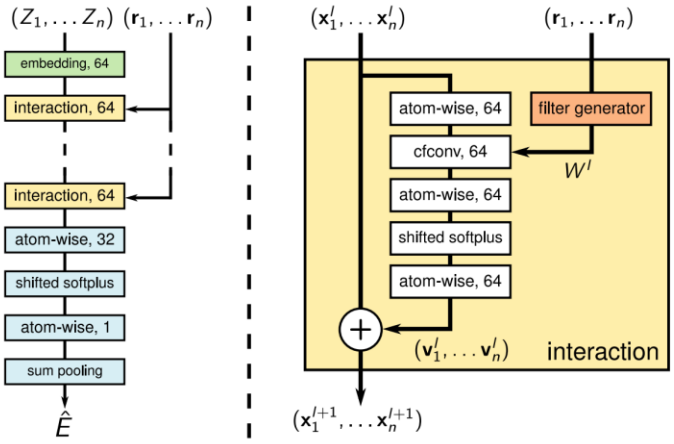


Figure 3.5 Illustrations of the SchNet architecture (left) and interaction blocks (right) with atom embedding in green, interaction blocks in yellow, and property prediction network in blue. For each parameterized layer, the number of neurons is given. Reproduced with permission from [12].

3.2.6 Transfer Learning

Transfer learning is a pivotal ML technique that has gained substantial recognition for its efficacy in leveraging knowledge learned from one domain to improve performance in a related but different domain. Transfer learning operates on the principle that models trained on a source domain can be fine-tuned or adapted for a target domain, even when the target domain has limited labeled data.[13] This approach enables the efficient transfer of valuable knowledge, representations, and features learned during the training of a source model to the target task. Transfer learning can manifest in various ways, including domain adaptation, where models are adapted to a new distribution of data, and feature extraction, where pre-trained models serve as feature extractors for downstream tasks. By reusing and building upon learned knowledge, transfer learning significantly reduces the need for extensive labeled data in the target domain, enhances model generalization, and accelerates the development of robust ML systems. This technique has found widespread applications across domains such as natural language processing, computer vision, and scientific research,[14-17] where data scarcity or the cost of collecting labeled data often poses challenges. Transfer learning stands as a testament to the power of leveraging prior knowledge to address new and complex ML challenges efficiently.

It unfolds in three consequential phases: pre-training, model adaptation, and fine-tuning as shown in **Figure 3.6**. Initially, a ML model is trained on the source dataset which is large, comprehensive, and available. This dataset could encompass a broad range of materials and their associated properties. The primary aim of this step is to enable the model to identify and understand the underlying relationships and patterns between different material properties and characteristics. This phase equips the model with a fundamental understanding of the materials' behaviour. Following the pre-training, the model is tailored to the specific materials science problem under investigation. This customization can be achieved by altering the model's architecture, such as adding specially designed new layers, modifying existing layers, or making other adaptations to the model's structure. This results in an adapted model that can recognize patterns relevant to the target problem. The adapted model is then further trained on the target dataset, which is more specific and relevant to the research problem at hand, but generally much smaller in size. This stage involves fine-tuning the model's parameters to optimize its performance for the specific task. Given that the model has already gained significant knowledge from the pre-training stage, it can perform more efficiently on the target dataset, requiring fewer data and resources to achieve robust results.

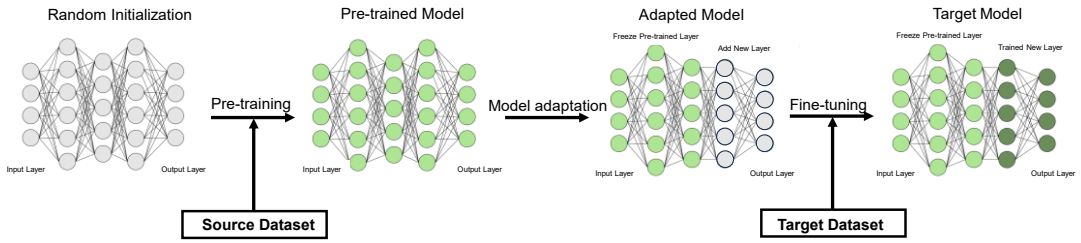


Figure 3.6 Schematic of the transfer learning protocol

3.3 Simulation Calculation

Computational approaches that operate across various size and time scales are crucial for understanding the properties of materials at both the microscopic and macroscopic levels. As illustrated in **Figure 3.7**, these methods include sub-atomic scale techniques such as

first-principles density functional theory (DFT), which can provide insights into the electronic structure of materials.[18] At the atomistic level, classical molecular dynamics (MD) simulations, which rely on force fields, are used to study the physical movements of atoms and molecules. For larger scale phenomena that involve the structural behavior of materials, finite element analysis (FEA) is used, which is particularly beneficial for process simulation and engineering design. In this thesis, the DFT method and MD methods are employed.

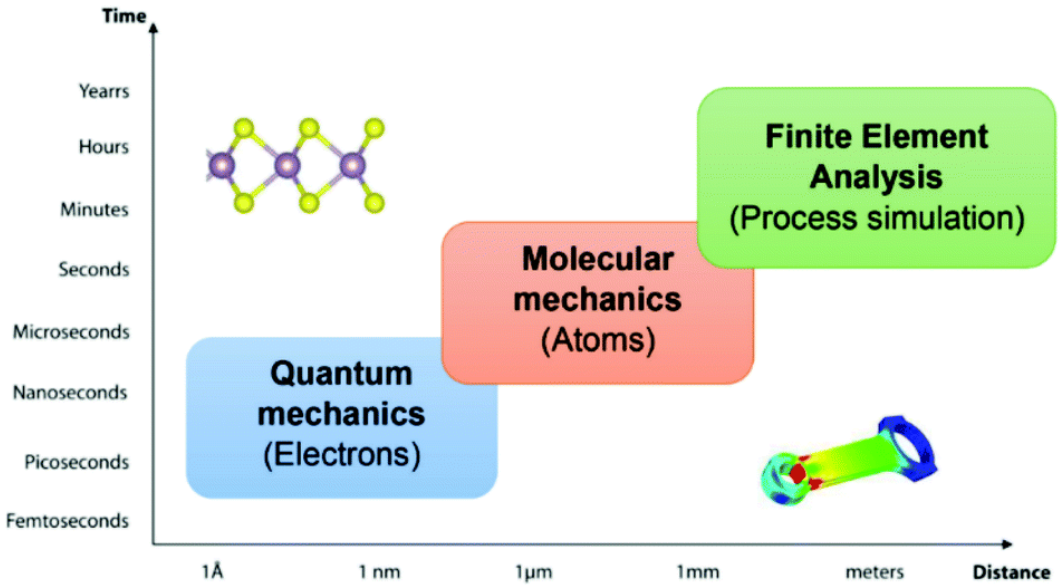


Figure 3.7 Space and time scale in computational materials science. Reproduced with permission from [18].

3.3.1 Density Functional Theory

The Born-Oppenheimer (B-O)[19] approximation and the Hartree-Fock (H-F)[20] method were introduced as simplifications for multi-particle systems. The H-F method transforms a multi-particle system into a one-electron system, yet it does not consider electron-electron correlation. In response to this limitation, the DFT was developed to address inaccuracies associated with the H-F method. DFT has its origins in the Thomas-Fermi model proposed in 1927,[21, 22] with foundational concepts introduced by P. Hohenberg and W. Kohn in 1964.[23] DFT relies on two fundamental theorems:

1. Ground state properties of interacting many-particle systems with a common external potential depend solely on the electron density distribution of the ground state, which is non-degenerate.

2 The correct density distribution, $n(r)$, minimizes the energy, yielding the ground state energy of the system, $E[n(r)]$.

DFT primarily employs the electron density function, as opposed to traditional wave functions, to represent ground state physical properties. In 1965, W. Kohn and L. J. Sham formulated the Kohn-Sham (K-S) equation,[24] which incorporates an interaction-free functional and consolidates errors into a single term. This equation serves as the foundation for modern electronic structure calculations, encompassing atoms, molecules, and condensed matter. The K-S equation is expressed as follows:

$$\left(-\frac{1}{2}\nabla^2 + \upsilon_{\text{eff}}\left[n^{(k)}(\boldsymbol{r})\right]\right)\psi_i(\boldsymbol{r}) = \epsilon_i\psi_i(\boldsymbol{r})$$

Where $\upsilon_{\text{eff}}(\boldsymbol{r})$ is the effective potential determined by electron density $n^{(k)}(\boldsymbol{r})$, $\psi_i(\boldsymbol{r})$ represents the wave function, and ϵ_i denotes the orbital energy corresponding to $\psi_i(\boldsymbol{r})$.

$$\upsilon_{\text{eff}}(\boldsymbol{r}) = \upsilon_{\text{ext}}(\boldsymbol{r}) + \upsilon_H\left[n^{(k)}(\boldsymbol{r})\right] + \upsilon_{xc}\left[n^{(k)}(\boldsymbol{r})\right]$$

The effective potential, $\upsilon_{\text{eff}}(\boldsymbol{r})$, is defined by above equation and depends on the given electron density. The first term, $\upsilon_{\text{ext}}(\boldsymbol{r})$, signifies the Coulomb potential, acting as a static external potential within the B-O approximation. The second term, $\upsilon_H\left[n^{(k)}(\boldsymbol{r})\right]$, accounts for the classical Coulomb potential originating from neighboring electrons, often referred to as the Hartree Potential. It approximates electron-electron interactions within the system. The third term, $\upsilon_{xc}\left[n^{(k)}(\boldsymbol{r})\right]$, corresponds to the exchange-correlation potential, encompassing exchange and correlation interactions. The Pauli Exchange Principle dictates that electrons with parallel spins cannot occupy the same location, resulting in effective repulsion. Correlation interaction, also a consequence of the Pauli Exchange Principle, arises from correlated motion between electrons with anti-parallel spins, driven by coulombic repulsion.

The general workflow of DFT calculations involves an iterative process to find the minimum energy point by iterating trial densities, $n^{(k)}$. Initially, an electron density estimate, $n^{(k=1)}(\boldsymbol{r})$, is derived from an initial geometry guess. Subsequently, the effective potential, $v_{\text{eff}}(\boldsymbol{r})$, is computed and used to solve the Kohn-Sham equation. This results in a new electron density, $n(\boldsymbol{r})$, which, when it matches $n^{(k=1)}(\boldsymbol{r})$, signifies self-consistency. The total energy and forces are then computed. If convergence is achieved, the calculation concludes; otherwise, the process continues with model reconstruction.

3.3.2 Molecular Dynamics

MD simulation relies on relies on solving Newton's equations of motion to predict the trajectories of individual atoms or molecules in a system.[25] By numerically integrating these equations, it becomes possible to simulate the dynamic evolution of a molecular system over time. This simulation allows researchers to observe how particles interact, move, and arrange themselves, proving valuable insights into a wide range of phenomena. The flowchart shown in **Figure 3.8** illustrates the subsequent processes to perform MD simulation.[26]

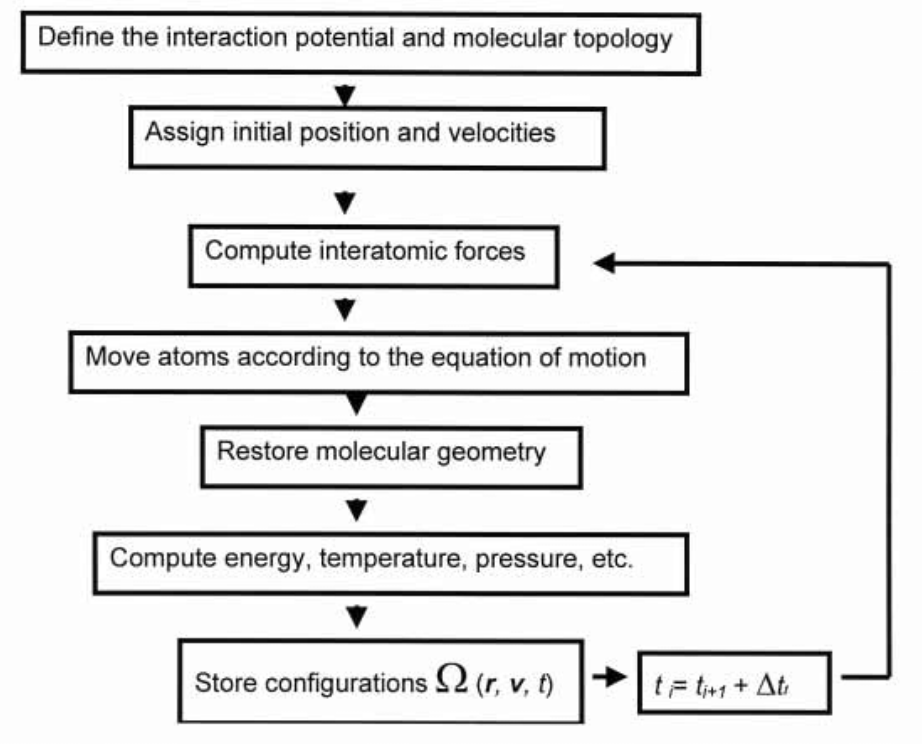


Figure 3.8 Scheme of the molecular dynamics simulation procedure. Reproduced with permission from [26].

A force field is a set of mathematical equations and parameters that describe the interactions and forces between atoms or molecules in a system. It includes terms for bonded interactions (e.g., covalent bonds, angles, dihedrals) and non-bonded interactions (e.g., van der Waals forces and electrostatic interactions). Force fields are essential for calculating the forces acting on particles during simulation. To accurately capture the intermolecular forces, such simulations often incorporate the Lennard-Jones (LJ) potential,[27] a classical force field model describing both attractive van der Waals and repulsive Pauli forces between particles. To start an MD simulation, an initial configuration of the system is required, which includes the positions and velocities of all particles. This configuration can be based on experimental data, previous simulations, or generated using specific algorithms. MD simulations are typically carried out in periodic boundary conditions, where a simulation box is replicated infinitely in all directions. This prevents edge effects and allows for the simulation of bulk properties. Alternatively, other boundary conditions, such as fixed boundaries or surfaces, can be used for specific studies.

Researchers choose an ensemble, such as the NVE (constant number of particles, constant volume, constant energy), NVT (constant number of particles, constant volume, constant temperature), or NPT (constant number of particles, constant pressure, constant temperature), to control the system's thermodynamic properties during the simulation. After the simulation, extensive analysis is performed to extract meaningful information from the generated trajectories. This includes calculating properties such as temperature, pressure, density, diffusion coefficients, radial distribution functions, and more.

References

[1] Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, 8 (31.10), 5281

[2] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, 32 (7), 1466-1474.

[3] Zhong, S.; Zhang, Y.; Zhang, H. Machine learning-assisted QSAR models on contaminant reactivity toward four oxidants: combining small data sets and knowledge transfer. *Environmental Science & Technology* **2021**, 56 (1), 681-692.

[4] Venkatraman, V. FP-ADMET: a compendium of fingerprint-based ADMET prediction models. *Journal of Cheminformatics* **2021**, 13 (1), 1-12.

[5] Müller, K.-R.; Tangermann, M.; Dornhege, G.; Krauledat, M.; Curio, G.; Blankertz, B. Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods* **2008**, 167 (1), 82-90.

[6] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **2002**, 46, 389-422.

[7] Ali, P. J. M.; Faraj, R. H.; Koya, E.; Ali, P. J. M.; Faraj, R. H. Data normalization and standardization: a technical report. *Machine Learning Technical Reports* **2014**, 1 (1), 1-6.

[8] Rani, A.; Kumar, N.; Kumar, J.; Sinha, N. K. Machine learning for soil moisture assessment. In *Deep Learning for Sustainable Agriculture*, **2022**; Academic Press: pp 143-168.

[9] Laudato, G.; Oliveto, R.; Scalabrino, S.; Colavita, A. R.; De Vito, L.; Picariello, F.; Tudosa, I. Identification of R-peak occurrences in compressed ECG signals. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, **2020**; IEEE: pp 1-6.

[10] Raizada, R. D.; Lee, Y.-S. Smoothness without smoothing: why Gaussian naive Bayes is not naive for multi-subject searchlight studies. *PLoS One* **2013**, 8 (7), e69566.

[11] Misra, S.; Li, H.; He, J. Noninvasive fracture characterization based on the classification of sonic wave travel times. *Machine Learning for Subsurface Characterization* **2020**, 4, 243-287.

[12] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, 148 (24).

[13] Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Communications* **2017**, 8 (1), 13890.

[14] Feng, S.; Zhou, H.; Dong, H. Application of deep transfer learning to predicting crystal structures of inorganic substances. *Computational Materials Science* **2021**, 195, 110476.

[15] Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature Communications* **2019**, 10 (1), 5316.

[16] Kong, S.; Guevarra, D.; Gomes, C. P.; Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Applied Physics Reviews* **2021**, 8, 021409.

[17] Li, X.; Zhang, Y.; Zhao, H.; Burkhart, C.; Brinson, L. C.; Chen, W. A transfer learning approach for microstructure reconstruction and structure-property predictions. *Scientific Reports* **2018**, 8 (1), 13461.

[18] Yengejeh, S. I.; Kazemi, S. A.; Wen, W.; Wang, Y. Multiscale numerical simulation of in-plane mechanical properties of two-dimensional monolayers. *RSC Advances* **2021**, 11 (33), 20232-20247.

[19] Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, 389 (20), 457-484.

[20] Hartree, D. R. The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods. In *Mathematical Proceedings of the Cambridge Philosophical Society*, **1928**; Cambridge University Press: Vol. 24, pp 89-110.

[21] Thomas, L. H. The calculation of atomic fields. In *Mathematical Proceedings of the Cambridge Philosophical Society*, **1927**; Cambridge University Press: Vol. 23, pp 542-548.

[22] Fermi, E. Un metodo statistico per la determinazione di alcune priorieta dell'atome. *Rend. Accad. Naz. Lincei* **1927**, 6 (602-607), 32.

[23] Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Physical Review* **1964**, 136 (3B), B864.

[24] Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review* **1965**, 140 (4A), A1133.

[25] Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms To Applications*; Elsevier, **2023**.

[26] Grigera, J. Molecular dynamics simulation for ligand-receptor studies. Carbohydrates interactions in aqueous solutions. *Current Pharmaceutical Design* **2002**, 8 (17), 1579-1604.

[27] Lennard, J.; Jones, I. On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature. *Proceedings of the Royal Society of London. Series A, containing papers of a mathematical and physical character* **1924**, 106 (738), 441-462.

Chapter 4*

Machine Learning Assisted High-Throughput Screening for Elastomers

This chapter discusses the integration of high-throughput screening (HTS) with machine learning (ML) to accelerate the discovery of new elastomers, using new structure-based multilevel (SM) descriptors. These descriptors are hierarchically organized to capture both the micro-level and macro-level structures of elastomers. The SM-Morgan Fingerprint (SM-MF), a type of SM descriptor, is showcased for its ability to enable ML models to predict the toughness of elastomers with an impressive accuracy. Additionally, the chapter elaborates on the development of an HTS pipeline that leverages the SM descriptors to rapidly screen for elastomers with desired mechanical properties. The effectiveness and generality of SM descriptors is also validated by their application in creating HTS pipelines for different properties, like critical strain or Young's modulus.

---

*This section published as **Siyan Deng**, Chao Chen, Ke Li, Xi Chen, Kelin Xia*, and Shuzhou Li*. Structure-Based Multilevel Descriptors for High-throughput Screening of Elastomers. *The Journal of Physical Chemistry B* (2023). Reproduced with permission. Copyright (2023) American Chemical Society

4.1 Introduction

Elastomers are a class of polymers that have a small Young's modulus and a large critical strain compared to other materials. They are a class of polymeric materials that can undergo large deformation under force and quickly return to their near-initial state after the withdrawal of the external force.

As the realm of flexible electronics rapidly evolves, there is an increasing demand for materials that can accommodate novel device architectures and functionalities. Traditional rigid and inflexible materials are increasingly being replaced or complemented by softer, stretchable counterparts to enable innovations such as wearable electronics, implantable medical devices, and flexible displays. In this context, elastomers stand at the forefront of potential solutions, offering intrinsic stretchability and adaptability. Elastomers that possess specific mechanical properties, such as high tensile strength, durability, and resistance to fatigue, can significantly enhance the longevity and robustness of these devices. Discovering new elastomers tailored for electronic devices is of paramount importance. Such advancements not only enable the creation of devices that can conform, bend, and stretch without compromising performance but also open the door to entirely new device concepts previously considered unfeasible. The exploration and identification of new elastomers are critical for advancing electronic devices, as these materials hold the potential to break through existing design constraints and pave the way for a new generation of electronics.

Historically, the discovery of new elastomers has largely been governed by a trial-and-error approach. This method, although foundational in many breakthroughs, is inherently time-consuming and resource-intensive, often requiring extensive laboratory work to synthesize, test, and validate potential materials. Furthermore, the vast chemical space and countless combinations of monomers and crosslinkers make it nearly impossible to exhaustively explore all potential elastomers through conventional methods. Recently, Machine Learning (ML) has emerged as a promising alternative in this endeavor. Leveraging computational power and data-driven algorithms, ML can predict material

properties, optimize synthesis pathways, and identify promising candidates with high accuracy.[1-4] Instead of laboriously testing each potential compound in the lab, researchers can now use ML models to narrow down the most promising candidates, thereby streamlining the discovery process. Thus, integrating ML into elastomer research not only expedites the discovery of novel materials but also provides a more systematic and comprehensive approach to exploring the vast landscape of potential elastomers.

In recent years, the application of ML methodologies to discover new materials has seen widespread adoption across various fields.[5-14] These techniques, harnessing the power of data-driven algorithms, have expedited the identification and optimization of materials with desirable properties. However, there have been limited reports on applying this computational approach for elastomer discovery. A primary reason for this gap is the challenges associated with elastomer descriptors. To date, accurately predicting the mechanical properties of elastomers requires descriptors derived from molecular structures, simulations, and/or experimental data.[15-19] Descriptors based on molecular structures typically originate from the monomers comprising the elastomers, providing insight into their physicochemical features and stoichiometry.[15-19] Descriptors based on numerical simulation, such as electronic parameters (e.g., HOMO/LUMO gap, polarizability) from density functional theory,[16, 17] thermodynamic parameters from thermodynamic models,[16] elongation simulation data from molecular dynamics,[19] and chain architecture from Monte Carlo simulations,[18] provide crucial information about intermolecular interactions. Descriptors based on experiments, such as FT-IR absorbance, can be utilized to gauge the degree of crosslinking.[16, 18] Although these descriptors, when combined, provide a comprehensive characterization of elastomers, their dependence on simulation and experimental data limits the applicability of HTS in elastomer system, as acquiring experimental data or conducting complex simulations for all candidates of interest is often not feasible. The elastomer descriptors derived solely from molecular structure have rarely been reported. Therefore, it is desired to develop accurate and efficient descriptors that are solely derived from the molecular structure of elastomers to enable fast and accurate prediction of their mechanical properties. Successful development of structure-based

descriptors for elastomers could expand the applicability of HTS in polymer research and greatly enhance the efficiency and speed of materials discovery in this field.

In this study, we propose a novel set of elastomer descriptors called structure-based multilevel (SM) descriptors. The SM descriptors offer a computationally efficient and universally applicable approach as they rely solely on the molecular structure of elastomers. Specifically, SM descriptors combine simplified dimer representation (SDR) descriptors, sparse descriptors based on soft segment (SS) mass, ratios of different blocks, and sparse descriptors based on polymer mass. Among these components, the SDR descriptors are capable of incorporating information from up to three monomers while retaining the connectivity information between different units. The ML model, using SM-Morgan Fingerprint (SM-MF) descriptors, one of our SM descriptors, predicts the toughness of polydimethylsiloxane (PDMS)-based elastomers with an impressive accuracy of 0.91, which is good enough for binary classification. Based on the model, an ML-assisted High-Throughput Screening (HTS) pipeline was successfully constructed to rapidly screen elastomers with targeted toughness. The generality of SM descriptors was also demonstrated by developing ML-assisted HTS pipelines for identifying elastomers with targeted critical strain and Young's modulus, achieving respective ML model accuracies of 0.89 and 0.87. The hierarchical structure of SM descriptors provides a comprehensive description of both local and global structure of elastomers, from the local soft and hard segment structure to the global polymer structure. The low computational cost and ease of use of SM descriptors are expected to substantially enhance HTS capabilities toward the successful discovery of novel materials.

4.2 Methods

4.2.1 Data Collection

In this study, a dataset of PDMS-based elastomers was compiled from previously published literature,[20-38] with a total of 76 entries. The elastomers were divided into two groups based on the number of modified units incorporated into the PDMS as shown in **Figure 4.1a**. Group 1 comprised elastomers with one type of unit incorporated into the PDMS, while

Group 2 comprised elastomers with two types of units incorporated into the PDMS. The modified units were crosslinked to form the hard segment (HS), while PDMS served as the SS. The Group 1 and Group 2 include 47 and 29 entries, respectively, all with their respective molecular structure and key mechanical properties. The mechanical properties in the dataset were toughness, critical strain, and Young's modulus.

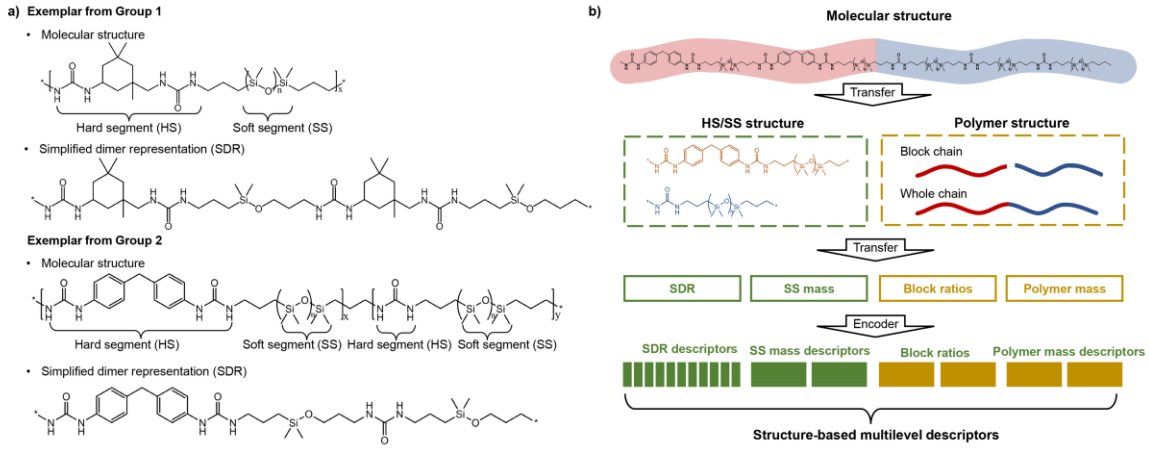


Figure 4.1 (a) Molecular structure and simplified dimer representation (SDR) of exemplars from Group 1 and Group 2. Group 1 incorporates one type of modified unit into PDMS, while Group 2 incorporates two types of modified units into PDMS. The modified units serve as the hard segment (HS), crosslinking to form the network structure, while the PDMS serves as the soft segment (SS). (b) Schematic representation of the structure-based multilevel (SM) descriptors.

4.2.2 Feature Engineering

SM descriptors are structure-based descriptors used to accurately describe the structure of elastomers by combining vital factors that determine their properties, including the local HS structure, SS structure, and overall polymer structure, as shown in **Figure 4.1b**.

To describe the local HS/SS structure of elastomers, conventional monomer representations can be challenging as elastomers often include multiple monomers. To address this issue, a simplified dimer representation (SDR) was developed as shown in **Figure 4.1a**. In the SDR, the SS is simplified to a polymerization degree of one, while key

information regarding the atom types, chemical surrounding of each atom, backbone structure, and the position of pendant groups was effectively retained. Notably, the SDR can encapsulate information from up to three monomers (one from the SS and a maximum of two from the HS) while maintaining the connectivity information between the different units. The SS mass complements the SDR by providing additional information about the local SS structure. The overall polymer structure, including the chain length of different blocks and the whole chain length, was described using block ratios and polymer mass.

To calculate the descriptors, the SDR was encoded into vectors using four different sets of descriptors: Morgan Fingerprint, PubChem Fingerprint, RDKit descriptors, and PaDel descriptors. Morgan Fingerprint and RDKit descriptors were generated using the RDKit software, while PubChem Fingerprint and PaDel descriptors were generated using the PaDel software.[39] The SS mass was encoded using a feature discretization technique to produce sparse descriptors that denoted the range of SS mass.[40] The block ratios were used directly as the descriptors, and the polymer mass was encoded into sparse polymer mass descriptors, which denoted the range of polymer mass. Four different sets of structure-based multilevel (SM) descriptors were obtained, namely SM-MF, SM-PF, SM-R, and SM-P, where MF, PF, R and P denoted Morgan Fingerprint, PubChem Fingerprint, Rdkit descriptors, PaDel descriptors, respectively.

To reduce the dimensionality of the descriptor space and avoid overfitting, various dimension reduction techniques were employed, including low variance filters, high correlation filters, and recursive feature elimination (RFE) algorithm.[41] Descriptors with low variance were removed as they provide less discriminatory power than those with high variance. Similarly, descriptors that exhibited a linear correlation coefficient higher than a specific threshold were randomly removed until only one of them remained in the descriptor set. The RFE algorithm was used to iteratively remove the least important descriptors until the desired set of descriptors was obtained.

4.2.3 Model Development

To predict the mechanical properties of elastomers based on their structure, ML models were developed using different algorithms and descriptors. The algorithms included Gaussian naïve Bayes (GNB), support vector machine classifiers (SVC),[42] adaptive boosting (AdaBoost),[43] and random forest (RF),[44] while the descriptors included SM-MF, SM-PF, SM-R, SM-P. The decision to utilize these four distinct algorithms was primarily driven by concerns related to overfitting and the need for model interpretability. Given the limited size of our dataset, overfitting becomes a significant concern. Our selected algorithms, namely SVC, GNB, AdaBoost, and RF, demonstrate resistance against overfitting.[44-47] Beyond their robustness, these algorithms provide invaluable insights into feature significance, enhancing model interpretability. The elastomers with a toughness greater than 2 MPa were classified as "durable" elastomers, and those that did not meet this criteria as "brittle" elastomers. This threshold was chosen to ensure a balance between the two classes.

To evaluate the performance of each model, we used five metrics: accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curves. Accuracy quantifies the model's overall correctness, while precision evaluates the ratio of true positives to all predicted positive instances. Recall assesses the proportion of true positives among all actual positive cases, and the F1 score serves as a harmonized mean of precision and recall, effectively balancing the two metrics. For these four metrics, a commonly used rule of thumb for evaluating classification model is as follows: a score of 0.9-1 indicates excellent performance, 0.8-0.9 represents good performance, 0.7-0.8 signifies fair performance, 0.6-0.7 denotes poor performance, and a score below 0.6 suggests a failure in classification.[48] ROC curves illustrate the trade-off between true positive rate and false positive rate at varying classification thresholds. Generally, an area under curve (AUC) of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.[49]

To optimize the performance of the models, we employed five-fold cross-validation (CV), which involved iterative withholding of a subset of the training data to evaluate the model's generalizability. We evaluated the performance of each model using cross-validation and

selected the models that exhibited the highest cross-validation performance. All modelling was performed in Python using the Scikit-Learn library, a widely used machine learning library.

4.2.4 High-throughput Screening

To enable HTS of elastomers, a candidate dataset was created as the search space. The dataset was constructed by combining 16 unique modified units selected from our collected dataset (m1 o m16), two types of SS mass, one type of block ratios (set at 0.5 for all candidates), and two types of polymer mass. By combining all possible combinations of these factors, a candidate dataset comprising 460 entries was generated.

The most accurate model in our study, selected based on the performance metrics from the cross-validation analysis in section 2.3, was applied to predict the toughness of each polymer in the candidate dataset. This allowed us to screen for potential “durable” elastomers that met our desired criteria. By using this approach, we aimed to rapidly identify the most promising candidate materials for further experimental validation, thereby accelerating the development of novel elastomers with desirable properties.

4.2.5 Generality of SM descriptors

To assess the generality of our SM descriptors in predicting additional mechanical properties, their performance was examined on critical strain and Young's modulus. Elastomers with a critical strain exceeding 650% were categorized as "stretchable" elastomers. In contrast, those that did not meet this threshold were classified as "non-stretchable" elastomers. Similarly, elastomers with a Young's modulus below 1.5 MPa were considered "flexible", while those exceeding this value were deemed "rigid" elastomers.

The same four SM descriptors (SM-MF, SM-PF, SM-R, and SM-P) and four algorithms (GNB, SVC, AdaBoost, and RF) as in Section 4.2.3 was used to select the optimal models

for each mechanical property. To evaluate the performance of the models, the same five metrics (accuracy, precision, recall, F1 score, and ROC curves) were used. The most accurate models in our study for critical strain and Young's modulus were selected based on their cross-validation performance.

The most accurate models were then applied to predict the mechanical properties of each polymer in the candidate dataset, allowing us to conduct HTS for identifying potential elastomers that met our desired criteria for critical strain or Young's modulus. This approach enabled us to assess the generality of our SM descriptors across different mechanical properties and identify elastomer candidates with desirable properties beyond toughness.

4.3 Results and Discussion

4.3.1 Effectiveness of SM Descriptors

To assess the effectiveness of SM descriptors in predicting the toughness of elastomers and compare their performance with the current best structure-based elastomer descriptors (Control), we employed four distinct ML algorithms (GNB, SVC, AdaBoost, and RF) and trained them with five different descriptors（SM-MF, SM-PF, SM-R, SM-P, and Control). To distinguish the models trained with specific algorithms and descriptors, the "algorithm/descriptor" model notation was adopted. For instance, the model trained using the GNB algorithm and SM-MF descriptor is represented as the GNB/SM-MF model.

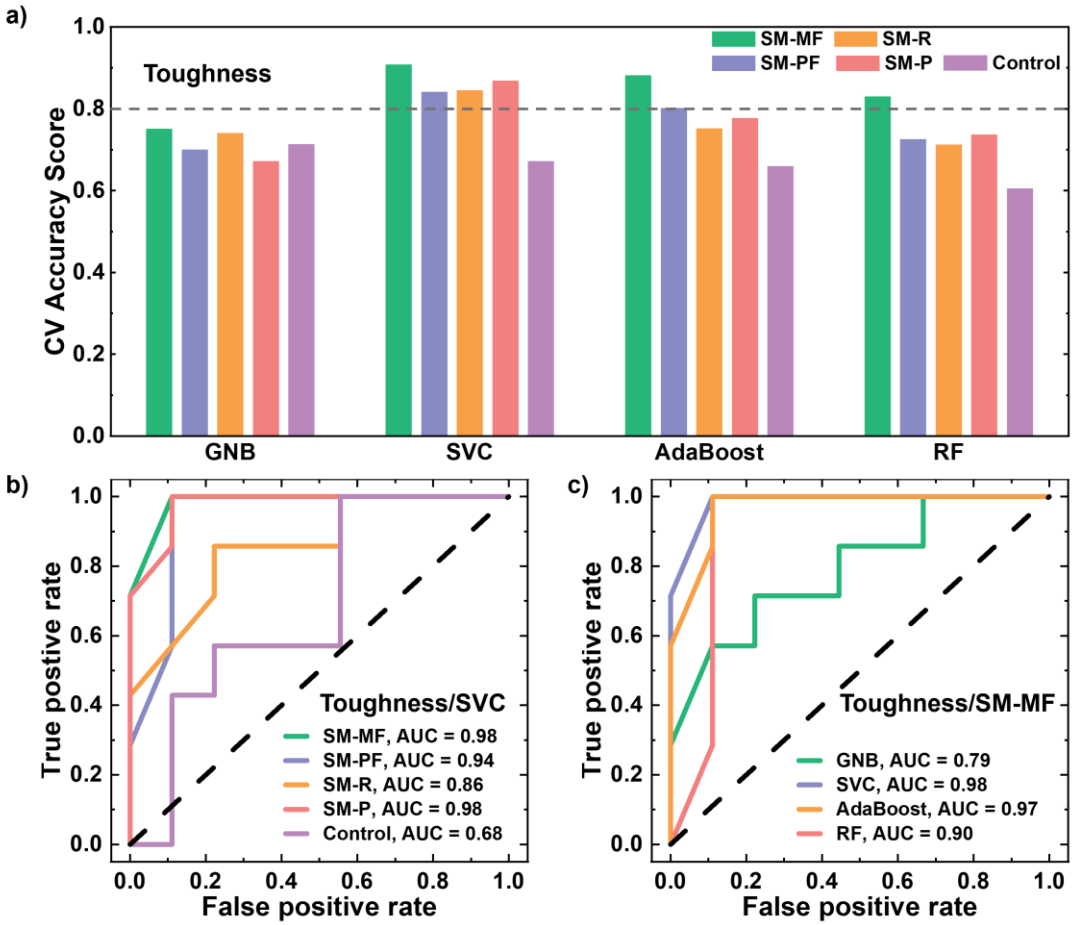


Figure 4.2 Model performance using different descriptors and algorithms. (a) Five-fold cross-validation accuracy scores for models trained using five different descriptors (SM-MF, SM-PF, SM-R, SM-P, and Control) and four different algorithms (GNB, SVC, AdaBoost, and RF). The grey dashed line represents the accuracy of 0.8. (b) ROC curves of models trained using the SVC algorithm and five different descriptors. (c) ROC curves of models trained using the SM-MF descriptor and four different algorithms.

Figure 4.2a displays the five-fold CV accuracy scores for the twenty models. Our primary metric was accuracy, while precision, recall, and F1 scores for these twenty models can be found in **Table 4.1**. The results indicate that the SM-MF descriptor consistently outperformed other descriptors across all four algorithms. The SVC/SM-MF model achieved the highest accuracy of 0.91. Other models using the SM-MF descriptor generally displayed strong performance, though the GNB/SM-MF model had a relatively lower

accuracy of 0.76. Notably, the AdaBoost/SM-MF model achieved an accuracy score of 0.88, and the RF/SM-MF model scored 0.83. The SM-PF descriptor yielded mixed results: the SVC/SM-PF and AdaBoost/SM-PF models achieved relatively high accuracy scores above 0.8, while the other two models had lower scores. Models trained with the SM-R descriptor had lower accuracy scores, with the highest score of 0.84 achieved by the SVC/SM-R model. The SM-P descriptor-based models performed moderately well, with the SVC/SM-P model exhibiting the highest accuracy score of 0.87. The Control descriptor-based models generated significantly lower accuracy scores compared to models using SM descriptors. The best Control descriptor-based model was the GNB/Control model with an accuracy score of 0.71.

Table 4.1 The performance of models predicting toughness.

Descriptor	Algorithm	Accuracy	Precision	Recall	F1
SM-MF	SVC	0.91	0.92	0.93	0.91
SM-MF	GNB	0.75	0.79	0.79	0.76
SM-MF	RF	0.83	0.87	0.79	0.81
SM-MF	AdaBoost	0.88	0.92	0.88	0.87
SM-PF	SVC	0.84	0.83	0.95	0.87
SM-PF	GNB	0.70	0.68	0.90	0.76
SM-PF	RF	0.73	0.71	0.93	0.76
SM-PF	AdaBoost	0.80	0.83	0.88	0.82
SM-R	SVC	0.84	0.84	0.93	0.86
SM-R	GNB	0.74	0.68	1.00	0.80
SM-R	RF	0.71	0.81	0.69	0.69
SM-R	AdaBoost	0.75	0.83	0.93	0.88
SM-P	SVC	0.87	0.86	0.89	0.87
SM-P	GNB	0.67	0.73	0.71	0.66
SM-P	RF	0.74	0.79	0.66	0.64
SM-P	AdaBoost	0.78	0.86	0.79	0.78
CONTROL	SVC	0.67	0.69	0.84	0.72
CONTROL	GNB	0.71	0.72	0.89	0.74
CONTROL	RF	0.61	0.69	0.57	0.57
CONTROL	AdaBoost	0.66	0.63	0.91	0.73

Figure 4.2b presents the ROC curves for the models trained with SVC algorithm and different descriptors. The results revealed that both SVC/SM-MF and SVC/SM-P models produced high ROC AUC scores of 0.98, indicating that these descriptors, when combined with the SVC algorithm, lead to excellent classification performance. Specifically, this suggests that there was a 98% chance that the models could correctly distinguish a "durable" elastomer from a "brittle" one based on the ordering of the predicted ratings. The SVC/SM-PF model achieved a moderately high ROC AUC score of 0.94, whereas the SVC/SM-R descriptor had a lower score of 0.86. It is noteworthy that, although the SVC/SM-PF model scores lower in accuracy compared to other SM descriptors, it achieves a relatively high AUC. This suggests that although the SVC/SM-PF model might make more false predictions at the current threshold, it has the commendable capability to distinguish between classes across different thresholds. The SVC/Control model had the lowest ROC AUC score of 0.68, indicating poor classification performance. **Figure 4.2c** further evaluated the ROC AUC scores of the SM-MF trained models using different algorithms. The SVC/SM-MF model achieved the highest ROC AUC score of 0.98, closely followed by the AdaBoost/SM-MF model with a score of 0.97. The RF/SM-MF model had a moderately high score of 0.90, while the GNB/SM-MF model scored the lowest at 0.79.

Among all models, the SVC/SM-MF model achieved the highest accuracy score of 0.91, indicating its outstanding predictive capabilities compared to the other models. This model also achieved a remarkable ROC AUC score of 0.98, further validating its superior performance. The high performance demonstrates that this model has excellent discriminatory ability, effectively distinguishing between “durable” and “brittle” elastomers.

These results emphasize the significance of descriptor selection in ML models. The overall performance of the SM descriptors (SM-MF, SM-PF, SM-R, and SM-P) was notably higher than that of the Control descriptors in terms of both accuracy and ROC AUC scores. This demonstrates the effectiveness of SM descriptors in capturing crucial molecular features impacting mechanical properties and highlights the limitations of the Control descriptor in representing the relevant molecular features. The superior performance of the

SM-MF descriptor can be attributed to its intrinsic ability to characterize the chemical environment surrounding an atom. The intermolecular forces play an important role in determining toughness. For instance, weaker forces like van der Waals interactions might permit greater deformation, but can be easily overcome, resulting in material fractures. On the other hand, more robust forces such as hydrogen bonds serve as sacrificial interactions for dissipating energy and thereby enhance toughness.[50] Because SM-MF provides a rich, detailed representation of the molecular environment, it can capture these subtle interactions. On the contrary, while the SM-PF descriptor also attempts to portray specific atomic surroundings, it does so through predefined structural patterns. This can limit its ability to capture the nuances of atomic interactions in the same depth as the SM-MF. The SM-R and SM-P descriptors, which are predominantly based on distinct molecular physical and chemical properties, might not capture the entirety of a molecule's atomic environment. While they provide valuable insights into specific attributes, they could not offer a comprehensive view to understand the subtle interactions.

Among the algorithms, the SVC algorithm consistently produced high accuracy and ROC AUC scores, indicating its suitability for these classification tasks with a small data. The superiority of the SVC algorithm in our dataset can be attributed to its foundational principle, which focus on maximizing the inter-class decision boundary or margin. Its inherent robustness against data noise, coupled with a built-in regularization parameter that harmoniously balances margin maximization and classification error minimization, renders it exceptionally effective.[51]

To determine whether the exceptional performance of the SM descriptor arises from its hierarchical structure, its performance was compared with that of structure-based local-level (SL) descriptors and structure-based polymer-level (SP) descriptors. The SL descriptors comprise SDR descriptors and SS mass descriptors, while the SP descriptors include block ratios and polymer mass descriptors. **Figure 4.3a** presents the evaluation metrics, including accuracy, precision, recall, and F1 score, for the models trained with SVC algorithm and these descriptors. The SVC/SM model achieved scores greater than 0.9 across all metrics, while the SVC/SL and SVC/SP models scored below 0.8 and 0.75,

respectively, for all evaluation metrics. **Figure 4.3b** presents the ROC curves for these three models, and we notice that the AUC scores for the SL-trained (0.74) and SP-trained (0.60) models are significantly lower than that of the SM-trained model (0.98).

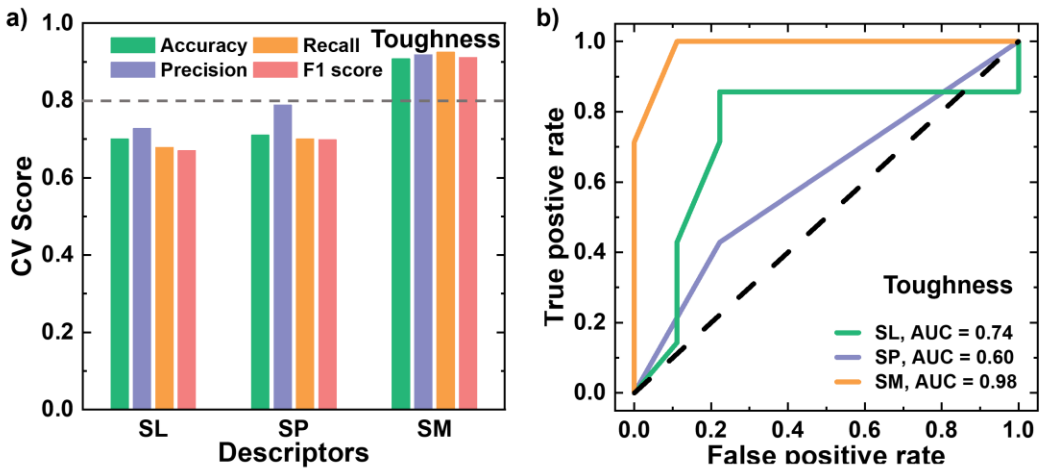


Figure 4.3 Performance comparison for different level descriptor. (a) Evaluation metrics, including accuracy, precision, recall, and F1 score, for models trained using different level descriptor: SL, SP, and SM, with the SVC algorithm. The SM descriptor employed here is SM-MF descriptor. The grey dashed line represents the accuracy of 0.8. (b) ROC curves for the models trained using the different level descriptors.

These results reveals that the SM descriptors demonstrate remarkable superiority across all evaluation metrics when compared to the SL and SP descriptors. Our observation suggests that the hierarchically organized structure of the SM descriptors plays a crucial role in their outstanding performance. The SL descriptor, focusing on local-level descriptors, achieves reasonable performance but is limited by its inability to capture the full complexity of the elastomer system. Conversely, the SP descriptor, emphasizing polymer-level descriptors, performs relatively poorly. By integrating both local-level and polymer-level features, the SM descriptors provide a more comprehensive representation of the structural information of the elastomers. This combination results in a richer and more informative feature set, leading to enhanced performance across all evaluation metrics. The findings underscore

the significance of the hierarchical organization of descriptors in optimizing the predictive capabilities of models for complex elastomer systems.

4.3.2 High-throughput Screening

The most accurate model in our study, the SVC/SM-MF model, was applied to the candidate dataset to establish an HTS pipeline for identifying the potential "durable" elastomers with the desired toughness. **Figure 4.4** illustrates the probability of candidates being classified as "durable" elastomers, allowing us to examine the influence of both local HS/SS structures and global polymer structure on toughness. In the heatmap, each 2*2 grid represents a unique combination of modified units, with individual squares within the grid corresponding to specific combinations of SS mass and polymer mass. Group 1, located on the diagonal of the heatmap, incorporates a single type of modified unit into PDMS, while Group 2, located off the diagonal, introduces two types of modified units into PDMS. From the 460 candidates evaluated, the top 20 candidates with the highest probability of being classified as "durable" elastomers were identified, with their details available in the **Table 4.2**.

Table 4.2 Top-20 candidates with the highest probability of being classified as "durable" elastomers

Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
m4	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)ccn3)c2)CC(C)(C)C1	S0	0.5	0.5	P1
m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
m4	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)cc1	S0	0.5	0.5	P1

Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
m4	m15	*CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC1CCC(CC2CCC(NC(=O)NNC(=O)C(=O)NCCCO[Si](C)(C)CCCNC(=O)Nc3ccc(Cc4ccc(NC(=O)N*)cc4)cc3)CC2)CC1	S0	0.5	0.5	P1
m4	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)cc1	S0	0.5	0.5	P1
m15	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC2CCC(CC3CCC(NC(=O)NNC(=O)C(=O)N*)CC3)CC2)cc1	S0	0.5	0.5	P1
m1	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
m1	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1	S0	0.5	0.5	P1
m1	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)NC2CCC(CC3CCC(NC(=O)N*)CC3)CC2)cc1	S0	0.5	0.5	P1
m1	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)NC2CCC(CC3CCC(NC(=O)N*)CC3)CC2)cc1	S0	0.5	0.5	P1
m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
m5	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)ccn3)c2)CC(C)(C)C1	S0	0.5	0.5	P1
m5	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)Nc2cc(NC(=O)N*)ccc2C)cc1	S0	0.5	0.5	P1
m5	m15	*CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC1CCC(CC2CCC(NC(=O)NNC(=O)C(=O)NCCCO[Si](C)(C)CCCNC(=O)Nc3cc(NC(=O)N*)ccc3C)CC2)CC1	S0	0.5	0.5	P1

Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
m5	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)Nc2cc(NC(=O)N*)ccc2C)cc1	S0	0.5	0.5	P1
m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
m4	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
m4	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)cc1	S1	0.5	0.5	P1
m4	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)cc1	S1	0.5	0.5	P1
m15	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC2CCC(CC3CCC(NC(=O)NNC(=O)C(=O)N*)CC3)CC2)cc1	S1	0.5	0.5	P1

For a particular modified unit combination, qualitative guidance on mass selection is presented. For instance, when considering the m3-m10 combination, the highest likelihood of being classified as a "durable" elastomer is exhibited when the SS mass is within the S0 range and the polymer mass is in the P1 range. This suggests that, for this specific modified unit combination, high toughness is highly probable when the SS mass is relatively small and the polymer mass is relatively large.

Upon examining the entire heatmap, it is observed that some grids display a uniformly dark color, such as the grid corresponding to the m4-m9 combination. All squares within this grid exhibit a consistent dark shade, suggesting that this modified unit combination consistently forms "durable" elastomers across a wide mass range. This observation indicates that the local HS structure of this combination is beneficial for enhancing toughness. Specifically, the abundance of hydrogen bond donors and acceptors in m4 and

m9 promotes the formation of plentiful hydrogen bonds, which are further strengthened through cooperation with the π-π stacking interaction. Strengthening hydrogen bonds, which serve as energy-dissipating sacrificial bonds, is advantageous for enhancing toughness. In contrast, some grids feature a uniformly light color, such as the grid corresponding to the m3-m10 combination, where all squares, except one, display a consistent light shade. This suggests that, for this modified unit combination, the local HS structure has a negligible effect on enhancing toughness. Although hydrogen bonds can be form between m3 and m10, their strength is diminished by the increased mobility of the hexane chain. To verify the robustness of the prediction results, molecular dynamic (MD) simulations were conducted for these two combinations, m4-m9 and m3-m10. The simulated stress-strain curves, embedded in **Figure 4.4**, indicate that the toughness of m4-m9 combination is around 10% higher than that of m3-m10 combination. These simulation results reveals that the m4-m9 combination is better suited for forming elastomers with higher toughness than m3-m10. Such consistency further validates the effectiveness of the developed model. Taking into account the probability of being classified as "durable" elastomers for each combination, we identified the top 15 modified unit combinations demonstrating favorable properties for forming "durable" elastomers among the 115 evaluated combinations. These combinations are documented in the **Table 4.3**

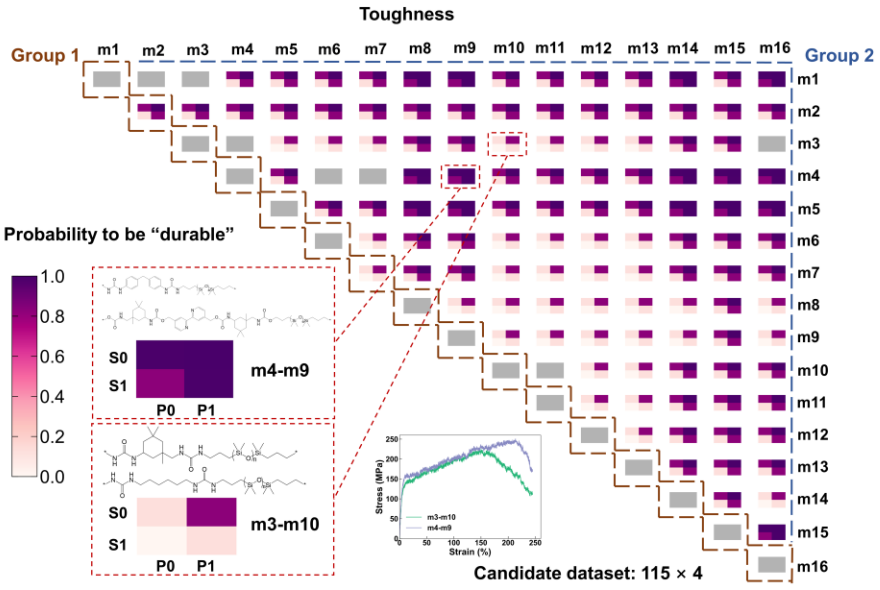


Figure 4.4 Probability of candidate dataset with 460 entries being classified as "durable" elastomers. Each 2*2 grid represents a unique combination of modified unit, with each small square within a grid representing a distinct combination of SS mass and polymer mass. The color of each square indicates the probability of the corresponding combination of SS mass, polymer mass, and modified units being classified as "durable" elastomers. The grey square indicates that the corresponding modified unit combination was included in the training dataset. The embedded image displays simulated stress-strain curves for combinations m4-m9 and m3-m10.

Table 4.3 Top-15 modified unit combinations demonstrating favorable properties for forming "durable" elastomers

Modified unit 1	Modified unit 2	SMILES of SDR
m4	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)ccn3)c2)CC(C)(C)C1
m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1
m4	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)cc1

Modified unit 1	Modified unit 2	SMILES of SDR
m4	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)cc1
m5	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)Nc2cc(NC(=O)N*)ccc2C)cc1
m15	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC2CCC(CC3CCC(NC(=O)NNC(=O)C(=O)N*)CC3)CC2)cc1
m1	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1
m1	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1
m1	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)NC2CCC(CC3CCC(NC(=O)N*)CC3)CC2)cc1
m1	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)NC2CCC(CC3CCC(NC(=O)N*)CC3)CC2)cc1
m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1
m5	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)ccn3)c2)CC(C)(C)C1
m5	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCNC(=O)Nc2cc(NC(=O)N*)ccc2C)cc1
m5	m15	*CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC1CCC(CC2CCC(NC(=O)NNC(=O)C(=O)NCCCO[Si](C)(C)CCCNC(=O)Nc3cc(NC(=O)N*)ccc3C)CC2)CC1
m4	m15	*CCCO[Si](C)(C)CCCNC(=O)C(=O)NNC(=O)NC1CCC(CC2CCC(NC(=O)NNC(=O)C(=O)NCCCO[Si](C)(C)CCCNC(=O)Nc3ccc(Cc4ccc(NC(=O)N*)cc4)cc3)CC2)CC1

4.3.3 Generality of SM Descriptors

To assess the generality of the SM descriptors, the performance of SM-trained models was evaluated in predicting other mechanical properties, including critical strain and Young's

modulus. We employed the same methodology as that used for predicting toughness to select optimal models for predicting critical strain and Young's modulus. Both optimal models for predicting these two properties use the SVC algorithm and SM-R descriptors. The five-fold CV scores of the optimal models for predicting different mechanical properties are presented in **Figure 4.5a**. The scores of these models were consistently above 0.8, regardless of the evaluation metric used. The ROC curves of the three models shown in **Figure 4.5b** also demonstrate their strong performance, with all AUCs exceeding 0.80. In particular, for the models predicting toughness and critical strain, the AUCs exceed 0.9.

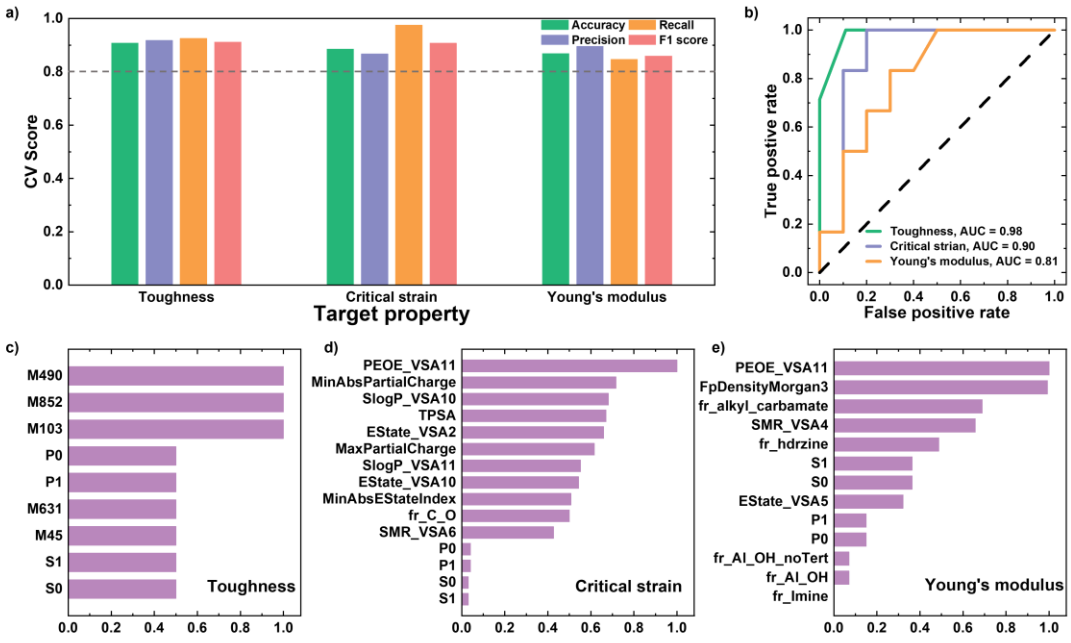


Figure 4.5 Performance and feature importance of models predicting different mechanical properties of elastomers. (a) Evaluation metrics, including accuracy, precision, recall, and F1 score, for models predicting toughness, critical strain, and Young's modulus. The grey dashed line represents the accuracy of 0.8. (b) ROC curves for the models predicting toughness, critical strain, and Young's modulus. (c) Feature importance of the model predicting toughness. (d) Feature importance of the model predicting critical strain. (e) Feature importance of the model predicting Young's modulus.

Overall, the SM-trained models demonstrated strong categorizing abilities for various mechanical properties, including toughness, critical strain, and Young's modulus. The

robust performance validates the generality and feasibility of SM descriptors in constructing effective models.

To gain deeper insight into the key features associated with various mechanical properties, the feature importance was examined for different models, as depicted in **Figure 4.5c-e**. For toughness, as illustrated in **Figure 4.5c**, the analysis highlights three Morgan Fingerprints containing -NH- groups as the most crucial features (refer to **Figure 4.6**). These fragments promote hydrogen bond formation between HS chains, introducing sacrifice bonds that improve the toughness of elastomers through energy dissipative mechanisms.[50] The SS mass and polymer mass descriptors contributed equally, suggesting that adjusting the elastomer toughness requires consideration of both SS and polymer mass. Regarding critical strain, as shown in **Figure 4.5d**, the descriptor representing the sum of surface area with a specific partial charge range (PEOE_VSA11) emerged as the most significant contributor. Apart from the descriptor representing the number of carbonyl groups (fr_C_O) and mass descriptors, the remaining descriptors were directly or indirectly related to electrostatic potential, influencing the intermolecular interactions, which subsequently determine network structures and strength of elastomers.[52] Our analysis found mass descriptors to be least important, suggesting that the focus should be on the HS structure rather than mass when adjusting critical strain. For Young's modulus, as illustrated in **Figure 4.5e**, the descriptors representing the sum of surface area with specific partial charge range (PEOE_VSA11) and molecular similarity (FpDensityMorgan3) were identified as the most important contributors. Besides mass descriptors, the other descriptors pertain to electrostatic potential, which govern intermolecular interactions and ultimately influence network structures and elastomers' Young's modulus . The SS mass descriptors ranked higher than the polymer mass descriptors in the model predicting Young's modulus. This can be attributed to the increased material flexibility resulting from a higher proportion of SS, which are not crosslinked with other components and can move relatively freely. Overall, local SDR descriptors consistently ranked as the most critical features across all models. This finding aligns with previous research emphasizing the significance of local HS structure in determining the mechanical properties of elastomers,

as the aggregation force of HS domains can control their crystal state and microphase separation.[53]

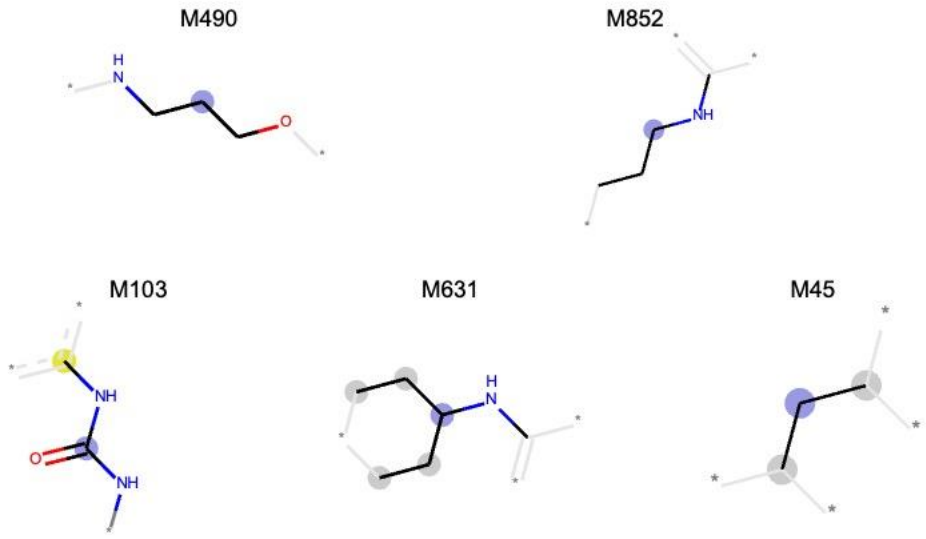


Figure 4.6 Visualization of Morgan Fingerprints

The alignment of the models with mechanical mechanisms reinforces their reliability. Moreover, the feature importance of the models provides a crucial aspect for enhancing their interpretability. By understanding which features are most significant in predicting mechanical properties, we can gain insights into the underlying mechanisms governing the materials' behaviour.

Leveraging the optimal models for predicting critical strain and Young's modulus, HTS pipelines were constructed for identifying potential elastomers with desired critical strain or Young's modulus. **Figure 4.7** displays the probability of each candidate being classified as "stretchable" or "flexible" elastomers. Out of the 460 candidates evaluated, the top 20 candidates with the highest probability for each category were identified, and the detailed results are available in the **Table 4.4**.

Table 4.4 The top 20 candidates with the highest probability of being classified as "stretchable" and "flexible" elastomers, respectively.

Target	Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
Critical Strain	m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
Critical Strain	m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
Critical Strain	m3	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
Critical Strain	m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Critical Strain	m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Critical Strain	m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P0
Critical Strain	m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P0
Critical Strain	m2	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
Critical Strain	m3	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Critical Strain	m3	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P0

Target	Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
Critical Strain	m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P0
Critical Strain	m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P0
Critical Strain	m1	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
Critical Strain	m8	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)Nc3ccc(SSc4ccc(NC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)cc4)cc3)CC(C)(C)C2)cc1	S0	0.5	0.5	P1
Critical Strain	m2	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Critical Strain	m8	m10	*CCCO[Si](C)(C)CCCNC(=O)NCCCCCCCNC(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)O*)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P1
Critical Strain	m2	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S0	0.5	0.5	P0
Critical Strain	m3	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P0
Critical Strain	m1	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Critical Strain	m3	m5	*CCCO[Si](C)(C)CCCNC(=O)Nc1cc(NC(=O)NCCCO[Si](C)(C)CCCNC(=O)NCC2(C)CC(NC(=O)N*)CC(C)(C)C2)ccc1C	S0	0.5	0.5	P1
Young's modulus	m8	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NCC5(C)CC(NC(=O)Nc6ccc(SSc7ccc(NC(=O)NC8CC(C)(C)CC(C)(CNC(=O)O*)C8)cc7)cc6)CC(C)(C)C5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1

Target	Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
Young's modulus	m2	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m9	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)OCc3ccnc(-c4cc(COC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)ccn4)c3)CC(C)(C)C2)cc1	S1	0.5	0.5	P1
Young's modulus	m3	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m9	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)OCc3ccnc(-c4cc(COC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)ccn4)c3)CC(C)(C)C2)cc1	S1	0.5	0.5	P1
Young's modulus	m1	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m5	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m3	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)NCC2(C)CC(NC(=O)N*)CC(C)(C)C2)cc1	S1	0.5	0.5	P1
Young's modulus	m4	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m2	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m8	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)Nc3ccc(SSc4ccc(NC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)cc4)cc3)CC(C)(C)C2)cc1	S1	0.5	0.5	P1
Young's modulus	m8	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NCC5(C)CC(NC(=O)Nc6ccc(SSc7ccc(NC(=O)NC8CC(C)(C)CC(C)(CNC(=O)O*)C8)cc7)cc6)CC(C)(C)C5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P0

Target	Modified unit 1	Modified unit 2	SMILES of SDR	SS mass range	Block ratio 1	Block ratio 2	Polymer mass range
Young's modulus	m2	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NC2CCC(CC3CCC(NC(=O)O*)CC3)CC2)cc1	S1	0.5	0.5	P1
Young's modulus	m9	m10	*CCCO[Si](C)(C)CCCNC(=O)NCCCCCCCNC(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)O*)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m2	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P0
Young's modulus	m9	m11	*CCCO[Si](C)(C)CCCNC(=O)C(Cc1ccccc1)NC(=O)NCCCCCCNC(=O)NC(Cc1ccccc1)C(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)O*)C4)ccn3)c2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m9	m12	*CCCO[Si](C)(C)CCCNC(=O)c1ccc(-c2ccc(C(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC3(C)CC(NC(=O)OCc4ccnc(-c5cc(COC(=O)NC6CC(C)(C)CC(C)(CNC(=O)O*)C6)ccn5)c4)CC(C)(C)C3)cn2)nc1	S1	0.5	0.5	P1
Young's modulus	m8	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)Nc3ccc(SSc4ccc(NC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)cc4)cc3)CC(C)(C)C2)cc1	S1	0.5	0.5	P1
Young's modulus	m3	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)cc3)cc2)CC(C)(C)C1	S1	0.5	0.5	P1
Young's modulus	m9	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)OCc3ccnc(-c4cc(COC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)ccn4)c3)CC(C)(C)C2)cc1	S1	0.5	0.5	P0

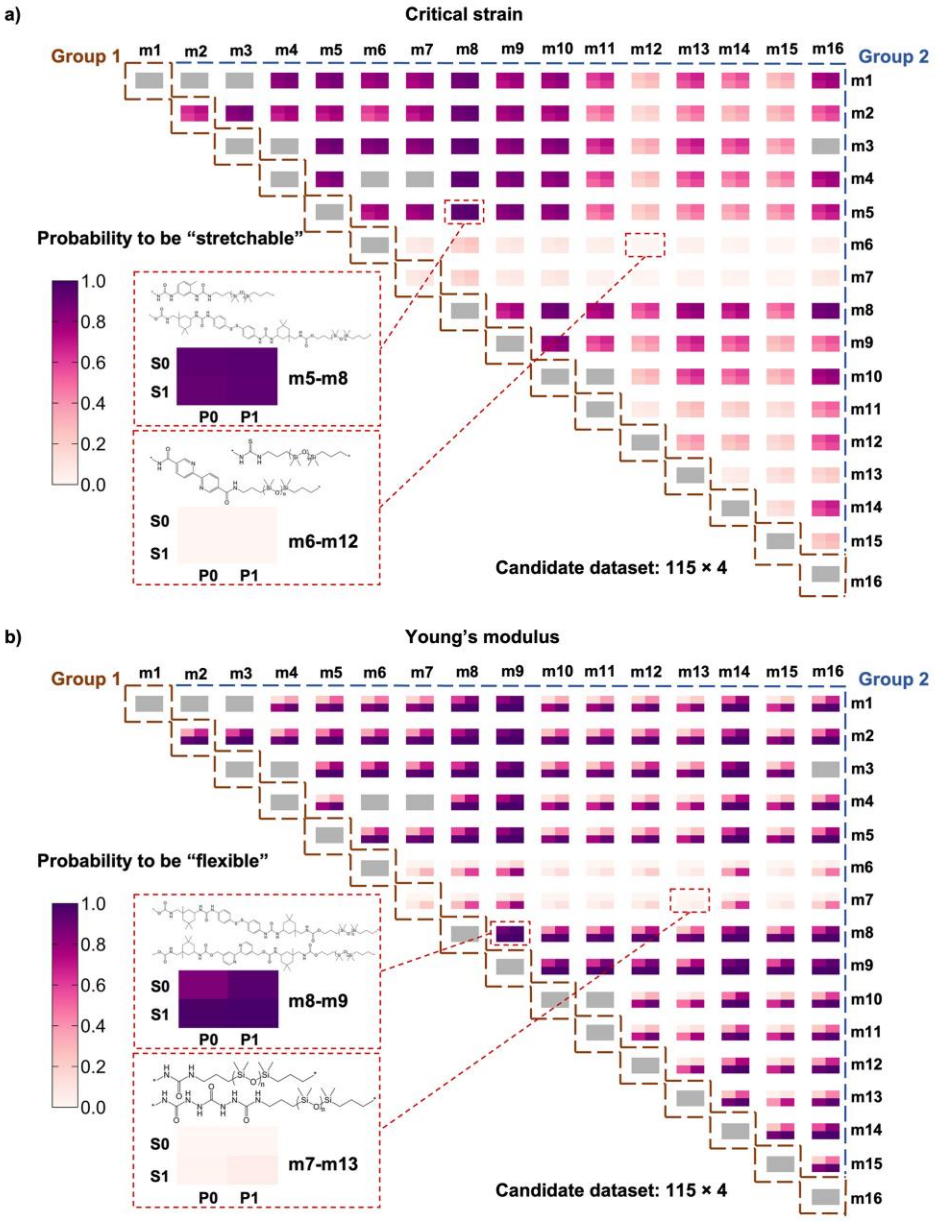


Figure 4.7 Probability of candidate dataset with 460 entries being classified as "stretchable" elastomers in terms of critical strain (a) and “flexible” elastomers in terms of Young's modulus (b). Each 2*2 grid represents a unique combination of modified unit, with each small square within a grid representing a distinct combination of SS mass and polymer mass. The color of each square indicates the probability of the corresponding combination of SS mass, polymer mass, and modified units being classified as "stretchable" elastomers or “flexible” elastomers, respectively. The grey square indicates that the corresponding modified unit combination was included in the training dataset.

Figure 4.7 provides qualitative guidance on mass selection for specific modified unit combinations by examining the corresponding grids. Furthermore, we identified the most and least favourable modified unit combinations for forming "stretchable" elastomers in terms of critical strain, which were m5-m8 combination and m6-m12 combination, respectively. Similarly, we identified the most and least favourable modified unit combinations for forming "flexible" elastomers in terms of Young's modulus, which were m8-m9 combination and m7-m13 combination, respectively. Based on our analysis of 115 modified unit combinations, we compiled a list of the top 15 combinations with favourable properties for forming "stretchable" and "flexible" elastomers, which are available in the **Table 4.5**.

Table 4.5 Top-15 modified unit combinations demonstrating favorable properties for forming "stretchable" and "flexible" elastomers, respectively.

Target	Modified unit 1	Modified unit 2	SMILES of SDR
Critical strain	m5	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)cc3)cc2)CC(C)(C)C1
Critical strain	m4	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)cc3)cc2)CC(C)(C)C1
Critical strain	m3	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)cc3)cc2)CC(C)(C)C1
Critical strain	m2	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1
Critical strain	m1	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1
Critical strain	m8	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)Nc3ccc(SSc4ccc(NC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)cc4)cc3)CC(C)(C)C2)cc1
Critical strain	m8	m10	*CCCO[Si](C)(C)CCCNC(=O)NCCCCCCCNC(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)O*)C4)cc3)cc2)CC(C)(C)C1
Critical strain	m3	m5	*CCCO[Si](C)(C)CCCNC(=O)Nc1cc(NC(=O)NCCCO[Si](C)(C)CCCNC(=O)NCC2(C)CC(NC(=O)N*)CC(C)(C)C2)ccc1C

Target	Modified unit 1	Modified unit 2	SMILES of SDR
Critical strain	m3	m6	*CCCO[Si](C)(C)CCCNC(=S)NCCCO[Si](C)(C)CCCNC(=O)NCC1(C)CC(NC(=O)N*)CC(C)(C)C1
Critical strain	m5	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)ccn3)c2)CC(C)(C)C1
Critical strain	m4	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)ccn3)c2)CC(C)(C)C1
Critical strain	m2	m3	*CCCO[Si](C)(C)CCCNC(=O)NCC1(C)CC(NC(=O)NCCCO[Si](C)(C)CCCOC(=O)NC2CCC(CC3CCC(NC(=O)O*)CC3)CC2)CC(C)(C)C1
Critical strain	m1	m5	*CCCO[Si](C)(C)CCCNC(=O)Nc1cc(NC(=O)NCCCO[Si](C)(C)CCCNC(=O)NC2CCC(CC3CCC(NC(=O)N*)CC3)CC2)ccc1C
Critical strain	m1	m4	*CCCO[Si](C)(C)CCCNC(=O)Nc1ccc(Cc2ccc(NC(=O)NCCCO[Si](C)(C)CCCNC(=O)NC3CCC(CC4CCC(NC(=O)N*)CC4)CC3)cc2)cc1
Critical strain	m4	m5	*CCCO[Si](C)(C)CCCNC(=O)Nc1cc(NC(=O)NCCCO[Si](C)(C)CCCNC(=O)Nc2ccc(Cc3ccc(NC(=O)N*)cc3)cc2)ccc1C
Young's modulus	m8	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NCC5(C)CC(NC(=O)Nc6ccc(SSc7ccc(NC(=O)NC8CC(C)(C)CC(C)(CNC(=O)O*)C8)cc7)cc6)CC(C)(C)C5)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m2	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m9	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)OCc3ccnc(-c4cc(COC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)ccn4)c3)CC(C)(C)C2)cc1
Young's modulus	m3	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NCC5(C)CC(NC(=O)N*)CC(C)(C)C5)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m9	m16	*CCCO[Si](C)(C)CCC/N=C/c1ccc(/C=N/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)OCc3ccnc(-c4cc(COC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)ccn4)c3)CC(C)(C)C2)cc1
Young's modulus	m1	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)NC5CCC(CC6CCC(NC(=O)N*)CC6)CC5)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m5	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5cc(NC(=O)N*)ccc5C)C4)ccn3)c2)CC(C)(C)C1

Target	Modified unit 1	Modified unit 2	SMILES of SDR
Young's modulus	m3	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCNC(=O)NCC2(C)CC(NC(=O)N*)CC(C)(C)C2)cc1
Young's modulus	m4	m9	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCNC(=O)Nc5ccc(Cc6ccc(NC(=O)N*)cc6)cc5)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m2	m8	*CCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)Nc2ccc(SSc3ccc(NC(=O)NC4CC(C)(C)CC(C)(CNC(=O)OCCCO[Si](C)(C)CCCOC(=O)NC5CCC(CC6CCC(NC(=O)O*)CC6)CC5)C4)cc3)cc2)CC(C)(C)C1
Young's modulus	m8	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NCC2(C)CC(NC(=O)Nc3ccc(SSc4ccc(NC(=O)NC5CC(C)(C)CC(C)(CNC(=O)O*)C5)cc4)cc3)CC(C)(C)C2)cc1
Young's modulus	m2	m14	*CCCO[Si](C)(C)CCCOCCOc1ccc(/C=N/NC(=O)N/N=C/CCCO[Si](C)(C)CCCOC(=O)NC2CCC(CC3CCC(NC(=O)O*)CC3)CC2)cc1
Young's modulus	m9	m10	*CCCO[Si](C)(C)CCCNC(=O)NCCCCCCNC(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)O*)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m9	m11	*CCCO[Si](C)(C)CCCNC(=O)C(Cc1ccccc1)NC(=O)NCCCCCCNC(=O)NC(Cc1ccccc1)C(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC1(C)CC(NC(=O)OCc2ccnc(-c3cc(COC(=O)NC4CC(C)(C)CC(C)(CNC(=O)O*)C4)ccn3)c2)CC(C)(C)C1
Young's modulus	m9	m12	*CCCO[Si](C)(C)CCCNC(=O)c1ccc(-c2ccc(C(=O)NCCCO[Si](C)(C)CCCOC(=O)NCC3(C)CC(NC(=O)OCc4ccnc(-c5cc(COC(=O)NC6CC(C)(C)CC(C)(CNC(=O)O*)C6)ccn5)c4)CC(C)(C)C3)cn2)nc1

4.3.4 Comparison of the SM Descriptor with Other Descriptor

After verifying the effectiveness and generality of our SM descriptors, we conducted a comparative analysis between our SM descriptor with four existing descriptors for elastomers. **Table 4.6** summarizes the comparison in terms of the descriptor sources, modelling approach, target properties, and their limitation in HTS.

Existing descriptors (Des-1 to Des-4) integrate molecular structure, simulation data, and/or experimental data to characterize elastomers. These descriptors enable the construction of regression models that can predict numerical values with high precision. However, the

computational simulations require specialized expertise and high-performance computing resources, while collecting experimental data is often time-consuming and costly, resulting in resource-intensive descriptors. Furthermore, certain experimental data, such as FT-IR absorbance, can only be obtained after material synthesis, making them inapplicable to unsynthesized candidate datasets. Consequently, the practicality of these descriptors in HTS applications is limited. In terms of target properties, existing descriptors focus on various combinations of mechanical properties. Des-1 estimates stress at break, critical strain, and toughness; Des-2 predicts the strain-stress curve; Des-3 is limited to predicting Young's modulus; and Des-4 predicts Young's modulus, critical strain, and tensile strength.

In contrast, our SM descriptor relies solely on molecular structure, enabling the development of classification models designed to predict categorical values. Due to their universal availability, SM descriptors are ideal for HTS applications, where a large number of potential candidates must be rapidly evaluated to identify promising materials. Our descriptor could be used to predict toughness, critical strain, and Young's modulus.

Given the trade-offs associated with each descriptor type, it is essential for researchers to carefully consider the specific objectives and constraints of their projects when selecting the most appropriate descriptors. The differences in target properties highlight the diverse goals and objectives of each study, as well as the potential for our descriptor to complement existing descriptors in predicting a broader range of mechanical properties.

Table 4.6 Comparison of the SM descriptor with other existing descriptors for elastomers

Descriptor	Source	Modelling Approach	Target Property	Limitations in HTS	Reference
SM	•Molecular structure	Classification	Toughness; Critical strain; Young's modulus	None	Our work
Des-1	•Molecular structure; •Simulation data (Monte Carlo simulations); •Experimental data (FT-IR absorbances, solubility parameter)	Regression	Stress at break; Critical strain; Toughness	Resource-intensive; data unavailable pre-synthesis	18[18]
Des-2	•Molecular structure; •Simulation data (Molecular Dynamics Simulations: COGNAC)	Regression	Strain-stress curve	Resource-intensive	19[19]
Des-3	•Molecular structure; •Simulation data (DFT and thermodynamic models); •Experimental data (FT-IR absorbances)	Regression	Young's modulus	Resource-intensive; data unavailable pre-synthesis	16[16]
Des-4	•Molecular structure; •Experimental data (Processing setting, measurements)	Regression	Young's modulus; Critical strain; Tensile strength	Resource-intensive; data unavailable pre-synthesis	15[15]

4.4 Conclusions

In this study, we developed a novel set of elastomer descriptors, termed SM descriptors, solely derived from the molecular structure of elastomers. By combining SDR descriptors, sparse descriptors based on soft SS mass, ratios of different blocks, and sparse descriptors based on polymer mass, a set of descriptors were constructed to provide a comprehensive description of both local and global structure of elastomers. Using the SM descriptors, ML models were trained to predict the toughness, critical strain, and Young's modulus of PDMS-based elastomers. Our models achieved impressive accuracy scores of 0.91, 0.89, and 0.87, respectively, which demonstrates the effectiveness and generality of the SM descriptors in capturing the important molecular features that affect mechanical properties. Based on the success of the ML models, ML-assisted HTS pipelines were constructed to rapidly screen elastomers with targeted mechanical properties. The significance of our study lies in creating a computationally efficient, universally applicable approach for

predicting the properties of elastomers. This approach significantly reduces the time and cost required for materials discovery, and offers an important contribution to the ongoing efforts to develop more efficient and accurate methods for materials discovery. In future research, the refinement of existing SM descriptors could further enhance the accuracy and applicability of SM-based models. Moreover, the application of SM descriptors to other materials systems could offer a promising approach for predicting the properties of a wide range of materials.

References

[1] Cravero, F.; Martínez, M. J.; Ponzoni, I.; Diaz, M. F. Computational modelling of mechanical properties for new polymeric materials with high molecular weight. *Chemometrics and Intelligent Laboratory Systems* **2019**, 193, 103851.

[2] Pronobis, W.; Tkatchenko, A.; Müller, K.-R. Many-body descriptors for predicting molecular properties with machine learning: Analysis of pairwise and three-body interactions in molecules. *Journal of Chemical Theory and Computation* **2018**, 14 (6), 2991-3003.

[3] Chaabene, W. B.; Flah, M.; Nehdi, M. L. Machine learning prediction of mechanical properties of concrete: Critical review. *Construction and Building Materials* **2020**, 260, 119889.

[4] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, 559 (7715), 547-555.

[5] Giles, S. A.; Sengupta, D.; Broderick, S. R.; Rajan, K. Machine-learning-based intelligent framework for discovering refractory high-entropy alloys with improved high-temperature yield strength. *npj Computational Materials* **2022**, 8 (1), 235.

[6] Shi, Z.; Yang, W.; Deng, X.; Cai, C.; Yan, Y.; Liang, H.; Liu, Z.; Qiao, Z. Machine-learning-assisted high-throughput computational screening of high performance metal–organic frameworks. *Molecular Systems Design & Engineering* **2020**, 5 (4), 725-742.

[7] Chong, Y.; Huo, Y.; Jiang, S.; Wang, X.; Zhang, B.; Liu, T.; Chen, X.; Han, T.; Smith, P. E. S.; Wang, S. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proceedings of the National Academy of Sciences* **2023**, 120 (20), e2220789120.

[8] Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO2 capture. *The Journal of Physical Chemistry Letters* **2014**, 5 (17), 3056-3060.

[9] He, B.; Chi, S.; Ye, A.; Mi, P.; Zhang, L.; Pu, B.; Zou, Z.; Ran, Y.; Zhao, Q.; Wang, D. High-throughput screening platform for solid electrolytes combining hierarchical ion-transport prediction algorithms. *Scientific Data* **2020**, 7 (1), 151.

[10] Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **2016**, 15 (10), 1120-1127.

[11] Ren, J.-C.; Zhou, J.; Butch, C. J.; Ding, Z.; Li, S.; Zhao, Y.; Liu, W. Predicting single-phase solid solutions in as-sputtered high entropy alloys: High-throughput screening with machine-learning model. *Journal of Materials Science & Technology* **2023**, 138, 70-79.

[12] Nguyen, T. N.; Nhat, T. T. P.; Takimoto, K.; Thakur, A.; Nishimura, S.; Ohyama, J.; Miyazato, I.; Takahashi, L.; Fujima, J.; Takahashi, K. High-throughput experimentation and catalyst informatics for oxidative coupling of methane. *ACS Catalysis* **2019**, 10 (2), 921-932.

[13] Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *The Journal of Physical Chemistry Letters* **2019**, 10 (15), 4401-4408.

[14] Chang, Y.-J.; Jui, C.-Y.; Lee, W.-J.; Yeh, A.-C. Prediction of the composition and hardness of high-entropy alloys by machine learning. *Jom* **2019**, 71, 3433-3442.

[15] Ding, F.; Liu, L.-Y.; Liu, T.-L.; Li, Y.-Q.; Li, J.-P.; Sun, Z.-Y. Predicting the mechanical properties of polyurethane elastomers using machine learning. *Chinese Journal of Polymer Science* **2023**, 41 (3), 422-431.

[16] Pugar, J. A.; Gang, C.; Huang, C.; Haider, K. W.; Washburn, N. R. Predicting Young's modulus of linear polyurethane and polyurethane–polyurea elastomers: Bridging length scales with physicochemical modeling and machine learning. *ACS Applied Materials & Interfaces* **2022**, 14 (14), 16568-16581.

[17] Pugar, J. A.; Childs, C. M.; Huang, C.; Haider, K. W.; Washburn, N. R. Elucidating the physicochemical basis of the glass transition temperature in linear polyurethane elastomers with machine learning. *The Journal of Physical Chemistry B* **2020**, 124 (43), 9722-9733.

[18] Menon, A.; Thompson-Colón, J. A.; Washburn, N. R. Hierarchical machine learning model for mechanical property predictions of polyurethane elastomers from small datasets. *Frontiers in Materials* **2019**, 6, 87.

[19] Doi, H.; Takahashi, K. Z.; Yasuoka, H.; Fukuda, J.-i.; Aoyagi, T. Regression analysis for predicting the elasticity of liquid crystal elastomers. *Scientific Reports* **2022**, 12 (1), 19788.

[20] Yan, H.; Dai, S.; Chen, Y.; Ding, J.; Yuan, N. A high stretchable and self–healing silicone rubber with double reversible bonds. *ChemistrySelect* **2019**, 4 (36), 10719-10725.

[21] Döhler, D.; Kang, J.; Cooper, C. B.; Tok, J. B.-H.; Rupp, H.; Binder, W. H.; Bao, Z. Tuning the self-healing response of poly (dimethylsiloxane)-based elastomers. *ACS Applied Polymer Materials* **2020**, 2 (9), 4127-4139.

[22] Yang, Z.; Li, H.; Zhang, L.; Lai, X.; Zeng, X. Highly stretchable, transparent and room-temperature self-healable polydimethylsiloxane elastomer for bending sensor. *Journal of Colloid and Interface Science* **2020**, 570, 1-10.

[23] Buckwalter, D. J.; Zhang, M.; Inglefield Jr, D. L.; Moore, R. B.; Long, T. E. Synthesis and characterization of siloxane-containing poly (urea oxamide) segmented copolymers. *Polymer* **2013**, 54 (18), 4849-4857.

[24] Zha, R. H.; de Waal, B. F.; Lutz, M.; Teunissen, A. J.; Meijer, E. End groups of functionalized siloxane oligomers direct block-copolymeric or liquid-crystalline self-assembly behavior. *Journal of the American Chemical Society* **2016**, 138 (17), 5693-5698.

[25] Ślęczkowski, M. L.; Meijer, E.; Palmans, A. R. Cooperative folding of linear poly (dimethyl siloxane) s via supramolecular interactions. *Macromolecular Rapid Communications* **2017**, 38 (24), 1700566.

[26] Colombani, O.; Barioz, C.; Bouteiller, L.; Chanéac, C.; Fompérie, L.; Lortie, F.; Montès, H. Attempt toward 1D cross-linked thermoplastic elastomers: structure and mechanical properties of a new system. *Macromolecules* **2005**, 38 (5), 1752-1759.

[27] Roy, N.; Buhler, E.; Lehn, J. M. Double dynamic self-healing polymers: supramolecular and covalent dynamic polymers based on the bis-iminocarbohydrazide motif. *Polymer International* **2014**, 63 (8), 1400-1405.

[28] Roy, N.; Buhler, E.; Lehn, J. M. The tris-urea motif and its incorporation into polydimethylsiloxane-based supramolecular materials presenting self-healing features. *Chemistry–A European Journal* **2013**, 19 (27), 8814-8820.

[29] Cui, J.; Daniel, D.; Grinthal, A.; Lin, K.; Aizenberg, J. Dynamic polymer systems with self-regulated secretion for the control of surface properties and material healing. *Nature Materials* **2015**, 14 (8), 790-795.

[30] Liu, Y.; Zhang, K.; Sun, J.; Yuan, J.; Yang, Z.; Gao, C.; Wu, Y. A type of hydrogen bond cross-linked silicone rubber with the thermal-induced self-healing properties based on the nonisocyanate reaction. *Industrial & Engineering Chemistry Research* **2019**, 58 (47), 21452-21458.

[31] Lamers, B. A.; Ślęczkowski, M. L.; Wouters, F.; Engels, T. A.; Meijer, E.; Palmans, A. R. Tuning polymer properties of non-covalent crosslinked PDMS by varying supramolecular interaction strength. *Polymer Chemistry* **2020**, 11 (16), 2847-2854.

[32] Rao, Y.-L.; Chortos, A.; Pfattner, R.; Lissel, F.; Chiu, Y.-C.; Feig, V.; Xu, J.; Kurosawa, T.; Gu, X.; Wang, C. Stretchable self-healing polymeric dielectrics cross-linked through metal–ligand coordination. *Journal of the American Chemical Society* **2016**, 138 (18), 6020-6027.

[33] Tazawa, S.; Shimojima, A.; Maeda, T.; Hotta, A. Thermoplastic Polydimethylsiloxane with L-phenylalanine-based Hydrogen-bond Networks. *Journal of Applied Polymer Science* **2018**, 135 (24), 45419.

[34] Guo, H.; Han, Y.; Zhao, W.; Yang, J.; Zhang, L. Universally autonomous self-healing elastomer with high stretchability. *Nature Communications* **2020**, 11 (1), 2037.

[35] Xu, J.; Chen, P.; Wu, J.; Hu, P.; Fu, Y.; Jiang, W.; Fu, J. Notch-insensitive, ultrastretchable, efficient self-healing supramolecular polymers constructed from multiphase active hydrogen bonds for electronic applications. *Chemistry of Materials* **2019**, 31 (19), 7951-7961.

[36] Kang, J.; Son, D.; Wang, G. J. N.; Liu, Y.; Lopez, J.; Kim, Y.; Oh, J. Y.; Katsumata, T.; Mun, J.; Lee, Y. Tough and water-insensitive self-healing elastomer for robust electronic skin. *Advanced Materials* **2018**, 30 (13), 1706846.

[37] Chen, H.; Koh, J. J.; Liu, M.; Li, P.; Fan, X.; Liu, S.; Yeo, J. C.; Tan, Y.; Tee, B. C.; He, C. Super tough and self-healable poly (dimethylsiloxane) elastomer via hydrogen bonding association and its applications as triboelectric nanogenerators. *ACS Applied Materials & Interfaces* **2020**, 12 (28), 31975-31983.

[38] Wu, X.; Wang, J.; Huang, J.; Yang, S. Robust, stretchable, and self-healable supramolecular elastomers synergistically cross-linked by hydrogen bonds and coordination bonds. *ACS Applied Materials & Interfaces* **2019**, 11 (7), 7387-7396.

[39] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, 32 (7), 1466-1474.

[40] Liu, H.; Hussain, F.; Tan, C. L.; Dash, M. Discretization: An enabling technique. *Data Mining and Knowledge Discovery* **2002**, 6, 393-423.

[41] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **2002**, 46, 389-422.

[42] Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2011**, 2 (3), 1-27.

[43] Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **1997**, 55 (1), 119-139.

[44] Breiman, L. Random forests. *Machine Learning* **2001**, 45, 5-32.

[45] Rätsch, G.; Onoda, T.; Müller, K.-R. Soft margins for AdaBoost. *Machine Learning* **2001**, 42, 287-320.

[46] Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A. J. Machine learning algorithm validation with a limited sample size. *PloS one* **2019**, 14 (11), e0224365.

[47] Byfield, R.; Weng, R.; Miller, M.; Xie, Y.; Su, J.-W.; Lin, J. Realtime classification of hand motions using electromyography collected from minimal electrodes for robotic control. *International Journal of Robotics and Control* **2021**, 3, 13.

[48] Natale, V.; Fabbri, M.; Tonetti, L.; Martoni, M. Psychometric goodness of the mini sleep questionnaire. *Psychiatry and Clinical Neurosciences* **2014**, 68 (7), 568-573.

[49] Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* **2010**, 5 (9), 1315-1316.

[50] Song, P.; Wang, H. High-performance polymeric materials through hydrogen-bond cross-linking. *Advanced Materials* **2020**, 32 (18), 1901244.

[51] Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, 20, 273-297.

[52] Sakai, T.; Matsunaga, T.; Yamamoto, Y.; Ito, C.; Yoshida, R.; Suzuki, S.; Sasaki, N.; Shibayama, M.; Chung, U.-i. Design and fabrication of a high-strength hydrogel with ideally homogeneous network structure from tetrahedron-like macromonomers. *Macromolecules* **2008**, 41 (14), 5379-5384.

[53] Kojio, K.; Nozaki, S.; Takahara, A.; Yamasaki, S. Influence of chemical structure of hard segments on physical properties of polyurethane elastomers: A review. *Journal of Polymer Research* **2020**, 27, 1-13.

Chapter 5

Deep Neural Network Assisted High-Throughput Screening for Organic Semiconductors

This chapter focus on predicting electronic properties of conjugated oligomers by employing transfer learning techniques. It gives a detailed description of how the source dataset is curated to train the base model and how the target dataset is prepared for the transfer learning model's training. The performance of both direct learning models and transfer learning models is thoroughly evaluated. Moreover, the chapter highlights the discovery of organic photovoltaic materials facilitated by high-throughput screening, employing the predictions of the transfer learning models.

5.1 Introduction

Organic semiconductors (OSCs) are increasingly becoming the focus of attention in the field of electronics, thanks to their potential use in devices such as organic field-effect transistors,[1-4] photovoltaic cells,[5-7] and light-emitting diodes.[8-10] Their desirable attributes, which include low weight, flexibility, and the simplicity of manufacturing processes, are driving this growing interest. Among OSCs, conjugated oligomers have emerged as a compelling class of materials for their adjustable electronic properties and molecular precision. These materials consist of a defined number of repeating units with extended π-conjugation, facilitating electronic interactions that confer semiconducting behavior.[11] The conjugated oligomers find a unique middle ground between small molecules and polymers. They blend the solution processability of polymers, which is ideal for printing techniques, with the reproducible device performance that stems from the precise structures of small molecules. Moreover, the properties of conjugated oligomers can be finely tuned by modifying their molecular structure, which includes variations in chain length,[12, 13] the addition of diverse side groups,[14] or the incorporation of heteroatoms.[15, 16] Consequently, the accurate prediction of these oligomers' electronic properties becomes crucial in designing next-generation semiconductors with specific functionalities to meet the increasing demand for enhanced electronic devices. However, the prediction of electronic properties for these oligomers is challenging. While traditional computational methods offer accuracy,[17] they are resource-intensive, particularly for larger oligomers, leaving a void in our capacity for quick assessment and design of new OSCs.

Machine Learning (ML), especially its recent advances, presents an opportunity for rapid prediction of conjugated oligomers' electronic properties. However, the prediction of conjugated oligomers has been a task previously known to be difficult for neural networks due to the sensitivity of size.[18] SchNet, a deep neural network (DNN), has shown remarkable performance in predicting various electronic properties of conjugated oligomers.[19] Nonetheless, its efficacy has been mostly confined to predicting the properties of specific conjugated oligomers with short chains. Besides that, DNN often struggle with limited data, unable to capture the full spectrum of features affecting the properties.[20-22]

Transfer learning, however, has emerged as a solution, allowing the prediction of longer-chain conjugated oligomers by tuning models pre-trained on short-chain data.[23] This approach adapts existing models to new tasks swiftly and accurately. However, the predictive scope of transfer learning models has been limited to predict specific types of conjugated oligomers comprising the same monomers with different configuration, with the prediction of oligomers with different monomers and degrees of polymerization still underexplored.

In this study, we aim to bridge a critical gap by employing transfer learning to predict the electronic properties of conjugated oligomers, varying in monomers and chain lengths. Utilizing pre-trained models from the PubChemQC dataset,[24] our method aims for high-precision predictions while drastically cutting down on computational resources and time. We've implemented transfer learning within Graph Neural Network (GNN) frameworks, enhancing the precision in predicting key electronic properties such as Highest Occupied Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO), and the HOMO-LUMO gap. The transfer learning approach has shown remarkable improvements in prediction accuracy, with Mean Absolute Errors (MAE) for these properties dropping from 1.34, 0.68, 0.71 to 0.74, 0.46, 0.54, respectively. Based on the transfer learning models, we have developed a high-throughput screening (HTS) pipeline, specifically tailored for identifying promising candidates for organic photovoltaic (OPV) materials. This pipeline has successfully highlighted a significant number of potential oligomers as potential materials for photovoltaic applications. To further substantiate our findings, we conducted computational simulations that corroborated the accuracy of our machine learning models. This research not only overcomes the challenges posed by limited data availability but also marks a significant stride in the discovery of new OPV materials. It underscores the pivotal role and effectiveness of ML in modern material discovery processes.

5.2 Methods

5.2.1 Dataset Preparation

The PubChemQC dataset,[24] recognized as one of the largest quantum chemistry databases, is based on first-principles methods. It encompasses an extensive collection of over three million molecules, each accompanied by comprehensive data on their fundamental electronic structures. This dataset originates from the extensive PubChem project, which compiles molecules synthesized by researchers and industry professionals across the scientific and chemical manufacturing sectors, highlighting key compounds that are crucial for the fields of chemistry and materials science. The PubChemQC dataset is renowned for its wide variety of molecular structures. The electronic configurations of these molecules have been determined using density functional theory (DFT) at the B3LYP/6-31G* level.[25-27] This approach is highly esteemed for its consistent and reliable calculations across a broad spectrum of molecules. It effectively balances computational intensity with the accuracy of results, making it a valuable method in the field of computational chemistry. While other quantum chemistry databases exist, many use lower-accuracy methods like PM6. The high precision of B3LYP/6-31G* ensures that the dataset is of superior quality, making it particularly suitable for training models aimed at high-fidelity property prediction.

For this study, we selected a specific subset of the PubChemQC dataset to serve as our source dataset. Our selection criteria were twofold: the molecules were required to have more than six double bonds, and the energy gap between their HOMO and LUMO states had to be less than 6 eV. These criteria were deliberately chosen to enhance the similarity between the source and the targeted dataset of our study, thereby supplying our base model with more relevant information from the source dataset. The size of the PubChemQC-100 dataset, with approximately 100,000 data points, is ideal for training robust pre-trained models. This volume of data is large enough to capture a wide variety of molecular features and interactions, enabling the deep learning model to learn complex patterns effectively. A well-trained model on such a dataset can generalize better to new, unseen data, which is crucial for the high-throughput screening process. This deliberate selection strategy led to the compilation of a final dataset called PubChemQC-100k, which consists of 106,429 molecules, thus optimizing the balance between computational efficiency and learning effectiveness for our analytical endeavors. Despite its large size, the PubChemQC-100

dataset is not excessively large to cause impractically long training times. The dataset size strikes a balance, being large enough to ensure robustness and diversity in the training process while remaining manageable in terms of computational resources and time required for training. This efficiency is critical for iterative model development and fine-tuning, allowing for faster experimentation and optimization cycles.

In this study, we meticulously assembled a target dataset of conjugated oligomers using the DFT method. The construction began by sourcing monomer structures from an array of peer-reviewed publications. These monomers were encoded in SMILES format,[28] a line notation for describing the structure of chemical species using short ASCII strings. The asterisk (*) sign was employed to denote the points of connection. While existing software such as PolyMaS[29] offers the convenience of transforming monomers into oligomers for specific degrees of polymerization, challenges often arise when the monomers include cyclic structures, as illustrated in **Figure 5.1**. To address this issue, we engineered a bespoke software tool utilizing the RDKit[30] cheminformatics library within the Python environment. Our custom tool deftly interprets SMILES strings with junction points and translates them into oligomers of desired polymerization degrees, effectively overcoming the hurdles associated with cyclic monomer structures. By leveraging the SMILES representation of monomers alongside our tailor-made software, we succeeded in generating oligomers across a spectrum of polymerization degrees. We incorporated 137 unique monomers into our analysis, selecting polymerization degrees that spanned from 4 to 10. The SMILES representation of these unique monomers is shown in **Table 5.1**. This rigorous process yielded an extensive collection of 959 oligomeric SMILES representations, setting the stage for subsequent quantum chemical calculations and further material property assessments.

The selection of these model molecules was guided by two primary principles: diversity and proven effectiveness in prior research. The diversity of the model molecules was a crucial factor in our selection process. We aimed to include a wide range of chemical structures and properties to ensure that our model could generalize well across different types of molecules. This diversity was achieved by selecting monomers with various

functional groups, chain lengths, and electronic properties. By doing so, we ensured that the training dataset encompassed a broad spectrum of molecular characteristics, which is essential for developing a robust and versatile predictive model. The inclusion of diverse molecules helps the model learn intricate structure-property relationships and improves its ability to predict the properties of novel molecules accurately. Another important selection criterion was the prior successful use of these monomers in identifying potential ground-state triplet polymers. This prior research provides a solid empirical foundation, indicating that these monomers are promising candidates for further investigation. The proven effectiveness of these monomers in previous work validates their relevance and increases the likelihood of achieving meaningful and actionable results in our current study.

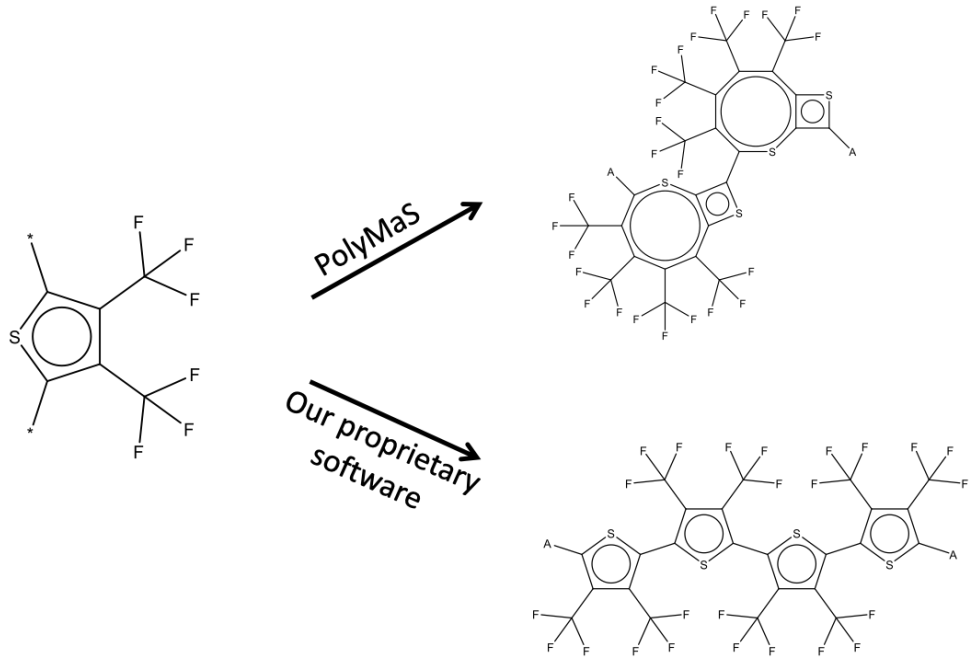


Figure 5.1 The comparison between PolyMaS software and our proprietary software

After generating the oligomeric SMILES representations, we progressed to convert them into three-dimensional (3D) configuration. This crucial step was performed using the Open Babel[31] software with the "-gen3d" command-line option, which is adept at translating chemical string representations into spatial structures. Open Babel initiated the 3D

configuration by applying structural rules and predefined ring templates, leading to a preliminary geometric model. Then this model underwent an optimization process governed by the MMFF94 force field,[32] a method used to refine the geometry of molecular models by minimizing the energy according to molecular mechanics principles. The refined 3D conformations thus provided a reliable foundation for the subsequent quantum chemical analysis, which was undertaken using the established B3LYP/6-31G* level of theory. This level, consistent with the source dataset, facilitates the balance between computational demand and the accuracy of the results. To precisely evaluate the electronic properties of the molecules, the ORCA computational software suite was employed.[33] Using the DFT method, we determined the key electronic parameters: the HOMO, LUMO, and the HOMO-LUMO gap.

Despite our initial target of 959 molecules, the final count of data points available for analysis was 610. This discrepancy arose due to non-convergence issues encountered with some molecules during the calculation process. As a result, the dataset, denoted as CO-610, comprises 610 molecular entries with complete and converged electronic property data.

Table 5.1 SMILES representation of the monomers comprising oligomers in CO-610

ID	SMILES
0	*c1sc(CC)c(*)c1CC
1	*/C=C/c1ccc2c(c1*)C(=O)OC2
2	*/C=C/c1sc(*)c2c1CCC[C@H]2C
3	*/C=C/c1sc(*)c2c1CCC[C@H]2C=C
4	*/C=C/c1ccc(*)c2nonc12
5	*c1cc2[nH]c3c(*)c(O)sc3c2s1
6	*/C=C/C1=CC=C(*)S1(=O)=O
7	*/C=C/c1c(OC)sc(C#N)c1*
8	*/C=C/N(CC)c1cscc1N(*)CC
9	*C1=C2C(=O)NC(*)=C2C(=O)N1
10	*c1c(*)c2c(c3c1CCS3)SCC2
11	*c1sc(*)c2cnncc12
12	*c1cc(C)c(*)s1
13	*C1=c2ccccc2=C(*)C1=O
14	*/C=C/c1cc(C(*)=C)sc1C
15	*/C=C/c1csc(C(C)=O)c1*

ID	SMILES
16	*c1sc(*)c2cc(F)c(F)cc12
17	*/C=C/c1csc2c(CC)c3scc(*)c3c(CC)c12
18	*c1sc2nc3c(c(CC)c2c1*)CCS3
19	*/C=C/c1[nH]c(*)cc1C=C
20	*c1oc(*)c(F)c1F
21	*c1sc(*)c2c1SCCCS2
22	*/C=C/c1csc(C(=O)C(F)(F)F)c1*
23	*/C=C/C1=C(*)C=CC1=O
24	*/C=C/c1sc(*)c2cc(S)ccc12
25	*/C=C/c1ccc(*)c2cccnc12
26	*c1ccc(*)cc1
27	*c1sc(*)c(OC)c1N
28	*c1sc(*)c2cc(C(C)=O)ccc12
29	*c1c2cccoc-2c(*)c1C
30	*C1=NC2=CC(*)=NC2=C1
31	*C#Cc1ccc(*)s1
32	*C=C*
33	*c1cc2[nH]c3cc(*)sc3c2s1
34	*/C=C/Nc1c(OC)csc1*
35	*c1sc(*)c2c1CC[C@@H](S)C2
36	*/C=C/c1sc(*)c2c1CCC[C@H]2C(=O)OC
37	*/C=C/c1sc(*)cc1C(=O)C(F)(F)F
38	*/N=N/C(*)=C
39	*/C=C/c1sc(*)c2ocnc12
40	*C#Cc1ccc(C#C*)s1
41	*/C=C/c1cc(*)sc1O
42	*C1=CC(=S)N(*)C1=S
43	*c1sc(*)c2c(C#N)ccc(C#N)c12
44	*c1sc(*)c2c1OCCCO2
45	*c1sc(*)c2c1C[C@H](F)[C@@H](F)C2
46	*/C=C/c1sc(*)c2cc(OC)c(OC)cc12
47	*c1cnc(*)c2nccnc12
48	*/C=C/c1ccc2c(c1)C(CC)(CC)c1cc(*)ccc1-2
49	*c1sc(*)c2c(C)cccc12
50	*/C=C/c1cc(C(C)=O)c(*)o1
51	*c1nc[nH]c1*
52	*/C=C/c1ncc(*)o1
53	*c1sc(*)c(C(F)(F)F)c1C(F)(F)F
54	*/C=C/c1sc2c(c1*)CC2
55	*c1c2c(=O)oc(=O)c2c(*)c2c(=O)sc(=O)c12

ID	SMILES
56	*/C=C/C1=C(*)C(=O)NC1=O
57	*/C=C/c1sc(*)c2c(C(=O)C(F)(F)F)cccc12
58	*/C=C/c1sc(*)c2c1C[C@H](OC)[C@@H](OC)C2
59	*/C=C/c1csc(/C=C/*)c1
60	*C#Cc1ccc(*)[nH]1
61	*c1sc(*)c2nc(CC)cnc12
62	*c1sc(*)c(C#N)c1N
63	*c1csc2nc3scc(*)c3c(CC)c12
64	*c1sc(*)c2c1SCS2
65	*/C=C/c1c(OC)sc(OC)c1*
66	*c1cc(OCC)c(OCC)c(OCC)c1*
67	*c1sc(*)c2c1CCC[C@H]2O
68	*/C=C/c1sc(*)c2cc(C=C)c(C)cc12
69	*/C=C/c1sc(*)c2c1OCC(=O)CS2
70	*C1=S(OC)C(OC)=Cc2csc(*)c21
71	*c1sc(*)c2sc(C(=O)OCC)cc12
72	*/C=C/c1c(CC)sc(*)c1CC
73	*/C=C/c1sc(*)c2c1SSCO2
74	*/C=C/c1sc(*)cc1C=O
75	*/C=C/c1cccc(*)c1CC
76	*/C=C/N1C=C2C(=O)SC(*)=C2C1=O
77	*c1ccc(*)c2nn(CC)nc12
78	*/C=C/c1c(*)sc(OC)c1C(C)=O
79	*C1=C(*)C(=O)C=C1
80	*NC1=C/C(=C2\C=CC(N)=C2)C=C1*
81	*/C=C/C1=CC(=S)N(*)C1=S
82	*c1cc2oc(*)c(C(N)=O)c2o1
83	*/C=C/c1cc2sc(*)cc2s1
84	*c1sc(*)c2c(C(=O)OC)cccc12
85	*/C=C/c1cc(C(=O)C(F)(F)F)c(*)o1
86	*c1sc(*)c2c1C(=O)c1ccccc1C2=O
87	*c1sc2c(c1*)C(CC)(CC)[C@@H]1CCS[C@@H]21
88	*/C=C/c1sc(*)cc1C#N
89	*/C=C/c1sc(*)c2ncoc12
90	*c1sc(*)c2c1C(=O)CC2=O
91	*/C(C)=C\c1c(CC)cc(*)cc1CC
92	*/C=C/c1sc(*)c(C)c1C
93	*c1cc(*)c(OCC)cc1OCC
94	*/C=C/c1cc(C#N)c(*)o1
95	*/C=C/c1sc(*)c2c1OCCO2

ID	SMILES
96	*c1cc2cc3occc3c(*)c2s1
97	*C1=c2ccccc2=C(*)C1=C
98	*/C=C/c1sc(*)c2c1SCC(=O)CS2
99	*c1sc(*)c2c1NCO2
100	*c1cc2nccnc2cc1*
101	*c1sc(*)c2c(F)c(F)c(F)c(F)c12
102	*c1sc(*)c2cc(S)c(O)cc12
103	*c1c(C#N)sc(OC)c1*
104	*c1c(OC)sc(OC)c1*
105	*/C=C/c1ccc(*)c2nsnc12
106	*/C=C/c1oc2cc(*)oc2c1C(N)=O
107	*/C=C/c1sc(*)c2c1[C@H](O)CCC2
108	*/C=N/Nc1ccc(*)cc1
109	*c1sc(*)c2c(F)c(C(=O)CC)sc12
110	*C1=C2C(=O)OC(*)=C2C(=O)O1
111	*/C=C/Sc1cscc1S*
112	*c1sc(*)c2c1C(=O)NC2=O
113	*/C=C/c1sc(*)c2sc(=O)sc12
114	*c1sc(*)c2c(S)ccc(O)c12
115	*/C=C/c1cc(C)cc(*)c1C
116	*/C=C/c1sc(*)c2c1[C@H](O)CC[C@H]2C(=O)O
117	*/C=C/c1oc(*)c(OC)c1C#N
118	*c1c(C#N)sc(C(F)(F)F)c1*
119	*/C=C/c1cc(OCC)c(OCC)c(OCC)c1*
120	*/C=C/c1c2cc(*)scc-2c2sccc12
121	*c1sc(*)c2scnc12
122	*/C=C/C1=Cc2c(csc2*)S1(=O)=O
123	*/C=C/c1sc(*)c2c(C#N)cccc12
124	*c1cc2c(s1)-c1sc(*)cc1C2
125	*/C=C/c1cccc(F)c1*
126	*/C=C/c1sc(*)c2c1CC(=O)C(=O)C2
127	*/C=C/c1ccc(*)cc1
128	*/C=C/c1sc(*)c2sc(=O)c(=O)sc12
129	*c1sc(*)c2c1OCC(=O)CO2
130	*C1=C/C(=C2/C=C(*)C=C2C#N)C(C#N)=C1
131	*/C=C/C1=C(*)c2cccc3cccc1c23
132	*/C=C/c1c(C#N)sc(C(F)(F)F)c1*
133	*/C=C/c1[nH]c(*)c2nccnc12
134	*c1ccc2c(c1)C(C)(C)c1cc(*)ccc1-2
135	*/C=C/N1CCN(C(=O)O)c2csc(*)c21

ID	SMILES
136	*c1sc(*)c2c1OCCO2

To this point in our research, we have successfully established two distinct datasets: the source dataset, known as PubChemQC-100k, and the target dataset, labeled CO-610. For both datasets, the inputs consist of 3D molecular configurations, which are intricately detailed representations of each molecule's spatial arrangement. The outputs are critical electronic properties—specifically, HOMO, LUMO, and the HOMO-LUMO gap. These properties are pivotal for understanding the electronic behaviour of the molecules and predicting their potential applications in various fields, including materials science and photovoltaic device engineering.

5.2.2 Model Development

The methodological framework of our study is characterized by the strategic application of transfer learning to construct robust predictive models, a process depicted in **Figure 5.2**. This transfer learning paradigm is meticulously structured into three stages: pre-training, model adaptation, and fine-tuning, each critical to the eventual success of the predictive analysis.

In the pre-training phase, a ML model, specifically a GNN model known as SchNet,[34] is trained on the extensive source dataset, PubChemQC-100k. This diverse dataset acts as a rich repository of molecular information, allowing the model to discern intricate patterns and correlations amongst a multitude of material properties. The pre-training is designed to impart the model with a versatile understanding of material behaviours and interactions. SchNet is selected for its demonstrated efficacy in capturing a spectrum of electronic properties across various materials, bypassing the need for handcrafted descriptors. This pre-training process involves iteratively refining the model over 1000 epochs to ensure a comprehensive learning experience.

In the subsequent stage of our methodology, the model undergoes a critical customization process to align with the unique challenges posed by our materials science inquiry. This adaptation is accomplished by the incorporation of a newly designed layer—an interaction block—into the existing model structure. This additional interaction block is initialized with random parameters, serving as a fresh component within the model to capture the interactions specific to the conjugated oligomers under study. Crucially, while this new block is allowed to evolve during training, the parameters of the pre-existing layers leading up to the interaction block are held constant, or 'frozen'. This approach ensures that the foundational knowledge previously acquired during the pre-training phase is preserved and not overwritten during the subsequent learning process. By implementing such strategic modifications, we effectively transition the pre-trained model into an adapted model, primed for the targeted task of predicting the electronic properties of conjugated oligomers systems. The resulting adapted model is thus a sophisticated analytical tool, tailored to the exigencies of the specialized domain of materials science.

The adapted model is then further trained on the target dataset, the CO-610 data set. This stage involves fine-tuning the model's parameters to optimize its performance for the specific task: to predict the electronic properties of conjugated oligomers. Given that the model has already gained significant knowledge from the pre-training stage, it can perform more efficiently on the target dataset, requiring fewer data and resources to achieve robust results. The weights of SchNet are then fine-tuned for up to 3000 epochs using data from the CO-610. This fine-tuning process is crucial for calibrating the model's parameters to enhance its predictive accuracy for this particular class of materials.

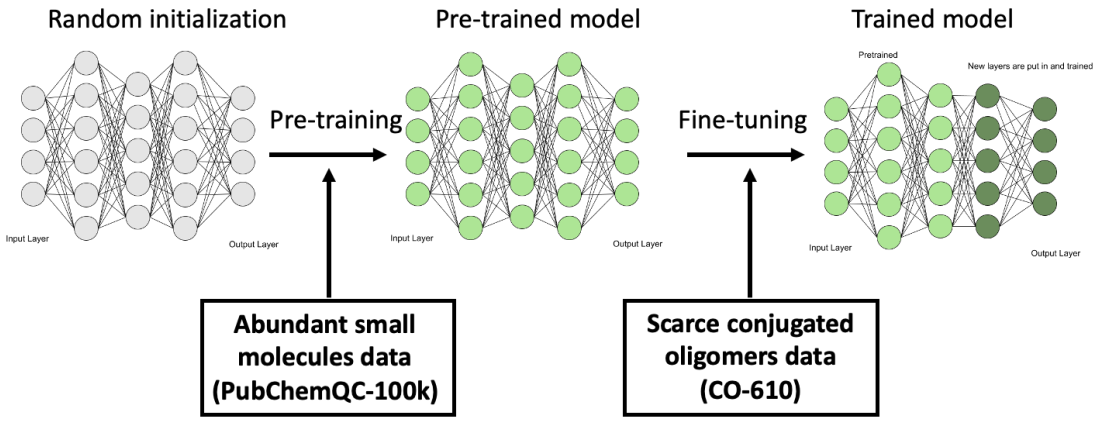


Figure 5.2 Schematic of the transfer learning protocol used in this work

A key feature in the prediction of electronic properties was the establishment of a cut-off distance of 5 Å for the models, ensuring the consideration of interactions within this specified range. The embedding dimension within the SchNet architecture was configured to 128, a setting that allows for a robust representation of molecular features across the network's seven interaction blocks. The batch size for the training process was standardized at 64, a size that balances computational load and learning stability. During the optimization of the pre-trained model, we utilized the Adam optimization algorithm with an initial learning rate of 0.0001. This rate was selected to facilitate a gradual and stable convergence of the model's weights during the learning process. For the target model, which was fine-tuned to our specific dataset of conjugated oligomers, we opted for a reduced learning rate of 0.00001 to refine the pre-learned weights without drastic alterations. During the final phase of model training, a subset of 400 data points, representing the conjugated molecules, is utilized for training purposes. This is supplemented by a validation set comprising 100 data points, while the evaluation of the model's predictive capacity is conducted on a test set of 110 data points. This methodology allows for a comprehensive assessment of the model's performance.

To confirm the enhancements attributed to the transfer learning approach, a comparative analysis is performed with SchNet models trained exclusively on the 400 data points without the benefit of transfer learning. This comparative training extends up to 4000 epochs to ensure model robustness. These models, referred to as SchNet-D, serve as a

benchmark against the transfer learning-augmented models, designated as SchNet-T. The comparison between SchNet-D and SchNet-T models provides insights into the efficacy of transfer learning in improving model performance for the prediction of electronic properties in conjugated molecular systems.

5.2.3 High-throughput Screening

In addition to the development of models, our methodology encompassed the creation of a candidate dataset aimed at identifying potential target materials with the desired electronic properties for photovoltaic applications. This candidate dataset was generated by synthesizing oligomers with polymerization degrees ranging from 4 to 10, using monomers referenced from previously published literature. The 3D geometries of these oligomers were meticulously optimized using the Open Babel software, employing the MMFF94 force field to ensure accuracy in the subsequent computational predictions.

The optimized SchNet-T models were then deployed to predict the HOMO, LUMO, and HOMO-LUMO gap of these oligomers within the candidate dataset. The photovoltaic materials chosen for this study had to fulfil rigorous specifications, including HOMO levels ranging from -6.5 to -4.9 eV, LUMO levels between -3.0 to -4.5 eV, and a band gap of 1.1 to 2.0 eV. These criteria are essential to guarantee efficient absorption of visible light, a key factor in achieving high photovoltaic efficiency.

To rigorously verify the robustness of our predictive model and the reliability of the selected conjugated oligomers, we conducted computational simulations to determine their electronic properties. These simulations were performed using the ORCA computational software suite, which is widely respected for its accurate electronic structure calculations. The comprehensive comparison of our model's predictions with the calculated electronic properties of the oligomers confirmed the reliability of our selection process. This validation step was crucial in demonstrating the efficacy of our HTS method, which significantly reduces computational costs and accelerates the discovery of promising materials for photovoltaic technology.

5.3 Results and Discussions

5.3.1 Analysis of the Dataset

Figure 5.3 presents a comparative analysis highlighting the differences in molecular size distributions and the chemical diversity among the datasets. For the PubChemQC-100k dataset, the size diversity is evident, with a histogram that peaks sharply for molecules with fewer atoms, suggesting a large number of small molecules within this dataset. This indicates that the majority of the PubChemQC-100k dataset is composed of relatively simple molecules, with the frequency of molecules decreasing as the number of atoms increases, highlighting a skew towards smaller molecular structures. For the CO-610 dataset, the size distribution is more expansive, with a wider spread of atom counts. This suggests that the dataset contains a more varied range of molecular sizes, which is essential for studying the effects of molecular size on electronic properties, particularly in the context of conjugated oligomers where the length of the polymer chain can significantly influence electronic behaviour.

Moreover, the PubChemQC-100k dataset showcases a more diverse chemical composition, with a wider range of elements such as silicon (Si), phosphorus (P), and chlorine (Cl). This diversity is indicative of the broad scope of the PubChemQC-100k database, capturing a vast array of chemical space which is beneficial for ML models that rely on a diverse training set to improve prediction accuracy. In contrast, the CO-610 dataset demonstrates a chemical diversity that is more restricted, focusing on elements predominantly found in organic electronic materials, such as carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S). This deliberate restriction aims to concentrate on the elements most prevalent and significant in organic electronic materials.

The comparative analysis between the two datasets depicted in these histograms underscores the broader generalizability of the PubChemQC-100k dataset versus the focused specificity of the CO-610 dataset.

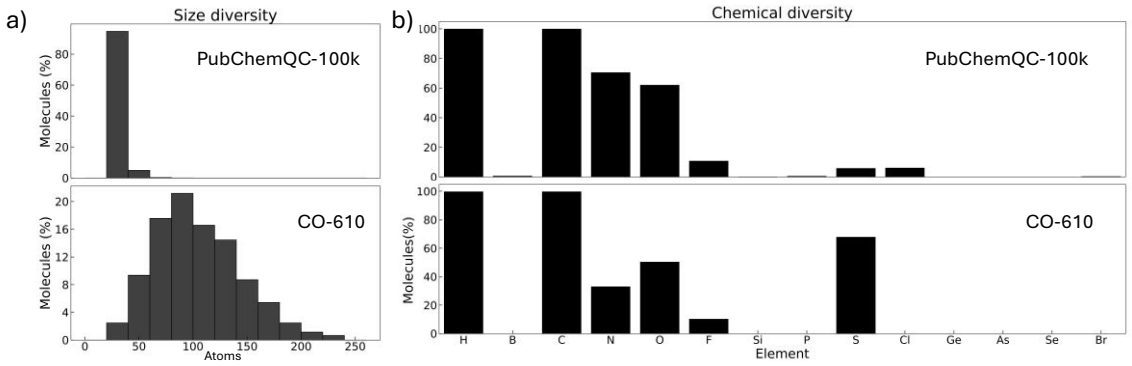


Figure 5.3 Comparative analysis of the PubChemQC-100k and CO-610 datasets. (a) The size diversity within each dataset. (b) The chemical diversity present in the datasets.

Additionally, we closely examined the size sensitivity of conjugated oligomers within the OG-610 dataset, focusing on how their electronic properties evolve with varying degrees of polymerization. **Figure 5.4** illustrate a systematic investigation into the electronic properties of a series of oligomers, each column representing oligomers derived from a unique monomer but at varying degrees of polymerization, ranging from 4 to 10. The plots use the oligomers with a polymerization degree of 7 as a reference point to anchor the comparison. In the HOMO deviation plot, the data points form a gradient pattern as the polymerization degree shifts from the reference, illustrating a consistent alteration in the HOMO energy levels relative to this central degree of polymerization. The trend suggests that as the number of monomeric units changes, there is a significant impact on the HOMO energy, which is a crucial factor in determining a material's electronic properties. The LUMO deviation plot similarly displays a variance in energy levels as the polymerization degree diverges from the reference value of 7. Each column reflects the LUMO energy shift for oligomers with a different monomeric composition, underlining the influence of both monomers and polymerization degree on electronic structure. The band gap deviation plot reinforces the relationship between polymerization degree and electronic property modulation. It shows a clear reduction in the energy gap with the increase of polymerization degree. This visual representation captures the essence of conjugation effects, where the length of the oligomer chain directly affects the electron delocalization and hence the optical and electronic behavior of the material. Collectively, these plots provide a

comprehensive overview of how varying both the monomeric unit and the degree of polymerization can fine-tune the electronic characteristics of oligomers, offering valuable insights for the design of tailored materials for advanced electronic and photonic applications.

The intricacies of size sensitivity in conjugated oligomers present a formidable obstacle for ML models aimed at predicting their electronic properties. The complexity arises from several factors: as oligomers increase in size, the conformational space grows exponentially, resulting in a myriad of possible structural arrangements. Each arrangement can have a profound impact on the electronic properties, such as HOMO and LUMO levels. This is due to the fact that with larger oligomer chains, more complex interactions come into play, including variations in electron delocalization, which can significantly influence their electronic behavior. Furthermore, as the oligomers grow, phenomena such as quantum confinement and edge effects can introduce additional variations in the electronic structure. These effects often lead to non-linear and unpredictable changes in the electronic properties, making it a challenge to model these behaviors accurately.

For ML algorithms to successfully predict the electronic properties of these materials, they must be exposed to and learn from datasets that are rich and diverse enough to capture the full breadth of these size-dependent effects. Training models on such datasets ensures that they can account for the intricate patterns of electron behavior and the various factors influencing them. Only then can the algorithms be expected to deliver reliable predictions for the electronic properties of conjugated oligomers, spanning the entire range of molecular sizes encountered in practical applications.

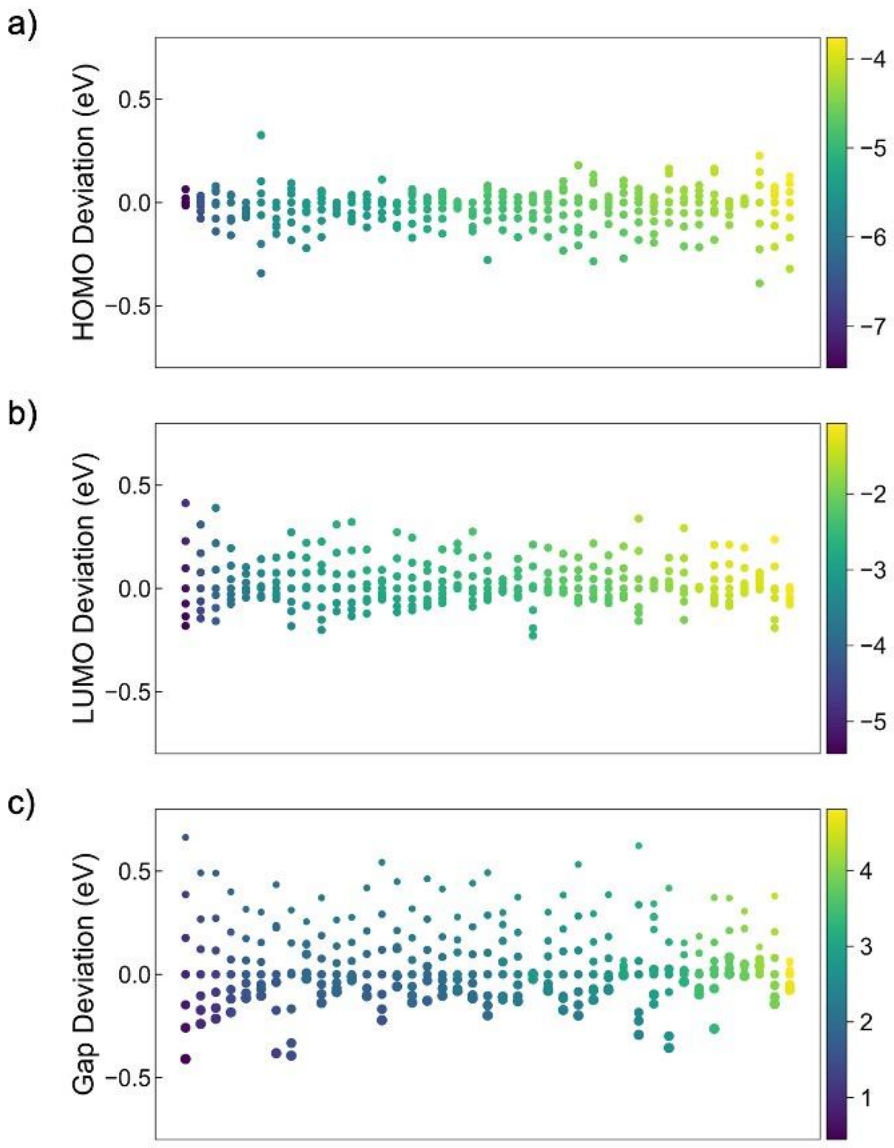


Figure 5.4 (a) HOMO Deviation: Displays oligomers of different polymerization degrees (4-10) with the same monomer, benchmarked to degree 7. Color indicates actual HOMO values. (b) LUMO Deviation: Columns represent oligomers with degrees of polymerization (4-10), using degree 7 as reference. Color shows actual LUMO values. (c) Gap Deviation: Shows oligomers across polymerization degrees (4-10), referenced to degree 7. Color denotes actual energy gap values.

5.3.2 Model Performance

Figure 5.5 illustrates the comparative performance of two models, SchNet-D and SchNet-T, in predicting the electronic properties of oligomers across a range of polymerization degrees. The figure visualizes the MAE for the HOMO energy, LUMO energy, and the HOMO-LUMO gap, demonstrating that while SchNet-D provides acceptable accuracy within the mid-range of polymerization degrees (6 to 8), its accuracy declines when predicting properties for oligomers at the higher and lower ends of the polymerization spectrum. This drop in performance of the SchNet-D model is indicative of the edge effect, where predictions for data points at the distribution's boundaries are less accurate due to the model's overfitting to the more common central data trends.[35] This overfitting limits the model's ability to generalize to less frequent, more diverse oligomer configurations that occur at the edges of the dataset's polymerization degree range.

In contrast, SchNet-T, which employs transfer learning, shows improved performance across the full range of polymerization degrees, as seen in the lower MAE values. Transfer learning leverages knowledge from pre-trained models on large and diverse datasets to enhance the model's ability to handle edge cases effectively. This approach broadens the model's applicability and predictive accuracy beyond the core range, addressing the shortcomings of direct learning models and improving generalization to a wider array of oligomer structures.

When we examine the impact of transfer learning across different molecular properties, we observe that the enhancement in prediction accuracy for different properties is different. It is evident that the accuracy enhancements are most substantial for HOMO energy level predictions. Although the improvements in predicting LUMO energy levels are also clear, they are not as pronounced as those observed for HOMO levels. Moreover, while the HOMO-LUMO gap predictions have benefitted from the transfer learning approach, the level of improvement is less substantial compared to that of the HOMO and LUMO energies. This variation in the extent of enhancement across these properties can be traced back to the level of similarity in energy distributions between the source dataset and the

target dataset. A greater similarity in HOMO energy distributions leads to more significant improvements due to transfer learning. In contrast, the LUMO energies and HOMO-LUMO gaps, which exhibit greater variances between the datasets, require more intricate model adjustments to enhance prediction accuracy. The transfer learning model has to navigate these differences, applying sophisticated fine-tuning to reconcile the discrepancies and sharpen its predictive precision.

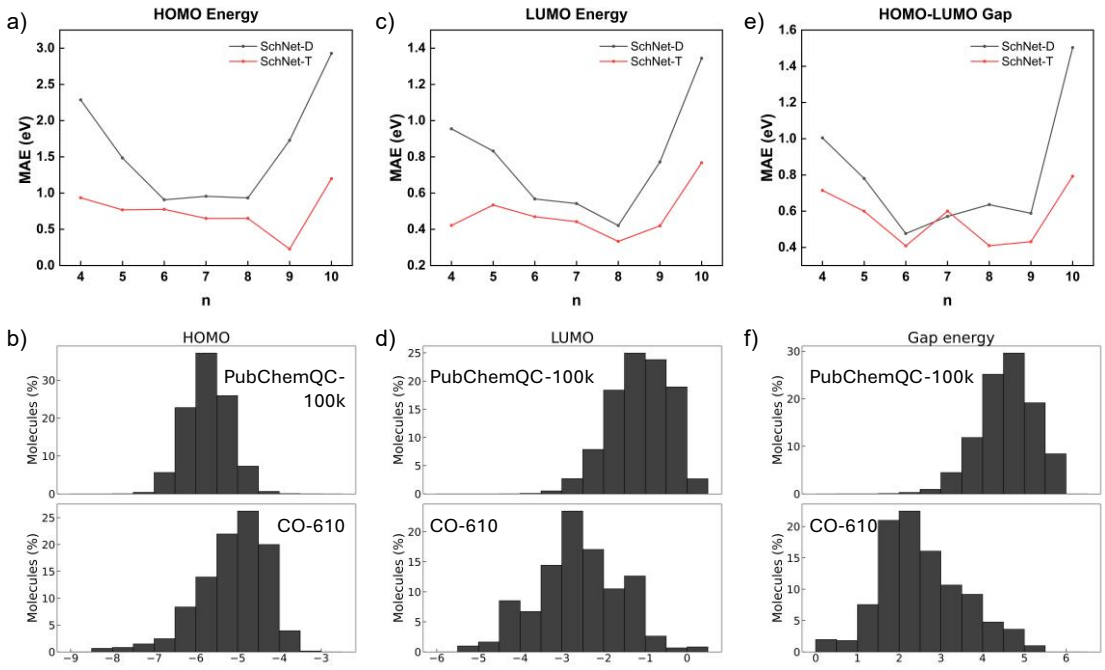


Figure 5.5 Comparative model accuracy and dataset distribution. Graphs (a), (c), and (e) depict the mean absolute error (MAE) for HOMO, LUMO, and HOMO-LUMO gap energy levels, respectively, comparing predictions by SchNet-D (black) with SchNet-T (red) models across varying polymerization degrees. Histograms (b), (d), and (f) show the energy level distribution for HOMO, LUMO, and HOMO-LUMO gap within the PubChemQC-100k (upper histograms) and CO-610 (lower histograms) datasets.

Figure 5.6 offers a comparative visualization of the electronic property predictions from the SchNet-D and SchNet-T models against the calculated values. The scatter plots reveal that the SchNet-D model has a noticeable variance in the HOMO level predictions, indicating a gap in the model's predictive accuracy for this property, as the data points deviate from the ideal one-to-one correlation line. The performance of the SchNet-D model

shows improvement in predicting LUMO levels and the energy gap, as the data points are more closely aligned with the calculated values, suggesting a better grasp of these electronic properties.

Conversely, the SchNet-T models demonstrate superior predictive prowess overall. By leveraging the extensive knowledge encapsulated in the source dataset, these models show enhanced learning capabilities and the ability to generalize better when confronted with new data. This is seen in the tighter clustering of data points around the line of perfect agreement. However, the plots also show outliers, hinting at the SchNet-T models' limitations. While the transfer learning approach yields a stronger average performance, likely due to its generalizability across diverse data, it may sometimes compromise the ability to capture the nuanced characteristics of the new dataset, which could lead to occasional prediction anomalies.

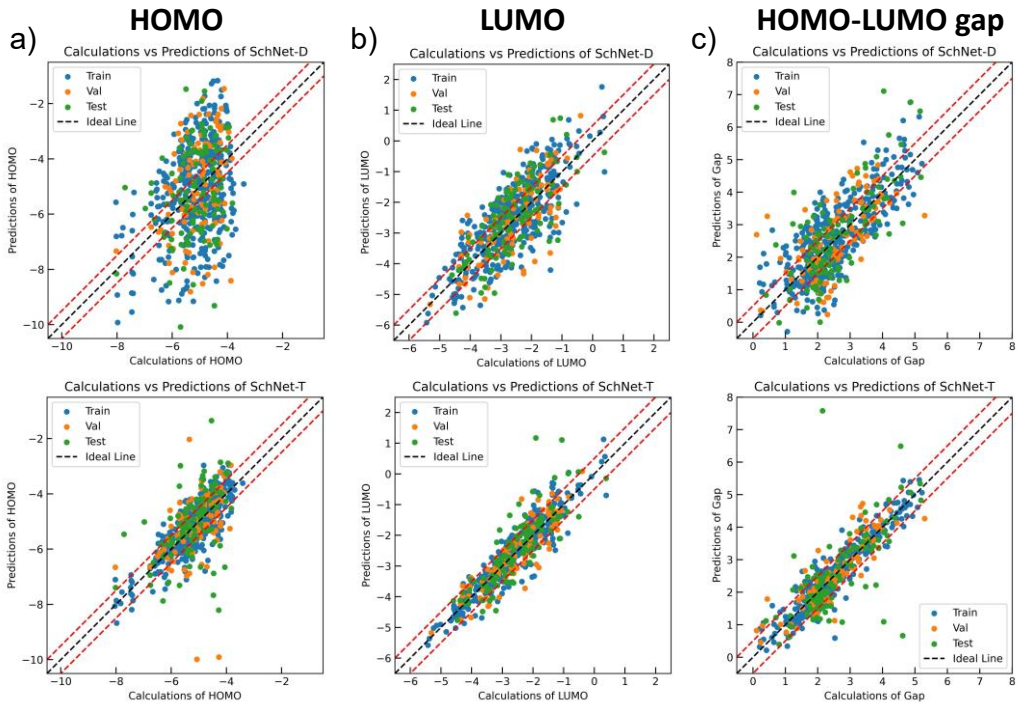


Figure 5.6 Predictive performance of SchNet-D and SchNet-T models. The correlation between the calculated values and predicted (a) HOMO, (b) LUMO, and (c) HOMO-LUMO gap using SchNet-D (top) and SchNet-T (bottom) models.

5.3.3 High-throughput Screening

The dataset for this study was compiled from a collection of monomers cited in prior research. Spanning a polymerization degree from 4 to 10, we utilized 530 distinct monomers to generate a candidate pool of 3,710 oligomers. Within this extensive candidate dataset, **Figure 5.7** showcases a select group of 85 oligomers identified as promising for photovoltaic applications based on their alignment with the criteria for photovoltaic material properties. These oligomers exhibit HOMO energy levels ranging from -6.5 to -4.9 eV, LUMO levels between -3.0 to -4.5 eV, and an energy gap spanning 1.1 to 2.0 eV. The SMILES representations, along with the predicted and calculated properties for these screened oligomers, are listed in **Tables 5.2** and **5.3**, respectively.

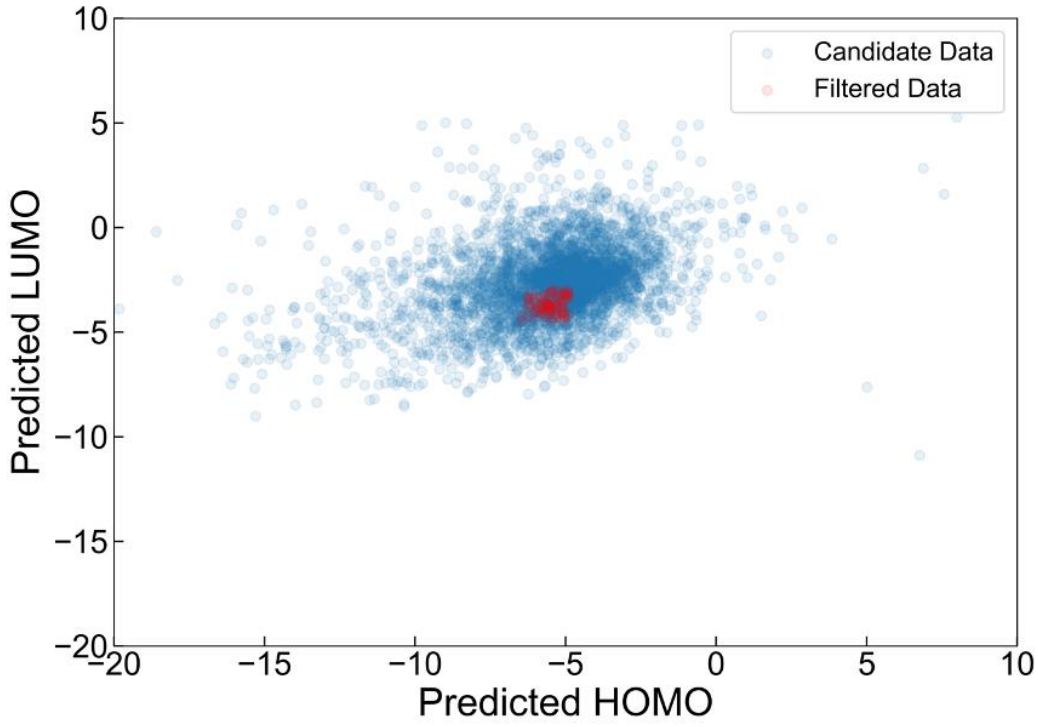


Figure 5.7 Scatter plot of predicted HOMO and LUMO levels. The plot visualizes the predicted HOMO and LUMO energy levels for a candidate dataset of 3,710 oligomers, with 85 oligomers highlighted (red stars) as promising candidates for photovoltaic applications.

Table 5.2 SMILES of the screened oligomers

Candidate ID	N	SMILES
272	5	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6scc7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1
21	7	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9ccc(C(F)(F)F)cc89)c8ccc(C(F)(F)F)cc78)c7ccc(C(F)(F)F)cc67)c6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1
515	8	Fc1ccc(F)c2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9scc%10c(F)ccc(F)c9%10)c9c(F)ccc(F)c89)c8c(F)ccc(F)c78)c7c(F)ccc(F)c67)c6c(F)ccc(F)c56)c5c(F)ccc(F)c45)c4c(F)ccc(F)c34)scc12
290	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc%10ccc(F)cc9%10)c9ccc(F)cc89)c8ccc(F)cc78)c7ccc(F)cc67)c6ccc(F)cc56)c5ccc(F)cc45)c4ccc(F)cc34)c3ccc(F)cc23)c2ccc(F)cc12
467	9	c1cc2nsnc2c(-c2cnc(-c3cnc(-c4cnc(-c5cnc(-c6cnc(-c7cnc(-c8cnc(-c9cncc%10nsnc9%10)c9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)n1
267	9	c1cc2c(s1)-c1sc(-c3cc4c(s3)-c3sc(-c5cc6c(s5)-c5sc(-c7cc8c(s7)-c7sc(-c9cc%10c(s9)-c9sc(-c%11cc%12c(s%11)-c%11sc(-c%13cc%14c(s%13)-c%13sc(-c%15cc%16c(s%15)-c%15sc(-c%17cc%18c(s%17)-c%17scnc%17C%18)nc%15C%16)nc%13C%14)nc%11C%12)nc9C%10)nc7C8)nc5C6)nc3C4)nc1C2
433	5	C=Cc1c(F)c(F)c(C=Cc2c(F)c(F)c(C=Cc3c(F)c(F)c(C=Cc4c(F)c(F)c(C=Cc5c(F)c(F)cc6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)c2nsnc12
272	4	N#Cc1cc2c(-c3sc(-c4sc(-c5scc6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1
87	5	N#Cc1csc(-c2sc(-c3sc(-c4sc(-c5scc(C#N)c5C#N)c(C#N)c4C#N)c(C#N)c3C#N)c(C#N)c2C#N)c1C#N
430	9	N#Cc1nc2cc3nsnc3c(-c3c4nsnc4c(-c4c5nsnc5c(-c5c6nsnc6c(-c6c7nsnc7c(-c7c8nsnc8c(-c8c9nsnc9c(-c9c%10nsnc%10c(-c%10c%11nsnc%11cc%11nc(C#N)c(C#N)nc%10%11)c%10nc(C#N)c(C#N)nc9%10)c9nc(C#N)c(C#N)nc89)c8nc(C#N)c(C#N)nc78)c7nc(C#N)c(C#N)nc67)c6nc(C#N)c(C#N)nc56)c5nc(C#N)c(C#N)nc45)c4nc(C#N)c(C#N)nc34)c2nc1C#N
267	8	c1cc2c(s1)-c1sc(-c3cc4c(s3)-c3sc(-c5cc6c(s5)-c5sc(-c7cc8c(s7)-c7sc(-c9cc%10c(s9)-c9sc(-c%11cc%12c(s%11)-c%11sc(-c%13cc%14c(s%13)-c%13sc(-c%15cc%16c(s%15)-c%15scnc%15C%16)nc%13C%14)nc%11C%12)nc9C%10)nc7C8)nc5C6)nc3C4)nc1C2
267	10	c1cc2c(s1)-c1sc(-c3cc4c(s3)-c3sc(-c5cc6c(s5)-c5sc(-c7cc8c(s7)-c7sc(-c9cc%10c(s9)-c9sc(-c%11cc%12c(s%11)-c%11sc(-c%13cc%14c(s%13)-c%13sc(-c%15cc%16c(s%15)-c%15sc(-c%17cc%18c(s%17)-c%17sc(-c%19cc%20c(s%19)-c%19scnc%19C%20)nc%17C%18)nc%15C%16)nc%13C%14)nc%11C%12)nc9C%10)nc7C8)nc5C6)nc3C4)nc1C2
295	5	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6scc7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1
409	9	C=CC1=CC=C(C=CC2=CC=C(C=CC3=CC=C(C=CC4=CC=C(C=CC5=CC=C(C=CC6=CC=C(C=CC7=CC=C(C=CC8=CC=C(C=CC9=CC=CC9=S)C8=S)C7=S)C6=S)C5=S)C4=S)C3=S)C2=S)C1=S
87	6	N#Cc1csc(-c2sc(-c3sc(-c4sc(-c5sc(-c6scc(C#N)c6C#N)c(C#N)c5C#N)c(C#N)c4C#N)c(C#N)c3C#N)c(C#N)c2C#N)c1C#N
430	10	N#Cc1nc2cc3nsnc3c(-c3c4nsnc4c(-c4c5nsnc5c(-c5c6nsnc6c(-c6c7nsnc7c(-c7c8nsnc8c(-c8c9nsnc9c(-c9c%10nsnc%10c(-c%10c%11nsnc%11c(-c%11c%12nsnc%12cc%12nc(C#N)c(C#N)nc%11%12)c%11nc(C#N)c(C#N)nc%10%11)c%10nc(C#N)c(C#N)nc9%10)c9nc(C#N)c(C#N)nc89)c8nc(C#N)c(C#N)nc78)c7nc(C#N)c(C#N)nc67)c6nc(C#N)c(C#N)nc56)c5nc(C#N)c(C#N)nc45)c4nc(C#N)c(C#N)nc34)c2nc1C#N

Candidate ID	N	SMILES
21	6	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8ccc(C(F)(F)F)cc78)c7ccc(C(F)(F)F)cc67)c6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1
515	7	Fc1ccc(F)c2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9c(F)ccc(F)c89)c8c(F)ccc(F)c78)c7c(F)ccc(F)c67)c6c(F)ccc(F)c56)c5c(F)ccc(F)c45)c4c(F)ccc(F)c34)scc12
508	4	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4scc5ccccc45)c4ccccc34)c3ccccc23)c2ccccc12
117	8	C=CC=Cc1cnc(C=CC=Cc2cnc(C=CC=Cc3cnc(C=CC=Cc4cnc(C=CC=Cc5cnc(C=CC=Cc6cnc(C=CC=Cc7cnc(C=CC=Cc8cncc8C)c7C)c6C)c5C)c4C)c3C)c2C)c1C
103	5	C=Cc1ccc(-c2ccc(-c3ccc(C=Cc4ccc(-c5ccc(-c6ccc(C=Cc7ccc(-c8ccc(-c9ccc(C=Cc%10ccc(-c%11ccc(-c%12ccc(C=Cc%13ccc(-c%14ccc(-c%15cccs%15)s%14)s%13)s%12)s%11)s%10)s9)s8)s7)s6)s5)s4)s3)s2)s1
416	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc(O)c8C#N)c(O)c7C#N)c(O)c6C#N)c(O)c5C#N)c(O)c4C#N)c(O)c3C#N)c(O)c2C#N)c(O)c1C#N
33	6	O=C1c2ccccc2-c2ccc(-c3ccc4c(c3)C(=O)c3cc(-c5ccc6c(c5)C(=O)c5cc(-c7ccc8c(c7)C(=O)c7cc(-c9ccc%10c(c9)C(=O)c9cc(-c%11ccc%12c(c%11)C(=O)c%11ccccc%11-%12)ccc9-%10)ccc7-8)ccc5-6)ccc3-4)cc21
139	6	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6scc7nccnc67)c6nccnc56)c5nccnc45)c4nccnc34)c3nccnc23)c2nccnc12
14	4	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4scc5c(F)ccc(F)c45)c4c(F)ccc(F)c34)c3c(F)ccc(F)c23)c2c(F)ccc(F)c12
301	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc9cccc(F)c89)c8cccc(F)c78)c7cccc(F)c67)c6cccc(F)c56)c5cccc(F)c45)c4cccc(F)c34)c3cccc(F)c23)c2cccc(F)c12
187	10	c1cc2sc(-c3cc4sc(-c5cc6sc(-c7cc8sc(-c9cc%10sc(-c%11cc%12sc(-c%13cc%14sc(-c%15cc%16sc(-c%17cc%18sc(-c%19cc%20sccc%20n%19)cc%18n%17)cc%16n%15)cc%14n%13)cc%12n%11)cc%10n9)cc8n7)cc6n5)cc4n3)cc2n1
328	5	C=CC#Cc1ccc(C=CC#Cc2ccc(C=CC#Cc3ccc(C=CC#Cc4ccc(C=CC#Cc5cccn5)n4)n3)n2)n1
470	8	C=Cc1oc(C=Cc2oc(C=Cc3oc(C=Cc4oc(C=Cc5oc(C=Cc6oc(C=Cc7oc(C=Cc8occ9c8C(=O)NC9=O)c8c7C(=O)NC8=O)c7c6C(=O)NC7=O)c6c5C(=O)NC6=O)c5c4C(=O)NC5=O)c4c3C(=O)NC4=O)c3c2C(=O)NC3=O)c2c1C(=O)NC2=O
394	10	CC(=O)c1cnc(-c2nc(-c3nc(-c4nc(-c5nc(-c6nc(-c7nc(-c8nc(-c9nc(-c%10nccc%10C(C)=O)cc9C(C)=O)cc8C(C)=O)cc7C(C)=O)cc6C(C)=O)cc5C(C)=O)cc4C(C)=O)cc3C(C)=O)cc2C(C)=O)c1
21	8	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9scc%10ccc(C(F)(F)F)cc9%10)c9ccc(C(F)(F)F)cc89)c8ccc(C(F)(F)F)cc78)c7ccc(C(F)(F)F)cc67)c6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1
188	8	C=Cc1cc2c(n1)C=C(C=Cc1cc3c(n1)C=C(C=Cc1cc4c(n1)C=C(C=Cc1cc5c(n1)C=C(C=Cc1cc6c(n1)C=C(C=Cc1cc7c(n1)C=C(C=Cc1cc8c(n1)C=C(C=Cc1cc9c(n1)C=CC9)C8)C7)C6)C5)C4)C3)C2
267	7	c1cc2c(s1)-c1sc(-c3cc4c(s3)-c3sc(-c5cc6c(s5)-c5sc(-c7cc8c(s7)-c7sc(-c9cc%10c(s9)-c9sc(-c%11cc%12c(s%11)-c%11sc(-c%13cc%14c(s%13)-c%13scnc%13C%14)nc%11C%12)nc9C%10)nc7C8)nc5C6)nc3C4)nc1C2
495	9	NCc1nc2csc(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11nc(CN)c(CN)nc%10%11)c%10nc(CN)c(CN)nc9%10)c9nc(CN)c(CN)nc89)c8nc(CN)c(CN)nc78)c7nc(CN)c(CN)nc67)c6nc(CN)c(CN)nc56)c5nc(CN)c(CN)nc45)c4nc(CN)c(CN)nc34)c2nc1CN
312	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc%10cc(C(F)(F)F)ccc9%10)c9cc(C(F)(F)F)ccc89)c8cc(C(F)(F)F)ccc78)c7cc(C(F)(F)F)ccc67)c6cc(C(F)(F)F)ccc56)c5cc(C(F)(F)F)ccc45)c4cc(C(F)(F)F)ccc34)c3cc(C(F)(F)F)ccc23)c2cc(C(F)(F)F)ccc12
204	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc9c8SCSO9)c8c7SCSO8)c7c6SCSO7)c6c5SCSO6)c5c4SCSO5)c4c3SCSO4)c3c2SCSO3)c2c1SCSO2

Candidate ID	N	SMILES
104	5	C=Cc1ccnc1C=Cc1ccc(C=Cc2ccnc2C=Cc2ccc(C=Cc3ccnc3C=Cc3ccc(C=Cc4ccnc4C=Cc4ccc(C=Cc5ccnc5C=Cc5cccn5)n4)n3)n2)n1
112	10	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9sc(C=Cc%10scc%11c%10OCCSO%11)c%10c9OCCSO%10)c9c8OCCSO9)c8c7OCCSO8)c7c6OCCSO7)c6c5OCCSO6)c5c4OCCSO5)c4c3OCCSO4)c3c2OCCSO3)c2c1OCCSO2
515	10	Fc1ccc(F)c2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10sc(-c%11scc%12c(F)ccc(F)c%11%12)c%11c(F)ccc(F)c%10%11)c%10c(F)ccc(F)c9%10)c9c(F)ccc(F)c89)c8c(F)ccc(F)c78)c7c(F)ccc(F)c67)c6c(F)ccc(F)c56)c5c(F)ccc(F)c45)c4c(F)ccc(F)c34)scc12
294	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc8nc(CC)cnc78)c7nc(CC)cnc67)c6nc(CC)cnc56)c5nc(CC)cnc45)c4nc(CC)cnc34)c3nc(CC)cnc23)c2nc(CC)cnc12
409	10	C=CC1=CC=C(C=CC2=CC=C(C=CC3=CC=C(C=CC4=CC=C(C=CC5=CC=C(C=CC6=CC=C(C=CC7=CC=C(C=CC8=CC=C(C=CC9=CC=C(C=CC%10=CC=CC%10=S)C9=S)C8=S)C7=S)C6=S)C5=S)C4=S)C3=S)C2=S)C1=S
236	5	O=C1c2csc(-c3sc(-c4sc(-c5sc(-c6scc7c6C(=O)N(C(F)(F)F)C7=O)c6c5C(=O)N(C(F)(F)F)C6=O)c5c4C(=O)N(C(F)(F)F)C5=O)c4c3C(=O)N(C(F)(F)F)C4=O)c2C(=O)N1C(F)(F)F
26	9	C=Cc1sc(F)c2c1SC=C(C=Cc1sc(F)c3c1SC=C(C=Cc1sc(F)c4c1SC=C(C=Cc1sc(F)c5c1SC=C(C=Cc1sc(F)c6c1SC=C(C=Cc1sc(F)c7c1SC=C(C=Cc1sc(F)c8c1SC=C(C=Cc1sc(F)c9c1SC=C(C=Cc1sc(F)c%10c1SC=CO%10)O9)O8)O7)O6)O5)O4)O3)O2
515	9	Fc1ccc(F)c2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11c(F)ccc(F)c%10%11)c%10c(F)ccc(F)c9%10)c9c(F)ccc(F)c89)c8c(F)ccc(F)c78)c7c(F)ccc(F)c67)c6c(F)ccc(F)c56)c5c(F)ccc(F)c45)c4c(F)ccc(F)c34)scc12
21	4	FC(F)(F)c1ccc2c(-c3sc(-c4sc(-c5scc6ccc(C(F)(F)F)cc56)c5ccc(C(F)(F)F)cc45)c4ccc(C(F)(F)F)cc34)scc2c1
210	4	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4scc5nc(C#N)cc45)c4nc(C#N)cc34)c3nc(C#N)cc23)c2nc(C#N)cc12
295	9	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11cncnc%10%11)c%10cncnc9%10)c9cncnc89)c8cncnc78)c7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1
186	6	C=CC1=C2C(=O)NC(C=CC3=C4C(=O)NC(C=CC5=C6C(=O)NC(C=CC7=C8C(=O)NC(C=CC9=C%10C(=O)NC(C=CC%11=C%12C(=O)NC=C%12C(=O)O%11)=C%10C(=O)O9)=C8C(=O)O7)=C6C(=O)O5)=C4C(=O)O3)=C2C(=O)O1
470	4	C=Cc1oc(C=Cc2oc(C=Cc3oc(C=Cc4occ5c4C(=O)NC5=O)c4c3C(=O)NC4=O)c3c2C(=O)NC3=O)c2c1C(=O)NC2=O
98	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc8c7CCCS8(=O)=O)c7c6CCCS7(=O)=O)c6c5CCCS6(=O)=O)c5c4CCCS5(=O)=O)c4c3CCCS4(=O)=O)c3c2CCCS3(=O)=O)c2c1CCCS2(=O)=O
33	5	O=C1c2ccccc2-c2ccc(-c3ccc4c(c3)C(=O)c3cc(-c5ccc6c(c5)C(=O)c5cc(-c7ccc8c(c7)C(=O)c7cc(-c9ccc%10c(c9)C(=O)c9ccccc9-%10)ccc7-8)ccc5-6)ccc3-4)cc21
32	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc%10c9C(=O)CCC%10=O)c9c8C(=O)CCC9=O)c8c7C(=O)CCC8=O)c7c6C(=O)CCC7=O)c6c5C(=O)CCC6=O)c5c4C(=O)CCC5=O)c4c3C(=O)CCC4=O)c3c2C(=O)CCC3=O)c2c1C(=O)CCC2=O
394	9	CC(=O)c1cnc(-c2nc(-c3nc(-c4nc(-c5nc(-c6nc(-c7nc(-c8nc(-c9nccc9C(C)=O)cc8C(C)=O)cc7C(C)=O)cc6C(C)=O)cc5C(C)=O)cc4C(C)=O)cc3C(C)=O)cc2C(C)=O)c1
272	6	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1
301	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc%10cccc(F)c9%10)c9cccc(F)c89)c8cccc(F)c78)c7cccc(F)c67)c6cccc(F)c56)c5cccc(F)c45)c4cccc(F)c34)c3cccc(F)c23)c2cccc(F)c12
301	4	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4scc5cccc(F)c45)c4cccc(F)c34)c3cccc(F)c23)c2cccc(F)c12
433	10	C=Cc1c(F)c(F)c(C=Cc2c(F)c(F)c(C=Cc3c(F)c(F)c(C=Cc4c(F)c(F)c(C=Cc5c(F)c(F)c(C=Cc6c(F)c(F)c(C=Cc7c(F)c(F)c(C=Cc8c(F)c(F)c(C=Cc9c(F)c(F)c(C=Cc%10c(F)c(F

Candidate ID	N	SMILES
		)cc%11nsnc%10%11)c%10nsnc9%10)c9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)c2nsnc12
467	8	c1cc2nsnc2c(-c2cnc(-c3cnc(-c4cnc(-c5cnc(-c6cnc(-c7cnc(-c8cncc9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)n1
139	4	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4scc5nccnc45)c4nccnc34)c3nccnc23)c2nccnc12
139	5	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5scc6nccnc56)c5nccnc45)c4nccnc34)c3nccnc23)c2nccnc12
72	8	O=C(O)c1coc(-c2oc(-c3oc(-c4oc(-c5oc(-c6oc(-c7oc(-c8occc8C(=O)O)cc7C(=O)O)cc6C(=O)O)cc5C(=O)O)cc4C(=O)O)cc3C(=O)O)cc2C(=O)O)c1
229	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc8nc(C)cnc78)c7nc(C)cnc67)c6nc(C)cnc56)c5nc(C)cnc45)c4nc(C)cnc34)c3nc(C)cnc23)c2nc(C)cnc12
409	8	C=CC1=CC=C(C=CC2=CC=C(C=CC3=CC=C(C=CC4=CC=C(C=CC5=CC=C(C=CC6=CC=C(C=CC7=CC=C(C=CC8=CC=CC8=S)C7=S)C6=S)C5=S)C4=S)C3=S)C2=S)C1=S
409	7	C=CC1=CC=C(C=CC2=CC=C(C=CC3=CC=C(C=CC4=CC=C(C=CC5=CC=C(C=CC6=CC=C(C=CC7=CC=CC7=S)C6=S)C5=S)C4=S)C3=S)C2=S)C1=S
272	7	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8scc9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1
295	6	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8cncnc78)c7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1
98	6	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6scc7c6CCCS7(=O)=O)c6c5CCCS6(=O)=O)c5c4CCCS5(=O)=O)c4c3CCCS4(=O)=O)c3c2CCCS3(=O)=O)c2c1CCCS2(=O)=O
430	8	N#Cc1nc2cc3nsnc3c(-c3c4nsnc4c(-c4c5nsnc5c(-c5c6nsnc6c(-c6c7nsnc7c(-c7c8nsnc8c(-c8c9nsnc9c(-c9c%10nsnc%10cc%10nc(C#N)c(C#N)nc9%10)c9nc(C#N)c(C#N)nc89)c8nc(C#N)c(C#N)nc78)c7nc(C#N)c(C#N)nc67)c6nc(C#N)c(C#N)nc56)c5nc(C#N)c(C#N)nc45)c4nc(C#N)c(C#N)nc34)c2nc1C#N
117	7	C=CC=Cc1cnc(C=CC=Cc2cnc(C=CC=Cc3cnc(C=CC=Cc4cnc(C=CC=Cc5cnc(C=CC=Cc6cnc(C=CC=Cc7cncc7C)c6C)c5C)c4C)c3C)c2C)c1C
204	9	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8sc(C=Cc9scc%10c9SCSO%10)c9c8SCSO9)c8c7SCSO8)c7c6SCSO7)c6c5SCSO6)c5c4SCSO5)c4c3SCSO4)c3c2SCSO3)c2c1SCSO2
272	8	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9scc%10cc(C#N)oc9%10)c9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1
236	6	O=C1c2csc(-c3sc(-c4sc(-c5sc(-c6sc(-c7scc8c7C(=O)N(C(F)(F)F)C8=O)c7c6C(=O)N(C(F)(F)F)C7=O)c6c5C(=O)N(C(F)(F)F)C6=O)c5c4C(=O)N(C(F)(F)F)C5=O)c4c3C(=O)N(C(F)(F)F)C4=O)c2C(=O)N1C(F)(F)F
311	4	N#Cc1coc(-c2oc(-c3oc(-c4occc4C#N)cc3C#N)cc2C#N)c1
204	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc8c7SCSO8)c7c6SCSO7)c6c5SCSO6)c5c4SCSO5)c4c3SCSO4)c3c2SCSO3)c2c1SCSO2
312	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc9cc(C(F)(F)F)ccc89)c8cc(C(F)(F)F)ccc78)c7cc(C(F)(F)F)ccc67)c6cc(C(F)(F)F)ccc56)c5cc(C(F)(F)F)ccc45)c4cc(C(F)(F)F)ccc34)c3cc(C(F)(F)F)ccc23)c2cc(C(F)(F)F)ccc12
295	10	c1ncc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10sc(-c%11scc%12cncnc%11%12)c%11cncnc%10%11)c%10cncnc9%10)c9cncnc89)c8cncnc78)c7cncnc67)c6cncnc56)c5cncnc45)c4cncnc34)scc2n1
416	7	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7scc(O)c7C#N)c(O)c6C#N)c(O)c5C#N)c(O)c4C#N)c(O)c3C#N)c(O)c2C#N)c(O)c1C#N
328	6	C=CC#Cc1ccc(C=CC#Cc2ccc(C=CC#Cc3ccc(C=CC#Cc4ccc(C=CC#Cc5ccc(C=CC#Cc6cccn6)n5)n4)n3)n2)n1
32	8	C=Cc1sc(C=Cc2sc(C=Cc3sc(C=Cc4sc(C=Cc5sc(C=Cc6sc(C=Cc7sc(C=Cc8scc9c8C(=O)CCC9=O)c8c7C(=O)CCC8=O)c7c6C(=O)CCC7=O)c6c5C(=O)CCC6=O)c5c4C(=O)CCC5=O)c4c3C(=O)CCC4=O)c3c2C(=O)CCC3=O)c2c1C(=O)CCC2=O

Candidate ID	N	SMILES
269	4	c1sc(-c2sc(-c3sc(-c4scc5c4N=S=N5)c4c3N=S=N4)c3c2N=S=N3)c2c1N=S=N2
350	10	COc1coc(-c2oc(-c3oc(-c4oc(-c5oc(-c6oc(-c7oc(-c8oc(-c9oc(-c%10occ(C#N)c%10OC)c(C#N)c9OC)c(C#N)c8OC)c(C#N)c7OC)c(C#N)c6OC)c(C#N)c5OC)c(C#N)c4OC)c(C#N)c3OC)c(C#N)c2OC)c1C#N
142	10	Fc1cccc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10sc(-c%11scc%12cccc(F)c%11%12)c%11cccc(F)c%10%11)c%10cccc(F)c9%10)c9cccc(F)c89)c8cccc(F)c78)c7cccc(F)c67)c6cccc(F)c56)c5cccc(F)c45)c4cccc(F)c34)scc12
272	9	N#Cc1cc2c(-c3sc(-c4sc(-c5sc(-c6sc(-c7sc(-c8sc(-c9sc(-c%10scc%11cc(C#N)oc%10%11)c%10cc(C#N)oc9%10)c9cc(C#N)oc89)c8cc(C#N)oc78)c7cc(C#N)oc67)c6cc(C#N)oc56)c5cc(C#N)oc45)c4cc(C#N)oc34)scc2o1
311	5	N#Cc1coc(-c2oc(-c3oc(-c4oc(-c5occc5C#N)cc4C#N)cc3C#N)cc2C#N)c1
467	10	c1cc2nsnc2c(-c2cnc(-c3cnc(-c4cnc(-c5cnc(-c6cnc(-c7cnc(-c8cnc(-c9cnc(-c%10cncc%11nsnc%10%11)c%10nsnc9%10)c9nsnc89)c8nsnc78)c7nsnc67)c6nsnc56)c5nsnc45)c4nsnc34)c3nsnc23)n1

Table 5.3 Predicted and calculated electronic properties of the screened oligomers

Candidate ID	N	Predicted HOMO	Predicted LUMO	Predicted Gap	Calculated HOMO	Calculated LUMO	Calculated Gap
272	5	-5.856	-3.817	1.336	-5.332	-3.570	1.761
21	7	-5.286	-3.297	1.610	-5.036	-3.169	1.867
515	8	-5.522	-3.193	1.677	-4.992	-2.589	2.403
290	9	-4.932	-4.191	1.356	-4.040	-3.314	0.726
467	9	-5.726	-3.617	1.837	-6.066	-3.542	2.523
267	9	-5.750	-3.901	1.694	-4.543	-2.682	1.861
433	5	-5.235	-3.313	1.196	-5.442	-3.381	2.061
272	4	-6.264	-3.824	1.984	-5.405	-3.381	2.024
87	5	-6.236	-3.416	1.646	-7.352	-4.175	3.177
430	9	-5.088	-4.097	1.456	-7.135	-4.783	2.352
267	8	-5.325	-3.565	1.484	-4.553	-2.663	1.889
267	10	-6.259	-4.264	1.958	-4.534	-2.697	1.837
295	5	-5.473	-3.098	1.529	-5.280	-3.504	1.776
409	9	-5.944	-3.978	1.280	-4.581	-4.521	0.061
87	6	-6.266	-3.280	1.168	-7.349	-4.281	3.068
430	10	-5.467	-4.456	1.689	-6.591	-5.304	1.287
21	6	-5.170	-3.189	1.573	-5.103	-3.089	2.014
515	7	-5.042	-3.047	1.793	-4.898	-2.608	2.290
508	4	-4.936	-3.145	1.608	-4.097	-2.827	1.270
117	8	-5.594	-3.645	1.812	-4.746	-3.512	1.233
103	5	-5.650	-3.695	1.241	-4.542	-2.519	2.024
416	8	-5.839	-3.343	1.551	-5.358	-3.484	1.874
33	6	-5.636	-4.273	1.693	-5.685	-2.609	3.076
139	6	-5.544	-3.906	1.832	-4.263	-3.240	1.024
14	4	-5.284	-3.744	1.224	-4.368	-3.087	1.280

Candidate ID	N	Predicted HOMO	Predicted LUMO	Predicted Gap	Calculated HOMO	Calculated LUMO	Calculated Gap
301	8	-5.027	-4.206	1.423	-5.018	-1.974	3.044
187	10	-5.022	-4.033	1.406	-5.710	-4.435	1.276
328	5	-5.157	-3.762	1.530	-5.287	-3.951	1.515
470	8	-5.116	-3.782	1.915	-5.608	-3.516	2.091
394	10	-5.789	-3.285	1.810	-6.360	-4.716	1.644
21	8	-5.815	-3.635	1.777	-5.100	-3.209	1.891
188	8	-5.030	-4.472	1.713	-4.948	-3.598	1.350
267	7	-4.979	-3.269	1.309	-4.567	-2.639	1.928
495	9	-6.212	-3.650	1.317	-4.926	-3.291	1.636
312	9	-5.540	-3.784	1.797	-4.384	-3.634	0.749
204	8	-5.685	-4.226	1.414	-4.359	-2.589	1.770
104	5	-5.043	-4.335	1.721	-5.481	-3.792	1.690
112	10	-5.328	-4.146	1.349	-4.152	-2.224	1.928
515	10	-6.170	-3.753	1.967	-4.954	-2.645	2.309
294	7	-5.320	-4.160	1.889	-4.081	-3.233	0.849
409	10	-6.460	-4.225	1.459	-4.964	-3.912	1.052
236	5	-6.236	-3.399	1.877	-6.715	-3.806	2.909
26	9	-5.289	-4.263	1.880	-5.022	-2.115	2.907
515	9	-5.677	-3.234	2.000	-4.890	-2.643	2.247
21	4	-6.126	-3.194	1.937	-5.160	-2.864	2.296
210	4	-5.698	-3.469	1.129	-6.384	-4.856	1.528
295	9	-5.061	-3.561	1.342	-4.910	-4.021	0.889
186	6	-6.007	-4.037	1.604	-5.375	-4.428	0.947
470	4	-5.286	-3.016	1.531	-5.749	-3.325	2.425
98	7	-5.167	-3.585	1.964	-5.119	-3.108	2.011
33	5	-5.096	-3.942	1.530	-5.699	-2.586	3.113
32	9	-5.564	-3.813	1.977	-5.621	-3.103	2.518
394	9	-5.440	-3.021	1.690	-5.570	-3.210	1.675
272	6	-5.608	-3.725	1.208	-5.275	-3.706	1.569
301	9	-5.323	-4.311	1.515	-3.866	-3.294	0.572
301	4	-4.906	-3.108	1.360	-4.229	-2.952	1.276
433	10	-5.095	-4.125	1.244	-5.330	-3.510	1.819
467	8	-6.039	-3.542	1.806	-6.086	-3.523	2.563
139	4	-5.611	-3.665	1.893	-4.462	-3.090	1.372
139	5	-5.316	-3.771	1.761	-4.364	-3.196	1.169
72	8	-5.510	-4.072	1.866	-5.968	-3.288	2.680
229	7	-4.945	-3.255	1.676	-4.075	-3.112	0.963
409	8	-5.485	-3.705	1.294	-4.587	-4.517	0.070

Candidate ID	N	Predicted HOMO	Predicted LUMO	Predicted Gap	Calculated HOMO	Calculated LUMO	Calculated Gap
409	7	-5.100	-3.463	1.264	-4.594	-4.512	0.082
272	7	-5.485	-3.842	1.246	-5.193	-3.770	1.423
295	6	-4.937	-3.010	1.436	-5.213	-3.626	1.587
98	6	-5.051	-3.207	1.912	-5.107	-3.055	2.052
430	8	-5.322	-3.968	1.244	-7.107	-4.754	2.352
117	7	-5.034	-3.168	1.547	-5.287	-3.951	1.515
204	9	-6.466	-4.486	1.774	-4.332	-2.608	1.724
272	8	-5.780	-4.109	1.258	-5.157	-3.859	1.298
236	6	-6.021	-3.368	1.821	-6.741	-3.875	2.866
311	4	-5.593	-3.978	1.994	-6.125	-2.863	3.262
204	7	-5.498	-3.865	1.222	-4.391	-2.564	1.827
312	8	-4.920	-3.403	1.428	-5.537	-2.547	2.990
295	10	-5.529	-3.710	1.472	-5.203	-3.600	1.602
416	7	-5.429	-3.026	1.310	-5.371	-3.434	1.937
328	6	-5.861	-4.037	1.696	-5.394	-4.749	0.645
32	8	-4.903	-3.289	1.663	-5.420	-3.365	2.056
269	4	-5.341	-3.557	1.351	-4.563	-3.757	0.806
350	10	-5.859	-3.859	1.443	-4.832	-2.786	2.047
142	10	-5.527	-3.068	1.846	-4.640	-2.528	2.112
272	9	-6.093	-4.329	1.176	-5.140	-3.939	1.200
311	5	-5.929	-3.850	1.559	-6.067	-3.034	3.033
467	10	-5.580	-3.683	1.871	-6.054	-3.555	2.499

To validate the credibility of the oligomers identified by our HTS process, we conducted computational simulations using the ORCA software, renowned for its precise electronic structure calculations. These simulations were crucial in verifying whether the selected oligomers conformed to the requisite energy levels for effective photovoltaic materials. The results from these simulations are quantitatively summarized by the MAE and the Root Mean Square Error (RMSE) between the predicted and calculated electronic properties as shown in **Table 5.4**. For the HOMO levels, the MAE was found to be 0.70 eV and the RMSE was 0.72 eV, indicating a close proximity between the predicted and simulated values. Similar accuracy was observed for the LUMO levels, with an MAE of 0.66 eV and an RMSE of 0.70 eV. The HOMO-LUMO gap predictions also showed a high degree of precision, with an MAE of 0.59 eV and an RMSE of 0.55 eV.

These low error values underscore the effectiveness of the ML models in forecasting the electronic properties of the screened oligomers. The consistency of the predictions with the computational results not only validates the screening process but also demonstrates the potential of the selected oligomers in photovoltaic material development.

Table 5.4 MAE and RMSE of screened conjugated oligomers.

	HOMO	LUMO	HOMO-LUMO gap
MAE	0.70	0.66	0.59
RMSE	0.72	0.70	0.55

5.4 Conclusion

In this study, we have addressed the challenge of data scarcity for conjugated oligomers by applying transfer learning within GNN frameworks to predict their electronic properties. Our results clearly demonstrate the superiority of transfer learning models over direct learning models, evidenced by a marked improvement in the prediction accuracy for critical electronic properties such as HOMO, LUMO, and the HOMO-LUMO gap. The MAE for these properties saw a significant reduction from 1.34, 0.68, 0.71 to 0.74, 0.46, 0.54, respectively, when using transfer learning approaches. Utilizing these enhanced models, we have established a HTS pipeline aiming at identifying potential candidates for OPV materials that exhibit the desired electronic properties. From our candidate dataset, a substantial number of potential oligomers were identified as promising materials for OPV applications. The validity of our screening has been substantiated through computational simulations, which confirmed the precision of our ML models. A notable advantage of this computational method is its efficiency, requiring significantly less time than traditional simulation techniques. The success of this research overcomes the challenges posed by limited data, paving the way for discovering new, promising materials for OPV and showcasing the potential of ML to revolutionize the field of material science.

References

[1] Chua, L.-L.; Zaumseil, J.; Chang, J.-F.; Ou, E. C.-W.; Ho, P. K.-H.; Sirringhaus, H.; Friend, R. H. General observation of n-type field-effect behaviour in organic semiconductors. *Nature* **2005**, 434 (7030), 194-199.

[2] Di, C. a.; Zhang, F.; Zhu, D. Multi-functional integration of organic field-effect transistors (OFETs): advances and perspectives. *Advanced Materials* **2013**, 25 (3), 313-330.

[3] Mei, J.; Diao, Y.; Appleton, A. L.; Fang, L.; Bao, Z. Integrated materials design of organic semiconductors for field-effect transistors. *Journal of the American Chemical Society* **2013**, 135 (18), 6724-6746.

[4] Riera-Galindo, S.; Leonardi, F.; Pfattner, R.; Mas-Torrent, M. Organic semiconductor/polymer blend films for organic field-effect transistors. *Advanced Materials Technologies* **2019**, 4 (9), 1900104.

[5] Bernede, J. Organic photovoltaic cells: History, principle and techniques. *Journal of the Chilean Chemical Society* **2008**, 53 (3), 1549-1564.

[6] Hains, A. W.; Liang, Z.; Woodhouse, M. A.; Gregg, B. A. Molecular semiconductors in organic photovoltaic cells. *Chemical Reviews* **2010**, 110 (11), 6689-6735.

[7] Nunzi, J.-M. Organic photovoltaic materials and devices. *Comptes Rendus Physique* **2002**, 3 (4), 523-542.

[8] Capelli, R.; Toffanin, S.; Generali, G.; Usta, H.; Facchetti, A.; Muccini, M. Organic light-emitting transistors with an efficiency that outperforms the equivalent light-emitting diodes. *Nature Materials* **2010**, 9 (6), 496-503.

[9] Pfeiffer, M.; Leo, K.; Zhou, X.; Huang, J.; Hofmann, M.; Werner, A.; Blochwitz-Nimoth, J. Doped organic semiconductors: Physics and application in light emitting diodes. *Organic Electronics* **2003**, 4 (2-3), 89-103.

[10] Qin, Z.; Gao, H.; Dong, H.; Hu, W. Organic light-emitting transistors entering a new development stage. *Advanced Materials* **2021**, 33 (31), 2007149.

[11] Brédas, J.-L.; Beljonne, D.; Coropceanu, V.; Cornil, J. Charge-transfer and energy-transfer processes in π-conjugated oligomers and polymers: a molecular picture. *Chemical Reviews* **2004**, 104 (11), 4971-5004.

[12] Gaylord, B. S.; Wang, S.; Heeger, A. J.; Bazan, G. C. Water-soluble conjugated oligomers: effect of chain length and aggregation on photoluminescence-quenching efficiencies. *Journal of the American Chemical Society* **2001**, 123 (26), 6417-6418.

[13] Meier, H.; Stalmach, U.; Kolshorn, H. Effective conjugation length and UV/vis spectra of oligomers. *Acta Polymerica* **1997**, 48 (9), 379-384.

[14] Pan, C.; Zhao, C.; Takeuchi, M.; Sugiyasu, K. Conjugated oligomers and polymers sheathed with designer side chains. *Chemistry–An Asian Journal* **2015**, 10 (9), 1820-1835.

[15] Hutchison, G. R.; Ratner, M. A.; Marks, T. J. Hopping transport in conductive heterocyclic oligomers: reorganization energies and substituent effects. *Journal of the American Chemical Society* **2005**, 127 (7), 2339-2350.

[16] Hutchison, G. R.; Ratner, M. A.; Marks, T. J. Intermolecular charge transfer between heterocyclic oligomers. Effects of heteroatom and molecular packing on hopping transport in organic semiconductors. *Journal of the American Chemical Society* **2005**, 127 (48), 16866-16881.

[17] Roldao, J. C.; Oliveira, E. F.; Milián-Medina, B.; Gierschner, J.; Roca-Sanjuán, D. Accurate calculation of excited-state absorption for small-to-medium-sized conjugated oligomers: multiconfigurational treatment vs quadratic response TD-DFT. *Journal of Chemical Theory and Computation* **2022**, 18 (9), 5449-5458.

[18] Liu, Z.; Lin, L.; Jia, Q.; Cheng, Z.; Jiang, Y.; Guo, Y.; Ma, J. Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *Journal of Chemical Information and Modeling* **2021**, 61 (3), 1066-1082.

[19] Lu, C.; Liu, Q.; Sun, Q.; Hsieh, C.-Y.; Zhang, S.; Shi, L.; Lee, C.-K. Deep learning for optoelectronic properties of organic semiconductors. *The Journal of Physical Chemistry C* **2020**, 124 (13), 7048-7060.

[20] Wang, X.; Zhao, Y.; Pourpanah, F. Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics* **2020**, 11, 747-750.

[21] Chauhan, N. K.; Singh, K. A review on conventional machine learning vs deep learning. In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, **2018**; IEEE: pp 347-352.

[22] Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, **2017**; pp 843-852.

[23] Lee, C.-K.; Lu, C.; Yu, Y.; Sun, Q.; Hsieh, C.-Y.; Zhang, S.; Liu, Q.; Shi, L. Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers. *The Journal of Chemical Physics* **2021**, 154 (2).

[24] Nakata, M.; Shimazaki, T. PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling* **2017**, 57 (6), 1300-1308.

[25] Becke, A. D. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *The Journal of Chemical Physics* **1992**, 96 (3), 2155-2160.

[26] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of Physical Chemistry* **1994**, 98 (45), 11623-11627.

[27] Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review* **1965**, 140 (4A), A1133.

[28] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, 28 (1), 31-36.

[29] Schustik, S. A.; Cravero, F.; Martinez, M. J.; Ponzoni, I.; Diaz, M. F. PolyMaS: A new software to generate high molecular weight polymer macromolecules from repeating structural units. *Polimery* **2021**, 66 (5), 293-297.

[30] Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, 8, 31.

[31] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, 3 (1), 1-14.

[32] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996**, 17 (5-6), 490-519.

[33] Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, 2 (1), 73-78.

[34] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, 148 (24).

[35] Rice, L.; Wong, E.; Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, **2020**; PMLR: pp 8093-8104.

Chapter 6

Conclusions and Recommendations

This chapter begins with a summary of the research findings detailed in the previous chapters. A serious of solutions, including the introduction of novel descriptors and the incorporation of transfer learning technology, are designed to address the challenges in the application of machine learning to the field of flexible electronics. Finally, the chapter will present prospective avenues for future research, inspired by the results and conclusions drawn from this thesis.

6.1 Conclusion

In the field of flexible electronics, the significance of both active and substrate materials is paramount, as they are fundamental in determing the performance of the devices.[1] As the realm of flexible electronics broadens to cover a vast array of applications, the need for specialized materials that can conform to the diverse requirements of these devices intensifies. The expansion of Artificial Intelligence (AI) has catapulted machine learning (ML) to the forefront as a transformative tool in the search and development of innovative materials for flexible devices. A range of methodologies has been developed to address the challenges associated with applying ML to the development of materials for flexible devices.

Among these innovations are structure-based descriptors, specifically engineered to address the complexities of polymers, such as elastomers, which are characterized by their elaborate structural intricacies. These descriptors have provided a means to overcome the limitations posed by traditional polymer descriptors, allowing for more accurate modeling and prediction of material properties. In addition, transfer learning has emerged as a critical technique to address the challenges brought about by the scarcity of data, a situation that is particularly prevalent in the study of conjugated oligomers. By leveraging pre-existing models and datasets, transfer learning enables the extrapolation of knowledge to new materials, thus enhancing the capability of ML models to make accurate predictions even with limited data.

These methodological advancements are not only facilitating the discovery of novel materials suited for flexible electronics but also streamlining the process, making it more efficient and less resource-intensive. This progress holds great promise for the future of material discovery in flexible device technology.

6.1.1 Structure-based Multilevel Descriptors

In our research, we have innovated a novel set of elastomer descriptors, the structure-based multilevel (SM) descriptors, solely based on the intrinsic molecular structure. These descriptors integrate simplified dimer representation (SDR) descriptors, sparse descriptors based on the soft segments mass, the ratios of various block, and the sparse descriptors based on the polymer mass. This integration forms a comprehensive suite of descriptors that capture both the intricate local and the overarching global structures of elastomers.

The study showcases the remarkable predictive ability of SM descriptors by accurately predicting key mechanical properties of elastomers, including toughness, critical strain, and Young's modulus. Impressively, the ML models achieved accuracy scores of 0.91, 0.89, and 0.87, respectively, for these properties. This high level of accuracy demonstrates the efficacy of SM descriptors in capturing crucial molecular features that influence the mechanical behavior of elastomers.

Furthermore, the development of machine learning-assisted HTS pipelines, based on the foundation of SM descriptors, revolutionizes the elastomer materials discovery process. These pipelines enable rapid and targeted screening of elastomers with specific mechanical properties, significantly reducing the time and resources required for materials research. This research contributes substantially to the field of materials discovery by providing a computationally efficient and user-friendly tool for predicting and designing elastomer properties.

6.1.2 Transfer Learning for Optoelectronic Properties Prediction of Conjugated Oligomers

This thesis has demonstrated the significant impact of transfer learning technology in addressing the challenges posed by insufficient data for conjugated oligomers. By utilizing knowledge from the source dataset, PubChemQC-100k, we equipped the pre-trained model with the fundamental understanding of materials' patterns. Subsequently, fine-tuning with the target dataset, OG-600, allowed the transfer learning model to accurately predict various electronic properties of conjugated oligomers, including Highest Occupied

Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO), and HOMO-LUMO gap. Notably, this approach outperformed direct learning models trained exclusively on OG-600, reducing Mean Absolute Errors (MAE) significantly.

The practical implications of our research became evident when we applied the trained transfer learning models to screen potential photovoltaic materials. From a pool of 3710 conjugated oligomers with polymerization degrees ranging from 4 to 10, our model identified 85 potential photovoltaic materials that met the criteria for photovoltaic materials. The validity of our screening has been substantiated through computational simulations, which confirmed the precision of our ML models.

Beyond its immediate application to conjugated oligomers, it is important to note that the challenge of insufficient data is not exclusive to this class of materials. Our research underscores the broader potential of transfer learning technology in addressing data scarcity issues across various polymer categories. This methodology can undoubtedly be extended to enhance materials discovery in other domains facing similar data limitations.

6.1.3 Summary of the Thesis

Conclusively, the projects in this thesis aim to apply ML technology to help find potential materials for flexible electronics. For first project, the elastomers that satisfy the specific mechanical properties are identified through the machine learning-assisted high-throughput screening (HTS) pipeline. The success of ML models relying on the development of SM descriptors. For the second project, the conjugated oligomers that can be potential photovoltaic materials are identified through the deep learning-assisted HTS. The success of the models relies on the integration of transfer learning technology.

6.2 Recommendations for Future Works

6.2.1 Application of SM Descriptors in Polymers Beyond Elastomers

The innovative use of SM descriptors has already marked a significant advancement in understanding and predicting the properties of elastomers. Building on this success, a compelling area for future research is the exploration of the applicability of SM descriptors across a broader spectrum of polymers beyond elastomers. This adaptability lies in the inclusion of the SDR descriptor within the SM descriptor. By using tailored SDR descriptors, our method could be applied to various types of polymers and their specific properties, offering our method considerable flexibility.

We have initiated some preliminary attempts. We applied our HTS framework to polymeric surfactants. We constructed an HTS pipeline to identify the potential polymeric surfactants with desired critical micelle concentration (CMC). Our dataset comprised 130 polymeric surfactants sourced from previous studies.[2-9] We categorized surfactants with a CMC below 1 mmol/L as "high-efficient" and others as "low-efficient". Using the SVC algorithm and a variety of SM descriptors, including SM-MF, SM-PF, SM-R, SM-P, and a novel SM-HCM descriptor, we developed predictive models. For the SM-HCM descriptor, the SDR was encoded using HCM descriptors previously proposed by our team.[10] The performance of these models is depicted in **Figure 6.1**. Obviously, the SM-HCM model outperforms other models across various metrics. This superiority can be attributed to the molecular energy information of HCM descriptors, considering the energetically driven nature of the micellization process.

Our candidate dataset comprises 176 entries, constructed by combining block mass, polymer mass, hydrophilic groups, and hydrophobic groups. All candidates incorporate the same hydrophilic group as our collected dataset only contain one such hydrophilic group. Utilizing the most accurate model in our study, the SM-HCM model, we constructed a HTS pipeline to identify the potential "high-efficient" polymeric surfactants. **Figure 6.2** illustrates the probability of candidates being classified as "high-efficient" surfactant. In the heatmap, each grid represents individual candidate, and each column represents one combination of hydrophilic and hydrophobic group (from c1 to c11). Obviously, c9, c10, and c11 are the top three combinations demonstrating the highest probability for forming "high-efficient" surfactant. Details of these three combinations can be found in **Table 6.1**.

Notably, the investigation on polymeric surfactants, is part of our broader, ongoing research focusing on the CMC of surfactants. We are preparing another manuscript that includes this work and other work focusing on the small molecular surfactants to demonstrate a comprehensive and in-depth study in the surfactant domain.

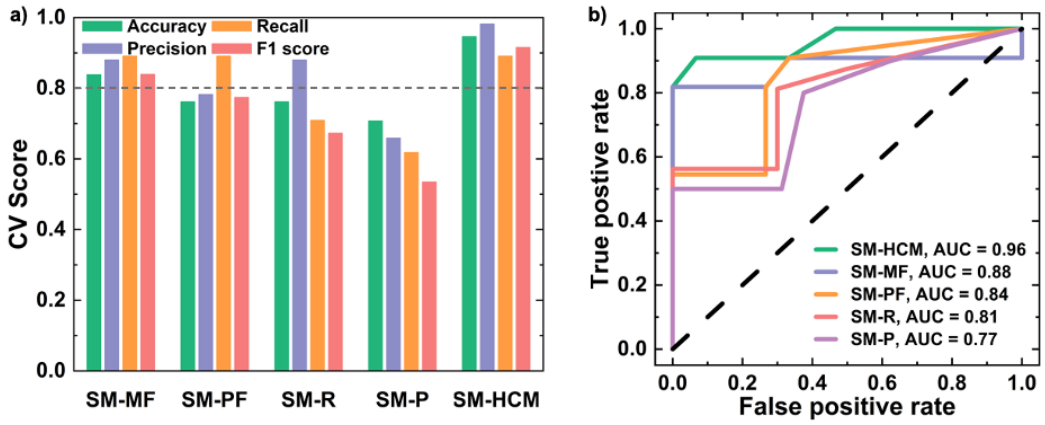


Figure 6.1. Model performance using different descriptors. (a) Five-fold cross-validation scores for models trained using five different descriptors (SM-MF, SM-PF, SM-R, SM-P, and SM-HCM) and SVC algorithm. The grey dashed line represents the accuracy of 0.8. (b) ROC curves of models trained using the five different descriptors.

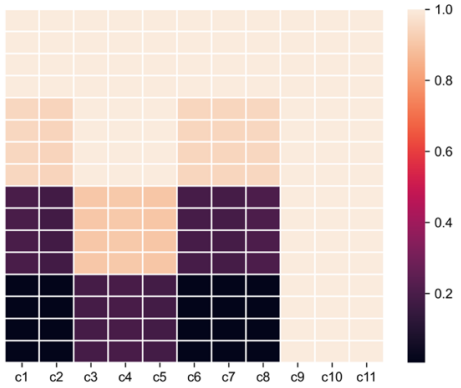
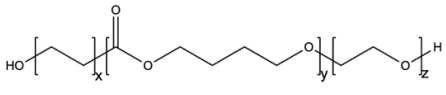
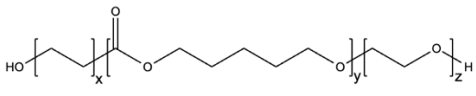
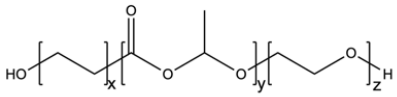


Figure 6.2 Probability of candidate dataset with 176 entries being classified as "high-efficient" surfactants. Each grid represents a candidate, and each column represents a distinct combination of hydrophilic and hydrophobic group. The color of each gird indicates the probability of the corresponding candidate being classified as "high-efficient" surfactant.

Table 6.1 The structure of the top three combinations.

Combinations	Structure
c9	
c10	
c11	

The application in polymeric surfactants demonstrates the extension of our HTS framework. By tailoring the SDR descriptors, like using the HCM descriptor in this case, we can address diverse polymers and their associated properties.

In summary, the future work should be directed towards the applicability of SM descriptors across a broader spectrum of polymers, emphasizing their potential to revolutionize our understanding and utilization of these materials. Such research not only promises to enhance our material science capabilities but also aligns with the broader goals of sustainable and innovative material development.

6.2.2 Knowledge-infused Algorithm Development

In this thesis, while traditional ML and deep learning approaches have made substantial contributions, they often operate primarily on data-driven principles, sometimes overlooking the rich domain-specific knowledge that exists in materials science. Future research could focus on developing and applying knowledge-infused algorithms, which incorporate domain-specific expertise, theoretical foundations, and empirical insights directly into the learning process.

We have initiated some preliminary attempts. Specifically, we have adapted the SchNet algorithm by incorporating an energy contribution factor, which call SchNetE. This algorithm utilizes a single-atom energy contribution (E) to provide a more accurate depiction of molecular systems. Within the SchNetE framework, a unique description of an atomistic system can be generated through a set of n atom sites with nuclear charges $Z = (Z_1, \dots, Z_n)$, single-atom energy contribution $E = (E_1, \dots, E_n)$, and positions $R = (r_1, \dots, r_n)$. The single-atom contribution E is a novel component introduced in SchNetE, replaceable with a single-atom volume contribution (V) or any other feature offering additional physical-chemical information. This adaptation endows SchNetE with an expanded descriptor set for the molecular system, allowing for a more accurate model. A graphical overview of the SchNetE algorithm is depicted in **Figure 6.3**.

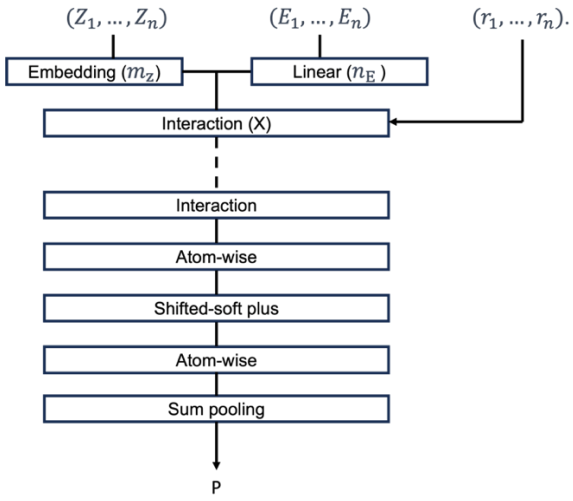


Figure 6.3 Overview of the SchNetE algorithm

6.2.3 Laboratory to Real-World Application Transition

Translating theoretical models into practical applications involves rigorous testing of materials under real-world conditions. This step is vital for verifying model predictions and identifying areas for improvement. Collaboration with industry partners is essential to understand practical constraints and requirements. Testing materials in actual product

environments and assessing scalability and economic feasibility are crucial to ensure the materials developed are not only high-performing but also viable for commercial use.

References

[1] Chen, X.; Rogers, J. A.; Lacour, S. P.; Hu, W.; Kim, D.-H. Materials chemistry in flexible electronics. *Chemical Society Reviews* **2019**, 48 (6), 1431-1433.

[2] Yang, Y.-W.; Yang, Z.; Zhou, Z.-K.; Attwood, D.; Booth, C. Association of triblock copolymers of ethylene oxide and butylene oxide in aqueous solution. A study of B n E m B n copolymers. *Macromolecules* **1996**, 29 (2), 670-680.

[3] Attwood, D.; Booth, C.; Yeates, S. G.; Chaibundit, C.; Ricardo, N. M. Block copolymers for drug solubilisation: Relative hydrophobicities of polyether and polyester micelle-core-forming blocks. *International Journal of Pharmaceutics* **2007**, 345 (1-2), 35-41.

[4] Zana, R.; Marques, C.; Johner, A. Dynamics of micelles of the triblock copolymers poly (ethylene oxide)–poly (propylene oxide)–poly (ethylene oxide) in aqueous solution. *Advances in Colloid and Interface Science* **2006**, 123, 345-351.

[5] Sakai, T.; Alexandridis, P. Mechanism of gold metal ion reduction, nanoparticle growth and size control in aqueous amphiphilic block copolymer solutions at ambient conditions. *The Journal of Physical Chemistry B* **2005**, 109 (16), 7766-7777.

[6] Croy, S.; Kwon, G. Polymeric micelles for drug delivery. *Current Pharmaceutical Design* **2006**, 12 (36), 4669-4684.

[7] Khimani, M.; Patel, H.; Patel, V.; Parekh, P.; Vekariya, R. L. Self-assembly of stimuli-responsive block copolymers in aqueous solutions: An overview. *Polymer Bulletin* **2020**, 77, 5783-5810.

[8] Yang, Z.; Pousia, E.; Heatley, F.; Price, C.; Booth, C.; Castelletto, V.; Hamley, I. W. Solubilization of alkylcyanobiphenyls in aqueous micellar solutions of a diblock copolymer of propylene oxide and ethylene oxide. *Langmuir* **2001**, 17 (7), 2106-2111.

[9] Kelarakis, A.; Havredaki, V.; Yu, G.-E.; Derici, L.; Booth, C. Temperature dependences of the critical micelle concentrations of diblock oxyethylene/oxybutylene copolymers. A case of athermal micellization. *Macromolecules* **1998**, 31 (3), 944-946.

[10] Chen, C.; Liu, D.; Deng, S.; Zhong, L.; Chan, S. H. Y.; Li, S.; Hng, H. H. Accurate machine learning models based on small dataset of energetic materials through spatial matrix featurization methods. *Journal of Energy Chemistry* **2021**, 63, 364-375.