UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos

Predicting Properties of Polymer Materials Using Machine Learning Methods

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Elaheh Kazemi Khasragh
M.Sc. in Materials Science and Engineering

Madrid, 2024

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos

Doctoral Degree in Engineering of Structures, Foundations and Materials

Predicting Properties of Polymer Materials Using Machine Learning Methods

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Elaheh Kazemi Khasragh
M.Sc. in Materials Science and Engineering

Under the supervision of:
Dr. Maciej Haranczyk
Dr. Carlos González

Madrid, 2024

Title: Predicting Properties of Polymer Materials Using Machine Learning Methods

Author: Elaheh Kazemi Khasragh

Doctoral Programme: Engineering of Structures, Foundations and Materials

Thesis Supervision:

Dr. Maciej Haranczyk, senior researcher, IMDEA Materials Institute, Madrid, Spain (Supervisor)

Dr. Carlos González, full professor, Technical University of Madrid, Madrid, Spain (Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

"This thesis has been partially supported by Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación under reference number PRE2019-090286"

Dedicated to my beloved family

Acknowledgement

I am profoundly grateful to a number of people whose unwavering encouragement and assistance have been instrumental in the completion of this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Maciej Haranczyk and Dr. Carlos González, for their invaluable guidance, insightful advice, and continuous mentorship throughout this journey. Your expertise and encouragement have been the pillars of my research.

A heartfelt thank you to Dr. Rocío Mercado Oropeza, Dr. Juan Pedro Fernández and Dr. David Garoz Gómez for their crucial assistance and contributions to my project. Your help and dedication have been indispensable, and I am deeply appreciative of your input.

I would also like to extend my sincere thanks to my incredible teammates, Miguel Hernández, Burcu Ozdemir, Sergei Zorkaltsev, Dr. Christina Schenk, and Dr. Phuong Thúy Vo. Your friendship, positivity, and unwavering enthusiasm have made this journey enjoyable and memorable. Thank you for bringing joy and happiness into our collaborative efforts. I would also like to thank Dr. Afshin Pendashteh for being a good friend. I am deeply grateful to everyone at IMDEA Materials for creating such a peaceful and inspiring environment to work in. Your collective efforts have provided me with the perfect setting to focus and thrive in my research.

To my family, words cannot express how grateful I am for your love, encouragement, and sacrifices. To my dear Maman, your boundless love and nurturing have been my driving force. To my hero Baba, thank you for being the best father in the world and for all the wisdom and strength you have imparted. To my only brother, Amir, thank you for always being there for me, for your friendship, for visiting, and for your understanding and support.

Lastly, but most importantly, to my beloved husband, Farid, your endless love, understanding, and patience have been my rock. Thank you for standing by me through every high and low, and for your unwavering belief in me. I could not have done this without you.

To everyone mentioned and to all those who have supported me in one way or another, thank you from the bottom of my heart. This achievement is as much yours as it is mine.

Abstract

Accurate prediction of polymer properties is essential for their effective application and development. However, traditional experimental methods are often prohibitively time-consuming and expensive. This thesis employs an integrated approach to predict various polymer properties, including mechanical, rheological, and thermal characteristics. By combining group interaction modelling, molecular dynamics methods, and machine learning, we predicted 14 distinct physical properties of both homo-polymers and binary copolymers with varying monomer compositions.

Group interaction modelling was utilized to calculate properties such as glass transition temperature, heat capacity, elastic modulus, linear thermal expansion, and Poisson's ratio for homo-polymers. The accuracy of this method was found to depend heavily on the precision of input parameters, such as the Debye temperature. A substantial portion of this study employed machine learning techniques to predict polymer properties. For homo-polymers with sufficiently large datasets, we applied random forest algorithms. Molecular descriptors derived from chemical structures were used to train the model. This approach yielded robust predictive performance, with squared correlation coefficient values ranging from 0.83 to 0.955 across six different properties of homo-polymers. Our results demonstrated that machine learning models outperformed group interaction modelling, highlighting the superior reliability of machine learning approaches.

One challenge in extending the machine learning approach was the lack of sufficient datasets. To address this, we employed transfer learning to predict properties of homo-polymers with smaller datasets. A neural network initially trained on heat capacity at constant pressure was adapted to predict four additional properties: specific heat capacity at constant volume, shear modulus, flexural strength, and dynamic viscosity. Despite the small dataset sizes (ranging from 13 to 18 samples), the transfer learning models achieved high accuracy, illustrating the effectiveness of transfer learning in leveraging limited data for diverse property predictions.

Further expanding our research, we explored the domain of copolymers, which present a broad range of chemical compositions and potential applications. By integrating molecular dynamics simulations with machine learning, we predicted seven physical properties of 140 binary copolymers with varying monomer compositions. Using random forest models with molecular descriptors and graph neural networks with graph representations, we found that random forest models excelled in predicting properties such as density and heat capacity at constant pressure and volume, while graph neural networks outperformed in predicting properties like volume thermal expansion, density, and bulk modulus. This dual approach underscores the importance of selecting appropriate molecular representations for accurate property predictions.

Collectively, these studies demonstrate the efficacy of combining machine learning, transfer learning, group interaction modelling, and molecular dynamics to predict polymer properties both accurately and efficiently. This thesis offers a comprehensive framework for accelerating polymer design and application through advanced computational methods, paving the way for future innovations in material science.

Resumen

La predicción precisa de las propiedades de los polímeros es esencial para su aplicación y desarrollo efectivo. Sin embargo, los métodos experimentales tradicionales suelen ser excesivamente lentos y costosos. Esta tesis emplea un enfoque integrado para predecir varias propiedades de los polímeros, incluidas las características mecánicas, reológicas y térmicas.

Combinando el modelado de interacciones de grupo, métodos de dinámica molecular y aprendizaje automático, hemos predicho 14 propiedades físicas diferentes de homo-polímeros y copolímeros binarios con diversas composiciones de monómeros.

Se utilizó el modelado de interacciones de grupo para calcular propiedades como la temperatura de transición vítrea, la capacidad calorífica, el módulo elástico, la expansión térmica lineal y el coeficiente de Poisson de los homo-polímeros. Se observó que la precisión de este método depende en gran medida de la exactitud de los parámetros de entrada, como la temperatura de Debye. Una parte sustancial de este estudio empleó técnicas de aprendizaje automático para predecir las propiedades de los polímeros. Para los homo-polímeros con bases de datos suficientemente grandes, aplicamos algoritmos de random forest. Se utilizaron descriptores moleculares derivados de las estructuras químicas para entrenar el modelo, logrando un rendimiento predictivo robusto, con valores del coeficiente de correlación cuadrado que oscilaban entre 0.83 y 0.955 en seis propiedades diferentes de los homo-polímeros. Nuestros resultados demostraron que los modelos de aprendizaje automático superaron al modelado de interacciones de grupo, subrayando la mayor fiabilidad de los enfoques de aprendizaje automático.

Un desafío en la extensión del aprendizaje automático fue la falta de bases de datos suficientes. Para resolver este problema, empleamos el aprendizaje por transferencia para predecir las propiedades de homo-polímeros con bases de datos más pequeñas. Una red neuronal inicialmente entrenada para la capacidad calorífica a presión constante se adaptó para predecir cuatro propiedades adicionales: capacidad calorífica a volumen constante, módulo de corte, resistencia a la flexión y viscosidad dinámica. A pesar del pequeño tamaño de los bases de datos (que oscilaban entre 13 y 18 muestras), los modelos de aprendizaje por transferencia lograron una alta precisión, ilustrando la efectividad del aprendizaje por transferencia para aprovechar datos limitados en la predicción de diversas propiedades.

Ampliando aún más nuestra investigación, exploramos el dominio de los copolímeros, que presentan una amplia gama de composiciones químicas y aplicaciones potenciales. Al integrar simulaciones de dinámica molecular con aprendizaje automático, predijimos siete propiedades físicas de 140 copolímeros binarios con diversas composiciones de monómeros. Utilizando modelos de random forest con descriptores moleculares y redes neuronales de grafos con representaciones de grafos, encontramos que los modelos de random forest sobresalieron en la predicción de propiedades como la densidad y la capacidad calorífica a presión y volumen constantes, mientras que las redes neuronales de grafos fueron superiores para predecir propiedades como la expansión térmica volumétrica, densidad y el módulo de compresibilidad. Este enfoque dual subraya la importancia de seleccionar representaciones moleculares adecuadas para una predicción precisa de propiedades.

En bases, estos estudios demuestran la eficacia de combinar aprendizaje automático, aprendizaje por transferencia, modelado de interacciones de grupo y dinámica molecular para predecir las propiedades de los polímeros de manera precisa y eficiente. Esta tesis ofrece un marco integral para acelerar el diseño y la aplicación de polímeros mediante métodos computacionales avanzados, allanando el camino para futuras innovaciones en la ciencia de materiales.

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

UPM	Universidad Politécnica de Madrid
GIM	Group Interaction Modelling
ML	Machine Learning
MD	Molecular Dynamics
T_g	Glass transition temperature
C_p	Heat capacity in constant pressure
E	Elastic modulus
α	Linear thermal expansion
ν	Poisson's ratio
TL	Transfer Learning
C_v	Heat capacity in constant volume
G	Shear modulus
σ	Flexural strength
η	Dynamic viscosity
ρ	Density
R_g	Radius of gyration
γ	Volume thermal expansion
K	Bulk modulus
RF	Random Forest
NN	Neural Network
MAE	Mean absolute error
MSE	Mean squared error
GNNs	Graph neural networks

R^2	Squared correlation coefficient
DSC	Differential scanning calorimetry
DMA	Dynamic mechanical analysis
D-MPNN	Directed message passing neural network
GAFF	General Amber force field
HF	Hartree-Fock
wD-MPNN	Weighted, directed message passing neural network
DFT	Density functional theory
GCNN	Graph convolutional neural networks
FFNN	Feed-forward neural network
$\hat{H}$	Hamiltonian operator
$\Psi(\mathbf{r})$	Wave function
θ	Debye temperature
GMM	Gaussian mixture models
AIS	Artificial immune systems
PVT	Pressure-volume-temperature
E_{coh}	Cohesive energy
$C_{v,\ skeletal}$	Heat capacity contributed by skeletal vibration
V	Volume
H_M	Mechanical energy
H_T	Thermal energy
H_C	Contribution energy
ν_1	Skeletal frequency
ν_E	Group vibrations frequencies
θ_E	Einstein temperature
$C_{v,\ group}$	Heat capacity contributed by group vibration

$N_{skeletal}$	Number of skeletal vibrations
N_{group}	Number of group vibrations
B_{gam}	Bulk modulus of the polymers in the glassy states
S	Gaussian function of loss peak
$\tan\Delta_g$	Cumulative loss tangent for the glass transition

Chapter 1

Introduction

1.1 The Role of Polymers in Industrial Applications

Polymers are long chain molecules composed of repeating structural units called monomers, which are bonded together. The word "polymer" was introduced by the Swedish chemist J. J. Berzelius. Polymeric materials have been successfully employed in a wide spectrum of applications that virtually encompass every human life. The increase in their production is driven, in addition to social factors, by the need to substitute traditional materials [Feldman, 2008; Khan and Roy, 2018; Nicholson, 1991; Young and Lovell, 2011]. They can be categorized in various ways, including by their origin, thermal behaviour, composition, and structure, as illustrated in Figure 1.1.

Polymers are macromolecules with a wide array of properties and structures, making them highly adaptable for numerous applications. From structural materials in construction to biodegradable packaging and cutting-edge medical uses, polymers demonstrate exceptional versatility, serving a broad range of industries globally. Figure 1.2 illustrates the diverse applications of polymers in key sectors such as aerospace, automotive, textiles, and electronics. By engineering these materials with specific attributes like strength, flexibility, conductivity, and biocompatibility, polymers have become integral to advancements in various fields. As technology progresses, the development and optimization of new polymers continue to expand their potential uses and enhance their performance in established industries [Audus and De Pablo, 2017; Khan and Roy, 2018; LLorca et al., 2011].

To leverage the vast usage of polymers in industry, there is a crucial need to precisely control the properties of these materials according to specific applications. Key properties such as mechanical strength, rheological behaviour, and thermal characteristics play pivotal roles in determining their suitability for different industrial applications. Mechanical properties are critical in sectors requiring materials to withstand varying loads and environmental conditions, such as construction and automotive manufacturing. Rheological properties, which govern how polymers flow under stress, are crucial in manufacturing processes like injection moulding and extrusion, ensuring optimal product quality and consistency. Thermal properties, including parameters such as specific heat capacity, thermal conductivity, and

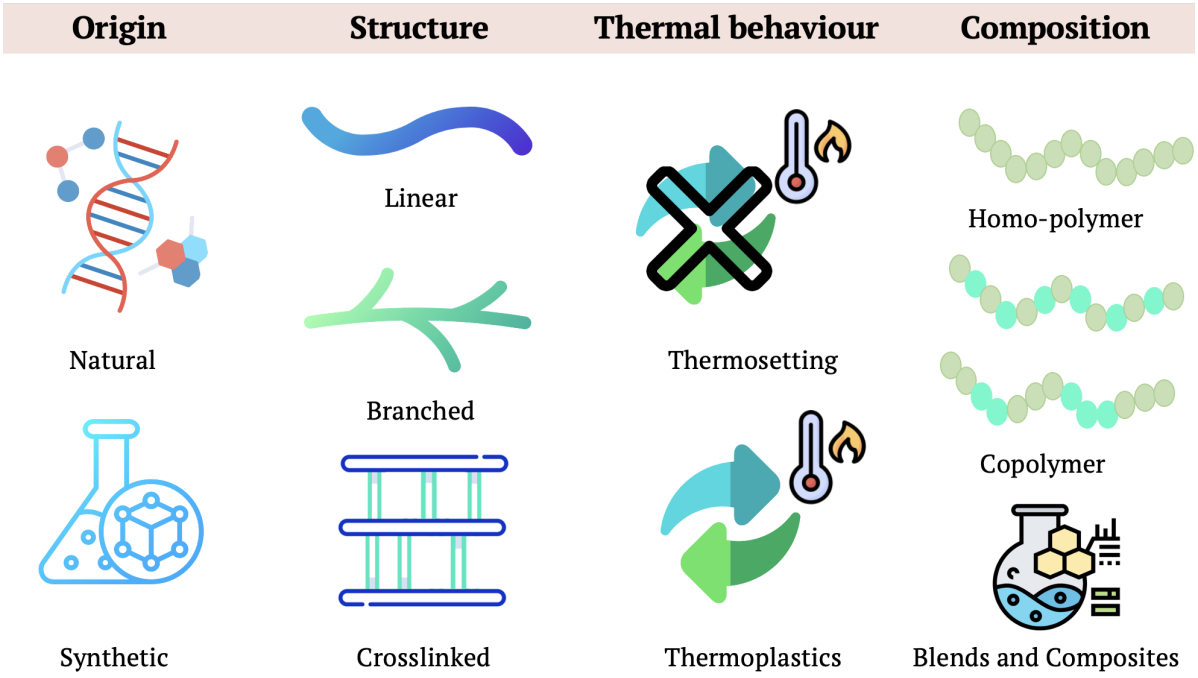

Figure 1.1: Schematic representation of polymer categorization based on various criteria.

Figure 1.2: Applications of polymers across various industries.

glass transition temperature, are vital in industries where materials must perform reliably across wide temperature ranges, such as electronics and aerospace.

These properties can be measured using advanced techniques such as differential scanning calorimetry (DSC), dynamic mechanical analysis (DMA), and rheometry [Brown, 2001; Kumar et al., 2019; Landel and Nielsen, 1993; Macosko, 1994; Menard and Menard, 2020; Sadiku-Agboola et al., 2011]. However, the complexity of polymer structures and behaviours poses challenges in accurate characterization and optimization. The inherent complexity of these materials, stemming from their varied properties, often obstructs traditional experimental design methods, which are typically resource-intensive and time-consuming. Moreover, variability in polymer composition from different sources adds another layer of difficulty to the characterization and optimization processes [Audus and De Pablo, 2017; Khan and Roy, 2018; LLorca et al., 2011].

In addition to the complexities associated with characterizing new systems, another bottleneck arises from the large chemical and configuration space accessible to polymers. Consider a linear copolymer chain composed of n potential monomers and m chemical moieties. The total number of unique sequences that can be expressed in this copolymer can be expressed as $m^{n/2}$. The division by 2 in the exponent accounts for eliminating the double-counting of polymer sequences, where a sequence and its reverse represent the same copolymer. For instance, in the case of an AB-type copolymer with two chemical moieties and a chain length of 50, the total combinations amount to 2^{49}, surpassing 10^{15}. Consider that many copolymers possess more than two chemical moieties and several hundreds of monomer units, so navigating through this expansive space is like searching for a needle in a haystack [Moosavi et al., 2020; Patra, 2021; Pilania, 2021; Sattari et al., 2021]. To address this complexity, efficient tools are essential for navigating through this chemical space and to minimize the number of property measurements needed to complete a design task within a reasonable time frame.

1.2 Role of Modeling

Over the past several decades, significant effort has been devoted to developing mathematical and numerical models for predicting material properties based on their microstructure and chemical structures. Establishing accurate links between these models and the properties of polymeric materials can simplify their design and application processes. By improving predictions from both microstructural data and chemical structure, we can enhance the efficiency and effectiveness of material development.

1.2.1 Physically based Multi-scale Models

Establishing a relation between the properties and microstructure of polymers is not a simple task. The difficulty arises from the wide range of length scales involved in the characterization of polymeric materials, as demonstrated in Figure 1.3 [Kanouté et al., 2009; Y. Li et al., 2013].

As shown in Figure 1.3, the typical length of a single monomer is a nanometer, and the size of each polymer chain is between 10 and 100 nanometers. Also, the size of the polymeric network is around 1 to 100 micrometres, and bulk polymeric material is 1mm. These multiple scales of

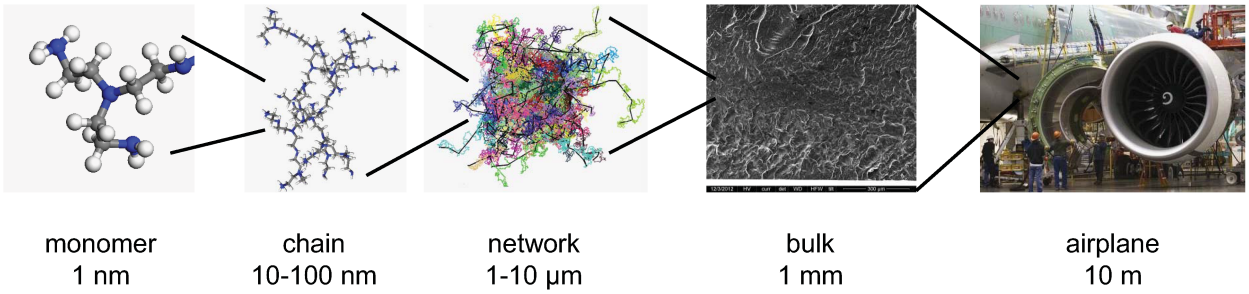

Figure 1.3: Hierarchical length scales for polymeric materials [Y. Li et al., 2013].

polymeric materials and their interdependence among each other in terms of system behaviour (i.e., bulk behaviour depends on the behaviour of individual polymer chains, and so forth) make it crucial to develop a multiscale modelling technique that can appropriately describe the hierarchy of scales if we wish to connect molecular structure with macroscopic mechanical properties [Gooneie et al., 2017; Kanouté et al., 2009; Y. Li et al., 2013]. Multiscale modelling concerns the derivation of equations, parameters, or simulation algorithms that describe behaviour at a given length scale based on the physics at a finer scale, provided that the fine-scale physics and structure are better understood than those at a coarser scale. The fine scale may include electrons, atoms, molecules, and their assemblies, or mesoscale structures such as phases or grains [Fish et al., 2021]. There are several common challenges in multiscale modelling, including selecting tools for programming and executing multiscale simulations. In particular, a large number of modelling tools capable of accurately predicting the formation, structure, and properties of materials are available nowadays. The quantum mechanics and density functional theory, molecular dynamics (thermodynamics, crystalline structure, defect interactions), Monte Carlo methods (kinetics), computational thermodynamics (phase diagrams), phase-field modelling (phase transformation and microstructure development), are advanced ways for multiscale modelling [LLorca et al., 2011]. In this research work, using the multiscale modeling technique, numerous computational methods and theories were developed primarily at smaller scales. These methods, which are demonstrated in Figure 1.4, were designed to predict and optimize the structures and properties of polymers across different scales, with a particular emphasis on the micro and molecular levels.

Electronic structure methods

The smallest scale in our research framework focuses on electronic structure methods. These methods are essential for analyzing the electronic properties and interactions at the atomic level. By employing methods such as density functional theory (DFT), researchers can gain detailed insights into the behaviour of electrons within molecules and solids. DFT does provide valuable information about electronic density, molecular orbitals, and potential energy surfaces. These are key outputs of DFT calculations and are crucial for understanding the behaviour of electronic systems. The method offers a computationally efficient alternative to traditional quantum mechanical approaches, particularly for systems with many electrons. The time-independent Schrödinger equation is a fundamental equation in quantum mechanics that describes how the quantum state of a physical system is determined by its energy. The

general form of the time-independent Schrödinger equation is presented in equation 1.1.

$$\hat{H}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{1.1}$$

where $\hat{H}$ is the Hamiltonian operator, representing the total energy of the system, $\Psi(\mathbf{r})$ is the wave function, which contains information about the probability distribution of particles, and E is the energy eigenvalue associated with the wave function.

While equation 1.1 provides a complete description of the system, solving it directly for systems with many interacting electrons is impractical due to the exponential scaling of the wave function with the number of electrons. DFT simplifies the problem by shifting the focus from the many-electron wave function to the electron density. Instead of solving for the full wave function, which scales exponentially with the number of electrons, DFT allows researchers to solve a set of equations based on electron density, which depends only on three spatial coordinates, regardless of the number of electrons.

The foundation of density functional theory is built upon two crucial mathematical theorems established by Hohenberg and Kohn [Hohenberg and Kohn, 1964b], along with a set of equations derived by Kohn and Sham [Kohn and Sham, 1965] in the mid-1960s.

The first theorem, established by Hohenberg and Kohn, asserts that the ground-state energy obtained from the Schrödinger equation is a unique functional of the electron density. This means that there is a one-to-one correspondence between the ground-state wave function and the ground-state electron density. To fully understand the significance of this theorem, it's important to grasp the concept of a "functional."

A functional is a concept closely related to a function. While a function takes one or more variables and assigns a single value to them, a functional operates on a function and returns a single value. For example, consider the function $f(x) = x^2 + 1$. In contrast, a functional might look like this:

$$F[f] = \int_{-1}^{1} f(x)\, dx$$

Here, $F[f]$ is a functional that integrates the function $f(x)$ over a certain domain. If we apply this functional to the function $f(x) = x^2 + 1$, the result is $F[f] = \frac{8}{3}$. The first theorem by Hohenberg and Kohn can therefore be restated: the ground-state energy E is a functional of the electron density $n(\mathbf{r})$. This relationship is the basis for what is called density functional theory. Another way to express the result of Hohenberg and Kohn's theorem is that the ground-state electron density uniquely determines all properties of the system, including both the energy and the wave function of the ground state. This result is significant because it allows us to approach the problem of solving the Schrödinger equation by focusing on the electron density, which is a function of just three spatial variables, instead of the wave function, which depends on 3N variables for a system with N electrons. In practical terms, for a nanocluster of 100 Pd atoms, this reduces the problem from one involving over 23,000 dimensions to one involving just three dimensions.

However, while the first Hohenberg-Kohn theorem guarantees the existence of a functional that can be used to determine the ground-state energy from the electron density, it does not specify the form of this functional. The second Hohenberg-Kohn theorem provides further insight by stating that the correct electron density is the one that minimizes the total energy functional. If the exact form of this functional were known, we could determine the ground-state energy by adjusting the electron density until the energy is minimized, thereby solving for the true electron density. In practice, this variational principle is used with approximate functionals to find solutions.

One effective way to express the functional described by the Hohenberg-Kohn theorem is through the single-electron wave functions, $\Psi_i(\mathbf{r})$. The energy functional can be represented as:

$$E[\Psi_i] = E_{\text{known}}[\Psi_i] + E_{\text{XC}}[\Psi_i] \tag{1.2}$$

In the decomposition of the functional, we can separate it into two main components: $E_{\text{known}}[\Psi_i]$, which includes terms that can be expressed in a straightforward analytical form, and E_{XC}, which represents the remaining contributions. The "known" terms comprise four distinct components, as outlined in equation 1.3.

These components, in sequence, are the kinetic energy of electrons, the Coulomb interactions between electrons and nuclei, the Coulomb interactions among pairs of electrons, and the Coulomb interactions among pairs of nuclei. The term $E_{\text{XC}}[\Psi_i]$, known as the exchange-correlation functional, captures all the quantum mechanical effects that are not addressed by the "known" terms.

$$E_{\text{known}}[\Psi_i] = -\frac{\hbar^2}{2m}\sum_i \int \Psi_i^*(\mathbf{r})\nabla^2\Psi_i \, d^3\mathbf{r} + \int V(\mathbf{r})\, n(\mathbf{r})\, d^3\mathbf{r} + \frac{e^2}{2}\int\int \frac{n(\mathbf{r})\, n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3\mathbf{r}\, d^3\mathbf{r}' + E_{\text{ion}} \tag{1.3}$$

For the moment, let's assume that we can define the exchange-correlation energy functional in a practical manner. What then is involved in finding the minimum energy solution for the total energy functional? Up to this point, nothing ensures that this process is simpler than the challenging task of solving the Schrödinger equation for the entire wave function.

This challenge was addressed by Kohn and Sham, who demonstrated that the problem of determining the correct electron density can be reformulated. Specifically, they showed that this task can be approached by solving a set of equations where each equation pertains to an individual electron.

The Kohn-Sham equations, which are detailed in equation 1.4, represent this approach.

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathrm{r}) + V_H(\mathrm{r}) + V_{XC}(\mathrm{r})\right]\Psi_i(\mathbf{r}) = \epsilon_i\Psi_i(\mathbf{r}) \tag{1.4}$$

These equations bear a resemblance to equation 1.1, but with a notable distinction. Unlike

the full Schrödinger equation, the Kohn-Sham equations do not include the many-body summations. This simplification arises because the Kohn-Sham approach involves single-electron wave functions, $\Psi_i(\mathbf{r})$, which depend only on three spatial coordinates.

In the Kohn-Sham equations, the left-hand side features three types of potentials: V, $V_H(\mathbf{r})$, and V_{XC}. The potential V is included in the "known" part of the total energy functional described earlier (equation 1.3). This potential represents the interaction between an electron and the surrounding atomic nuclei.

The potential $V_H(\mathbf{r})$, known as the Hartree potential, is defined as follows:

$$V_H(\mathbf{r}) = e^2 \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d^3\mathbf{r}' \tag{1.5}$$

This potential accounts for the Coulomb repulsion between the electron considered in one of the Kohn-Sham equations and the total electron density arising from all electrons in the system.

The Hartree potential, $V_H(\mathbf{r})$, includes an artefact known as the self-interaction term. This occurs because the electron described in the Kohn-Sham equation is also part of the overall electron density. Consequently, a portion of V_H reflects the Coulomb interaction between the electron and itself, which is unphysical.

To address this self-interaction issue and other related effects, these corrections are incorporated into the final potential used in the Kohn-Sham equations, denoted as $V_{XC}(\mathbf{r})$. This exchange-correlation potential captures both exchange and correlation effects. Formally, $V_{XC}(\mathbf{r})$ is defined as the functional derivative of the exchange-correlation energy functional with respect to the electron density:

$$V_{XC}(\mathbf{r}) = \frac{\delta E_{XC}(r)}{\delta(\mathbf{r})} \tag{1.6}$$

The mathematical concept of a functional derivative is more nuanced than the usual notion of a derivative of a function. Conceptually, however, it can be viewed similarly to a standard derivative. The notation δ is used in place of d to highlight that it is not identical to a conventional derivative.

If you sense that there might be a circular issue in our discussion of the Kohn-Sham equations, you're right. To solve these equations, we need the Hartree potential, which in turn depends on knowing the electron density. Yet, to determine the electron density, we need the single-electron wave functions, which are derived from solving the Kohn-Sham equations. To resolve this circular dependency, the problem is typically approached through an iterative process, as outlined in the following steps:

1. Start with an initial guess for the electron density, $n(\mathbf{r})$.
2. Solve the Kohn-Sham equations using this trial density to obtain the single-electron wave functions, $\Psi_i(\mathbf{r})$.

3. Compute a new electron density, $n_{KS}(\mathbf{r})$, from the wave functions obtained in the previous step using the equation:

$$n_{KS}(\mathbf{r}) = 2\sum_{i} \Psi_i(\mathbf{r})\Psi_i^*(\mathbf{r}) \tag{1.7}$$

   where $\Psi_i(\mathbf{r})$ are the Kohn-Sham orbitals and the factor of 2 accounts for spin degeneracy.

4. Compare the newly calculated density $n_{KS}(\mathbf{r})$ with the initial trial density $n(\mathbf{r})$. If they match, this density represents the ground-state electron density, which can then be used to calculate the total energy. If there is a discrepancy, update the trial electron density and repeat the process from step 2 [Dakin and Brown, 2006; Hohenberg and Kohn, 1964a; Kohn and Sham, 1965; Levy, 1979; Sholl and Steckel, 2022; Ziegler, 1991].

DFT is pivotal in predicting fundamental properties such as bond strengths, reaction mechanisms, and electronic transitions. This atomic-level understanding is crucial for designing and optimizing materials with specific desired characteristics. In materials science, these methods are used to explore and predict the properties of a wide range of materials, including polymers, metals, and semiconductors [Cramer, 2013; Shivaleela et al., n.d.; Szabo and Ostlund, 2012].

In the field of computational chemistry, DFT is a widely utilized method for predicting vibrational frequencies of molecules [Koch and Holthausen, 2015]. Ming [Wong, 1996] used DFT methods, specifically comparing several functionals, to calculate the vibrational frequencies of a set of 122 molecules.

In this research, we employed the electronic structure method to investigate the atomic vibrations in polymer monomers. We utilized the CAM-B3LYP functional, which is particularly suited for systems involving electronic excitations and long-range interactions. This functional, an advancement of the widely-used B3LYP, integrates a Coulomb-attenuating method (CAM) to better handle the separation of short-range and long-range electron interactions. The choice of CAM-B3LYP is crucial for achieving accurate predictions of electronic properties, which, in turn, affect the vibrational characteristics of the molecules [Yanai et al., 2004].

To accurately model the electronic structure of polymer monomers and predict their vibrational properties, we selected an appropriate basis set that balances computational efficiency with precision. Basis sets are collections of mathematical functions used to approximate the electron wavefunctions in atoms and molecules. In this research, we utilized the 6-311++G** basis set. This split-valence basis set includes additional polarization and diffuse functions, providing a more detailed representation of the electronic density compared to simpler basis sets. The 6-311++G** basis set enhances the accuracy of the calculated vibrational frequencies by effectively describing both the core and valence electron interactions and accounting for electron density extending beyond the core regions. This choice ensures that the calculated vibrational modes reflect the true electronic environment of the polymer monomers, thereby improving the reliability of our insights into their stability and dynamic behaviour [Buczek et al., 2016; Erdogdu et al., 2010; Yamamoto et al., 2017]. The vibrational modes of molecules, including polymer monomers, are determined by analyzing the second derivatives of the energy with respect to the nuclear coordinates. These second derivatives form the Hessian matrix [Galimberti and Milani, 2014], whose diagonalization yields the vibrational frequencies. The calculated frequencies offer insights into the stability of the polymer structure, as well as

the potential energy landscape governing the atomic motions within the monomer.

By analyzing the vibrational modes of polymer monomers, we can gain insights into their stability and dynamic behaviour, which are critical for understanding and optimizing the performance of the resulting polymer materials.

Molecular Dynamics

The next scale in our research framework is focused on molecular dynamics (MD) simulations. MD simulations are a powerful computational technique used to model the physical movements of atoms and molecules over time. For a system of N particles, Newton's second law is applied as follows:

$$m_i \frac{d^2 r_i}{dt^2} = F_i \tag{1.8}$$

where m_i is the mass of particle i, r_i is the position of particle i, and F_i is the force acting on particle i. The force F_i is typically derived from the potential energy $U(r_1, r_2, \ldots, r_N)$ of the system:

$$F_i = -\nabla_i U. \tag{1.9}$$

This method is particularly effective in capturing the dynamical behaviour and thermodynamic properties of materials at the atomic and molecular scales. By solving Newton's equations of motion for a system of interacting particles, MD simulations provide detailed insights into the time-dependent evolution of material structures and properties.

At the core of MD simulations lies the integration of Newton's equations of motion, which govern the trajectory of each atom in the system. This involves calculating the forces acting on each atom, derived from a potential energy function that characterizes the interactions within the system. The potential energy function, often referred to as a force field, is crucial as it approximates the electronic structure and dictates the accuracy of the simulation. A force field provides a mathematical framework that describes how the energy of a system depends on the positions of its constituent particles. In molecular dynamics, a force field is defined by an analytical expression for the interatomic potential energy, $U(r_1, r_2, \ldots, r_N)$, along with a set of parameters incorporated into this expression. These parameters are typically derived from quantum mechanical calculations, either ab initio or semi-empirical, or by fitting to experimental data such as neutron, X-ray, electron diffraction, NMR, infrared, Raman spectroscopy, and other techniques. In this context, molecules are modelled as assemblies of atoms connected by simplified elastic (harmonic) forces, with the force field providing an approximation of the true potential. This model is designed to be computationally efficient while sufficiently capturing the key properties of the system under study. Various force fields exist, differing in complexity and suitability for different types of systems. A typical force field expression may include terms that model interactions between atoms, as well as bonded and non-bonded interactions (1.10).

The equation encompasses various components that describe different interactions within a molecular system. The first four terms account for intramolecular or local contributions to the total energy, which include bond stretching, angle bending, dihedral angles, and improper torsions. The remaining two terms describe non-bonded interactions: the Lennard-Jones potential, which captures both repulsive and van der Waals forces, and the Coulombic interactions, which represent electrostatic forces between charged particles [González, 2011].

$$U = \sum_{\text{bonds}} \frac{1}{2} k_b (r - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{\text{torsions}} \frac{V_{\text{n}}}{2} \left[1 + \cos(n\phi - \gamma)\right] + \sum_{\text{improper}} V_{\text{imp}} + \sum_{\text{LJ}} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) + \sum_{\text{elec}} \frac{q_i q_j}{r_{ij}} \tag{1.10}$$

Classical force fields, such as AMBER [Cornell et al., 1996], CHARMM [MacKerell Jr et al., 1998], OPLS [Jorgensen et al., 1996], and General Amber force field (GAFF) [J. Wang et al., 2004], describe bonded interactions (such as bonds, angles, and torsions) and non-bonded interactions (like van der Waals and electrostatic forces) with empirical parameters tailored for specific classes of molecules.

MD simulations are instrumental in studying phenomena such as diffusion, phase transitions, and mechanical properties of materials. They allow researchers to observe how materials respond to various stimuli, such as temperature changes and mechanical forces. In the context of polymer research, MD simulations help in understanding the conformational dynamics, mechanical strength, and thermal behaviour of polymer chains and networks [Allen and Tildesley, 2017; Frenkel and Smit, 2000, 2023].

For polymers, MD simulations provide insights into the relationship between molecular structure and macroscopic properties. Researches focus on studying properties such as the density [Jang et al., 2024], glass transition temperature [Gupta et al., 2013; Jang et al., 2024; Mohammadi, Davoodi, et al., 2017; Soldera, 1998; Z. Wang et al., 2015], and mechanical properties [Deng, 2017; Han and Elliott, 2007; Jang et al., 2024; Mae et al., 2008] of polymer systems. In Chapter 2, Section 2.5, we discuss the comprehensive multistep process involved in applying MD simulations to investigate and calculate the properties of a polymeric system.

This can allow for the optimization of materials for specific applications such as packaging, electronics, and biomedical devices. Additionally, recent advancements in computational power and algorithms have enabled the simulation of increasingly large systems and longer timescales, providing more realistic models of polymer behaviour under various conditions.

Moreover, MD simulations contribute to understanding the self-assembly processes of block copolymers, the diffusion and transport properties of polymers in solution, and the interactions at polymer interfaces and surfaces. These insights are critical for developing new materials with tailored properties, such as high-strength composites, conductive polymers, and responsive materials.

Overall, MD simulations serve as a bridge between theoretical predictions and experimental observations, allowing researchers to validate hypotheses and explore scenarios that are

challenging to investigate experimentally. The ongoing development of more accurate force fields and enhanced sampling techniques continue to expand the scope and accuracy of MD simulations in materials science [Karniadakis et al., 2006; Karuth et al., 2021; Plimpton, 1995; Varshney et al., 2008].

Group Interaction Modeling

One approach to predicting material properties is based on defining parameters from the microstructure and then using a semi-empirical expression based on these parameters to calculate the properties. This method is called Group Interaction Modeling (GIM), as a microscale framework estimates engineering and thermomechanical properties of desired polymers as a function of temperature and strain rate [J. P. Foreman et al., 2006; J. Foreman et al., 2012].

GIM leverages the intermolecular interaction energies between groups of atoms in adjacent polymer chains to predict key engineering properties of polymers based on their chemical composition and molecular structure. These groups, typically corresponding to the mer units of the polymer, represent the smallest set of atoms that characterize the polymer's composition. The interaction energy predominantly involves van der Waals forces, as these weak interactions are most likely to be influenced by external energy fields, such as those encountered during mechanical deformation.

GIM relies on an energy balance framework that contrasts the negative potential energy from intermolecular forces with various positive thermodynamic energy parameters. This mean-field approach allows for the development of a simplified model that captures the essential structural factors influencing a polymer's engineering properties. Consequently, GIM serves as a model framework, within which the specific details of generalized energy terms can be explored to understand their impact on physical properties.

Even though the thermodynamic energy terms are treated as average values, fluctuation theory can be employed to account for variations in thermal energy, leading to a distribution of free volume within the polymer. This distribution influences the broadening of transition peaks as a function of temperature, providing insights into the thermal behaviour of the material [Porter, 1995].

The GIM method is particularly useful in the initial steps of designing new polymers due to its speed, cost-effectiveness, and ability to fill in gaps where experimental techniques are problematic such as experiments at temperatures lower than 25 °C. In addition, the method's meticulous physical basis allows the user to calculate numerous significant properties of the materials that would not be easy to measure otherwise. For instance, GIM can be used to predict the glass transition temperature, heat capacity, linear thermal expansion, elastic modulus, and Poisson's ratio, among other properties [J. P. Foreman et al., 2010]. This framework, introduced by Porter [Porter, 1995] was initially devoted to predicting linear polymers' properties. However, the method can be used to predict the mechanical behaviour of polymer networks such as epoxy resins. Application of GIM technique to predict the properties of the epoxy resins needs some modification to the original method of Porter because the interaction of the atoms in the crosslinked polymer is different compared with the

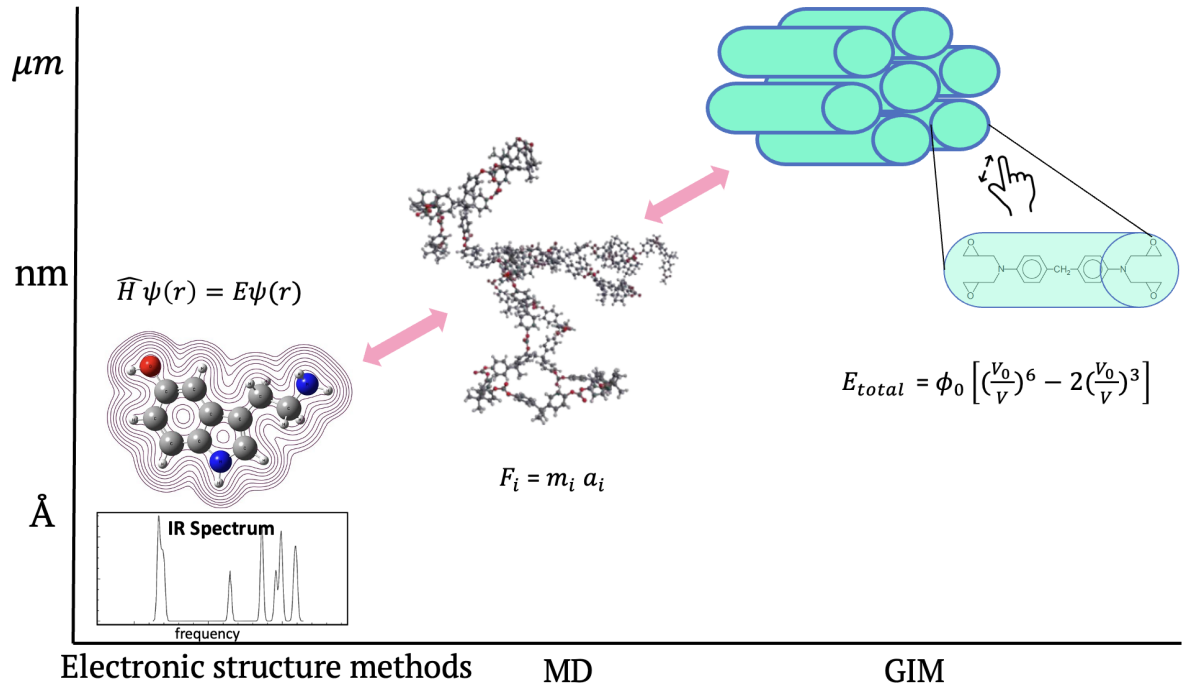

Figure 1.4: Framework of the research illustrating different methods at various length scales.

linear polymers. For this reason, the degrees of freedom, N, as an important parameter in the GIM method, were evaluated, and the incorporation of the resin crosslinking was shown by decreasing the value of N by three for each trifunctional branching site on the monomers [J. P. Foreman et al., 2008]. By this modification, Guest et al. [Guest et al., 2013] calculated the glass transition temperature (T_g) and stress-strain curves of the epoxy system and compared them with the results of the experimental tests. Although the prediction values of T_g have good agreement with experimental values, the predicted stress-strain curves do not fit well with the experimental results. Besides applying the GIM method to predict the mechanical properties of the polymer materials [J. P. Foreman et al., 2006; J. Foreman et al., 2012; Porter and Gould, 2009], this method was used in few studies to predict the properties of the polymer composites [J. P. Foreman et al., 2009; J. Foreman et al., 2010] which are based on epoxy systems. Although this prediction can help make quick decisions about the suitability of polymers for specific applications, there are difficulties involved. One main challenge is obtaining input parameters for the studied polymer. Some parameters can be determined using tables of group interaction based on the monomer structure, but others require experimental results or advanced simulations. For example, the Debye temperature requires studying the polymeric structure with atomistic codes that describe electronic configuration and atomic displacements.

In the context of polymer modelling, continuum approaches are employed to bridge the gap between molecular-level insights and macroscopic material behaviour. These continuum models are essential for translating detailed atomic-level simulations into practical engineering parameters, such as elastic modulus, glass transition temperature, and heat capacity, which are crucial for the design and analysis of polymer-based materials in real-world applications.

1.2.2 The fundamentals of machine learning

In the field of materials science, which encompasses various functional applications, machine learning (ML) has rapidly emerged as a powerful and versatile approach. Its ability to broadly explore the materials space and accurately predict material properties has made it an invaluable tool for discovering and optimizing new materials and phenomena [Alberi et al., 2018; Schleder et al., 2019; Schmidt et al., 2019].

Also, molecular modelling (especially molecular mechanics and dynamics) has become a standard method for calculating the properties of polymers, however, the time-temperature scales are not simply applied to actual engineering problems and the 'black box' character of numerical simulations of large assemblies of atoms is not easily translated into an analytical understanding of dynamic mechanical properties for polymer design and application. Besides, considering the cost of these simulation tools, it is not practical to use these molecular modelling methods to discover the properties of a significant number of polymers. Therefore, we use ML as a replacement tool that is cheaper and faster.

ML is a set of statistical methods that build predictive models by fitting functions to datasets, allowing the system to learn from and adapt to new data without explicit human intervention. It addresses the challenge of how to use computers that improve automatically through experience. These algorithms create models based on sample data, known as training data, to make predictions or decisions without being explicitly programmed for specific tasks. Machine learning is at the core of data science and artificial intelligence, and as an interdisciplinary field, it connects computer science and statistics. This method enhances the accessibility of low-cost computing and online data through algorithmic learning [El Naqa and Murphy, 2015; Jordan and Mitchell, 2015; Nilsson, 1996]. A variety of ML algorithms have been developed to address the diverse range of data and problem types encountered in various applications. In essence, machine learning algorithms can be seen as exploring a vast space of potential solutions, using training data to guide them towards finding an optimal program that maximizes performance metrics. ML algorithms can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labeled data, where the desired output is known. The algorithm learns to map input data to the correct output during training and can then make predictions on new, unseen data. The workflow of supervised machine learning algorithms is illustrated in Figure 1.5 [Alpaydin, 2020; Balachandran, 2019; Jordan and Mitchell, 2015; Mahesh, 2020].

Key algorithms in supervised machine learning include linear regression [Groß, 2003], logistic regression [Kleinbaum et al., 2002], decision trees [De Ville, 2013], random forest [Rigatti, 2017], support vector machines [Hearst et al., 1998], k-Nearest neighbors [Peterson, 2009], Naive bayes [Webb et al., 2010], and neural networks [Abdi et al., 1999]. Each of these algorithms in supervised machine learning can be applied to solve either regression or classification problems.

Unsupervised learning involves training a model on unlabeled data, where the algorithm explores the data and identifies patterns or structures without explicit guidance on what to look for. In unsupervised learning, key algorithms that can be utilized include K-means

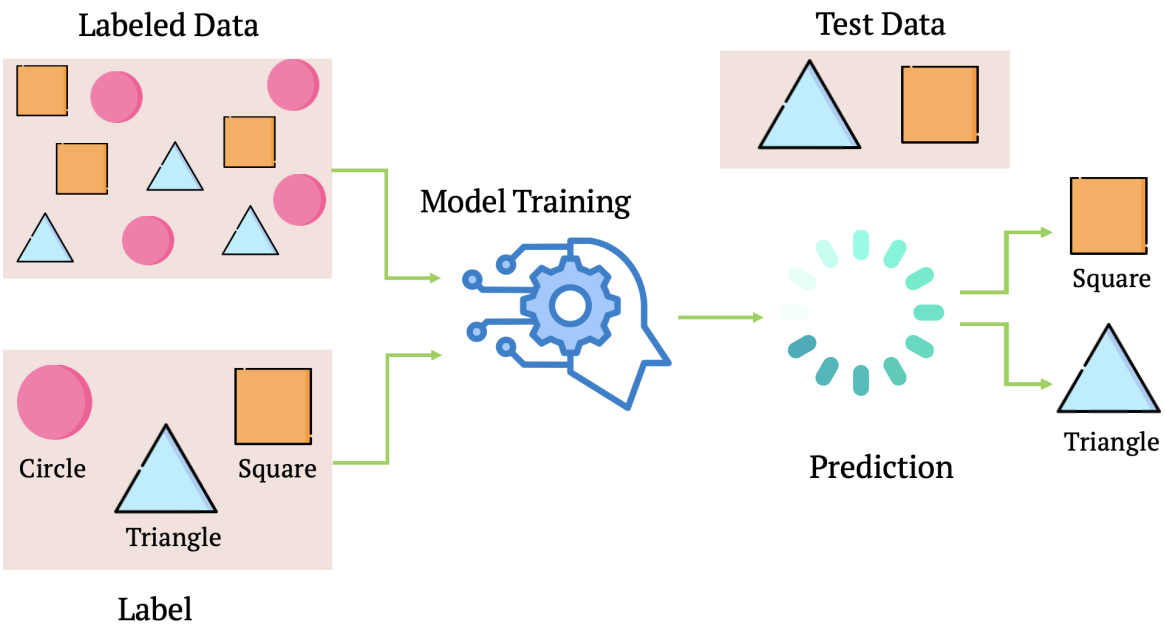

Figure 1.5: Supervised learning workflow.

clustering [Sinaga and Yang, 2020], hierarchical clustering [Nielsen and Nielsen, 2016], spectral clustering [Von Luxburg, 2007], Gaussian mixture models (GMM) [Reynolds et al., 2009], boosting [Ferreira and Figueiredo, 2012], and artificial immune systems (AIS) [Dasgupta, 2012], among others.

Reinforcement Learning involves an agent learning to make decisions by taking actions in an environment to maximize some notion of cumulative reward. It is widely used in applications like robotics, gaming, and autonomous systems [Thrun and Littman, 2000]. The workflow of reinforcement machine learning is illustrated in Figure 1.7. Some key algorithms used in reinforcement learning include Monte Carlo [Horowitz, 1991], Q-learning [Watkins and Dayan, 1992], deep deterministic policy gradient [H. Tan, 2021], asynchronous advantage actor-critic [Babaeizadeh et al., 2016].

In understanding machine learning algorithms, a fundamental scientific and practical objective is to theoretically assess their capabilities and the inherent challenges posed by different learning problems [Alpaydin, 2020; Balachandran, 2019; Jordan and Mitchell, 2015; Mahesh, 2020]. Regardless of the specific models or algorithms employed, practitioners in the field of machine learning often adhere to a common workflow. This process typically includes the following key steps:

- **Define the problem**: Clearly articulate the problem to be solved, identify the objectives, and determine the metrics for evaluating success.
- **Data collection and preparation**: Gather relevant data and prepare it by cleaning, normalizing, and pre-processing to ensure it is suitable for analysis.
- **Dataset construction**: Construct the dataset by dividing it into training, validation,

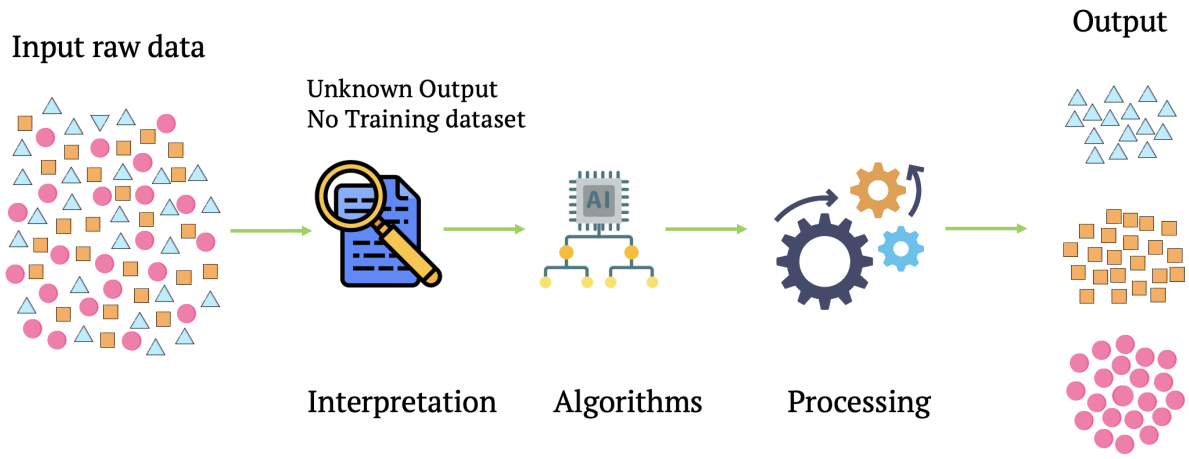

Figure 1.6: Unsupervised learning workflow.

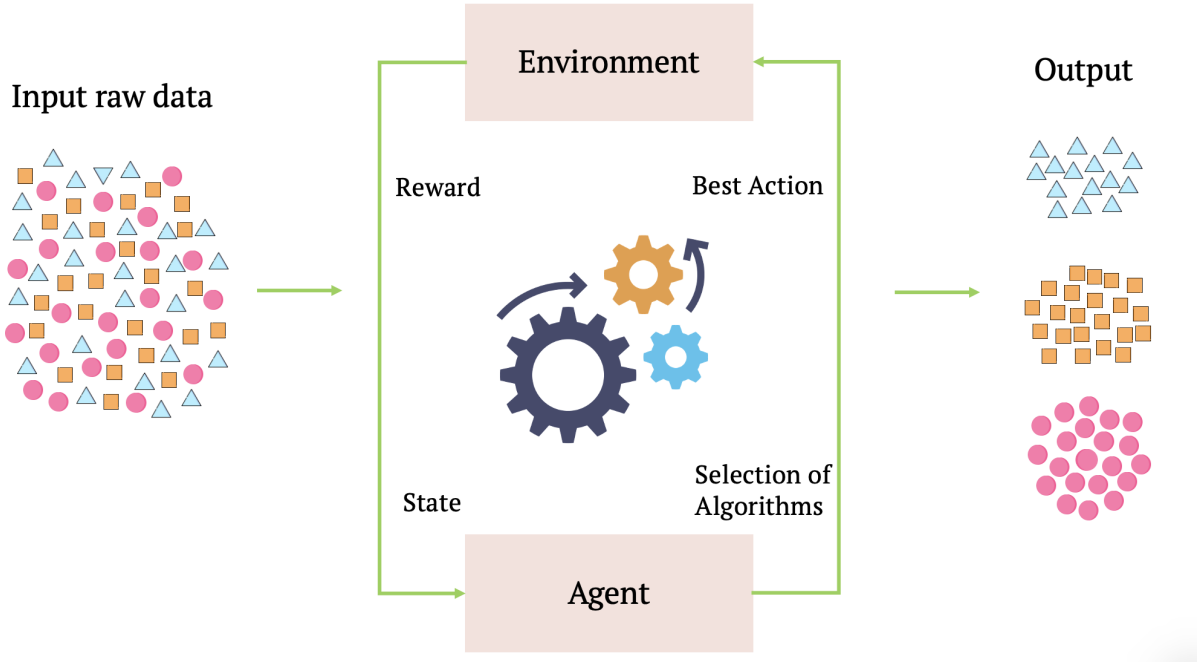

Figure 1.7: Reinforcement learning workflow.

and test sets, ensuring a robust setup for model training and evaluation.

- **Model training**: Select and train the appropriate machine learning model using the training data, optimizing the model's parameters and performance.
- **Model evaluation and use**: Evaluate the trained model using the test set to assess its performance based on predefined metrics. Once validated, the model is deployed and integrated into practical applications, with continuous monitoring and adjustments as needed.

This streamlined workflow provides a structured approach to developing and deploying machine learning solutions, ensuring that each stage is carefully executed to achieve the best possible outcomes.

In the process of developing ML models, understanding and addressing overfitting and underfitting are critical for achieving optimal performance. Overfitting occurs when a model learns not only the underlying patterns in the training data but also the noise and specificities that do not generalize to new, unseen data. This often happens when the model is too complex, with too many parameters relative to the amount of training data. As a result, the model may perform very well on the training data but poorly on the test data. In the following chapters, we will explore techniques such as regularization, cross-validation, and dropout. These methods are crucial for addressing overfitting by reducing excessive model complexity and enhancing the model's ability to generalize effectively to new, unseen data.

Underfitting, on the other hand, happens when a model is too simple to capture the underlying patterns in the data. This typically occurs when the model lacks sufficient complexity or is not trained adequately. An underfitted model will perform poorly on both the training and test data, as it fails to learn the relationships within the data. To address underfitting, one might consider using a more complex model, increasing the training duration, or incorporating additional features [El Naqa and Murphy, 2015].

Understanding and balancing these issues is essential for building effective machine learning models. Proper model evaluation and tuning help in identifying and addressing these problems, ensuring that the model achieves good performance on both training and unseen data.

1.2.3 Molecular representations in machine learning

To effectively integrate polymers into machine learning frameworks, we must first represent them in a format that is machine-readable. The representation of molecules has intrigued scientists since the nineteenth century. Classically, molecules are depicted using molecular formulas that show only bonds and atoms. While atomic information is readily available from the molecular formula, it does not provide details on how atoms are bonded or the molecular geometry. Therefore, it is crucial to focus on identifying the most suitable molecular representations for our objectives. This research investigates two approaches: 2D or 3D molecular descriptors, and graph-based representations.

Capturing information from chemical structural is facilitated by molecular descriptors, which are numerical values that characterize various properties of molecules. These descriptors can

represent physicochemical properties or be derived using algorithmic techniques applied to molecular structures. Descriptors are unique and ambiguous ways to represent a molecule and are widely used in cheminformatics. A diverse range of molecular descriptors have been developed and utilized across various applications. They can range from simple calculations, such as molecular weight, to more complex descriptors based on quantum mechanics [Cereto-Massagué et al., 2015; David et al., 2020; Leach and Gillet, 2007; Wiswesser, 1968].

A graph is defined as G = (V, E), where V is the set of nodes and E is the set of edges connecting pairs of nodes. In a molecular graph, V typically represents all atoms in a molecule, and E represents all bonds. The concept of molecular graph representation involves mapping the atoms and bonds of a molecule onto sets of nodes and edges. Any spatial relationships between nodes are encoded as attributes of nodes or edges, as nodes in a mathematical graph lack formal spatial positions; they only denote pairwise connections [Bondy, Murty, et al., 1976; David et al., 2020].

To translate a graph from an abstract mathematical concept to a concrete representation that can be processed on a computer, one typically converts the sets of nodes and edges into linear data structures, such as matrices or arrays. These linear data structures are essential to specify the connectivity of the nodes. This involves assigning an artificial ordering to the nodes for encoding a molecule, even though V and E are formally sets where the order of elements is irrelevant. The mapped information includes: (1) how the atoms in the molecule are connected, (2) the identity of the atoms, and (3) the identity of the bonds.

The connections between atoms are typically represented by an adjacency matrix A. For an element a_{ij} of A, $a_{ij} = 1$ indicates a bond between nodes v_i and v_j in the molecular graph G, whereas $a_{ij} = 0$ indicates no bond (see Figure 1.8(b). This adjacency matrix is also known as the connectivity matrix. Note that the adjacency matrix does not specify the type of bond between each pair of nodes.

The identity of the atoms is represented by a node features matrix X (see Figure 1.8(c)). Each row of X corresponds to a node v_i (i.e., an atom in the molecule) in G. This row is also referred to as the node feature vector x_i for that atom. The length of x_i corresponds to the number of atom features chosen to encode (e.g., a one-hot encoding of atom type and formal charge).

The identity of the bonds is represented by an edge features matrix E (see Figure 1.8(d)). Each row of E corresponds to an edge $e_{ij} = (v_i, v_j)$ in G, and is referred to as the edge feature vector e_{ij} for that edge. The length of e_{ij} corresponds to the number of edge features chosen to encode (e.g., a one-hot encoding of possible bond types: single, double, triple, aromatic).

Although one-hot encoding is common in AI applications, it is not necessary for various node and edge features. For instance, the node features matrix shown in Fig. 1c could have only three columns, using integers to represent the same three properties (atom type, formal charge, and number of implicit Hs) [David et al., 2020].

While graphs are inherently non-linear data structures composed of sets of nodes and edges, matrix representations of graphs are dependent on node order. The node order in a matrix representation is determined by a graph traversal algorithm (see Figure 1.9). For certain

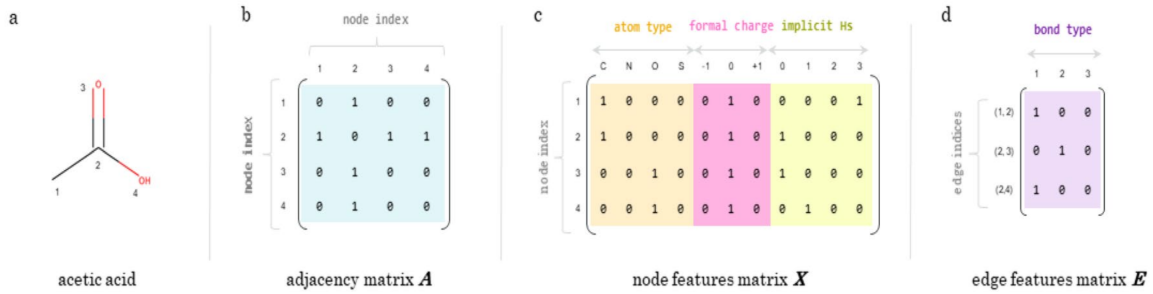

Figure 1.8: Graph representation example for acetic acid. (a) Graph representation of acetic acid with nodes numbered from one to four. (b) Adjacency matrix A for acetic acid with corresponding node ordering on the left. (c) Node features matrix X, showing a one-hot encoding of selected properties. (d) Edge features matrix E, where each edge feature vector is a one-hot encoding of bond types (single, double, or triple). "Implicit Hs" refers to the number of implicit hydrogens on a given node [David et al., 2020].

applications, it is crucial to generate the exact same representation for the same molecule consistently. Ensuring the same representation requires obtaining the same node order every time. This can be achieved using methods such as depth-first or breadth-first search to generate graph matrix representations. The graph traversal algorithm must include a consistent way to break ties when a node branches off, ensuring the same branch traversal order is always selected. Different software packages often distinguish themselves by how they handle these tie-breaking situations.

However, if consistency is not a priority (and for some deep learning applications, noisier data might be preferred), a random search can be utilized. The matrix representations discussed above are not the only way to represent graphs; there are multiple ways to represent the same information. For example, depending on the graph traversal algorithm used, the order of the rows in the adjacency matrix (or atom/bond block) will vary. Additionally, when dealing with molecular graphs, there is no single correct way to represent a molecule; the chosen representation must be appropriate for the specific task [David et al., 2020].

1.2.4 Machine learning applications in polymer research

Numerous works have explored various strategies for predicting polymer properties, using data-driven approaches. The successful utilization of neural networks [Velten et al., 2000; X.-L. Yu et al., 2008] and support vector regression [Pei et al., 2012] in predicting glass transition temperature, tribological properties, electronic and dielectric properties, storage modulus, damping, specific heat capacity, and mechanical characteristics, underscores the immense potential of ML methods in this domain [Bhowmik et al., 2021; Mannodi-Kanakkithodi et al., 2016].

Besides of all ML advantages, the limited availability and variety of data pose a challenge to fully exploiting the potential of ML. Addressing the issue of insufficient data can be solved with the aim of transfer learning (TL). This approach encompasses various methodologies that involve re-purposing a model trained on one task for another related task. By leveraging

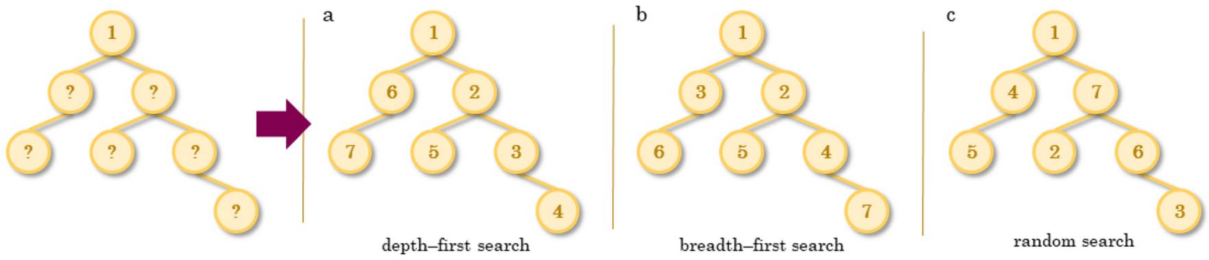

Figure 1.9: Graph traversal algorithms. Three common graph traversal algorithms are illustrated using a sample branched graph. The numbers indicate the order in which nodes are visited, starting from node 1. (a) Depth-first search (DFS) explores each branch of the graph as far as possible before backtracking to explore other branches from the last node. (b) Breadth-first search (BFS) explores all immediate neighbours of a node first, then proceeds to explore neighbours of those neighbours, continuing this process until the entire graph is traversed. (c) Random search visits nodes in an arbitrary sequence, disregarding their connections [David et al., 2020].

pre-trained models, TL enables efficient knowledge transfer and enhances performance on new tasks with limited data availability [C. Tan et al., 2018; Weiss et al., 2016]. TL has emerged as a powerful technique which can be very helpful in the field of materials science [Cubuk et al., 2019; Hutchinson et al., 2017; Kaya and Hajimirza, 2019]. However, this technique has been relatively under explored in the field of polymer characterization, with limited research studies available on this topic. Zhang et al. [Z. Zhang et al., 2022] utilize TL to successfully predict stress-strain curves of fiber reinforced polymer (FRP) composites fabricated via additive manufacturing. The results demonstrate the effectiveness of TL in achieving accurate predictions with limited training data, highlighting its potential in transforming the generation of stress-strain curves. Ma et al. [Ma et al., 2021] used the TL technique to enhance the accuracy of ML models to predict adhesive free energy in polymer-surface interactions trained on small data sets. The TL approach significantly improves prediction accuracy and has implications for inverse materials design.

Yamada et al. [Yamada et al., 2019] have developed some foundational models that serve as a basis for TL methods to predict material properties. They used these base models to predict the properties of inorganic crystalline materials, polymers, and small molecules. They examined these base models and transferred them to predict a couple of properties with the small size of dataset.

More recently, ML-based methods have made significant progress in screening polymer libraries. For instance, Alfaraj [AlFaraj et al., 2023] and Gurnani [Gurnani et al., 2023] demonstrated the effectiveness of ML in accelerating feature extraction and enabling the large-scale screening of massive polymer libraries.

To enable ML-driven predictions, researchers are particularly interested in creating extensive data sets, often in conjunction with simulation methods. For instance, Tao et al. [Tao et al., 2023] harnessed MD simulations to construct a vast data set on polymer fractional free volume for >6500 homopolymers and 1400 polyamides. Using ML models, they established composition-structure relationships, surpassing traditional group contribution theories with

efficient feed-forward neural network (FFNN) models.

A substantial number of different molecular representations have been developed to enhance system descriptions of polymers in the field of ML. One of the most important molecular representations are molecular graphs, and several studies have highlighted the use of graph neural networks (GNNs) in the field of polymer property prediction [Park et al., 2022; Wu et al., 2023]. Zeng et al. [Zeng et al., 2018] emphasized the significance of graph convolutional neural networks (GCNN) in predicting polymer properties, showing remarkable agreement with DFT results and outperforming other ML algorithms. Graph-based approaches remove the need for complex hand-crafted descriptors while maintaining prediction accuracy. Recently, Aldeghi et al. [Aldeghi and Coley, 2022] introduced a graph representation of molecular ensembles and a GNN architecture tailored for polymer property prediction. Their work demonstrated that by using this framework to model polymers, they were able to capture the relevant features that distinguish polymer materials from one another, outperforming traditional cheminformatics methodologies for polymer representation.

These studies in polymer property prediction showcase the transformative impact of ML, which has empowered researchers to leverage diverse techniques and expansive data sets for more accurate polymer property prediction. The integration of ML with simulation methods and advanced molecular representations signals a promising direction for advancing our understanding and predictive capabilities in polymer science.

1.3 Problem statement

Accurate prediction of polymer properties is essential for their effective application and development, making it crucial to find methods that facilitate such predictions to enhance the industrial use of polymers. This thesis addresses the challenges involved in identifying suitable methods to connect polymer structure to behaviour by exploring two innovative computational approaches: the Group Interaction Modeling (GIM) method and Machine Learning (ML).

Our research focuses on comparing the accuracy and reliability of GIM and ML in predicting six key thermal and mechanical properties of polymers. While GIM is effective for characterizing polymer properties, its predictive scope is limited and relies heavily on input parameters from Density Functional Theory (DFT). DFT calculations are resource-intensive and time-consuming, especially for large polymer systems, and require careful selection of basis sets and exchange-correlation functionals to achieve accurate predictions.

In contrast, ML offers an alternative but presents its own set of challenges. Data availability and quality are major issues, as high-quality datasets are scarce, complicating the training of robust ML models. Moreover, representing molecular structures in a way that captures all necessary information is challenging.

To overcome the primary challenge of data scarcity, we employed two strategies: Transfer Learning and generating our own dataset via molecular simulations. TL was used in a neural network model initially trained on the heat capacity at constant pressure (C_p), a property with ample available data. The model was then adapted to predict additional properties

such as specific heat capacity (C_v), shear modulus, flexural strength at yield, and dynamic viscosity.

Our research also delves into copolymer materials, specifically random, block, and alternate copolymers, driven by the goal of addressing a significant gap in reference data for these materials. The PolyInfo database, one of the largest polymer databases, reveals a scarcity of comprehensive data on critical properties such as bulk modulus, compressibility, and thermal expansion coefficients for copolymers. To bridge this gap, we used MD simulations to generate a rich dataset that includes properties like density, radius of gyration (R_g), specific heat capacities (C_p and C_v), bulk modulus, and volume expansion coefficients.

This dataset serves as the foundation for our subsequent data-driven approach. By integrating MD simulations with ML techniques, we aim to develop predictive models for optimizing copolymer performance based on the calculated properties. The dynamic nature of copolymers, influenced by factors such as composition and sequence, adds an additional layer of complexity to our research.

For example, the copolymer derived from poly(ethylene terephthalate)/poly(ethylene sebacate) demonstrates that block copolymers exhibit higher elastic properties and melting points compared to their random copolymer counterparts. Such examples underscore the significance of understanding how monomer sequence distribution affects various copolymer properties, including stiffness, tensile strength, and dielectric properties.

Traditional representations of molecules, such as diagrams with bonds and atoms, are insufficient for computational analysis in cheminformatics. We, therefore, explored alternative molecular representations that encapsulate all structural and chemical characteristics necessary for our simulations.

This thesis leverages the synergy between computational simulations and data-driven modelling to reshape the landscape of polymer science. By combining GIM, ML, TL, and MD, we aim to develop and validate computational methodologies for predicting polymer properties, thus accelerating polymer design and application through advanced computational methods.

1.3.1 Research objectives

This research aims to address these challenges by employing advanced machine learning techniques and molecular. The objectives of this study are as follows:

- Adopt a data-driven approach to present a more efficient and reliable method for predicting polymer properties, significantly expediting the material selection process.
- Highlight the importance of leveraging advanced computational techniques to ensure accurate parameter inputs that enhance the predictive capabilities of models in the field of polymer materials.
- Demonstrate the effectiveness of transfer learning by achieving high accuracy in predicting four additional polymer properties using relatively small datasets.
- Evaluate the model's performance using five different loss functions, emphasizing the

importance of selecting an appropriate loss function for accurate predictions.

- Showcase the benefits of transfer learning in efficiently predicting multiple properties of linear polymers, paving the way for advancements in various industrial and scientific applications.
- Provide an in-depth understanding of the molecular composition, configuration, and sequence distribution of copolymers.
- Utilize a diverse set of representations, notably graph-based representations, to achieve a comprehensive understanding of copolymer properties and behaviours.

1.3.2 Contributions and outline

In the current chapter (Chapter 1), we have provided a brief background covering the role of polymers in industrial applications and the models used in this work, such as DFT calculations, MD simulations, the group interaction model, and machine learning. We also discuss related work by other researchers, the motivation for this work, its potential impact, and the challenges and contributions. An outline of the thesis is provided as well.

Details about the MD simulations conducted, the dataset, and the machine learning methods and algorithms used are also presented in Chapter 2.

Chapters 3 to 5 present the results of our work. The original contributions in these chapters are derived from three publications, two of which are published [Kazemi-Khasragh, Blázquez, et al., 2024; Kazemi-Khasragh, Gonzaleza, and Haranczyk, 2024], and one is currently submitted.

Chapter 3 provides the results of ML methods and GIM methods in predicting various thermal and mechanical properties.

Chapter 4 demonstrates the power of transfer learning in predicting properties from small datasets.

Chapter 5 illustrates the effectiveness of using different molecular representations in ML by generating our dataset for copolymers with MD simulations.

The final chapter, 6, is dedicated to conclusion.

Chapter 2

Methods and Methodology

2.1 Definition of Key Properties

In this thesis, we focus on a set of thermal, mechanical, and rheological properties crucial for understanding polymeric materials' behaviour. These properties provide insights into how polymers respond to thermal changes, mechanical forces, and flow conditions.

Thermal properties describe the reaction of a material to changes in temperature. The glass transition temperature (T_g) is a critical parameter, marking the point at which a polymer transitions from a rigid, glassy state to a softer, rubbery state. The heat capacity at constant volume (C_v) and at constant pressure (C_p) are important as they measure the amount of heat required to raise the temperature of a material by one unit temperature under constant volume and constant pressure conditions, respectively [Mark et al., 2007]. These capacities offer insight into the thermal energy exchange capabilities of the material. The volume expansion coefficient (γ) quantifies the relative change in volume per degree of temperature change under constant pressure, while the linear expansion coefficient (α) measures the relative change in length under the same conditions [Ehrenstein et al., 2012]. For an isotropic expansion, the relationship can be expressed $\gamma = 3\alpha$.

The mechanical properties of polymers define how these materials respond to applied forces, which is essential for understanding their structural integrity and performance under various loading conditions. The elastic modulus (E), or Young's modulus, represents a material's stiffness by describing the ratio of uniaxial stress to strain within the elastic deformation region. Poisson's ratio (ν) offers insight into how a material deforms in different directions when subjected to axial stress, by measuring the ratio of transverse to axial strain. The shear modulus (G) measures a material's ability to resist shear deformation, reflecting the ratio of shear stress to shear strain. The bulk modulus (K), on the other hand, quantifies a material's resistance to uniform compression, indicating the ratio of volumetric stress to volumetric strain. Additionally, density (ρ) is a fundamental property that influences both mechanical and thermal behaviours, representing the mass per unit volume [Callister Jr and Rethwisch, 2020; Flory, 1953; Shaw and MacKnight, 2018; Van Krevelen and Te Nijenhuis, 2009].

Other structural properties of polymers, such as the radius of gyration (R_g), are considered.

The radius of gyration is a measure of the polymer's size, reflecting the distribution of monomer units around the centre of mass of the polymer chain. It provides insight into the polymer's conformational characteristics in solution and is crucial for understanding the polymer's behaviour in different environments [Flory, 1953].

Finally, the rheological properties describe how a material flows or deforms under applied forces, particularly in non-solid states. Dynamic viscosity (η) is the primary rheological property considered in this thesis, as it measures the resistance of a fluid to flow when subjected to shear strain rate. This parameter is essential for understanding the flow behaviour of polymer melts and solutions [Larson, 1999].

These properties are interconnected; for example, as temperature affects T_g, it subsequently influences the elastic modulus, viscosity, and expansion coefficients. Understanding these relationships allows for a more comprehensive prediction of polymer behaviour under varying conditions [Young and Lovell, 2011].

2.2 Experimental data curation

The extracted 1221 reference values for homo-polymers and co-polymers from reference values from databases, books, and handbooks [Brandrup et al., 1999; Mark et al., 2007; Otsuka et al., 2011; "PolyInfo", n.d.; Porter, 1995] were used for comparison the GIM and MD results. The details of data curation are provided in the Section 2.6.1.

2.3 Group Interaction Modeling: The model framework

The Group Interaction Modeling (GIM) [J. P. Foreman et al., 2010; J. Foreman et al., 2010, 2012] technique is employed to forecast the thermal, volumetric, and mechanical properties of polymers through the utilization of a mean-field potential function. By employing a contribution-based methodology, the GIM method computes the total energy of the system. The interactions between adjacent polymer chains are described by a potential function comprising various thermodynamic energy terms. This thermodynamic potential function serves as the equation of state for the system, as illustrated in equation (2.1).

$$E_{\text{total}} = \phi_0 \left[\left(\frac{V_0}{V} \right)^6 - 2 \left(\frac{V_0}{V} \right)^3 \right] = H_M + H_T + H_c - E_{\text{coh}} = -0.89 E_{\text{coh}} + H_T \tag{2.1}$$

The potential energy is based on a standard Lennard-Jones potential function where the cohesive energy E_{coh} is the molar quantity of the molecular potential well depth ϕ_0, V_0 is the volume with zero potential, and V is the volume. Also, the E_{total} contains mechanical H_M, thermal H_T, configurational contributions H_C, and cohesive energy E_{coh}. In the context of GIM techniques, the term H_M refers to the energy associated with mechanical deformation per interaction. H_T is based upon the molecular level temperature changes through heat capacity ((2.2)) where $C_{V,\text{skeletal}}$ is the heat capacity contributed by skeletal vibration.

H_C was estimated by the level of the cohesive energy for different conformations. The energy equation (equation (2.1)) was used by the GIM system to develop calculations to predict the structural properties of polymers. For this purpose, numerous parameters based on the representative mer unit were used as input parameters of the GIM method. The input parameters of the GIM method are represented in 2.1 [J. P. Foreman et al., 2008, 2009; J. Foreman et al., 2010, 2012; Porter and Gould, 2009].

$$H_T = \int_0^T C_{V,\text{skeletal}}\, dT \tag{2.2}$$

Table 2.1: Input parameters of GIM method.

Parameter	**Description**	**Units**
M	Molecular weight of mer unit	g/mol
V_W	Van der Waals volume of mer unit	m^3/mol
E_{coh}	Cohesive energy	J/mol
N	Degree of freedom per mer unit	Mer $unit^{-1}$
θ_1	Debye temperature	K
θ_E	Einstein temperature	K
L	Length of a mer unit in the chain axis	m or Å

The monomers' molecular weight, van der Waals volume, cohesive energy, and freedom degree can be calculated with the group contribution method. Each repeating unit in the chain of the polymers contains several functional groups, and the information of each functional group is tabulated in References [J. P. Foreman et al., 2009; Porter, 1995]. The other three parameters (the length of the monomer, Debye temperature, and Einstein temperature) can be obtained via molecular modelling. Once the molecular structure is generated, the length of each mer unit can be measured from the coordinates of atoms. The length of the monomer was calculated using the Avogadro software, an advanced molecular editor and visualizer known for its intuitive interface and versatility in handling molecular structures [Hanwell et al., 2012]. Avogadro provides a wide range of tools for designing and editing molecules, making it ideal for tasks such as constructing complex polymer structures. After designing the polymer molecules in Avogadro, their geometries were optimized using its energy minimization capabilities, which rely on a force field approach, as discussed in section 1.2.1. This method allowed for the accurate measurement of the length of each monomer. The graphical representation and view of the number of polymers designed by the Avogadro program are presented in Figure 2.1.

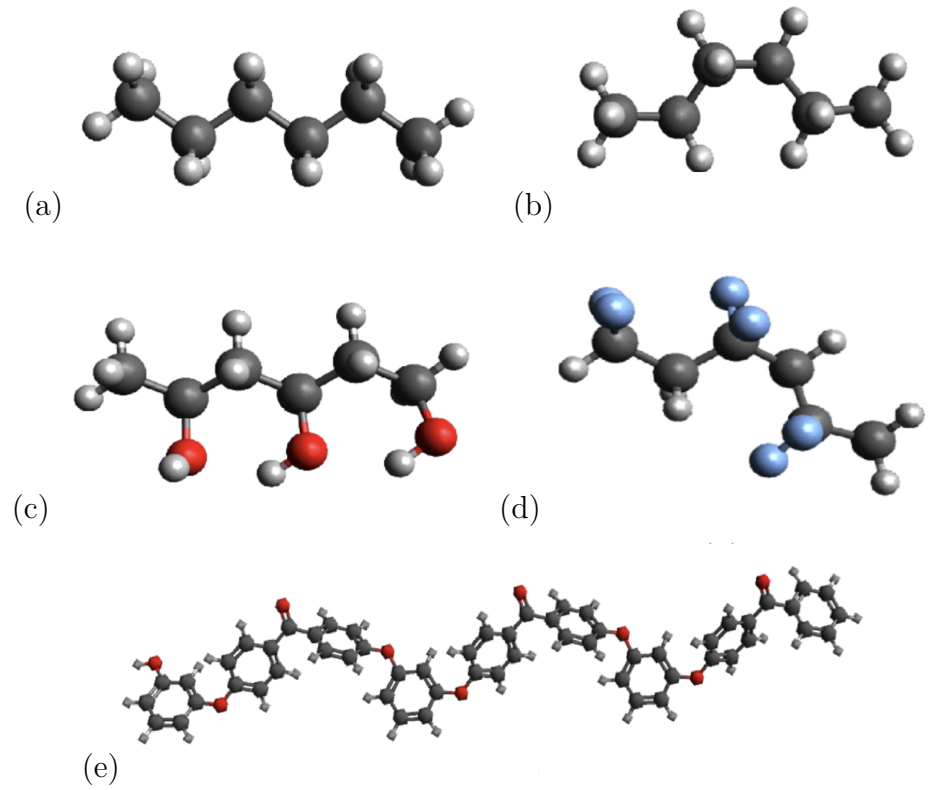

Figure 2.1: The graphical view of (a) polyethylene (trans), (b) polyethylene (gauche), (c) poly(vinyl alcohol), (d) polyvinylidene fluoride, and (e) polyether ether ketone with three repeated units. Results obtained with Avogadro.

Debye temperature and Einstein temperature are related to the vibration of atoms in the structure and can be predicted using electronic structure methods. Debye temperature, which is the maximum temperature related to the single normal vibration, can be estimated with the skeletal vibration. Note that the skeletal vibration is transmitted to the entire material. Einstein temperature (θ_E) is estimated by Group vibration [Porter, 1995]. As mentioned in 1 the vibrational frequencies of several polymers have been computed by density-functional theory (DFT) at gas-phase conditions. A functional of CAM-B3LYP was applied with the basis of 6-311++G** [Yamamoto et al., 2017] to the polymer trimers. The calculation was performed by the Gaussian 16 program [Frisch et al., 2004]. Gaussian 16 represents the latest advancement in the Gaussian series of software. It offers cutting-edge tools for electronic structure modelling based on the principles of quantum mechanics. This software can predict a wide range of molecular characteristics, including energies, structures, vibrational frequencies, and various properties of compounds and reactions across diverse chemical contexts.

In Gaussian, to calculate the vibration mode, second derivatives of the energy with respect to the Cartesian nuclear coordinates are determined. The necessary force constants to determine the frequencies are on Hessian matrix diagonalization [Galimberti and Milani, 2014]. All the required information to estimate the vibration modes, such as IR intensities, force constants, reduced masses, and so on, exists in the output of the Gaussian calculation.

To compute the Debye temperature (θ_1), the highest frequency associated with skeletal vibration mode should be chosen. According to the definition of skeletal vibration, we need to find the frequency with the maximum contribution of the backbone atoms during the vibration and the highest Infrared (IR) intensities to discover the highest skeletal vibration frequency. equation (2.3) is used to calculate Debye temperature from skeletal frequency (ν_1). On the other hand, the group vibrations frequencies (ν_E) are determined by vibrational spectra for each molecular unit, using the same output of Gaussian 16. Then, the Einstein temperature is determined with equation (2.4).

$$\theta_1 = \frac{h\nu_1}{k} \tag{2.3}$$

Debye temperature (θ_1) is associated with the highest frequency of the density of the vibrational states. Regarding equation (2.3), ν_1 represents the highest frequency of the vibrational states' density, and h and k are Planck's ($6.62607015 \times 10^{-34}$ J $\cdot$ s) and Boltzmann's (1.380649×10^{-23} J/K) constants, respectively. Group vibration frequency (ν_E), which can be determined with Infrared (IR) absorption and Raman scattering spectrum, was used to calculate the Einstein temperature (θ_E) by equation (2.4)

$$\theta_E = \frac{h\nu_E}{k} \tag{2.4}$$

With the assistance of input parameters in the GIM method, some physical features, such as heat capacity, thermal expansion, and glass transition temperature, among others, can be calculated [Ali and Rahaman, 2018; Yokota and Tsukushi, 2020; Yokota et al., 2020].

2.4 Group Interaction Modeling: Property prediction

Heat capacity is a principal property of the materials in the case of the thermodynamic evaluation because the entropy and the enthalpy can be calculated with it. Also, a profound investigation of this property which reflects molecular motion, reveals critical data about the vibrational state of molecules [Nishiyama et al., 2021; Yokota and Tsukushi, 2020; Yokota et al., 2020].

For calculating heat capacity at constant pressure ($C_{\mathrm{p,\ cal}}$), the Einstein and Debye models were used to calculate the heat capacity at constant volume. The value of $C_{\mathrm{p,\ cal}}$ was calculated by equation (2.5), where $C_{\mathrm{v,\ group}}$ is the heat capacity contributed by group vibration, and $C_{\mathrm{v,\ skeletal}}$ is the heat capacity contributed by skeletal vibrations.

$$C_{\mathrm{p,\ cal}} = C_{\mathrm{v,\ group}} + C_{\mathrm{v,\ skeletal}} \tag{2.5}$$

$C_{\mathrm{v,\ skeletal}}$ and $C_{\mathrm{v,\ group}}$ represent the heat capacity at the constant volume and are calculated by equation (2.6) and equation (2.7) [Czerniecka-Kubicka et al., 2015; Nishiyama et al., 2021; Pyda, Bartkowiak, and Wunderlich, 1998; Pyda et al., 2004, 2019; Roles et al., 1993; Thybring, 2014].

$$C_{\mathrm{V,skeletal}} = N_{\mathrm{sk}} R \left(\frac{\left(\frac{6.7T}{\theta_1}\right)^2}{1 + \left(\frac{6.7T}{\theta_1}\right)^2} \right) \tag{2.6}$$

$$C_{\mathrm{V,group}} = N_{\mathrm{group}} R \left(\frac{\left(\frac{\theta_E}{T}\right)^2 exp(\frac{\theta_E}{T})}{(exp(\frac{\theta_E}{T}) - 1)^2} \right) \tag{2.7}$$

The parameter N_{skeletal} in equation (2.6) indicates the number of skeletal vibrations and was used to assign the heat capacity contributed by skeletal vibrations, N_{group} indicates the number of group vibrations, and R is the universal constant of gas. N is the number of atoms in the monomer, and $3N$ is the total number of vibrational degrees of freedom ((2.8)). The equation below describes the relations of N, N_{skeletal}, and N_{group} [Czerniecka-Kubicka et al., 2015; Nishiyama et al., 2021; Pyda, Bartkowiak, and Wunderlich, 1998; Pyda, Boller, et al., 1998; Pyda et al., 2004, 2019; Roles et al., 1993; Thybring, 2014].

$$3N = N_{\mathrm{group}} + N_{\mathrm{skeletal}} \tag{2.8}$$

As a principal transition in polymers, the glass transition temperature T_g can affect the other physical performance of polymers. Although polymers have glassy behaviour before T_g, above T_g they are in the rubbery state [Domínguez, 2018]. The GIM method (equation (2.9)) can be used to determine T_g, and all the required input parameters, except Debye temperature, can be calculated with the group contribution method [J. P. Foreman et al., 2008]. In the equation below, the applied strain rate r is assumed to be equal to the angular frequency of 1 Hz in all the calculations of this study.

$$T_g = 0.224\theta_1 + \frac{0.0513E_{\mathrm{coh}}}{N} - 50 + \left(\frac{1280 + 50 \ln \theta_1}{\ln\left(\frac{2\pi\nu_1}{r}\right)} \right) \tag{2.9}$$

The coefficient of linear thermal expansion is an essential parameter and is defined as an alteration in dimension as a function of temperature. This parameter significantly affects the structural and mechanical design of materials. As it is revealed in equation (2.10), the linear thermal expansion coefficient depends on the skeletal heat capacity and cohesive energy [J. Foreman et al., 2010; M. Wang et al., 2013].

$$\alpha = \frac{1.38C_{V,\mathrm{skeletal}}}{3RE_{\mathrm{coh}}} \tag{2.10}$$

As mentioned before, using a set of equations in the GIM method, it is possible to progressively predict the mechanical properties of materials. K_γ represents the bulk modulus of the polymers in the glassy states and can be calculated by equation (2.11). Similarly, the bulk modulus (K) can be predicted with equation (2.12).

$$K_\gamma = 5.68 \frac{E_{\text{coh}}}{V_W} \times 10^6 \tag{2.11}$$

$$K(T) = K_\gamma \left(1 - \left(0.1 + \frac{0.09}{\frac{0.48\times\theta_1}{10^5 L K_{gam}} - 0.9}\right) \int_0^T 0.0067 \exp\left[\frac{(-T - T_g)^2}{2S^2}\right] dT\right) \tag{2.12}$$

Where (S) is the width of the Gaussian function of loss peak and is calculated by equation ((2.13)).

$$S = 6 \left[\frac{1280 + 50 \ln \theta_1}{50 \ln\left(\frac{2\pi\nu_1}{r}\right)}\right]^2 \tag{2.13}$$

The bulk modulus defined in equation (2.12) was used to predict the tensile modulus (E) by equation (2.14). Where $\tan\Delta_g$ is the cumulative loss tangent for the glass transition, calculated by (2.15).

$$E = \frac{K(T)}{(1 + \tan \Delta_g)^2} \tag{2.14}$$

$$\tan \Delta_g = 0.0085 \frac{E_{\text{coh}}}{N_c} \tag{2.15}$$

N_c in equation (2.15) represents the number of degrees of freedom in the chain backbone per structural unit. With the combination of the bulk modulus and elastic modulus, the Poisson's ratio (ν) can be predicted by equation (2.16) [J. Foreman et al., 2010; Guest et al., 2013; Porter and Gould, 2009], assuming isotropic behaviour of the polymer.

$$\nu = 0.5 \left(1 - \frac{E}{3K}\right) \tag{2.16}$$

2.5 Molecular Dynamics

We use MD simulations to model various copolymer properties at the molecular level. We use the Large-scale/ Molecular Massively Parallel Simulator (LAMMPS) [Thompson et al., 2022] through the RadonPy interface [Hayashi et al., 2022] to run MD simulations of copolymer chains using the General Amber force field (GAFF) [J. Wang et al., 2004].

We start with the generation of the initial unit cell structures. A polymer chain is constructed by connecting a repeating unit consisting of 2 monomers via the self-avoiding random walk algorithm [Hayashi et al., 2022]. We equilibrate the system through a meticulous 21-step compression/decompression equilibration protocol proposed by Larsen and co-workers [Larsen et al., 2011]. This protocol orchestrates temperature cycling from 600 K to 300 K, coupled with compression (50,000 atm) and decompression (1 atm) steps. Temperature and pressure

are regulated through NVT and NPT simulations using a Nosé–Hoover thermostat and barostat. After the 21-step equilibration protocol, NPT simulations are run for each system at 300 K and 1 atm until equilibrium is achieved (typically more than 5 ns). Following equilibrium, an extensive suite of copolymer properties is calculated, including density, radius of gyration (R_g), specific heat capacities at constant pressure (C_p) and constant volume (C_v), bulk modulus (K), volume expansion coefficient (γ), and linear expansion coefficient (α).

The density was determined by averaging the mass per unit volume over the equilibrated simulation time, as described in equation 2.17.

$$\rho = \frac{m}{\langle V \rangle} \tag{2.17}$$

Here, m represents the mass of the polymer system and V is the volume of the simulation box. The angle brackets, $\langle \cdot \rangle$, show time averaging.

The radius of gyration, R_g, which assesses the polymer's size and mass distribution, is calculated using equation 2.18.

$$R_g = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r_i - r_{cm})^2} \tag{2.18}$$

Here, N is the number of atoms, r_i represents the position of the atom i, and r_{cm} the center of mass of the polymer [Liu et al., 2017]. For specific heat capacities, the fluctuation formulas were utilized to derive the values at constant pressure and constant volume. The specific heat capacity at constant pressure, C_p, follows equation 2.19 [Allen and Tildesley, 2017; L. Li et al., 2008].

$$C_p = \frac{\langle \delta H^2 \rangle}{m k_B T^2} \tag{2.19}$$

where:

$$\langle \delta H^2 \rangle = \langle H^2 \rangle - \langle H \rangle^2 \tag{2.20}$$

In these equations, $\langle \delta H^2 \rangle$ represents the ensemble average of enthalpy fluctuations, where δH^2 is the variance of the system's enthalpy. The parameter k_B refers to the Boltzmann constant, which relates energy at the molecular level to temperature, and T represents the system's temperature. $\langle \delta H^2 \rangle$ defines the difference between $\langle H^2 \rangle$, the ensemble average of the square of the system's enthalpy, and $\langle H \rangle^2$, the square of the average enthalpy. This difference is the variance of the enthalpy fluctuations, which is key to understanding how the system behaves thermodynamically [Allen and Tildesley, 2017; L. Li et al., 2008].

For mechanical property evaluation, the bulk modulus, K, is computed using equation 2.21 [Riggleman et al., 2010].

$$K = \frac{k_B T \langle V \rangle}{\langle V^2 \rangle - \langle V \rangle^2} \tag{2.21}$$

where $\langle V \rangle$ is the average volume of the system, representing the mean volume observed during the simulation. $\langle V^2 \rangle$ is the average of the square of the volume. $\langle V \rangle^2$ is the square of the average volume, representing the squared mean volume. The term $\langle V^2 \rangle - \langle V \rangle^2$ represents the variance in volume, which quantifies the fluctuations around the average volume. The bulk modulus K is then calculated by dividing the product of the thermal energy $k_B T$ and the average volume $\langle V \rangle$ by this variance, providing a measure of the system's resistance to uniform compression [Allen and Tildesley, 2017].

The volumetric expansion coefficient, γ, is determined as shown in equation 2.22 [Allen and Tildesley, 2017].

$$\gamma = \frac{\langle \delta H . \delta V \rangle}{k_B T^2 \langle V \rangle} \tag{2.22}$$

The linear expansion coefficient, α, for isotropic systems was determined using the equation 2.23.

$$\alpha = \frac{\gamma}{3} \tag{2.23}$$

Lastly, the specific heat capacity at constant volume, C_v, is determined by equation 2.24 [Allen and Tildesley, 2017].

$$C_v = C_p - \frac{K \gamma T \langle \rangle}{m} \tag{2.24}$$

The calculated properties were compared with experimental data extracted from the PolyInfo database [Otsuka et al., 2011] to determine the accuracy of the calculations.

2.6 Machine learning

In this section, we delve into the application of ML techniques to model a diverse range of structural homo-polymers and copolymers, focusing on their various properties. The ML process employed in our study is structured into four key phases: data selection, descriptor calculation (or optimal data representation), model building and training, and performance evaluation. Each of these phases plays a crucial role in ensuring the accuracy and reliability of the predictive models. In the following sections, we will explore each of these steps in detail, beginning with the dataset selection and preparation.

2.6.1 Dataset

We extracted 1221 reference values from databases, books, and handbooks [Brandrup et al., 1999; Mark et al., 2007; Otsuka et al., 2011; "PolyInfo", n.d.; Porter, 1995]. The properties considered in this study include Debye temperature (θ_1), glass transition temperature (T_g), elastic modulus (E), Poisson's ratio (ν), heat capacity at constant volume (C_v), flexural strength (σ), shear modulus (G), and dynamic viscosity (η) for homo-polymers; density (ρ), bulk (K) modulus, and volume expansion coefficient (γ) for copolymers; and heat capacity at constant pressure (C_p) and linear expansion coefficient (α) for both homo-polymers and copolymers. The sources and size of these properties are listed in Table 2.3. As evident, data for C_v, σ are relatively sparse.

In addition, using MD simulations, we curated a dataset of 140 copolymers and their properties. For each copolymer, we computed its ρ, R_g, C_p, C_v, K, γ, and α. This resulted in a final dataset size of 980 for copolymers, comprising 92 random copolymers, 8 block copolymers, and 40 alternating copolymers, each using different types of monomers. In the general structure of the copolymers, there are 97 different and unique monomers. Specifically, monomer A consists of 45 unique monomers, while monomer B comprises 57 unique monomers.

These data were used to train ML models with different algorithms to predict 14 different properties of homo-polymers and copolymers. The copolymer data gathered from the PolyInfo database will be used to validate the MD results rather than for training the ML models. Instead, the copolymer data calculated using MD will serve as inputs for the ML models, although it's important to note that MD simulation data may not always be entirely accurate. For training, validation, and testing of models, the dataset was split into 80% for training, 10% for validation, and 10% for testing, using a random stratified split to ensure adequate representation for each polymer class.

2.6.2 Feature Engineering

Representation

After collecting the dataset, we represented each polymer building block using molecular descriptors. These Descriptors were computed from the SMILES (Simplified Molecular Input Line Entry Specification) notation [Weininger, 1988] of each polymer building block. SMILES expression is a line notation system using an ASCII string to represent the structure of a polymer. Figure 2.2 shows an example of a polymer's SMILES representation. The descriptors calculating using various tools, including Alvadesc [Mauri, 2020], Dragon ["Kode-Chemoinformatics", n.d.], RDKit ["RDKit, Open-Source Cheminformatics", n.d.], and PaDEL2 [Yap, 2011]. Using these tools, we generate over 1400 molecular descriptors, including 1D, 2D, and 3D descriptors. These descriptors offer valuable insights into the chemical and structural characteristics of the polymers, enabling the development of accurate prediction models for the target properties. Additionally, these descriptors have been previously demonstrated to effectively model the properties of interest in similar systems [Kazemi-Khasragh, Blázquez, et al., 2024; Khan and Roy, 2018; S. Wang et al., 2021].

To compute the descriptors, D, for a copolymer, we sum the monomer descriptors in proportion

Table 2.3: Summary of the dataset size for each polymer property, including references for the sources of the data

Property	Type of polymers	From	Size
θ_1	homo-polymer	[Porter, 1995]	140
T_g	homo-polymer	[Brandrup et al., 1999]	394
C_p	homo-polymer	[Mark et al., 2007]	124
E	homo-polymer	["PolyInfo", n.d.]	35
α	homo-polymer	[Mark et al., 2007]	293
ν	homo-polymer	[Otsuka et al., 2011]	29
C_v	homo-polymer	["PolyInfo", n.d.]	13
	homo-polymer	["PolyInfo", n.d.]	13
G	homo-polymer	["PolyInfo", n.d.]	18
η	homo-polymer	["PolyInfo", n.d.]	14
ρ	Copolymer	["PolyInfo", n.d.]	68
		MD	140
C_p	Copolymer	["PolyInfo", n.d.]	48
		MD	140
C_v	Copolymer	MD	140
R_g	Copolymer	MD	140
K	Copolymer	["PolyInfo", n.d.] and MD	14
		MD	140
α	Copolymer	["PolyInfo", n.d.] and MD	5
		MD	140
γ	Copolymer	["PolyInfo", n.d.] and MD	13
		MD	140

to their composition and relative proportion within the polymer chain. For example, for copolymers comprising two monomer components, this can be expressed as

$$D = C_1 \cdot D_1 + C_2 \cdot D_2, \tag{2.25}$$

where C_1 and C_2 represent the fraction length of each monomer in the copolymer, and D_1 and D_2 are sets of descriptors characterizing the molecular properties of the individual monomers. This approach enables us to create a vector representation for a copolymer from the descriptors of its individual building blocks.

The input descriptors for all ML models were standardized using min-max normalization before training. This step ensures that all features contribute equally to the model training process by scaling each feature to a given range, typically [0, 1] [Bishop and Nasrabadi, 2006].

Recognizing the limitations of traditional string-based representations, such as SMILES, in capturing the features of copolymer structures, we acknowledge the need to explore alternative

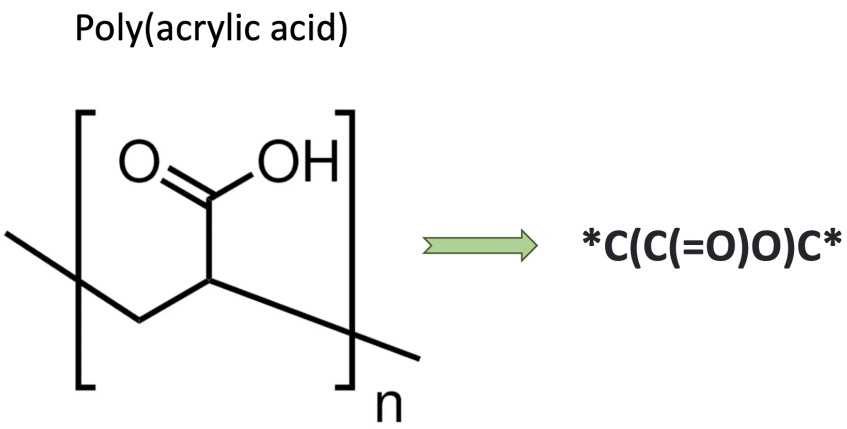

Figure 2.2: SMILES representation of poly(acrylic acid).

representational methods. Our research will encompass a comprehensive examination of various approaches that go beyond SMILES, aiming to better capture the distinctive features of copolymer structures.

One limitation of the descriptor fingerprints is that they cannot incode for differences in how the monomers are connected, e.g., whether in a random, block or alternating copolymer structure (Figure 2.3). To tackle this limitation we can use a graph representation of copolymer structure that allows us to encode more complex information about monomer connectivity [Aldeghi and Coley, 2022; Tao et al., 2022].

In this graph representation, atoms are represented as vertices and bonds as edges, with each edge assigned a weight between 0 and 1 that reflects the probability of the bond appearing in each repeating unit. Through the connection of disparate monomers by edges, we can capture not only the recurring patterns inherent in polymer chains but also the spectrum of potential structural arrangements.

In AB binary alternating copolymers, the chain sequence follows an A-B pattern. The two ends of the repeating unit are connected, so in our representation, all edges have a weight of 1. In random AB copolymers, a variety of arrangements can emerge, encompassing A-A, A-B, and B-B connections. In contrast, block copolymers maintain these patterns, yet A-A and B-B connections prevail over A-B connections in frequency. This discrepancy is reflected in the assigned weights. In the random configuration, uniform weights of 0.5 are assigned to A-A, B-B, and A-B connections. Conversely, in the block configuration, A-A and B-B connections are notably higher at 0.95, while A-B connections are markedly lower at 0.05.

Feature extraction method

Principal component analysis (PCA) is a commonly employed technique the dimensionality reduction. It transforms a set of features into a set of uncorrelated principal components (PCs), which correspond to the directions in the feature space along which the data exhibits the most significant variation. One can then choose a subset of PCs that explain the desired

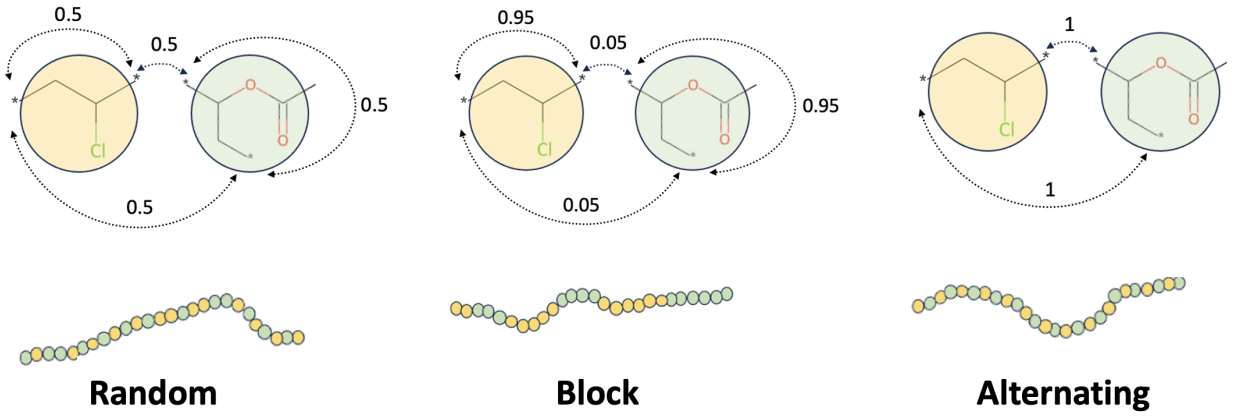

Figure 2.3: Schematic of the graph representation used in this work for random, block, and alternating copolymers.

threshold of the original data. The resulting principal components are orthogonal to each other and capture successively less variation in the data. One of the advantages of PCA is its ability to enhance the interpretability of the dataset. By reducing the dimensionality, the transformed data can be visualized and analyzed more easily. Additionally, employing PCA can lead to computational efficiency, as the reduced feature set requires less computational resources for model development [Cabestany et al., 2005; Jolliffe and Cadima, 2016; Ringnér, 2008].

That said, by acknowledging the capabilities of feature selection methods such as SHAP (SHapley Additive exPlanations) analysis [Babbar et al., 2024], we have deliberated and decided upon Principal Component Analysis (PCA) for feature extraction in our context. This decision stems from several reasons tailored to our objective of TL. Unlike feature selection methods which tailor features to the specific model being used, PCA provides a more generalized approach by transforming the entire feature space into orthogonal principal components. This characteristic is particularly advantageous for TL scenarios, where the aim is to adapt models trained on one task to perform effectively on a related but different task. Notably, employing feature selection often necessitates the inclusion of a larger number of features to capture the most relevant information, thereby increasing computational complexity. In contrast, PCA effectively reduces the dimensionality of the dataset while retaining essential variance information, thus facilitating computational efficiency and ensuring that the resulting feature representation remains robust across different tasks and datasets.

PCA is a popular technique for dimension reduction, but it does have some drawbacks, including the loss of interpretability, as it transforms original features into principal components that can be difficult to understand in terms of the original variables. PCA also assumes linear relationships among features, which may not capture complex, non-linear patterns in the data effectively. Additionally, PCA is sensitive to feature scaling, requiring careful standardization or normalization of data to avoid disproportionate influence from features with different scales [Jolliffe, 2002].

2.6.3 Loss function

In our study, an artificial neural network (ANN) model was formulated to predict C_p with Mean Squared Error (MSE) serving as the primary loss function. To assess the impact of alternative loss functions on model performance, we experimented with five distinct functions: Mean Absolute Error (MAE), Huber Loss, Wing Shape Loss, and a combined loss. The choice of loss function plays a crucial role in assessing the discrepancy between the predicted output and the true target values [Fernández-León et al., 2023]. MSE and MAE are defined as in equation (2.26) and equation (2.27).

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.26}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.27}$$

In the provided equations, y_i represents the expected value of the i-th observation, while $\hat{y}_i$ denotes the predicted value for the same observation. The parameter N is the total number of observations in the dataset. We further employed the Huber loss function to model which is defined in equation (2.28):

$$\text{Huber Loss} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}(y_i - \hat{y}_i)^2, & \text{if } |y_i - \hat{y}_i| \leq \delta \\ \delta(|y_i - \hat{y}_i| - \frac{1}{2}\delta), & \text{otherwise}\end{cases} \tag{2.28}$$

The loss is calculated as the average of the individual losses over all the samples. When the absolute difference between the true value y_i and the predicted value $\hat{y}_i$ is less than or equal to the threshold δ, the loss is computed as half of the squared difference. This corresponds to the squared loss (MSE) in this region. However, when the absolute difference exceeds the threshold, the loss is calculated as δ times the absolute difference minus half of δ squared. This introduces a linear loss instead of a quadratic loss, providing robustness to outliers. The choice of the threshold δ determines the transition point between the squared and linear loss regions and it is equal by one [Meyer, 2021; Y. Wang et al., 2019].

The Wing Shape loss function is a custom loss function used in training machine learning models (equation2.29). This function introduces two parameters: w and ϵ. The parameter w is a weight parameter that determines the threshold for the wing region. ϵ is a parameter that controls the smoothness of the loss function and can be adjusted based on the characteristics of the dataset to achieve the desired balance between robustness and sensitivity to deviations in the predictions. In this study w and ϵ are 5 and 1.5, respectively [Feng et al., 2018, 2020; Fernández-León et al., 2023].

$$\text{Wing Shape Loss} = \begin{cases} w \cdot \log\left(2 + \frac{|y_{\text{true}} - y_{\text{pred}}|}{\epsilon}\right), & \text{if } |y_{\text{true}} - y_{\text{pred}}| < w, \\ |y_{\text{true}} - y_{\text{pred}}| - c, & \text{otherwise,}\end{cases} \tag{2.29}$$

Where c is calculated by:

$$c = 1.0 - \log(1.0 + \frac{w}{\epsilon}) \tag{2.30}$$

The combined loss function is a weighted sum of multiple loss functions which are discussed above. The formula for the combined loss function can be written as:

$$\text{Combined Loss Function} = \frac{1}{4}(\text{MSE} + \text{Wing Shape Loss} + \text{Huber Loss} + \text{MAE}) \tag{2.31}$$

By incorporating multiple loss functions, the combined loss function aims to leverage the strengths of each individual loss function and provide flexibility and a more comprehensive measure of the model's performance.

2.6.4 Algorithms

In this study, a variety of regression-based machine learning algorithms were employed to model and predict the properties of homo-polymers and copolymers. The selection of algorithms was based on their capability to manage diverse datasets effectively, proven success in analogous scientific fields, dataset size, complexity of property prediction tasks, and other pertinent considerations. The primary goal was to develop models that could reliably predict multiple polymer properties based on their molecular representation, thereby advancing the understanding and application of polymer science.

Random forest

Random Forest (RF) as a ML algorithm was introduced by Breiman and Cutler [Breiman, 2001]. This algorithm is based on the decision tree and combines a group of tree predictors [Palmer et al., 2007]. Each decision tree is made in the forest by choosing the section from the initial dataset using the bootstrap method and electing the random number of all variables in each decision node. Figure 2.4 illustrates how multiple decision trees are combined to form a random forest. Each tree is trained on a random subset of the data and features, and the final prediction is made by averaging the predictions of all individual trees.

Random Forests [Breiman, 2001] were implemented using the RandomForestRegressor class of the scikit-learn library [Kramer and Kramer, 2016] for solving regression problems. The RandomForestRegressor class in scikit-learn is specifically designed for regression tasks, where the goal is to predict continuous numerical values rather than categorical outcomes. We use a RF [Breiman, 2001] as a baseline to predict the properties of polymers from the molecular descriptors. It allows us to efficiently uncover the relationships between the polymer's structural attributes and its properties and advances our understanding of polymer behaviour and performance. RFs have been previously shown to be effective in numerous predictive tasks involving molecules and materials [Kazemi-Khasragh, Blázquez, et al., 2024; Palmer et al., 2007; Xiao et al., 2023].

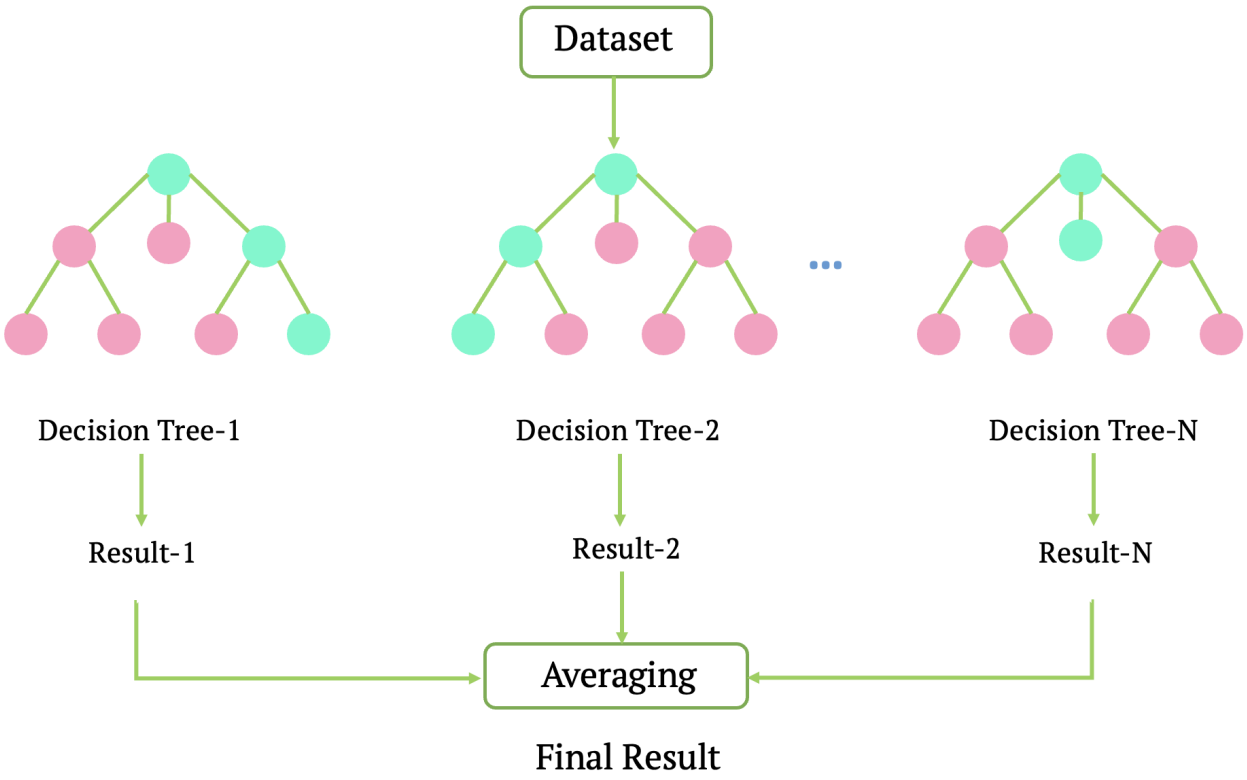

Figure 2.4: schematic architecture of a random forest.

We implemented RFs to predict Debye temperature, T_g, C_p, E, α, ν of homo-polymers and density, R_g, C_p, C_v, γ, α, and K of copolymers. We train the models using the mean-squared error (MSE) loss function. Hyperparameter optimization was performed for the RF models using the parameter grid search method [Pedregosa et al., 2011]. The parameters tuned included the number of estimators (n_estimators), which was varied among 30, 50, 100, 200, 500, 600, 700, 800 and 900. The n_estimators value indicates the number of decision trees used in RFs and controls the model's complexity and performance. The maximum depth of the trees was set to 10, 20, 80, 100,200, and 300, while the maximum number of features considered for splitting a node was either 'log2' or 'sqrt'. Additionally, the minimum number of samples required to split an internal node was adjusted between 2 and 12, and the minimum number of samples required to be at a leaf node was set between 1 and 5. For all other parameters, we used the default values. The hyperparameters used to train the models are listed in Table 2.5.

In addition to the standard RF models, we employed multi-task RF models to predict multiple properties simultaneously. Multitask learning is a form of inductive transfer where the model is trained on multiple related tasks at the same time, leveraging the commonalities and differences across tasks to improve learning efficiency and prediction accuracy. This approach is particularly advantageous in the context of polymer property prediction, as it allows the model to utilize the shared information between different properties, leading to more robust and generalizable predictions [Caruana, 1997; Ramsundar et al., 2015; Y. Zhang and Yang, 2021].

Table 2.5: Hyperparameters' values for the RF models

Model	I	II	III	IV	V
θ_1	600	100	2	1	'sqrt'
T_g	500	100	2	1	'sqrt'
C_p of homo-polymers	500	10	2	1	'sqrt'
E	900	300	2	1	'sqrt'
Linear expansion of homo-polymers	100	100	2	1	'sqrt'
Poisson ratio	100	100	2	1	'sqrt'
Density	100	20	2	1	'sqrt'
R_g	200	10	2	1	'log2'
C_p of copolymers	500	80	2	1	'sqrt'
C_v	500	20	2	1	'sqrt'
Bulk Modulus	250	20	12	1	'sqrt'
Linear expansion of copolymers	500	80	2	1	'sqrt'
Volume expansion	200	20	2	1	'sqrt'

[I] n_estimators
[II] max_depth
[III] min_samples_split
[IV] min_samples_leaf
[V] max_features

For the multitask RF models, we focused on predicting the following properties of copolymers derived from MD simulations: density, R_g, C_p, C_v, volume expansion, linear expansion, and bulk modulus.

To evaluate the importance of each descriptor, we also use the RF to compute the feature importance. We compute this using the mean decrease in impurity, as implemented in the scikit-learn [Kramer and Kramer, 2016] package.

Neural networks

Neural networks (NN) are a class of ML algorithms and inspired by the structure and function of the human brain, NN consist of layers of interconnected nodes, or neurons, where each connection has an associated weight. Weights are parameters that scale the input data to each neuron, and are obtained during the training process to minimize the error between the model predictions and the outcomes [Abdi et al., 1999; Anderson, 1995; Bishop, 1994]. Activation functions are mathematical functions applied to the output of each neuron. They introduce non-linearity into the model, enabling it to learn and represent complex patterns. Common activation functions include ReLU, which outputs the input directly if it is positive, and sigmoid or tanh functions, which map inputs to a bounded range. Activation functions are crucial for allowing the network to model intricate relationships within the data [Sharma et al., 2017].

Each neuron also has a bias, an additional parameter that helps shift the activation function. This shift allows the model to better fit the data by providing flexibility in the neuron's

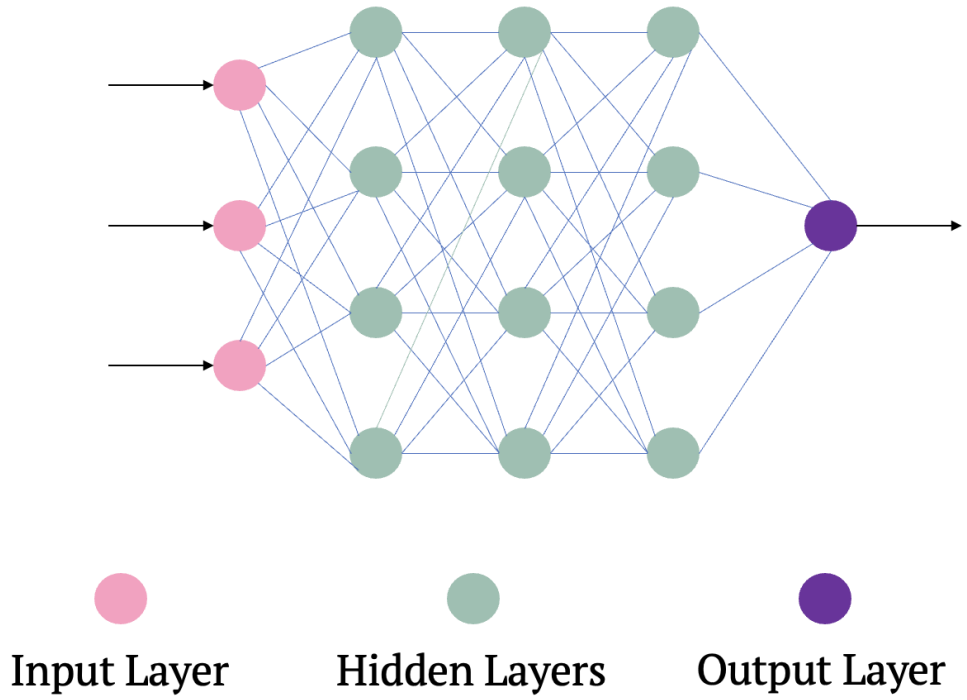

Figure 2.5: schematic architecture of neural networks.

output. By adjusting the bias, the network can more effectively learn and capture patterns in the data. The architecture of a neural network typically includes an input layer, one or more hidden layers, and an output layer (Figure 2.5). Each layer consists of a set of neurons, and the neurons in adjacent layers are fully or partially connected. The input layer receives the input data, and the output layer produces the final predictions. Hidden layers are responsible for transforming the input into meaningful patterns through a series of non-linear transformations.

In NN, dropout is a technique where nodes (neurons) in both the input and hidden layers are randomly deactivated during training. This process involves temporarily removing all connections to and from these dropped nodes, effectively altering the network's structure for each training iteration. The probability controls the fraction of neurons that are dropped out, creating a new, smaller network architecture each time. This approach helps to combat overfitting by ensuring that the network does not become overly reliant on specific neurons or their connections, thus promoting more robust learning and better generalization [Abdi et al., 1999; Anderson, 1995; Bishop, 1994].

A NN model is employed to predict the C_p of homo-polymers with five different loss functions (MAE, MSE, Hubber loss, wing shape, and combined loss function). The NN model was built using the Keras library with a sequential architecture. For this model hyperparameter tuning was performed using Optuna, a hyperparameter optimization framework [Akiba et al., 2019]. This process was conducted over 100 runs to optimize the model's performance. Optuna systematically explored to find the best learning rate, regularization parameters, and the best

combination for maximizing the model's accuracy. The learning rate controls the step size during the optimization process, determining how quickly or slowly the model's weights are updated. A smaller learning rate may result in more precise learning but could take longer to converge, while a larger learning rate might speed up the process but risks overshooting the optimal solution [Moreira and Fiesler, 1995]. Regularization parameters, on the other hand, are used to prevent overfitting by adding a penalty term to the loss function. This helps in controlling the model's complexity and encourages it to generalize better to unseen data [Goodfellow, 2016; Goodfellow et al., 2016].

The hyperparameters, such as epochs ranging from 500 to 1000, strike a balance between under-fitting and over-fitting, the number of hidden layers varying between 5 and 20 for architectural complexity. Additionally, we dynamically determined the number of units in each layer within the range of 100 to 900, allowing flexibility in capturing complex patterns.

It consisted of 15 dense layers, with the ReLU activation function applied to the hidden layers. The training dataset was randomly split into training and test sets, with a test size of 35%. The model was trained using the Adam optimizer and mean squared error (MSE) as the loss function. After building the NN model to predict C_p of homo-polymers, transfer learning (TL) was performed to further extend the predictive capabilities of the model. This method was used to predict other properties, namely C_v, σ, G, and η. As we mentioned previously, data for C_v, σ, G, and η are relatively sparse. So they are not suitable to make ML models. For the transfer learning process, the layers of the initial model were frozen and were transferred to the new models. By keeping the weights of the initial model unchanged, the focus was on training new layers dedicated to predicting the additional properties. To determine the number of layers to add during the transfer learning process, we conducted experimentation to strike a balance between model complexity and predictive performance. Through iterative testing, we evaluated the performance of the model with varying numbers of additional layers and observed the trade-offs between model complexity and computational efficiency. Ultimately, we found that adding five new layers provided a suitable level of complexity to capture the nuances of the additional properties while avoiding overfitting. So, five new layers were added to the end of the transfer layers with weights in the range of 20-100. A low learning rate was utilized, allowing the model to gradually adapt to the new property prediction tasks while retaining the valuable knowledge acquired from the initial model. Regarding the sensitivity of the transfer learning model to weight initialization, we observed that small variations in weight initialization could impact the convergence and stability of the training process. Specifically, subtle changes in the initial weights of the added layers affected the rate of convergence and the final performance metrics of the model. To mitigate this sensitivity, we employed techniques such as careful selection of initialization methods and fine-tuning of hyperparameters to ensure robustness and reproducibility across training runs.

These properties for prediction using transfer learning were selected based on their relevance and connection to the initial property, C_p, and their potential impact on material characterization. In addition to their relevance to the initial property, the chosen properties belong to different categories, namely mechanical, thermal, and rheological properties. This diverse selection was motivated by the interest in investigating the relationships and connections between these different aspects of the polymers' behaviour.

Another neural network architecture was utilized in this study to predict the properties of copolymers calculated using MD. This architecture is an extension of the directed message passing neural network (D-MPNN) implemented by Aldeghi [Aldeghi and Coley, 2022], known as the weighted, directed message passing neural network (wD-MPNN). Building upon the foundation of D-MPNNs, the wD-MPNN enhances the modelling of polymer structures by capturing more detailed structural information and improving the accuracy of property predictions. By weighting the edges based on the specific copolymer configuration and incorporating stoichiometry information, the wD-MPNN learns a more nuanced and effective representation of polymer structures compared to the D-MPNN.

The D-MPNN is a class of GNN that operates on molecular graphs. GNN is a type of NN designed to operate on graph-structured data, where nodes represent entities and edges represent relationships between them. The process begins with the assignment of feature vectors to each node and edge, describing the properties of the corresponding atoms and bonds. These feature vectors encode information such as atom type, formal charge, etc. Through iterative message passing, information is exchanged between neighbouring atoms via directed edges, with each edge transmitting a message to update the feature vectors of adjacent nodes. This iterative process allows information to propagate through the graph, refining the feature vectors based on directional relationships between atoms. After multiple iterations, the feature vectors are aggregated to produce a comprehensive representation of the molecule, which is subsequently fed into a feed-forward neural network for predicting molecular properties [Aldeghi and Coley, 2022; Flam-Shepherd et al., 2021; Gilmer et al., 2017; Heid and Green, 2021; Stokes et al., 2020; Yang et al., 2019].

The models were also optimized using the Optuna framework. Each wD-MPNN was tuned for 100 iterations using various parameters to optimize its performance according to the MSE objective function. The tuned parameters comprised of the number of epochs, sampled from [100, 250, 500]; the MPNN *depth* (number of message passing steps), adjusted between 3 and 10 layers; the dropout rate, sampled between 0 and 0.5; the learning rate, adjusted between 1e-3 and 1e-6; and the number of layers in the final FFNN, set between 1 and 10. The FFNN width was set between 500 to 2500, with the ReLU activation function used and no regularization applied. The optimizer used was Adam. Additionally, the objective in Optuna was to minimize the MSE of the model predictions. By doing so, we aimed to find the set of hyperparameters that would result in the most accurate predictions for our polymer property datasets. The model was trained using the MSE loss as implemented in scikit-learn. The aggregation function for message passing in the models was set to 'sum', and the number of warm-up epochs was set to 10. Three folds were used for cross-validation. Hyperparameters were used for wD-MPNN models outlined in Table 2.7. For all other parameters, we used the default values in the [Aldeghi and Coley, 2022] implementation.

2.6.5 Evaluation Metrics

Three statistics were calculated to validate and test the developed models made with the ML method: the squared correlation coefficient (R^2), Mean-squared-Error (MSE) and the Mean-Absolute-Error(MAE). It is possible to calculate R^2, MSE, and MAE by equation (2.32), and equation (2.33) and equation (2.34), respectively [Alpaydin, 2020; Gracheva et al.,

Table 2.7: Hyperparameters' values for the wD-MPNN models

Model	epochs	depth	dropout	learning rate	ffn_num_layers
ρ	500	6	0.5	1e-6	3
R_g	250	8	0.4	1e-6	4
C_p	500	6	0.5	1e-5	3
C_v	500	2	0	1e-6	3
K	250	5	0	1e-5	2
α	250	7	0.1	1e-6	5
γ	250	3	0.3	1e-5	3

2021].

$$R = \sqrt{\frac{\sum_{j=1}^{N}(y_{0,j} - \bar{y})(y_{p,j} - \bar{y})}{\sqrt{\sum_{j=1}^{N}(y_{0,j} - \bar{y})\sqrt{\sum_{j=1}^{N}(y_{p,j} - \bar{y})}}}} \tag{2.32}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_0 - y_p)^2 \tag{2.33}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_0 - y_p| \tag{2.34}$$

Where N, $y_{0,j}$, $y_{p,j}$, and $\bar{y}$ are the number of observations, measured values, predicted values, and mean predicted values of the properties, respectively (for $j = 1$ to N). The value of R^2 ranges from 0 to 1, where a higher number indicates better accuracy. For MAE and MSE, lower values indicate better accuracy.

Chapter 3

Facilitating polymer property prediction with ML and GIM

3.1 Group Interaction modelling results

The input parameters described in the Methodology section must be determined to predict the properties using the GIM method. Table 3.1 provides a compilation of GIM parameter values for several example polymers: polyethylene (PE), polypropylene (PP), polyvinyl alcohol (PVA), polyvinylidene fluoride (PVDF), polyether ether ketone (PEEK), and polybutene (PB-1). The monomer lengths were calculated using the Avogadro program, while other input parameters ($E_{\mathrm{coh}}, V_{\mathrm{w}}, N, N_{\mathrm{c}}$, and M) were estimated using the group contribution method.

Table 3.1: Input parameters of six selected polymers

Polymer	**Length (Å)**	$E_{\mathbf{coh}}$ **(J/mol)**	V_w **(cc/mol)**	**N**	N_c	**M (g/mol)**
PE	2.2	9000	20.5	4	4	28.0313
PP	2.4	13500	30.7	6	6	42.04695
PVA	2.5	22000	25.2	6	6	44.02621
PVDF	2.3	18000	20.46	4	4	64.01246
PEEK	17.1	105100	151.72	17	80	294.277
PB-1	2.6	18000	40.9	7	8	56.1

Vibrational spectroscopy of the polymer's monomer simulated in the gas phase has been used to estimate Debye temperature and Einstein temperature, which are input parameters of the GIM method. Table 3.3 shows the Debye temperature for several polymers ranging from 200 K to 600 K. The reference values of θ_1 from the book of David Porter ($\theta_{1,\mathrm{Porter}}$) are compared with the calculated ones. The calculated values are sorted in column 3 ($\theta_{1,\mathrm{simulation}}$), and the error versus the reference values are in column 4. Although the calculated θ_1 follows the reference behaviour, the discrepancies between the reference values ($\theta_{1,\mathrm{Porter}}$) and the

computed using electronic simulations ($\theta_{1,\text{simulation}}$) are noteworthy. Table 3.3 also includes the Debye temperature calculated with machine learning and the error versus the reference values. The following section analyzes the results of machine learning.

Table 3.3: Compare θ_1 predicted by ML and θ_1 calculated by simulation.

Polymer	$\theta_{1,\textbf{Porter}}$	$\theta_{1,\textbf{simulation}}$	Error*	$\theta_{1,\textbf{ML}}$	Error**
PE	316	344	8.8	380	20.2
PP	449	434	3.3	429	4.4
PVA	439	422	3.8	400	8.8
PVDF	209	227	8.6	261	24.8
PEEK	550	469	14.7	475	13.6
PB-1	389	394	1.2	374	3.8
Poly (ethylene isophthalate)	550	520	5.4	488	11.2
Poly (bisphenol-A terephthalate)	550	516	6.1	506	8
Poly (oxy-m-phenylene)	550	588	6.9	533	3.09
Poly (methyl-p-xylylene)	550	519	5.6	536	2.5
Poly (cyano-p-xylylene)	550	570	3.6	541	1.6
Poly (p-xylylene)	550	532	3.2	498	9.4
Poly (phenylene sulfide)	550	591	7.4	476	13.4
Poly (m-phenylene isophthalate)	550	576	4.7	535	2.7
Poly (oxy (2,6-dimethyl-1,4-phenylene))	550	552	0.3	496	9.8

* %Error ($\theta_{1,\text{simulation}}$, $\theta_{1,\text{Porter}}$)
** %Error ($\theta_{1,\text{ML}}$, $\theta_{1,\text{Porter}}$)

The input parameters of the GIM method shown in Table 3.1 are complemented with the calculated Debye temperature θ_1. Therefore, the polymers' thermal properties and mechanical behaviour have been analyzed by (2.5) - (2.15) as a function of temperature using the GIM method. The calculated values of the glass transition temperature, heat capacity at constant pressure, elastic modulus, linear thermal expansion, and Poisson's ratio for the several instance polymers are listed in Table 3.5. These properties have been calculated at room temperature.

Table 3.5: The properties of the selected polymers using the GIM method with $\theta_1$from simulation method.

Polymer	$\mathbf{T_g}$ **(K)**	$\mathbf{C_p}$ **(J/mol K)**	**E (MPa)**	α **[10^{-6} 1/K]**	ν
PE	221.3	64	1208	130	0.418
PP	241.6	70	1207	193.51	0.418
PVA	329.6	72	2479	276	0.41
PVDF	243.2	85	1627	245	0.43
PEEK	501.5	371	2669	360	0.47
PB-1	253.1	105	283	155	0.47

3.2 Machine learning results

The machine learning approach has been implemented and tested in 6 different models, on model per polymer property. In the beginning, an abundant number of descriptors were used to build the primary RF models. However, importance features analysis has been used to extract the more important and relevant descriptors and exclude the irrelevant ones to increase the model's performance. This analysis has been applied to all six machine learning models. After analyzing and finding the most relevant and important descriptors, new RF models were built to make the prediction. For this purpose, top descriptors from the importance features technique have been selected, and the latest forecast has been done. Table 3.7 shows the material properties of six selected polymers calculated using the described ML models. These results are in good agreement with the measured values for the selected polymers.

Table 3.7: The properties of the selected polymers using ML method

Polymer	T_g **(K)**	C_p **(J/mol · K)**	E **(MPa)**	α **[10^{-6} 1/K]**	ν
PE	236	40.8	1140	107.137	0.4590
PP	259	65.4	1065	205.45	0.4203
PVA	323	68.2	2733	226.99	0.3836
PVDF	258	76.0	1477	279.20	0.3646
PEEK	387	325.0	2239	380.03	0.4119
PB-1	237	116.0	482	111.90	0.4600

The performance of six RF models, each with selected descriptors, and sorted by their squared correlation coefficient of the training set (R^2_train set) and test set (R^2_test set), and MAE are analyzed in Table 3.9.

MAE values offer insight into the accuracy of our models in predicting different material properties. A lower MAE indicates that the model's predictions are, on average, closer to the true values, signifying higher accuracy. It's essential to note that the interpretation of MAE can vary depending on the specific properties under consideration. MAE is expressed in the same units as the properties being predicted. As a result, direct comparisons of MAE values across different properties can be challenging, primarily due to the inherent variations in units among these properties.

To address this challenge, we also consider the correlation coefficient (R^2), a crucial metric for assessing model performance. A high R^2 value signifies a strong correlation between the predicted and actual values, further indicating the model's accuracy. In our study, R^2 of the train set ranges from 0.83 to 0.958, and R^2 of the test set ranges from 0.85 to 0.967.

The model for the elastic modulus (ML_E) shows the lowest value of R^2 of the test set, 0.85 for the given property. Also, the MAE of this model is 229 MPa. On the contrary, the model for the heat capacity, ML_C_p, performs better than the other five models, with a mean absolute error of 12.1 J/mol $\cdot$ K and R^2 of the test set of 0.965.

Table 3.9: Performance of each RF model

Predicted Parameters	**R^2 train set**	**R^2 test set**	**MAE**
ML_θ_1	0.95	0.934	18.6 K
ML_T_g	0.93	0.83	13.1 K
ML_C_p	0.965	0.955	12.1 J/mol $\cdot$ K
ML_E	0.85	0.84	229 MPa
ML_α	0.91	0.90	2.1×10^{-5} [1/°C]

In Figure 3.1 the predicted values of various material properties using the RF method are compared to the expected (experimental) values for both the training and test sets. For Debye temperature 3.1(a), most of the predicted values are close to the expected values, except for a few outliers at 550 K. The expected values for Debye temperature were obtained from David Porter's book, with a value of θ_1 estimated at 550 K for polymers containing phenyl rings [J. Foreman et al., 2010]. Hence, the deviation of the model at 550 K is more significant than other temperatures, resulting in some data points deviating from the diagonal line on the plot.

For glass transition temperature Figure 3.1(b), the blue points representing the training set are closer to the expected = predicted line than the black points representing the test set. However, for heat capacity, linear thermal expansion, and Poisson ratio Figure 3.1(c, e, and f), most data points lie on the expected = predicted line, indicating good agreement between predicted and expected values.

In contrast, for elastic modulus 3.1(d), both the training and test set data points are scattered compared to the other five plots. The insets of 3.1 display each model's histogram of the MAE distribution. The mean absolute error histogram is a way to show the comparative forecasting performance. A model with higher accuracy will have a higher density of errors in the centre of the histogram (MAE = 0) and a lower error rate overall. Based on the results, the six RF models can predict material properties from chemical structures with reasonable accuracy, making them useful for predicting the behaviour of new materials.

As mentioned above, the important feature analysis is used to determine the relative importance of descriptors in predicting the target variable in the model. It helps to identify the features that have the most influence on the model's predictions. The plot of important descriptors and their corresponding importance scores are provided in Fig 3.2. This plot visually represents the relative importance of each feature in predicting the target variable. Moreover, the number and name of the applicated descriptors for each ML model are listed in Table 3.11. In the model to predict Debye temperature θ_1, five Continuous and Data-Driven Descriptors (CDDD) descriptors are used. CDDD descriptors are molecular structure descriptors that offer a unique, compact, and continuous vector representation for each compound [Winter et al., 2019].

"ChiA" descriptors as 2D matrix-based descriptors for prediction T_g reflect information about inter atomic distances, bond distances, ring types, planar and non-planar systems, and atom types. A small "ChiA" indicates a small inter atomic distance, which results in a low degree of freedom for rotation and leads to high T_g [X. Yu and Huang, 2017]. The third important descriptor for the ML_T_g model is the rotatable bond fraction (RBF). The number of rotatable bonds, which shows the stiffness of the chain, can affect the T_g, and a larger number of rotatable bonds will have a lower T_g. Also, SpMax_EA is an edge adjacency index derived from the H-depleted molecular graph and encodes the connectivity between graph edges. This descriptor leads to eigenvalue from the edge adjacency matrix weighted by bond order and reflects molecular shape [X. Yu and Huang, 2017] GATS1v, as a 2D autocorrelations descriptor, encodes the relative position of atoms or atom properties by calculating the separation between atom pairs in terms of the number of bonds [Sliwoski et al., 2016]. Although various factors affect the glass transition temperature of the polymers, rigidity and stiffness are important factors related to this property. Thus, these five descriptors can predict glass transition sufficiently [X. Yu and Huang, 2017].

SAtot, as a molecular descriptor, presents the total surface area [Poša et al., 2014]. The relation between the surface area and heat capacity C_p is obvious. Moreover, "Autocorr3d" descriptors exist in an important descriptor list of RF_C_p, ML_E, and ML_α models. These descriptors capture the patterns and relationships between atoms at different distances and orientations. These patterns can be indicative of the molecular characteristics, provide important 3D-structural information and translate the relative position of atoms to the atom properties [Sliwoski et al., 2016]. From a physical perspective, it is not easy to explain the exact effect of these descriptors on the properties. However, it is obvious that the position of the atoms has an extreme effect on the thermodynamic properties of the polymers, such as heat capacity, elastic modulus, and linear thermal expansion.

SpMAD_B(s) as a 2D matrix-based descriptor shows the spectral mean absolute deviation

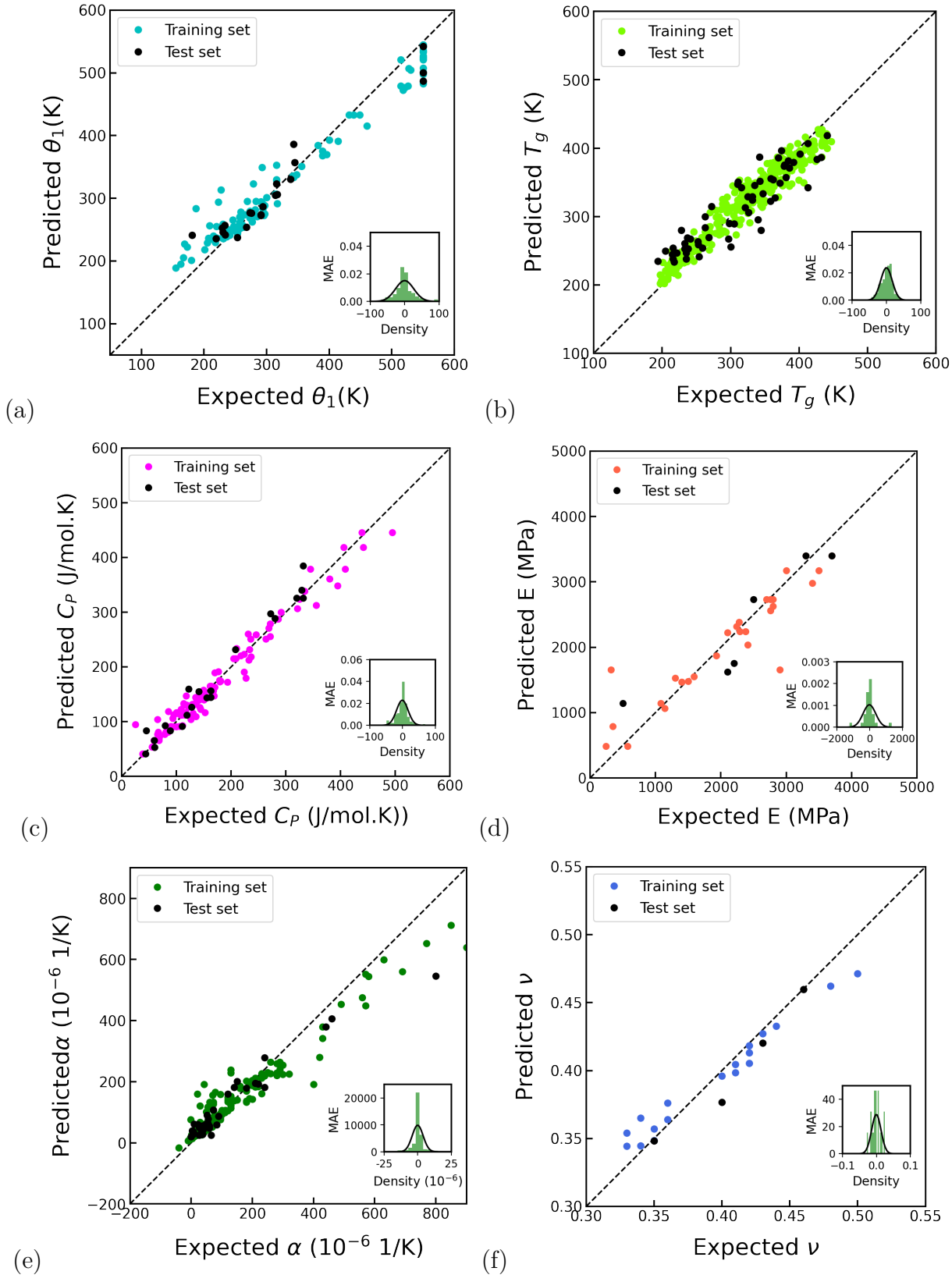

Figure 3.1: Expected vs. predicted values of (a) Debye temperature, (b) glass transition temperature, (c) heat capacity, (d) Elastic modulus, (e) linear thermal expansion, (f) Poisson ratio, and the distribution of MAE of each RF model.

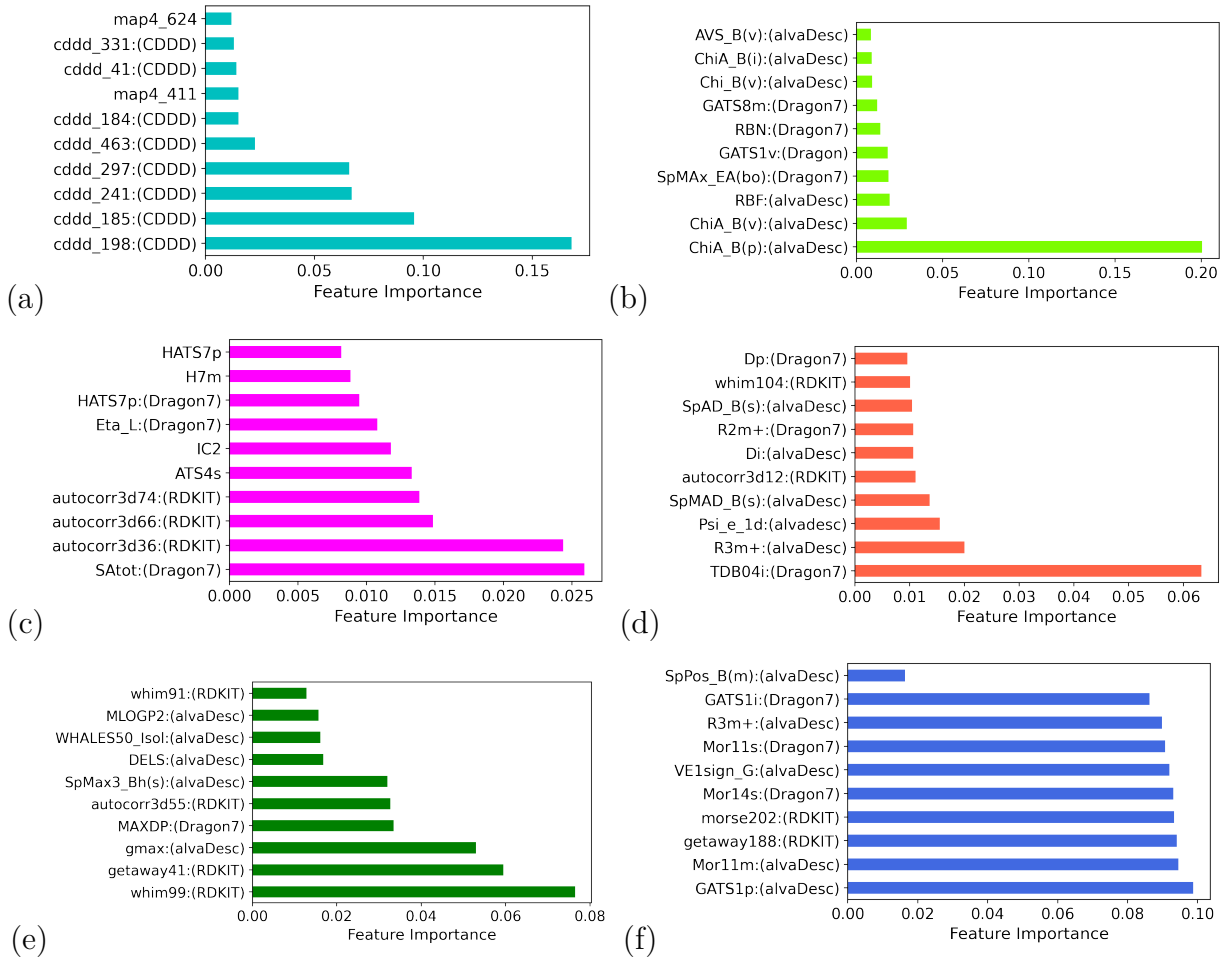

Figure 3.2: The plot of top 20 important descriptors and their corresponding importance scores for (a) Debye temperature, (b) glass transition temperature, (c) heat capacity, (d) Elastic modulus, (e) linear thermal expansion, (f) Poisson ratio model.

from the Burden matrix. The given information of the Burden matrix (atomic number of the atoms, type of the bonds, etc) is useful information that can affect the molecules' mechanical behaviour for prediction of the elastic modulus E. TDB04i is a 3D Topological distance-based descriptor and is the third important descriptor in ML_E model [Kowalewski and Ray, 2020]. These topological distance-based descriptors contain significant structural relations like conformation information. The conformation and crystallinity of polymers influence mechanical properties as well as modulus, as a more ordered and crystalline structure generally leads to a higher modulus, while a less ordered and amorphous structure typically results in a lower modulus. R3m+ relates the maximal autocorrelation of lag3 divided by mass and Psi_e_1d, is a 1D descriptor that belongs to the family of electron topological state indices [Sestraş et al., 2012]. These features are related to the molecular size, shape, and electronic properties of the polymer chains. Elastic modulus (E) is a measure of a material's stiffness and its ability to resist deformation under stress. It is influenced by the size, shape, and electronic properties of the polymer chains. Therefore, descriptors that capture these characteristics are critical for accurately predicting E. Table 3.11 reveals that the linear thermal expansion model relies on five descriptors, where the Weighted Holistic Invariant Molecular (WHIM) is the most influential. The WHIM descriptor encodes information about size, shape, symmetry, and atom distribution from the atomic coordinates. Additionally, GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors emerge as significant features in both ML_ α and ML_ν. Furthermore, Gmax is the maximum geometrical distance between the atoms, and MAXDP is the Maximum positive difference between the intrinsic states of the atoms of a molecule, which are also among the important descriptors [dos Reis et al., 2014].

For the ML_ α model, geometrical and topological factors, flexibility and stiffness of the bonds, and cohesive forces between molecules are considered the principal parameters [Gracheva et al., 2021]. GATS1p, Mor11m, and getaway188 are used to build the final version of the ML_ν model. Mor11m reflects the three-dimensional structure of the polymers, which is relevant in understanding their mechanical properties such as Poisson ratio [Faghihi et al., 2019]. GATS1p is a descriptor that takes into account the polarizability of molecules [Mauri, 2020], which is an important factor in determining the intermolecular forces that affect the Poisson ratio. Lastly, getaway188 is a descriptor that considers the geometrical and topological factors of molecules, which also play a crucial role in determining the Poisson ratio. By incorporating these descriptors into the ML model, it can better capture the complex relationships between the molecular structure and Poisson ratio, leading to more accurate predictions.

3.3 Comparison ML and GIM results

Machine learning and group contribution methods offer different approaches for predicting the properties of polymers. ML leverages data-driven models and algorithms to learn complex patterns and relationships between molecular features and properties. It can capture non-linear relationships and adapt to various polymer systems, making it suitable for predicting a wide range of properties. ML methods require large datasets for training and may exhibit high accuracy if trained on diverse and representative data. On the other hand, group contribution methods utilize predefined rules or parameters derived from empirical data. These methods simplify the prediction process by breaking down the polymer structure into functional groups

Table 3.11: Name and number of the descriptors in each ML model

Model	**Descriptors**
ML_θ_1	CDDD1_185, CDDD1_198, CDDD1_241, CDDD1_297, CDDD1_463
ML_T_g	ChiA_B (p, v), RBF, SpMAX_EA, GATS1v
ML_C_p	SAtot, autocorr3d36, 3d_66
ML_E	autocorr3d12, SpMAD_B(s), TDB04i, R3m+, Psi_e_1d
ML_α	whim99, getaway41, Gmax, autocorr3d55, MAXDP
ML_ν	GATS1p, Mor11m, getaway188

or fragments and estimating the contributions of these groups to the overall property. Group contribution methods are typically based on empirical correlations and can provide quick estimates of properties with limited data. However, they may have limitations in capturing the complexity and subtle variations in polymer systems, leading to lower prediction accuracy compared to ML approaches. The Debye temperature is estimated using Gaussian16 software for electronic structure calculations, and Table 3.3 compares ML models for several polymers. The values from the book of Porter are used as a reference to compute the error of the calculated values. Notably, the values calculated via Gaussian16 software in the gas phase show better accuracy than those calculated with ML models. Because the Debye temperature is an input parameter for the GIM method, the predicted properties by the GIM method can be calculated considering the Debye temperature from electronic structure calculations or ML models.

Figure 3.3 presents the percentage error of each method compared with experimental values.%Error (ML) is the percentage error of the ML prediction for each property. %Error (GIM/θ_1 from ML) and %Error (GIM/θ_1 from simulation) are the calculated values of properties which contributed to finding the error coming from the GIM model. The difference between GIM/θ_1 from ML and GIM/θ_1 from the simulation is that Debye temperature in the first model is predicted by ML, but in the second model, is computed by Gaussian software (electronic simulation method).

By comparing the errors of Figure 3.3, generally, the ML technique can predict the properties better than the other two GIM methods. ML is extracting the relation of the properties and chemical structure.

In simple words, ML was trained by patterns and relationships from large data sets. ML has discovered the complex connection between the input (descriptors) and the output (predicted properties) and inferred it by a fit model for each single data set. Chemical structure dictates the properties of the polymers, and the ML model can learn to associate them with each other without considering any semi-empirical or analytical model. The drawback of the ML models is the need for a large amount of experimental data and the definition of the relevant descriptors. By considering the percentage error of the linear thermal expansion, Poisson's ratio and elastic modulus of GIM/θ_1 from ML and GIM/θ_1 from simulation models are close between them. Hence, we conclude these properties are less sensitive to θ_1. However, the

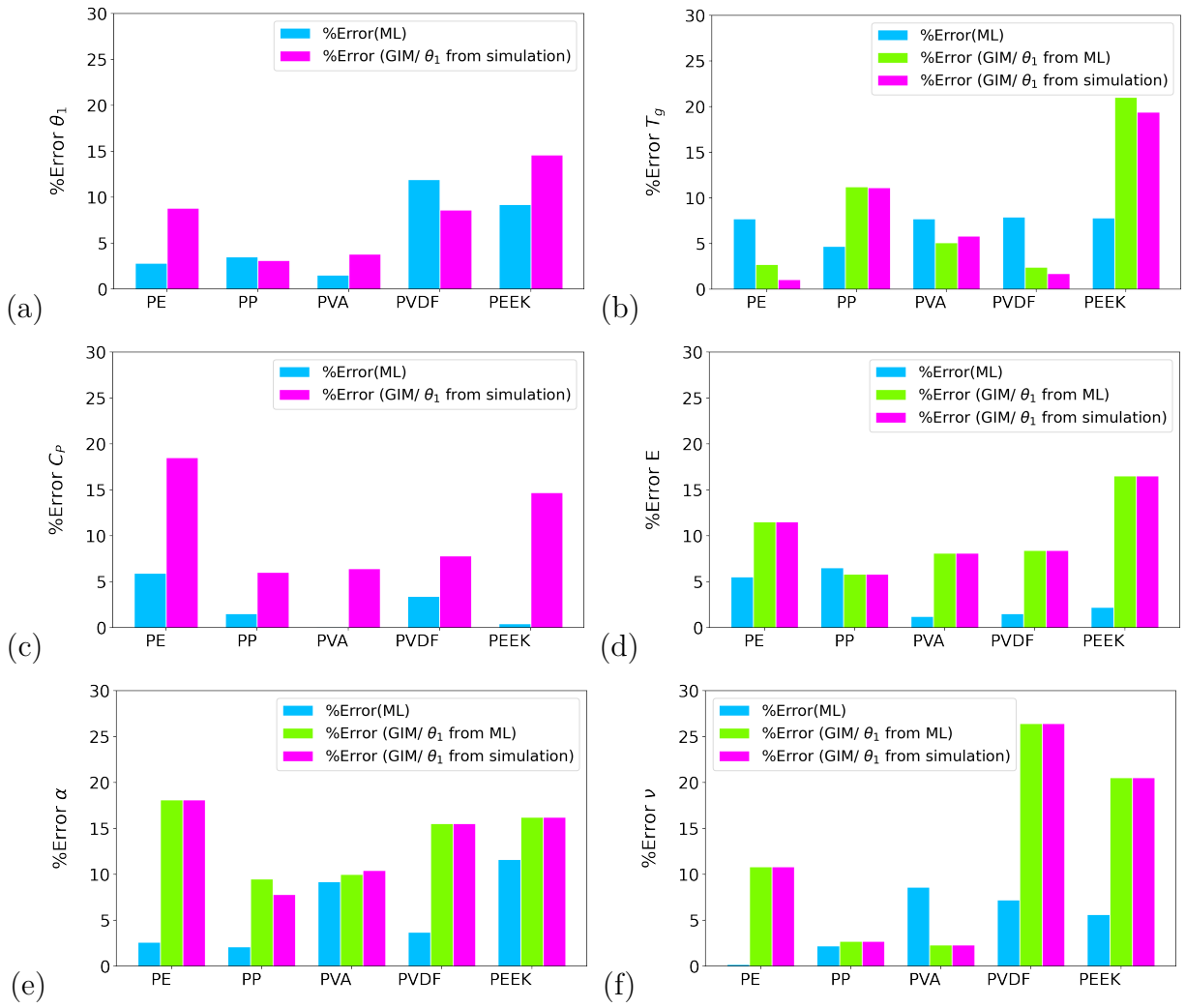

Figure 3.3: Error percentage of different polymers for (a) Debye temperature, (b) glass transition temperature, (c) heat capacity, (d) Elastic modulus, (e) linear thermal expansion, (f) Poisson ratio calculated by RF and GIM methods.

influence and effect of θ_1 on glass transition temperature calculated by equation (2.9) are significant.

Chapter 4

Toward Diverse Polymer Property Prediction Using Transfer Learning

4.1 Principle component analysis

Figure 4.1 demonstrates the explained variance ratio for 20 principal components (PCs). In the dataset, there are a total of more data points, each representing a principal component. However, for the purpose of visualization, only 20 principal components are shown in the figure depicting the explained variance ratio versus the number of principal components. Subsequently, a subset of 13 principal components was selected based on their cumulative explained variance ratios. Specifically, the cumulative explained variance ratios reached a threshold of approximately 0.7, indicating that these 13 principal components capture a significant portion of the overall variance in the dataset.

By examining the top contributing descriptors within the principal components, we uncovered details about their makeup. These descriptors offer a nuanced understanding of the molecular intricacies within the polymers. These descriptors highlight specific patterns, chemical structures, and how molecules are connected. For instance, Dragon7 descriptors, like B02[O-O], unravel specific structural patterns potentially related to oxygen-containing motifs. Other descriptors, such as PubchemFP582 and PubchemFP11, provide binary representations of chemical substructures based on PubChem data. RDKit descriptors, exemplified by morgan363 and maccs157, capture molecular arrangements and specific features. Additionally, topological descriptors like topological929 and topological308 shed light on the connectivity and spatial arrangement of atoms within the molecules. This comprehensive exploration of diverse descriptors enriches our grasp of the molecular composition and structural nuances inherent in the investigated polymers. It provides essential insights into the structural characteristics influencing the thermal and mechanical behaviour of the polymers.

PCA was selected for its ability to simplify the complexity of our data while retaining essential information, aligning well with our objective of transfer learning. This approach allows us to grasp overarching patterns across diverse properties, prioritizing broad understanding over specific implementation details in any particular context.

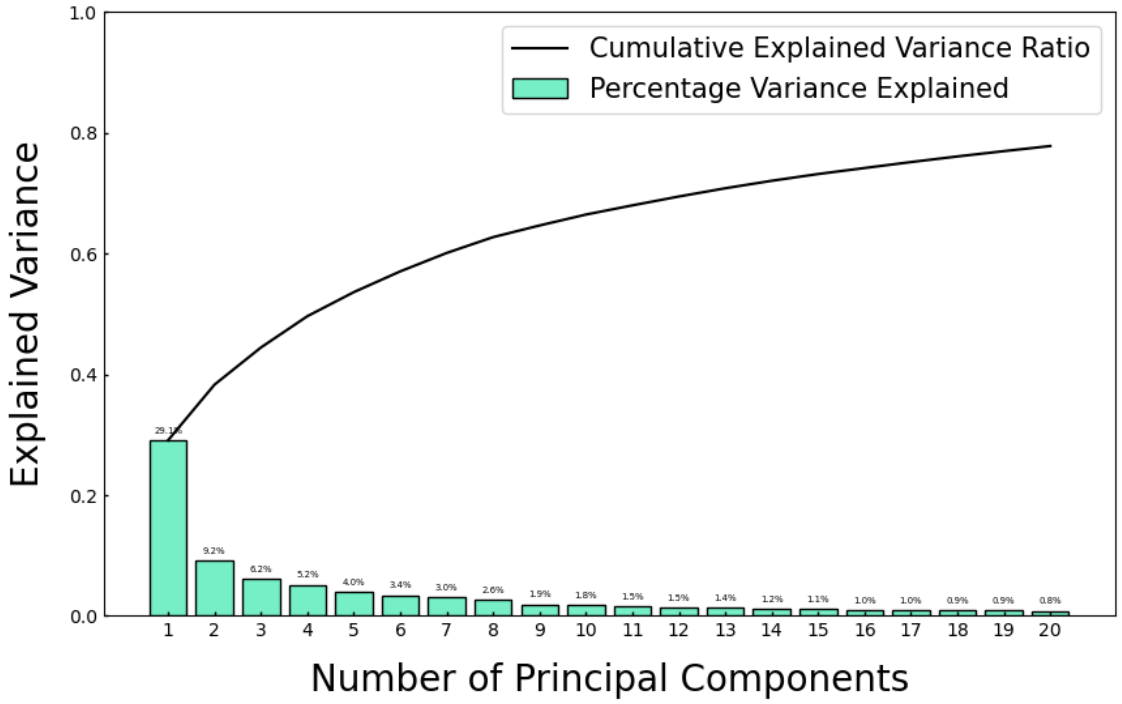

Figure 4.1: Explained variance ratio for 20 principal components.

The polymer chemical space is illustrated in Figure 4.2 using principal components obtained through PCA on the descriptors of all polymers. The first four principal components (PC1, PC2, PC3, and PC4) reveal the intricate relationships between polymers from various sets employed to predict different output properties, represented by distinct colours in the plot. Specifically, in Figure 4.2(a) through(d) respectively represent the relationships between PC1 and PC2, PC1 and PC3, PC1 and PC4, and PC2 and PC3.

The spread of the points in the plot suggests that, although differentiated by size, the datasets roughly occupy the same area of the chemical space. The absence of clustering or separation indicates that the descriptors used in the transfer learning model effectively captured common underlying factors influencing these properties. This observation supports the suitability of the transfer learning approach for predicting multiple properties simultaneously. The descriptors exhibit a degree of shared information among different properties, contributing to the comprehensive understanding of the polymer system. By utilizing the transfer learning model trained on the initial C_p prediction task and applying it to predict additional properties, we leverage the knowledge gained from the initial training and extend it to other related properties. This approach offers the advantage of exploiting the shared information and relationships among the properties, leading to improved predictions and a more comprehensive understanding of the polymer system.

In Figure 4.3, we present a comparison between the expected and predicted values of the neural network (NN) model for the C_p of polymers. The close alignment of the data points with a small deviation from the diagonal line indicates that the model performs well in predicting C_p. The proximity of the points to the diagonal line demonstrates the accuracy

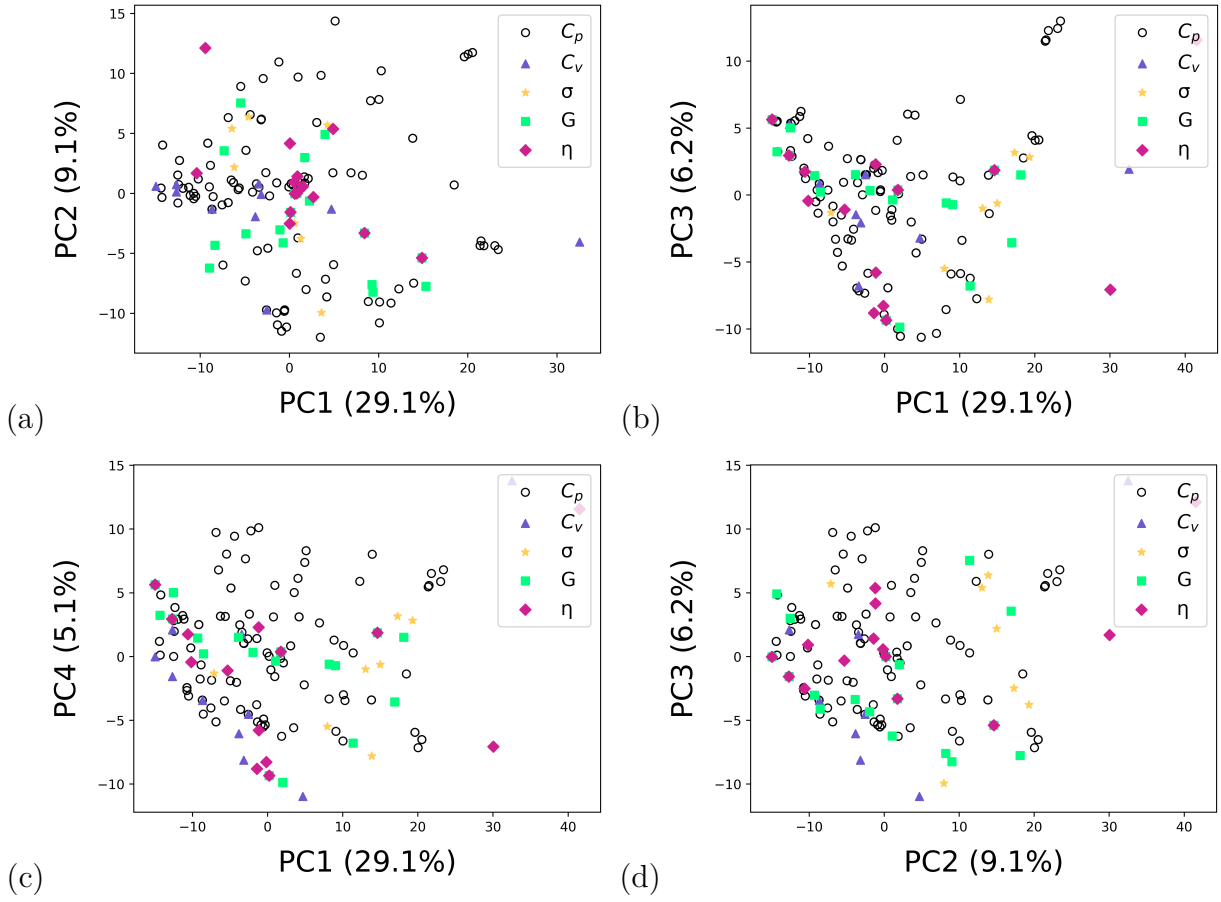

Figure 4.2: Comprehensive principal component analysis depicting relationships for predicting C_p, C_v, flexural strength, shear modulus, dynamic viscosity (a) PC1 vs. PC2, (b) PC1 vs. PC3, (c) PC1 vs. PC4, and (d) PC2 vs. PC3.

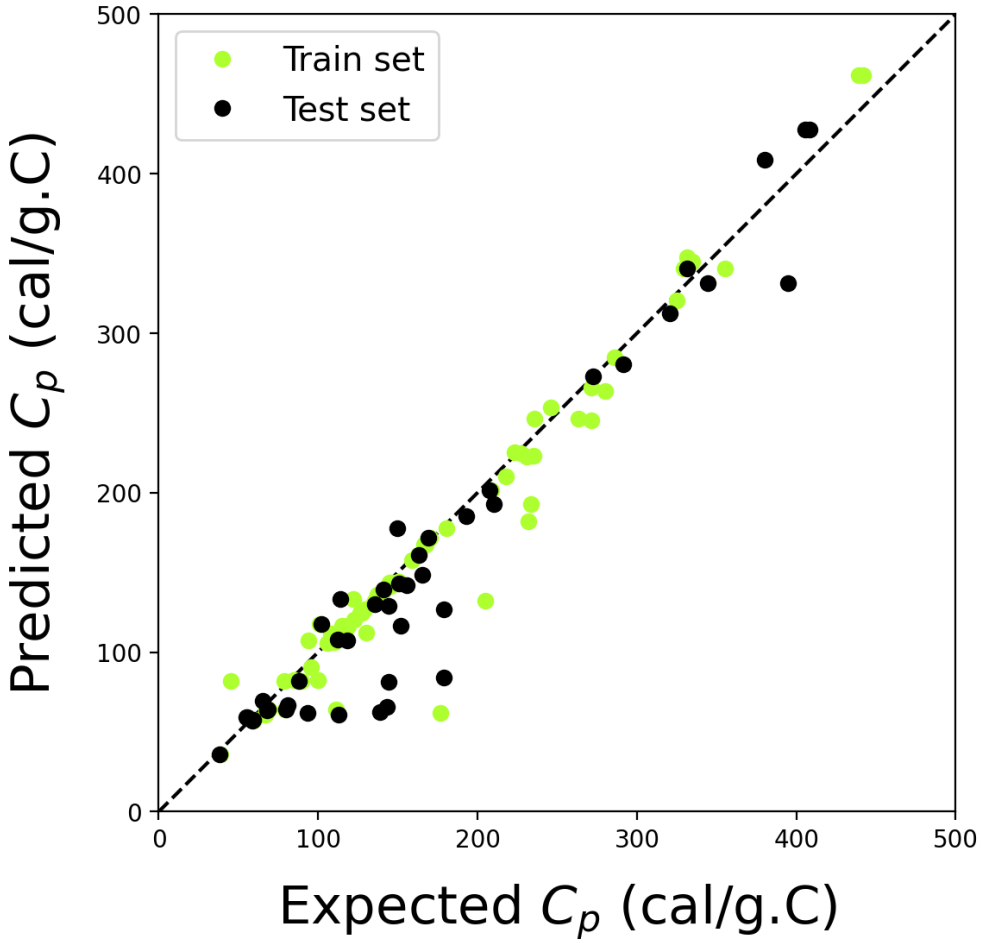

Figure 4.3: A comparison between the expected and predicted values of the NN model for C_p.

and precision of the model's predictions. These results prove the effectiveness of the NN model in estimating the heat capacity of polymers based on the selected descriptors.

4.2 Loss Function

The accuracy of the NN models constructed with five different loss functions (MSE, MAE, Huber Loss, wing shape loss, and combined loss) in the C_p are gathered in Table 4.1. Notably, it is crucial to consider the consistent units for C_p, which are expressed in calories per gram per degree Celsius (cal/g°C), when interpreting the values of the loss functions presented in Table 4.1.

When comparing the R^2 values, we observe that all the loss functions yield high values, indicating a good fit of the model to the training and testing data. The MSE loss function achieved an R^2 value of 0.962 on the training set and 0.93 on the testing set. Similarly, the MAE, Huber Loss, and Wing Shape Loss functions achieved R^2 values of 0.96, 0.967, and 0.967 on the training set and 0.948 on the testing set.

Also, the MSE and MAE statistics are used to evaluate the performance of models with different loss functions. The model utilizing the combined loss function showcased the most favourable results, yielding the lowest MSE (12.42) and MAE (15.86) values. Nevertheless, the Huber Loss and Wing Shape Loss functions displayed competitive performance as well, with MSE values of 12.47 and 12.32 and MAE values of 15.82 and 15.49, respectively. In the case of the model with the MSE loss function, the results are higher than other models.

Interestingly, the combined loss function, which incorporates a balanced combination of the MSE, MAE, Huber Loss, and Wing Shape Loss, resulted in the highest R^2 values of 0.97 on the training set and 0.95 on the testing set. This indicates that the combined loss function effectively captures the strengths of the individual loss functions, resulting in improved prediction accuracy.

Overall, based on the evaluation metrics and the results presented in Table 4.1, it can be concluded that the combined loss function outperforms the other individual loss functions in predicting the C_p. The combined loss function provides a robust and balanced approach, considering multiple aspects of the data, and yields more accurate predictions.

Table 4.1: Accuracy results of the NN trained model in predicting C_p with different loss function

Loss Function	**R^2_train set**	**R^2_test set**	**MSE**	**MAE**
MSE	0.962	0.933	16.10	19.07
MAE	0.96	0.948	12.42	15.86
Huber loss	0.967	0.948	12.47	15.82
wing shape	0.967	0.948	12.32	15.49
Combined	0.971	0.95	12.1	15.2

4.3 Transfer Learning

Once the base model to predict C_p has been built and the model's performance is approved in the previous section. In this section, we will discuss the effective utilization of transfer learning in four distinct scenarios, showcasing its successful applications. The visual representation in Figure 4.4 provides evidence of the transferability between C_v, flexural strength, shear modulus, dynamic viscosity and C_p. The transfer learning models demonstrate strong predictive capabilities, as evident from the expected versus predicted values depicted in Figure 4.4. The high accuracy achieved in predicting C_v, flexural strength, shear modulus, dynamic viscosity, and C_p further supports the effectiveness of the transfer learning approach. The performance metrics of the models are summarized in Table 4.3, highlighting their successful predictive performance. The high value of the R^2 low MAE corresponds to the high accuracy of the model. R^2 of the train set ranges from 0.98 to 0.91, and R^2 of the test set ranges from 0.89 to 0.83. The model for the dynamic viscosity shows the lower value of R^2 of the test set, 0.83 for the given property. On the contrary, the model for the C_v performs better than the other models, with the low mean absolute error and R^2 of the test set of 0.89.

For a clearer understanding and enhanced comprehension of how transfer learning enhances model performance, it is imperative to conduct comparisons between transfer learning and traditional NN models for each property. These comparisons revealed substantial enhancements in predictive accuracy through the integration of transfer learning methodologies. Specifically, the R^2 values for NN models for C_v, flexural strength (σ), shear modulus (G), and dynamic viscosity (η) were determined to be 0.58, 0.51, 0.44, and 0.38, respectively. These findings underscore the considerable improvements realized through the adoption of transfer learning methodologies.

Table 4.3: Performance of ML models built with transfer learning

Models	R^2**_train set**	R^2**_test set**	**MSE**
C_v	0.91	0.90	1.1×10^{-4}
Flexural strength	0.91	0.86	5.8×10^{-5}
Shear Modulus	0.98	0.87	2.4×10^{-2}
Dynamic Viscosity	0.94	0.83	10×10^{4}

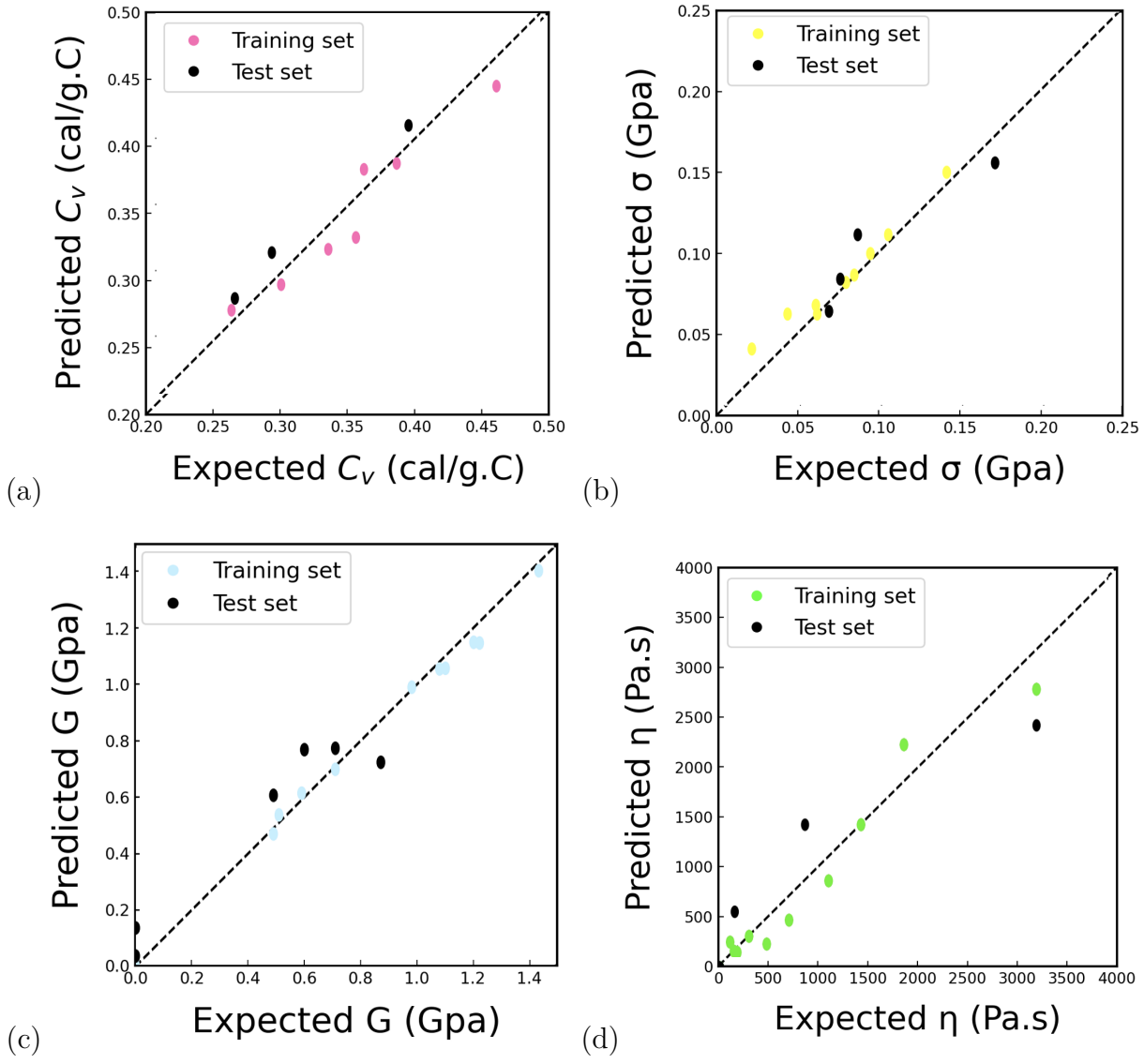

Figure 4.4: Expected and predicted values of the model to predict (a) C_v, (b) flexural strength, (c) shear modulus, and (d) dynamic viscosity.

As depicted in Figure 4.5, each plot represents the loss trajectory of a specific model over the course of training epochs. The x-axis denotes the training cycle, while the y-axis represents the corresponding loss values. Upon closer examination, several trends emerge from the visualization. The gradual decrease in loss across epochs suggests that the models are effectively optimizing their parameters during the transfer learning phase. Moreover, variations in the steepness of the loss curves among different models reflect variations in convergence rates and optimization efficiency. Interpreting these trends, we observe that the C_v model exhibits a more rapid convergence, as evidenced by its steeper loss curve, compared to other models. Conversely, the Shear modulus and Dynamic viscosity models demonstrate a more gradual decrease in loss, indicating a slower convergence rate.

These observations provide valuable insights into the optimization dynamics of transfer learning across diverse model architectures. Furthermore, they underscore the importance of model selection and parameter tuning in achieving optimal performance during the transfer learning process.

In summary, Figure 4.5 enriches our understanding of the transfer learning process and its impact on model optimization. The insights gleaned from this visualization contribute significantly to the discussion on the effectiveness and efficiency of transfer learning in our research context.

Given the limited size of our dataset, we employed the leave-one-out cross-validation (LOOCV) method to illustrate the capability of transfer learning (TL) in predicting outcomes within a confined data space. LOOCV, an extreme form of cross-validation, entails selecting only one sample for testing in each iteration while the remaining samples are utilized for training the model. This process continues until each sample has been tested once, and the final model is refined by averaging the LOOCV results [Chi et al., 2021]. Through LOOCV, we attained commendable R^2 scores: 0.89 for C_v, 0.82 for Flexural strength, 0.86 for Shear Modulus, and 0.78 for Dynamic Viscosity. These outcomes not only highlight the robustness of our TL methodology but also provide compelling evidence of its ability to generalize effectively within the confines of a small dataset. Such findings are instrumental in reinforcing the credibility and applicability of our transfer learning paradigm, offering valuable insights into its performance and potential utility in similar contexts.

Figure 4.6 illustrates the workflow for predicting polymer properties. The diagram shows the process of model training, layer freezing, and transfer learning to predict various properties of linear polymers, demonstrating improved accuracy compared to models trained without transfer learning.

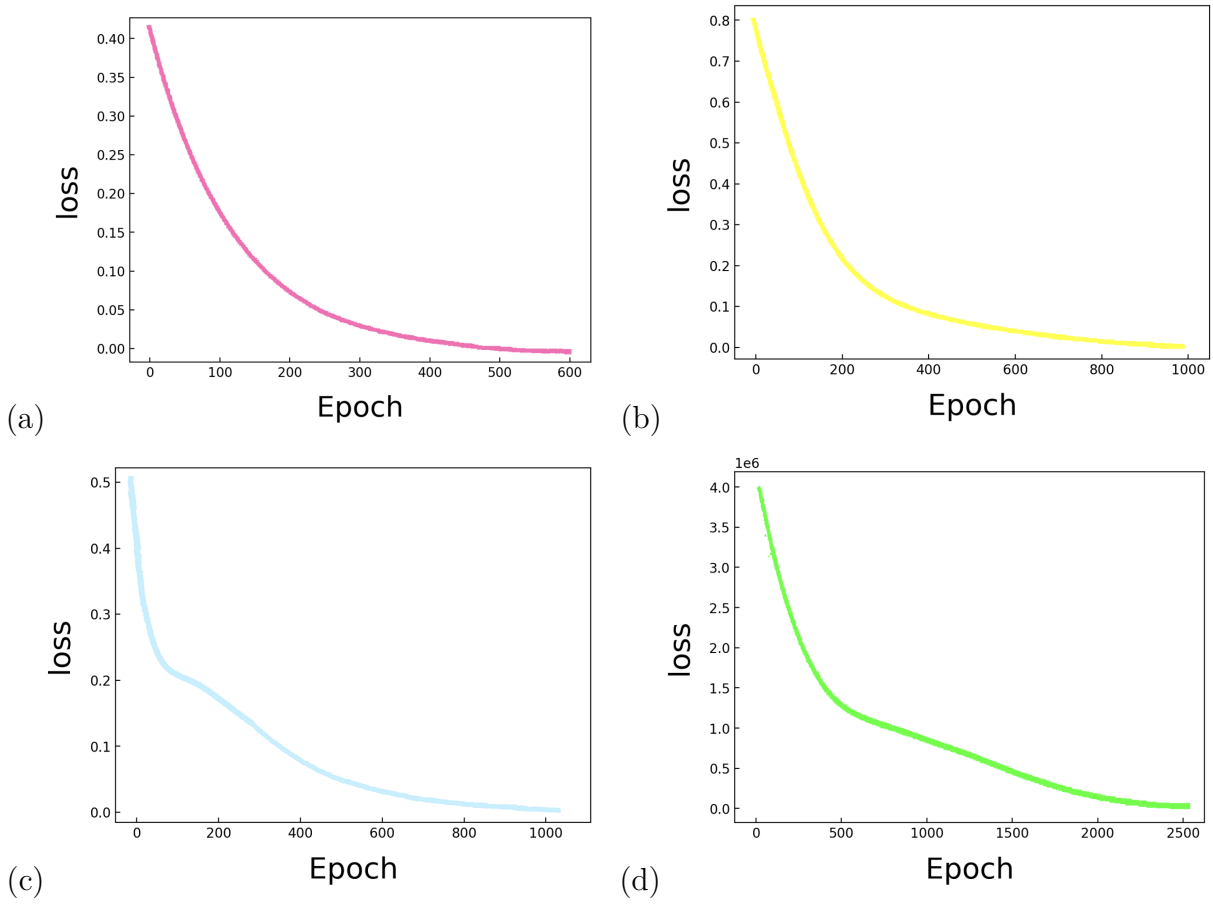

Figure 4.5: Loss as a function of the training cycle during the transfer learning process for the (a) C_v, (b) flexural strength, (c) shear modulus, and (d) dynamic viscosity models.

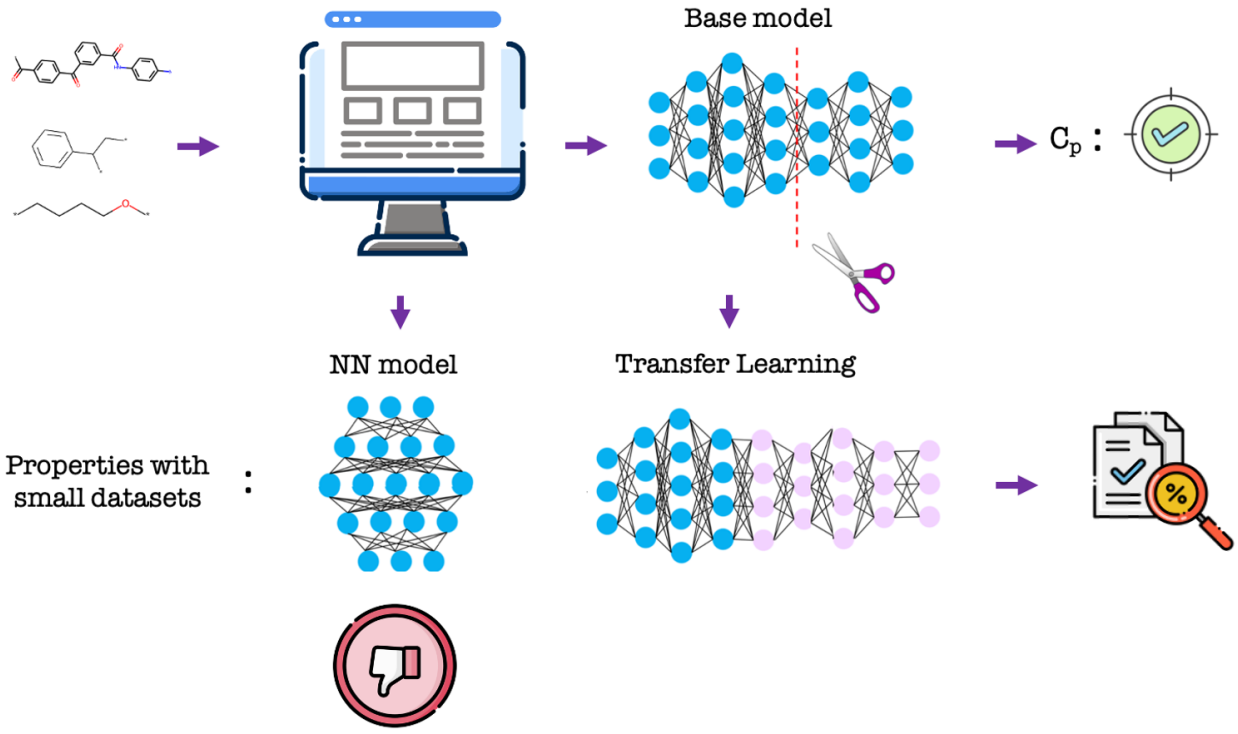

Figure 4.6: Transfer learning workflow for predicting polymer properties.

Chapter 5

Descriptor and Graph-based Molecular Representations in Prediction of Copolymer Properties using ML

5.1 Molecular dynamics predictions versus experimenet

Figure 5.1 illustrates a comprehensive comparison between simulated values and experimental values from the PolyInfoOtsuka et al., 2011 database of key thermodynamic properties for the copolymers under investigation. Additionally, Table 5.1 presents the R^2 and MSE values corresponding to the simulated properties, offering further insights into the quality of the simulations. This comparative analysis provides a valuable assessment of the simulation's performance in capturing thermodynamic behaviours, aiding in the validation and interpretation of the simulation results; we observe that for ρ, α, and γ, the R^2 values are all > 0.7, suggesting good agreement with experiment. K has a slightly less good agreement at R^2=0.61, and C_p has fine but not great agreement at R^2=0.53. We can see in Figure 5.1 that while the agreement is good for low (<3000 J Kg^{-1} K^{-1}) values of C_p, MD simulations are less accurate for larger values of C_p.

Table 5.1: Simulated properties and their corresponding R^2 and MSE values.

Simulated Properties	**R^2**	**MSE**
Density (ρ)	0.858	0.01
Bulk modulus (K)	0.691	0.22
Linear expansion coefficient (α)	0.965	6.4×10^{-9}
Volume expansion coefficient (γ)	0.705	3.4×10^{-8}

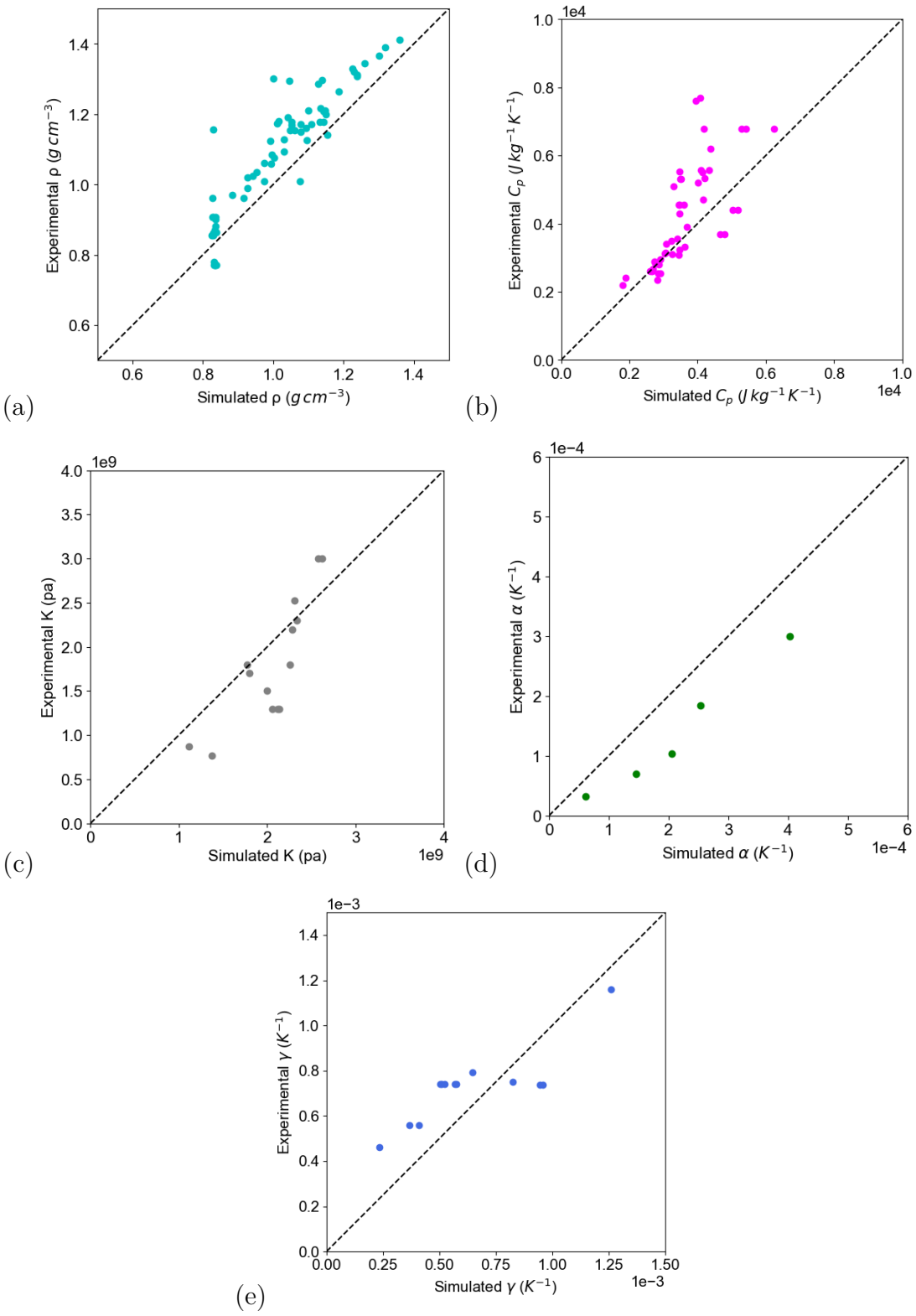

Figure 5.1: Expected and simulated values of the (a) density, (b) C_p, (c) bulk modulus, (d) linear expansion coefficient, and (e) volume expansion.

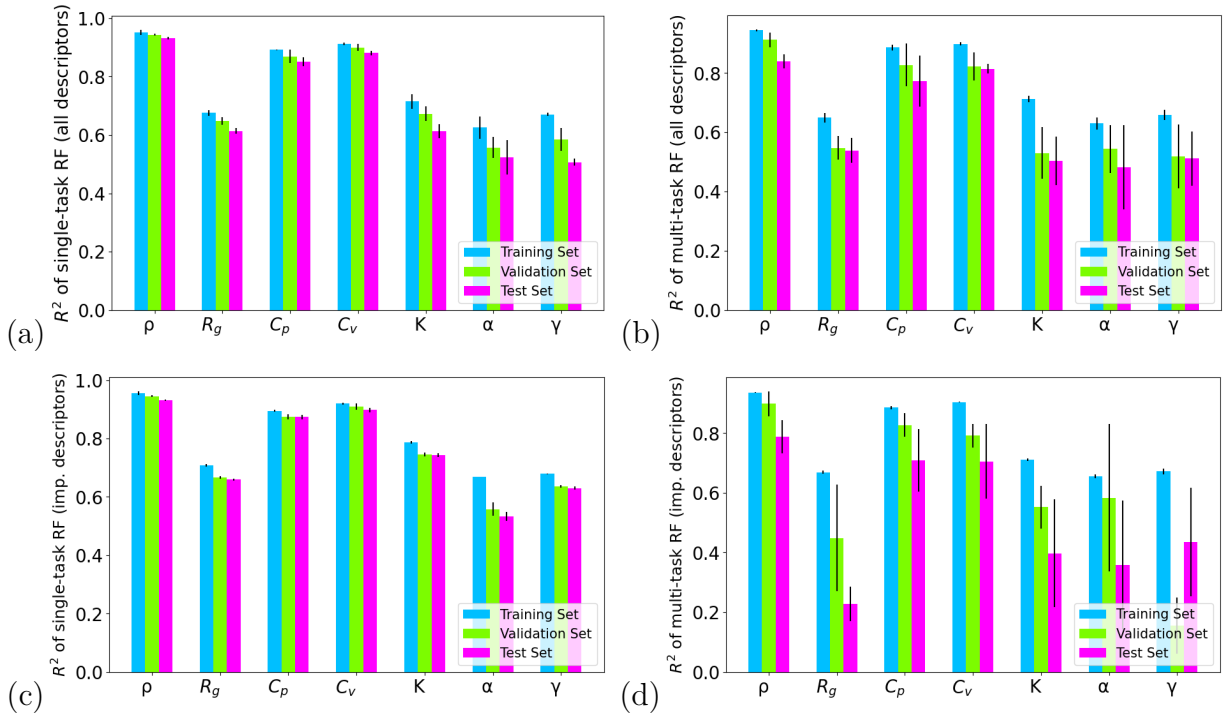

Figure 5.2: R^2 values of (a) single-task RF models using all descriptors, (b) multi-task RF models using all descriptors, (c) single-task RF models using 10 important descriptors, and (d) multi-task RF models using 10 important descriptors for the training, validation, and test sets across different properties.

5.2 Machine learning

5.2.1 Random forest model evaluation

In this section, we present results from testing the ability of single and multitask RF models, using molecular descriptors obtained from the PaDEL-Descriptor toolkit, to predict the properties of copolymers. The Figure 5.2(a) presents the R^2 values for the training, validation, and test sets across various properties, including density (ρ), C_p, C_v, bulk modulus (K), linear expansion coefficient (α), and volume expansion coefficient (γ).

The figure demonstrates the predictive performance of the single-task RF models developed using all calculated descriptors. Among the properties, density, C_p and C_v exhibit the highest R^2 values across all data sets, indicating strong predictive capabilities and suggesting that the RF models can accurately capture the underlying patterns for these properties. In contrast, the linear expansion coefficient and volume expansion coefficient show comparatively lower R^2 values, particularly in the validation and test sets, implying that the models struggle to generalize well for these properties. The bulk modulus and R_g present intermediate performance, with reasonable R^2 values but still lower than those of density, C_p and C_v. These results highlight the varying efficacy of the RF models depending on the specific property being predicted.

The Table 5.3 MSE values for the training, validation, and test sets across various properties.

These results provide a comprehensive evaluation of the model's predictive accuracy for each property.

Table 5.3: Performance of Single-task RF with all descriptors

Properties	**MSE of train set**	**MSE of val. set**	**MSE of test set**
ρ	1.62e-3 ± 2e-4	1.67e-3 ± 8e-4	3.2e-3 ± 5e-4
R_g	8.72 ± 0.23	9.014 ± 2.08	9.941 ± 1.6
C_p	1.14e+5 ± 3.1 e+3	1.31e+5 ± 3.7e+4	1.38e+5 ± 2.8e+4
C_v	5.8e+4 ± 5e+3	7.2e+4 ± 1.5e+4	8.5e+4 ±2.1e+4
K	2.1e+17 ± 1.4e+16	2.6e+17 ± 3.4e+16	3.3e+17 ± 8.2e+16
α	6.6e-09 ± 2.9e-10	9.6e-09 ± 2.7e-9	1.087e-09 ± 4e-9
γ	4.8e-08 ± 2e-09	6.9e-08 ± 2.1e-08	8.4e-08 ± 2.4e-08

In contrast to the single-task RF models, the multitask-RF models show reduced R^2 values for all properties. Figure 5.2(b) presents the R^2 values for the training, validation, and test sets across density, C_p, C_v, bulk modulus, linear expansion coefficient, and volume expansion coefficient.

The figure illustrates that the multitask-RF model generally achieves lower R^2 values compared to the single-task models across all properties. This reduction suggests that while multitask modelling may offer efficiencies in computation and feature utilization, it often sacrifices predictive accuracy compared to single-task models tailored to specific properties.

We utilized important feature analysis to reduce the number of descriptors. In Figure **??**, we present the top 3 most important descriptors for each model.

These analyses help streamline our RF models by focusing on the most influential descriptors, enhancing predictive performance and model interpretability. So, we retrained the RF models using only the top 10 most important descriptors coming from this analysis, and the updated results are presented in Figure 5.2(c).

Several molecular descriptors appear repeatedly across different physical properties of copolymers, highlighting their significance in capturing diverse molecular interactions and structural attributes. The descriptor ETA_EtaP_F_L is important in predicting properties such as density, C_p and C_v due to its ability to capture electronic and structural features within molecules. TDB3r, AATS0v and AATS0m are shared descriptors in C_p and C_v predictions,

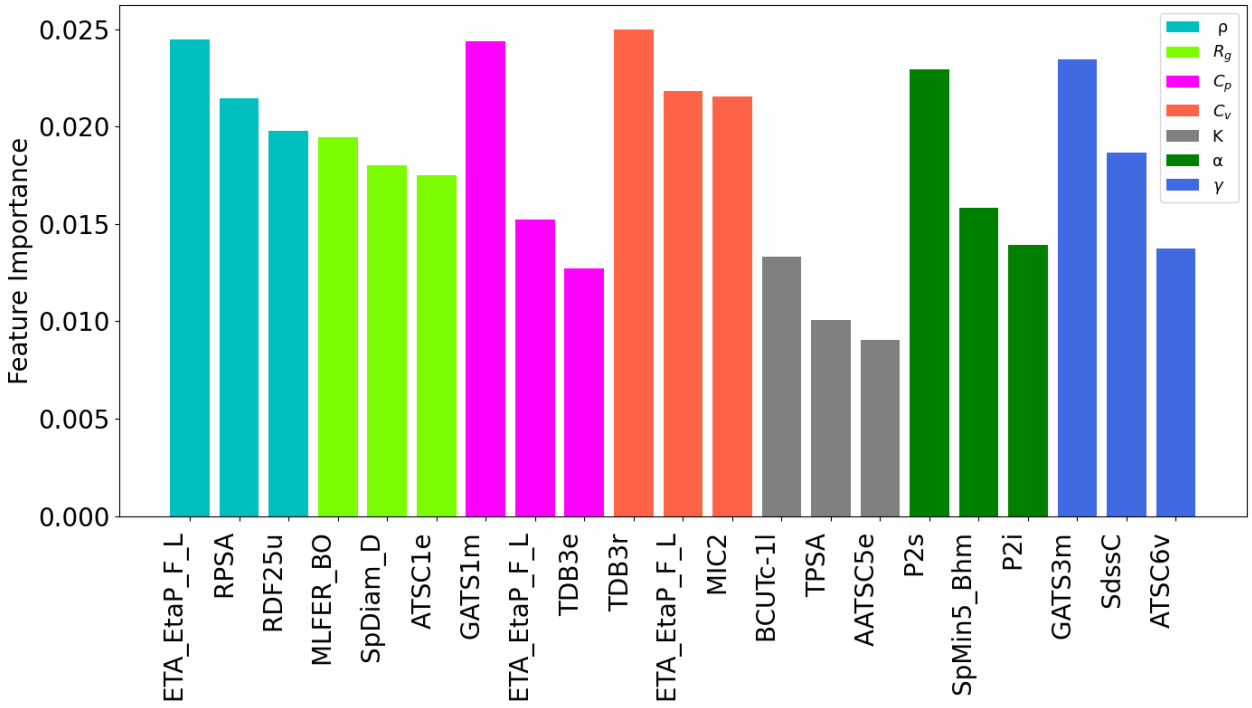

Figure 5.3: Important feature analysis showing the top 3 descriptors for each property (ρ, R_g, C_p, C_v, K, α, and γ.

emphasizing their ability to capture resonance and atomic charge distributions, respectively, crucial for understanding thermal and volumetric properties. TDB2s and SIC2 are significant in density, C_p, and bulk modulus predictions, illustrating their role in assessing electronic and topological complexities within copolymer structures. These descriptors capture molecular connectivity and surface characteristics that directly influence how copolymers pack, store thermal energy, and resist deformation under pressure.

These descriptors, including BCUTw_1h, meanI, MLFER_BO, SpDiam_D, ATSC1e, AATS6v, among others, highlight the diverse molecular characteristics crucial for accurate property prediction, underscoring their importance in copolymer design and optimization.

We constructed new single-task RF models with the 10 most important descriptors for each individual model to further refine our predictions. Figure 5.2(c) displays the R^2 values of these models.

Remarkably, all the new single-task RF models demonstrated an increase in R^2 values compared to previous iterations, illustrating how a reduced feature space can better learn meaningful variability in the data as opposed to also learning meaningless noise. Figure 5.2(d) illustrates the R^2 values of multitask-RF models predicting the same properties with 10 important descriptors. Interestingly, these models exhibit worse performance than multitask models utilizing all descriptors.

Table 5.5: Performance of single-task RF with important descriptors

Properties	MSE of train set	MSE of val. set	MSE of test set
ρ	1.4e-3 ± 8.9e-5	2.2e-3 ± 7.9e-4	2.9e-3 ± 1e-3
R_g	7.7 ± 0.0012	8.2 ± 1.6	12.3 ± 1.7
C_p	1.01e+5 ± 5.4 e+3	1.1e+5 ± 1.8e+4	1.16e+5 ± 2.8e+4
C_v	5.1e+4 ± 4.4e+4	6.1e+4 ± 2.1e+4	9.7e+4 ±1.4e+4
K	1.5e+17 ± 6.2e+15	2.2e+17 ± 3.4e+16	2.4e+17 ± 6.7e+16
α	5.6e-09 ± 4.4e-14	5.7e-09 ± 6.8e-10	6.6e-09 ± 6.4e-10
γ	4.9e-08 ± 1.1e-10	5.3e-08 ± 1.7e-09	5.4e-08 ± 1.5e-09

In summary, the study included single-task RF models using both all descriptors and a subset of important descriptors. It was found that reducing the number of descriptors generally enhanced model performance. When comparing multitask-RF models using all descriptors versus only important descriptors, the former consistently showed better performance. Overall, single-task RF models exhibited superior predictive capabilities across the properties studied.

5.2.2 Neural network model evaluation

In this section, we showcase the results from evaluating the performance of wD-MPNNs models that utilize molecular graph representations to predict the properties of copolymers. Figure 5.4 presents the R^2 values for the training, validation, and test sets of the wD-MPNN models, predicting the same properties as those predicted with the RF models. The results indicate that for some properties, the wD-MPNN models achieve good accuracy, highlighting the effectiveness of the graph representation. However, the error in this representation stems from the inherent randomness in the design of random copolymers, as the randomness can vary significantly across different molecular dynamics (MD) designs.

We compared the predictive power of RF models based on descriptors with wD-MPNN models based on graph representations. Figure 5.5 compares the R^2 values of the test sets for various models: single-task RF with all descriptors, single-task RF with important descriptors, multitask RF with all descriptors, multitask RF with important descriptors, and wD-MPNN. This comparison highlights the predictive power of RF models based on descriptors and wD-MPNN models based on graph representations in predicting different properties. wD-MPNN models by focusing on directed edges, the D-MPNN captures nuanced relational information between atoms, leading to improved accuracy in molecular property prediction tasks [Aldeghi

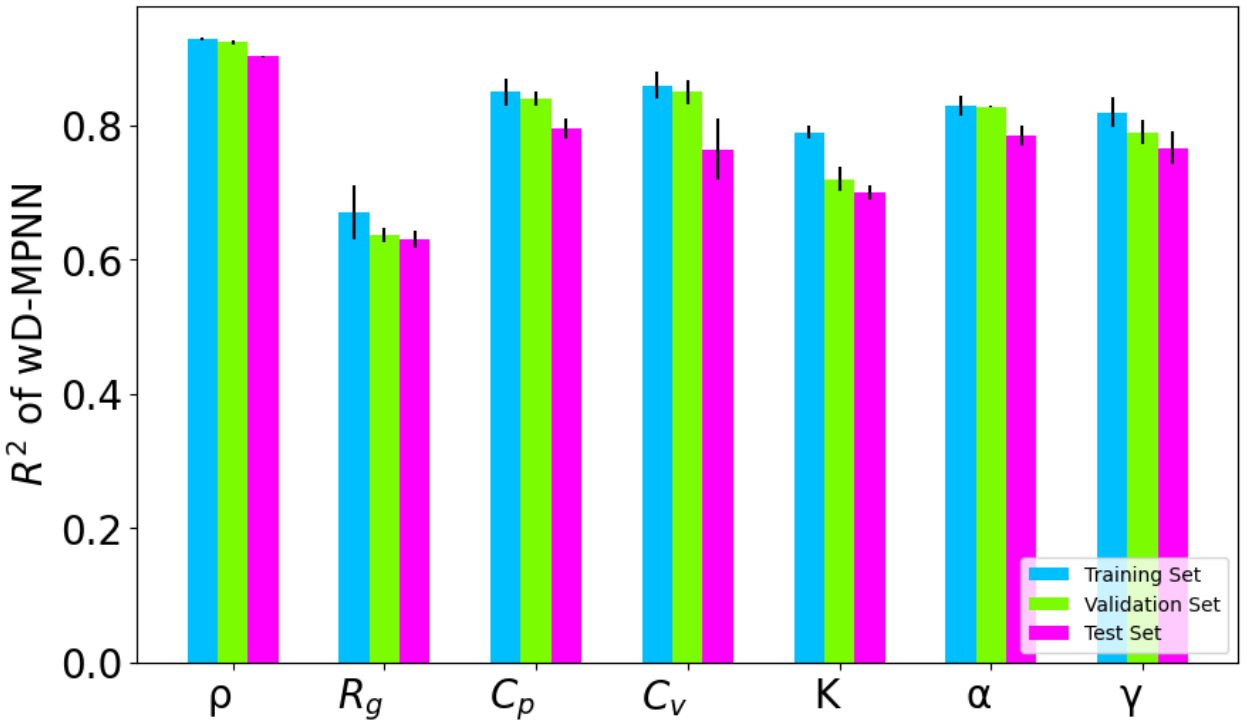

Figure 5.4: R^2 values of wD-MPNN models for the training, validation, and test sets across different properties.

and Coley, 2022; Flam-Shepherd et al., 2021; Gilmer et al., 2017; Heid and Green, 2021; Stokes et al., 2020; Yang et al., 2019].

5.2.3 Evaluating Random Forests versos Neural Networks

Here, we compare the performance of RF models and wD-MPNN models in predicting 7 different properties of copolymers. By leveraging different machine learning techniques, we aim to determine the most effective method for capturing the intricate relationships between the structural attributes of copolymers and their resulting properties. This evaluation provides insights into the strengths and weaknesses of each approach, guiding the selection of the most suitable model for future predictive tasks in polymer science.

The RF models demonstrate better predictive performance for properties such as density, C_p, and C_v. For instance, the R^2 for density is 0.93 with single-task RF with important descriptors, compared to 0.90 for wD-MPNN, showing an improvement of approximately 3.3%. Similarly, for C_p and C_v, the R^2 values are 0.87 and 0.89 for RF models, respectively, compared to 0.795 and 0.76 for wD-MPNN, resulting in improvements of about 9.4% and 17.1%, respectively. This can be attributed to the effectiveness of molecular descriptors in capturing the specific characteristics relevant to these properties, as well as the relatively straightforward relationships between these properties and the molecular structure.

Table 5.7: Performance of wD-MPNN models

Properties	**MSE of train set**	**MSE of val. set**	**MSE of test set**
ρ	1.9e-3 ± 2.3e-4	12.3e-4 ± 2.8e-4	2.9e-3 ± 4.4e-4
R_g	8.5 ± 0.18	9.4 ± 2.08	9.9 ± 2.06
C_p	1.2e+5 ± 2.1 e+3	1.38e+5 ± 2.7e+4	1.48e+5 ± 1.8e+4
C_v	5.7e+4 ± 2.1e+3	7.8e+4 ± 2.5e+3	8.7e+4 ±3.1e+3
K	1.7e+17 ± 2.4e+15	2.2e+17 ± 3.1e+15	2.3e+17 ± 8.8e+15
α	4.6e-09 ± 2.9e-10	5.3e-09 ± 6.7e-10	6.7e-09 ± 2.1e-10
γ	4.4e-08 ± 3.5e-09	5.7e-08 ± 1.1e-09	6.4e-08 ± 1.4e-09

On the other hand, the wD-MPNN models demonstrate superior predictive performance for properties such as α, γ, and K compared to the RF models. For example, the R^2 for γ is 0.76 for wD-MPNN, while it is 0.61 for single-task RF with important descriptors, representing an improvement of about 24.6%. Similarly, for α, the R^2 is 0.78 for wD-MPNN compared to 0.60 for single-task RF with important descriptors, resulting in an improvement of approximately 30%. In K, the improvement is smaller, with R^2 values of 0.70 for wD-MPNN compared to 0.691 for RF models, showing an improvement of about 1.3%.

Additionally, for properties like R_g, the RF model has a R^2 of 0.65, compared to 0.63 for wD-MPNN, demonstrating a minor improvement of about 3.1%.

These findings underscore the importance of selecting appropriate molecular representations for accurate property prediction. Descriptor-based RF models are particularly effective for properties that can be captured through specific molecular characteristics, while graph-based wD-MPNN models are better suited for properties influenced by complex structural interactions. By combining insights from both approaches, the design and optimization of copolymers can be significantly enhanced, allowing for the prioritization of candidates with favourable properties.

Figure 5.6 provides the workflow followed in this chapter to predict polymer properties. The workflow begins with data acquisition, where raw data is collected from MD simulations. This data is then used to train the Random Forest (RF) and Weighted Deep Multi-Physics Neural Network (WD-MPNN) models.

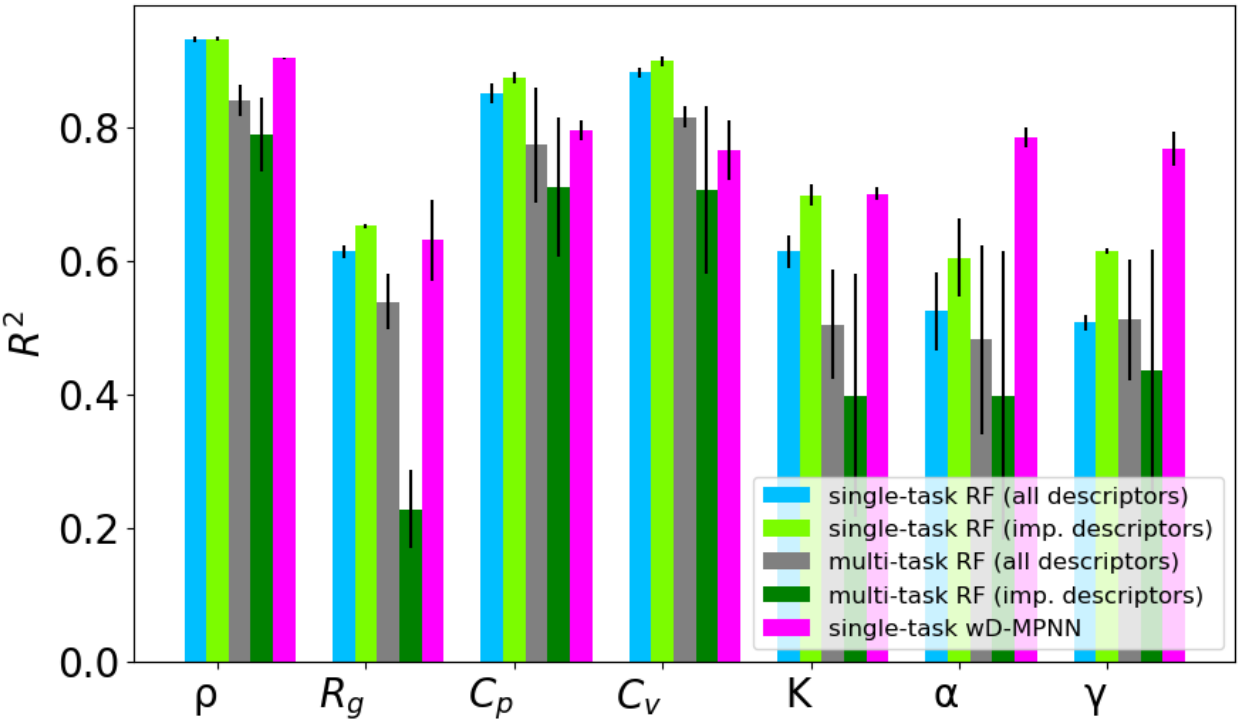

Figure 5.5: Comparison of R^2 values for the test sets across different models

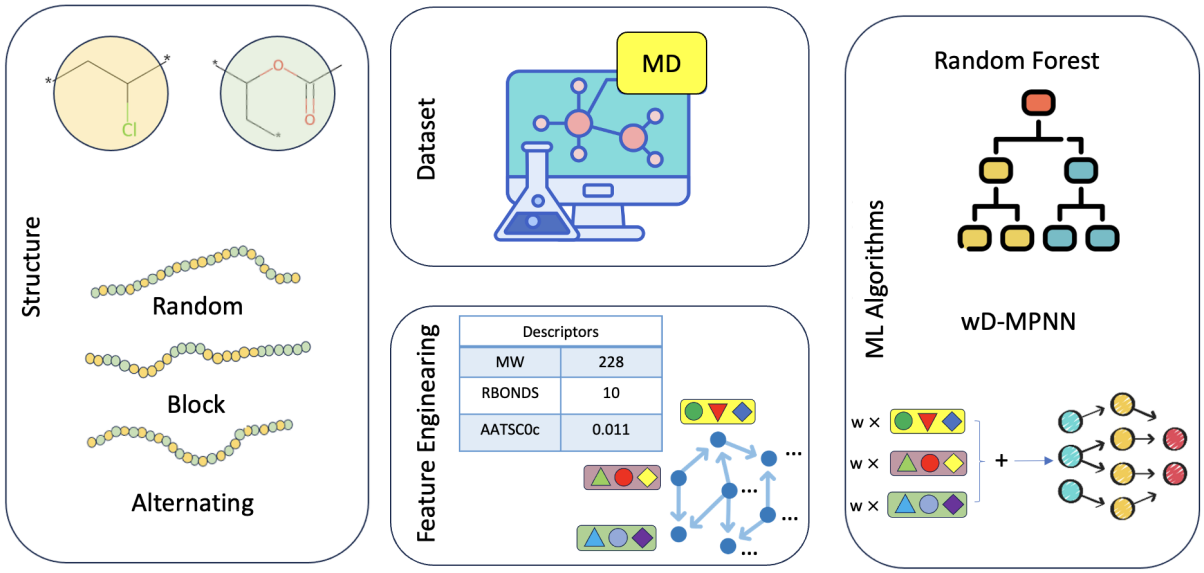

Figure 5.6: Workflow for predicting polymer properties.

Chapter 6

Conclusion

In conclusion, predicting the properties of polymers plays a crucial role in helping industries enhance product quality while reducing material usage. In this thesis, we aimed to evaluate methods for predicting various polymer properties based on their molecular structure and microstructure. We employed several computational techniques across different length scales to achieve this.

First, the GIM method as a microscale framework accurately predicted properties for over 114 polymers, including T_g, C_p, E, α, and ν. For comparison, the same properties were predicted using the ML approach with the RF algorithm. This model utilized fewer than five molecular descriptors, each relevant to the predicted properties. The best-performing model achieved high accuracy in predicting heat capacity, with a R^2 value of 0.955 and a low MAE of 12.1 J/mol·K. Conversely, the worst-performing model was for elastic modulus prediction, with an R^2 value of 0.84 and a relatively higher MAE of 229 MPa. Among these properties, the GIM method consistently achieved the highest prediction accuracy for glass transition temperature, while the prediction performance for linear thermal expansion was comparatively lower. Debye temperature was computed based on the IR simulated spectrum. Overall, the ML approach demonstrated higher accuracy levels than the GIM method.

For smaller datasets, the TL method was employed to enhance the applicability of ML techniques. The initial NN model, which served as the base model, used a dataset with 150 samples. PCA was first conducted to reduce the dimensionality of a dataset representing approximately 150 materials with 27,890 descriptors. Considering the cumulative explained variance ratio, a subset of 13 principal components was selected. The initial NN model utilizing these 13 principal components was constructed to predict the C_p of polymers, exhibiting good accuracy.

Additionally, we explored the performance of the NN model using five different loss functions: MSE, MAE, Huber Loss, Wing Shape Loss, and a combined loss function. Our findings indicate that the combined loss function outperformed the individual loss functions, highlighting the advantage of incorporating multiple criteria in the prediction process.

Transfer learning was successfully applied by freezing the layers of the initial NN model and transferring the knowledge to predict properties such as C_v, σ, G, and η. Through

PCA analysis, correlations and relationships among the predicted properties were observed, suggesting shared underlying factors captured by the selected descriptors. This supports the suitability of TL for the simultaneous prediction of multiple properties and emphasizes the relevance of these properties to one another.

Overall, our approach successfully leveraged PCA, NN modelling, and transfer learning to predict a diverse range of polymer properties. The combined loss function enhanced prediction accuracy, while the interrelated nature of the selected properties further facilitated successful transfer learning. By covering properties from different categories, our study offers valuable insights into the comprehensive characterization of polymers.

Moreover, in this study, we explored the predictive capabilities of RF models and wD-MPNN models for predicting seven different physical properties of copolymers. The dataset used to build these models was generated through MD simulations and validated against experimental data, yielding R^2 values of 0.85 for ρ, 0.53 for C_p, 0.965 for α, 0.691 for K, and 0.751 for γ.

Our analysis demonstrated that RF models, utilizing molecular descriptors, excel in predicting properties such as ρ, C_p, and C_v. In contrast, wD-MPNN models, leveraging graph representations, showed superior performance in predicting volume, linear expansion, and bulk modulus.

These findings underscore the importance of selecting appropriate molecular representations for accurate property prediction. Descriptor-based RF models are particularly effective for properties that can be captured through specific molecular characteristics. In contrast, graph-based wD-MPNN models are better suited for properties influenced by complex structural interactions. By combining insights from both approaches, the design and optimization of copolymers can be significantly enhanced, allowing for the prioritization of candidates with favourable properties.

References

Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Sage.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

Alberi, K., Nardelli, M. B., Zakutayev, A., Mitas, L., Curtarolo, S., Jain, A., Fornari, M., Marzari, N., Takeuchi, I., Green, M. L., et al. (2018). The 2019 materials by design roadmap. *Journal of Physics D: Applied Physics*, *52*(1), 013001.

Aldeghi, M., & Coley, C. W. (2022). A graph representation of molecular ensembles for polymer property prediction. *Chemical Science*, *13*(35), 10486–10498.

AlFaraj, Y., Mohapatra, S., Shieh, P., Husted, K., Ivanoff, D., Lloyd, E., Cooper, J., Dai, Y., Singhal, A., Moore, J., et al. (2023). A model ensemble approach enables data-driven property prediction for chemically deconstructable thermosets in the low data regime.

Ali, M. L., & Rahaman, M. Z. (2018). Investigation of different physical aspects such as structural, mechanical, optical properties and debye temperature of fe2scm (m= p and as) semiconductors: A dft-based first principles study. *International Journal of Modern Physics B*, *32*(10), 1850121.

Allen, M. P., & Tildesley, D. J. (2017). *Computer simulation of liquids*. Oxford university press.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Anderson, J. A. (1995). *An introduction to neural networks*. MIT press.

Audus, D., & De Pablo, J. (2017). Polymer informatics: Opportunities and challenges. acs macro lett 6: 1078–1082.

Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., & Kautz, J. (2016). Reinforcement learning through asynchronous advantage actor-critic on a gpu. *arXiv preprint arXiv:1611.06256*.

Babbar, A., Ragunathan, S., Mitra, D., Dutta, A., & Patra, T. K. (2024). Explainability and extrapolation of machine learning models for predicting the glass transition temperature of polymers. *Journal of Polymer Science*, *62*(6), 1175–1186.

Balachandran, P. V. (2019). Machine learning guided design of functional materials with targeted properties. *Computational Materials Science*, *164*, 82–90.

Bhowmik, R., Sihn, S., Pachter, R., & Vernon, J. P. (2021). Prediction of the specific heat of polymers from experimental data and machine learning methods. *Polymer*, *220*, 123558.

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments*, *65*(6), 1803–1832.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Bondy, J. A., Murty, U. S. R., et al. (1976). *Graph theory with applications* (Vol. 290). Macmillan London.

Brandrup, J., Immergut, E. H., Grulke, E. A., Abe, A., & Bloch, D. R. (1999). *Polymer handbook* (Vol. 89). Wiley New York.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brown, M. E. (2001). *Introduction to thermal analysis: Techniques and applications*. Springer.

Buczek, A., Kupka, T., Broda, M. A., & Żyła, A. (2016). Predicting the structure and vibrational frequencies of ethylene using harmonic and anharmonic approaches at the kohn–sham complete basis set limit. *Journal of molecular modeling*, *22*, 1–10.

Cabestany, J., Prieto, A., Sandoval, D. F., Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. *Lect. Notes Comput. Sci.*, *3512*, 758–770.

Callister Jr, W. D., & Rethwisch, D. G. (2020). *Materials science and engineering: An introduction*. John wiley & sons.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58–63.

Chi, M., Gargouri, R., Schrader, T., Damak, K., Maâlej, R., & Sierka, M. (2021). Atomistic descriptors for machine learning models of solubility parameters for small molecules and polymers. *Polymers*, *14*(1), 26.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules j. am. chem. soc. 1995, 117, 5179- 5197. *Journal of the American Chemical Society*, *118*(9), 2309–2309.

Cramer, C. J. (2013). *Essentials of computational chemistry: Theories and models*. John Wiley & Sons.

Cubuk, E. D., Sendek, A. D., & Reed, E. J. (2019). Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. *The Journal of chemical physics*, *150*(21), 214701.

Czerniecka-Kubicka, A., Zarzyka, I., Schliesser, J., Woodfield, B., & Pyda, M. (2015). Vibrational heat capacity of poly (n-isopropylacrylamide). *Polymer*, *63*, 108–115.

Dakin, J., & Brown, R. G. (2006). *Handbook of optoelectronics (two-volume set)*. CRC press.

Dasgupta, D. (2012). *Artificial immune systems and their applications*. Springer Science & Business Media.

David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in ai-driven drug discovery: A review and practical guide. *Journal of Cheminformatics*, *12*(1), 1–22.

De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448–455.

Deng, S. (2017). Tensile deformation of semi-crystalline polymers by molecular dynamics simulation. *Iranian Polymer Journal*, *26*, 903–911.

Domínguez, J. (2018). Rheology and curing process of thermosets. In *Thermosets* (pp. 115–146). Elsevier.

dos Reis, R. R., Sampaio, S. C., & de Melo, E. B. (2014). An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficients. *Water Research*, *53*, 191–199.

Ehrenstein, G. W., Riedel, G., & Trawiel, P. (2012). *Thermal analysis of plastics: Theory and practice*. Carl Hanser Verlag GmbH Co KG.

El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* Springer.

Erdogdu, Y., Unsalan, O., Amalanathan, M., & Joe, I. H. (2010). Infrared and raman spectra, vibrational assignment, nbo analysis and dft calculations of 6-aminoflavone. *Journal of Molecular Structure*, *980*(1-3), 24–30.

Faghihi, K., Safakish, M., Zebardast, T., Hajimahdi, Z., & Zarghi, A. (2019). Molecular docking and qsar study of 2-benzoxazolinone, quinazoline and diazocoumarin derivatives as anti-hiv-1 agents. *Iranian Journal of Pharmaceutical Research: IJPR*, *18*(3), 1253.

Feldman, D. (2008). Polymer history. *Designed monomers and polymers*, *11*(1), 1–15.

Feng, Z.-H., Kittler, J., Awais, M., Huber, P., & Wu, X.-J. (2018). Wing loss for robust facial landmark localisation with convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2235–2245.

Feng, Z.-H., Kittler, J., Awais, M., & Wu, X.-J. (2020). Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. *International Journal of Computer Vision*, *128*, 2126–2145.

Fernández-León, J., Keramati, K., Miguel, C., González, C., & Baumela, L. (2023). A deep encoder-decoder for surrogate modelling of liquid moulding of composites. *Engineering Applications of Artificial Intelligence*, *120*, 105945.

Ferreira, A. J., & Figueiredo, M. A. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning: Methods and applications*, 35–85.

Fish, J., Wagner, G. J., & Keten, S. (2021). Mesoscopic and multiscale modelling in materials. *Nature materials*, *20*(6), 774–786.

Flam-Shepherd, D., Wu, T. C., Friederich, P., & Aspuru-Guzik, A. (2021). Neural message passing on high order paths. *Machine Learning: Science and Technology*, *2*(4), 045009.

Flory, P. J. (1953). *Principles of polymer chemistry*. Cornell university press.

Foreman, J. P., Behzadi, S., Porter, D., & Jones, F. R. (2010). Multi-scale modelling of the effect of a viscoelastic matrix on the strength of a carbon fibre composite. *Philosophical Magazine*, *90*(31-32), 4227–4244.

Foreman, J. P., Behzadi, S., Tsampas, S. A., Porter, D., Curtis, P. T., & Jones, F. R. (2009). Rate dependent multiscale modelling of fibre reinforced composites. *Plastics, rubber and composites*, *38*(2-4), 67–71.

Foreman, J. P., Porter, D., Behzadi, S., & Jones, F. R. (2008). A model for the prediction of structure–property relations in cross-linked polymers. *Polymer*, *49*(25), 5588–5595.

Foreman, J. P., Porter, D., Behzadi, S., Travis, K. P., & Jones, F. R. (2006). Thermodynamic and mechanical properties of amine-cured epoxy resins using group interaction modelling. *Journal of materials science*, *41*, 6631–6638.

Foreman, J., Porter, D., Behzadi, S., Curtis, P., & Jones, F. (2010). Predicting the thermomechanical properties of an epoxy resin blend as a function of temperature and strain rate. *Composites Part A: Applied Science and Manufacturing*, *41*(9), 1072–1076.

Foreman, J., Porter, D., Pope, D., & Jones, F. (2012). Predicting the material properties of a polyurethane matrix (a composite within a composite). *ECCM15-15th European Conference on Composite Materials*, 24–28.

Frenkel, D., & Smit, B. (2000). Molecular simulation: From algorithms to applications.

Frenkel, D., & Smit, B. (2023). *Understanding molecular simulation: From algorithms to applications*. Elsevier.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery, Jr., A., J., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., ... Pople, J. A. (2004). *Gaussian 03*. Gaussian, Inc. Wallingford, CT, Gaussian, Inc.

Galimberti, D., & Milani, A. (2014). Crystal structure and vibrational spectra of poly (trimethylene terephthalate) from periodic density functional theory calculations. *The Journal of Physical Chemistry B*, *118*(7), 1954–1961.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *International conference on machine learning*, 1263–1272.

González, M. A. (2011). Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, *12*, 169–200.

Goodfellow, I. (2016). Deep learning.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Regularization for deep learning. *Deep learning*, 216–261.

Gooneie, A., Schuschnigg, S., & Holzer, C. (2017). A review of multiscale computational methods in polymeric materials. *Polymers*, *9*(1), 16.

Gracheva, E., Lambard, G., Samitsu, S., Sodeyama, K., & Nakata, A. (2021). Prediction of the coefficient of linear thermal expansion for the amorphous homopolymers based on chemical structure using machine learning. *Science and Technology of Advanced Materials: Methods*, *1*(1), 213–224.

Groß, J. (2003). *Linear regression* (Vol. 175). Springer Science & Business Media.

Guest, N. T., Tilbrook, D. A., Ogin, S. L., & Smith, P. A. (2013). Characterization and modeling of diglycidyl ether of bisphenol-a epoxy cured with aliphatic liquid amines. *Journal of Applied Polymer Science*, *130*(5), 3130–3141.

Gupta, J., Nunes, C., & Jonnalagadda, S. (2013). A molecular dynamics approach for predicting the glass transition temperature and plasticization effect in amorphous pharmaceuticals. *Molecular pharmaceutics*, *10*(11), 4136–4145.

Gurnani, R., Kuenneth, C., Toland, A., & Ramprasad, R. (2023). Polymer informatics at scale with multitask graph neural networks. *Chemistry of Materials*, *35*(4), 1560–1567.

Han, Y., & Elliott, J. (2007). Molecular dynamics simulations of the elastic properties of polymer/carbon nanotube composites. *Computational materials science*, *39*(2), 315–323.

Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., & Hutchison, G. R. (2012). Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics*, *4*, 1–17.

Hayashi, Y., Shiomi, J., Morikawa, J., & Yoshida, R. (2022). Radonpy: Automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Computational Materials*, *8*(1), 222.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18–28.

Heid, E., & Green, W. H. (2021). Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, *62*(9), 2101–2110.

Hohenberg, P., & Kohn, W. (1964a). Density functional theory (dft). *Phys. Rev*, *136*(1964), B864.

Hohenberg, P., & Kohn, W. (1964b). Inhomogeneous electron gas. *Physical review*, *136*(3B), B864.

Horowitz, A. M. (1991). A generalized guided monte carlo algorithm. *Physics Letters B*, *268*(2), 247–252.

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J., & Meredig, B. (2017). Overcoming data scarcity with transfer learning. *arXiv preprint arXiv:1711.05099*.

Jang, H., Ryu, D., Lee, W., Park, G., & Kim, J. (2024). Machine learning-based epoxy resin property prediction. *Molecular Systems Design & Engineering*.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

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

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the american chemical society*, *118*(45), 11225–11236.

Kanouté, P., Boso, D., Chaboche, J.-L., & Schrefler, B. (2009). Multiscale methods for composites: A review. *Archives of Computational Methods in Engineering*, *16*, 31–75.

Karniadakis, G., Beskok, A., & Aluru, N. (2006). *Microflows and nanoflows: Fundamentals and simulation* (Vol. 29). Springer Science & Business Media.

Karuth, A., Alesadi, A., Xia, W., & Rasulev, B. (2021). Predicting glass transition of amorphous polymers by application of cheminformatics and molecular dynamics simulations. *Polymer*, *218*, 123495.

Kaya, M., & Hajimirza, S. (2019). Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies. *Scientific reports*, *9*(1), 1–10.

Kazemi-Khasragh, E., Blázquez, J. P. F., Gómez, D. G., González, C., & Haranczyk, M. (2024). Facilitating polymer property prediction with machine learning and group interaction modelling methods. *International Journal of Solids and Structures*, *286*, 112547.

Kazemi-Khasragh, E., Gonzaleza, C., & Haranczyk, M. (2024). Toward diverse polymer property prediction using transfer learning. *arXiv preprint arXiv:2401.09139*.

Khan, P., & Roy, K. (2018). Qspr modelling for prediction of glass transition temperature of diverse polymers. *SAR and QSAR in Environmental Research*, *29*(12), 935–956.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. Springer.

Koch, W., & Holthausen, M. C. (2015). *A chemist's guide to density functional theory*. John Wiley & Sons.

Kode-Chemoinformatics [https://chm.kode-solutions.net/products_dragon.php]. (n.d.).

Kohn, W., & Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical review*, *140*(4A), A1133.

Kowalewski, J., & Ray, A. (2020). Predicting novel drugs for sars-cov-2 using machine learning from a> 10 million chemical space. *Heliyon*, *6*(8).

Kramer, O., & Kramer, O. (2016). Scikit-learn. *Machine learning for evolution strategies*, 45–53.

Kumar, A., Sharma, K., & Dixit, A. R. (2019). A review of the mechanical and thermal properties of graphene and its hybrid polymer nanocomposites for structural applications. *Journal of materials science*, *54*(8), 5992–6026.

Landel, R. F., & Nielsen, L. E. (1993). *Mechanical properties of polymers and composites*. CRC press.

Larsen, G. S., Lin, P., Hart, K. E., & Colina, C. M. (2011). Molecular simulations of pim-1-like polymers of intrinsic microporosity. *Macromolecules*, *44*(17), 6944–6951.

Larson, R. G. (1999). *The structure and rheology of complex fluids* (Vol. 150). Oxford university press New York.

Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. Springer.

Levy, M. (1979). Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proceedings of the National Academy of Sciences*, *76*(12), 6062–6065.

Li, L., Zhang, Y., Ma, H., & Yang, M. (2008). An investigation of molecular layering at the liquid-solid interface in nanofluids by molecular dynamics simulation. *Physics Letters A*, *372*(25), 4541–4544.

Li, Y., Abberton, B. C., Kröger, M., & Liu, W. K. (2013). Challenges in multiscale modeling of polymer dynamics. *Polymers*, *5*(2), 751–832.

Liu, P., Lu, J., Yu, H., Ren, N., Lockwood, F. E., & Wang, Q. J. (2017). Lubricant shear thinning behavior correlated with variation of radius of gyration via molecular dynamics simulations. *The Journal of chemical physics*, *147*(8).

LLorca, J., González, C., Molina-Aldareguía, J. M., Segurado, J., Seltzer, R., Sket, F., Rodríguez, M., Sádaba, S., Muñoz, R., & Canal, L. P. (2011). Multiscale modeling of composite materials: A roadmap towards virtual testing. *Advanced materials*, *23*(44), 5130–5147.

Ma, Z., Wang, S., Kim, M., Liu, K., Chen, C.-L., & Pan, W. (2021). Transfer learning of memory kernels for transferable coarse-graining of polymer dynamics. *Soft Matter*, *17*(24), 5864–5877.

MacKerell Jr, A. D., Bashford, D., Bellott, M., Dunbrack Jr, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, *102*(18), 3586–3616.

Macosko, C. W. (1994). Rheology principles. *Measurements and Applications*.

Mae, H., Omiya, M., & Kishimoto, K. (2008). Effects of strain rate and relaxation rate on elastic modulus of semi-crystalline polymer. *Transaction of the Japan Society for Computational Methods in Engineering*, *7*(2), 207–212.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Mannodi-Kanakkithodi, A., Pilania, G., & Ramprasad, R. (2016). Critical assessment of regression-based machine learning methods for polymer dielectrics. *Computational Materials Science*, *125*, 123–135.

Mark, J. E., et al. (2007). *Physical properties of polymers handbook* (Vol. 1076). Springer.

Mauri, A. (2020). Alvadesc: A tool to calculate and analyze molecular descriptors and fingerprints. *Ecotoxicological QSARs*, 801–820.

Menard, K. P., & Menard, N. (2020). *Dynamic mechanical analysis*. CRC press.

Meyer, G. P. (2021). An alternative probabilistic interpretation of the huber loss. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 5261–5269.

Mohammadi, M., Davoodi, J., et al. (2017). The glass transition temperature of pmma: A molecular dynamics study and comparison of various determination methods. *European Polymer Journal*, *91*, 121–133.

Moosavi, S. M., Jablonka, K. M., & Smit, B. (2020). The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society*, *142*(48), 20273–20287.

Moreira, M., & Fiesler, E. (1995). Neural networks with adaptive learning rate and momentum terms.

Nicholson, J. W. (1991). Etymology of 'polymers'. *Educ. Chem*, *28*, 70–71.

Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195–211.

Nilsson, N. J. (1996). Introduction to machine learning: An early draft of a proposed textbook.

Nishiyama, E., Yokota, M., & Tsukushi, I. (2021). Estimation of the configurational heat capacity of polyisobutylene, isobutane and 2, 2, 4-isomethylpentane above the glass transition temperature. *Polymer Journal*, *53*(9), 1031–1036.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011). Polyinfo: Polymer database for polymeric materials design. *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 22–29.

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. (2007). Random forest models to predict aqueous solubility. *Journal of chemical information and modeling*, *47*(1), 150–158.

Park, J., Shim, Y., Lee, F., Rammohan, A., Goyal, S., Shim, M., Jeong, C., & Kim, D. S. (2022). Prediction and interpretation of polymer properties using the graph convolutional network. *ACS Polymers Au*, *2*(4), 213–222.

Patra, T. K. (2021). Data-driven methods for accelerating polymer design. *ACS Polymers Au*, *2*(1), 8–26.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pei, J., Cai, C., Tang, J., Zhao, S., & Yuan, F. (2012). Prediction of the glass transition temperatures of styrenic copolymers by using support vector regression combined with particle swarm optimization. *Journal of Macromolecular Science, Part B*, *51*(7), 1437–1448.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Pilania, G. (2021). Machine learning in materials science: From explainable predictions to autonomous design. *Computational Materials Science*, *193*, 110360.

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, *117*(1), 1–19.

Polyinfo [Accessed: [Insert Date Accessed]]. (n.d.).

Porter, D. (1995). *Group interaction modelling of polymer properties*. CRC Press.

Porter, D., & Gould, P. J. (2009). Predictive nonlinear constitutive relations in polymers through loss history. *International journal of solids and structures*, *46*(9), 1981–1993.

Poša, M., Bjedov, S., Sebenji, A., & Sakač, M. (2014). Wittig reaction (with ethylidene triphenylphosphorane) of oxo-hydroxy derivatives of 5β-cholanic acid: Hydrophobicity, haemolytic potential and capacity of derived ethylidene derivatives for solubilisation of cholesterol. *Steroids*, *86*, 16–25.

Pyda, M., Bartkowiak, M., & Wunderlich, B. (1998). Computation of heat capacities of solids using a general tarasov equation. *Journal of thermal analysis and calorimetry*, *52*, 631–656.

Pyda, M., Boller, A., Grebowicz, J., Chuah, H., Lebedev, B., & Wunderlich, B. (1998). Heat capacity of poly (trimethylene terephthalate). *Journal of Polymer Science Part B: Polymer Physics*, *36*(14), 2499–2511.

Pyda, M., Bopp, R., & Wunderlich, B. (2004). Heat capacity of poly (lactic acid). *The Journal of Chemical Thermodynamics*, *36*(9), 731–742.

Pyda, M., Zawada, P., Drogon, A., Skotnicki, M., & Cebe, P. (2019). Vibrational heat capacity of collagen and collagen–water. *Journal of Thermal Analysis and Calorimetry*, *138*, 3389–3401.

Ramsundar, B., Liu, E., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., & Pande, V. (2015). Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*.

RDKit, Open-Source Cheminformatics [http://www.rdkit.org]. (n.d.).

Reynolds, D. A., et al. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31–39.

Riggleman, R. A., Douglas, J. F., & de Pablo, J. J. (2010). Antiplasticization and the elastic properties of glass-forming polymer liquids. *Soft Matter*, *6*(2), 292–304.

Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, *26*(3), 303–304.

Roles, K., Xenopoulos, A., & Wunderlich, B. (1993). Heat capacities of solid poly (amino acid) s. ii. the remaining polymers. *Biopolymers: Original Research on Biomolecules*, *33*(5), 753–768.

Sadiku-Agboola, O., Sadiku, E. R., Adegbola, A. T., Biotidara, O. F., et al. (2011). Rheological properties of polymers: Structure and morphology of molten polymer blends. *Materials Sciences and Applications*, *2*(01), 30.

Sattari, K., Xie, Y., & Lin, J. (2021). Data-driven algorithms for inverse design of polymers. *Soft Matter*, *17*(33), 7607–7622.

Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M., & Fazzio, A. (2019). From dft to machine learning: Recent approaches to materials science–a review. *Journal of Physics: Materials*, *2*(3), 032001.

Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, *5*(1), 83.

Sestraş, R. E., Jäntschi, L., & Bolboacă, S. D. (2012). Poisson parameters of antimicrobial activity: A quantitative structure-activity approach. *International Journal of Molecular Sciences*, *13*(4), 5207–5229.

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, *6*(12), 310–316.

Shaw, M. T., & MacKnight, W. J. (2018). *Introduction to polymer viscoelasticity*. John Wiley & Sons.

Shivaleela, B., Naziya, P., Naseem, F., Shivraj, G., YF, N., & SM, H. (n.d.). Density functional theory studies on radioactive isotope i-131 metaiodobenzylguanidinefor radionuclide therapy using gaussian 16.

Sholl, D. S., & Steckel, J. A. (2022). *Density functional theory: A practical introduction*. John Wiley & Sons.

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, *8*, 80716–80727.

Sliwoski, G., Mendenhall, J., & Meiler, J. (2016). Autocorrelation descriptor improvements for qsar: 2da_sign and 3da_sign. *Journal of computer-aided molecular design*, *30*, 209–217.

Soldera, A. (1998). Comparison between the glass transition temperatures of the two pmma tacticities: A molecular dynamics simulation point of view. *Macromolecular symposia*, *133*(1), 21–32.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, *180*(4), 688–702.

Szabo, A., & Ostlund, N. S. (2012). *Modern quantum chemistry: Introduction to advanced electronic structure theory*. Courier Corporation.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, 270–279.

Tan, H. (2021). Reinforcement learning with deep deterministic policy gradient. *2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, 82–85.

Tao, L., Byrnes, J., Varshney, V., & Li, Y. (2022). Machine learning strategies for the structure-property relationship of copolymers. *Iscience*, *25*(7).

Tao, L., He, J., Arbaugh, T., McCutcheon, J. R., & Li, Y. (2023). Machine learning prediction on the fractional free volume of polymer membranes. *Journal of Membrane Science*, *665*, 121131.

Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., et al. (2022). Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, *271*, 108171.

Thrun, S., & Littman, M. L. (2000). Reinforcement learning: An introduction. *AI Magazine*, *21*(1), 103–103.

Thybring, E. E. (2014). Explaining the heat capacity of wood constituents by molecular vibrations. *Journal of materials science*, *49*, 1317–1327.

Van Krevelen, D. W., & Te Nijenhuis, K. (2009). *Properties of polymers: Their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*. Elsevier.

Varshney, V., Patnaik, S. S., Roy, A. K., & Farmer, B. L. (2008). A molecular dynamics study of epoxy-based networks: Cross-linking procedure and prediction of molecular and material properties. *Macromolecules*, *41*(18), 6837–6842.

Velten, K., Reinicke, R., & Friedrich, K. (2000). Wear volume prediction with artificial neural networks. *Tribology International*, *33*(10), 731–736.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*, 395–416.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of computational chemistry*, *25*(9), 1157–1174.

Wang, M., Liao, Y., & Chen, D. (2013). Determination of linear thermal expansion coefficient of polymeric materials by infrared thermography. *Polymer testing*, *32*(2), 175–178.

Wang, S., Cheng, M., Zhou, L., Dai, Y., Dang, Y., & Ji, X. (2021). Qspr modelling for intrinsic viscosity in polymer–solvent combinations based on density functional theory. *SAR and QSAR in Environmental Research*, *32*(5), 379–393.

Wang, Y., Hu, S., & Wu, S. (2019). Object tracking based on huber loss function. *The Visual Computer*, *35*, 1641–1654.

Wang, Z., Lv, Q., Chen, S., Li, C., Sun, S., & Hu, S. (2015). Glass transition investigations on highly crosslinked epoxy resins by molecular dynamics simulations. *Molecular Simulation*, *41*(18), 1515–1527.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*, 279–292.

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naıve bayes. *Encyclopedia of machine learning*, *15*(1), 713–714.

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31–36.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1–40.

Winter, R., Montanari, F., Noé, F., & Clevert, D.-A. (2019). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, *10*(6), 1692–1701.

Wiswesser, W. J. (1968). 107 years of line-formula notations (1861-1968). *Journal of Chemical Documentation*, *8*(3), 146–150.

Wong, M. W. (1996). Vibrational frequency prediction using density functional theory. *Chemical Physics Letters*, *256*(4-5), 391–399.

Wu, X., Wang, H., Gong, Y., Fan, D., Ding, P., Li, Q., & Qian, Q. (2023). Graph neural networks for molecular and materials representation.

Xiao, J., Hobson, J., Ghosh, A., Haranczyk, M., & Wang, D.-Y. (2023). Flame retardant properties of metal hydroxide-based polymer composites: A machine learning approach. *Composites Communications*, *40*, 101593.

Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J., & Yoshida, R. (2019). Predicting materials properties with little data using shotgun transfer learning. *ACS central science*, *5*(10), 1717–1730.

Yamamoto, S., Miyada, M., Sato, H., Hoshina, H., & Ozaki, Y. (2017). Low-frequency vibrational modes of poly (glycolic acid) and thermal expansion of crystal lattice assigned on the basis of dft-spectral simulation aided with a fragment method. *The Journal of Physical Chemistry B*, *121*(5), 1128–1138.

Yanai, T., Tew, D. P., & Handy, N. C. (2004). A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chemical physics letters*, *393*(1-3), 51–57.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, *59*(8), 3370–3388.

Yap, C. W. (2011). Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, *32*(7), 1466–1474.

Yokota, M., Sugane, K., Tsukushi, I., & Shibata, M. (2020). Evaluation of the heat capacity of amorphous polymers composed of a carbon backbone below their glass transition temperature. *Polymer Journal*, *52*(7), 765–774.

Yokota, M., & Tsukushi, I. (2020). Heat capacities of polymer solids composed of polyesters and poly (oxide) s, evaluated below the glass transition temperature. *Polymer Journal*, *52*(9), 1103–1111.

Young, R. J., & Lovell, P. A. (2011). *Introduction to polymers*. CRC press.

Yu, X., & Huang, X. (2017). A quantitative relationship between t gs and chain segment structures of polystyrenes. *Polímeros*, *27*(1), 68–74.

Yu, X.-L., Yi, B., & Wang, X.-Y. (2008). Prediction of the glass transition temperatures for polymers with artificial neural network. *Journal of Theoretical and Computational Chemistry*, *7*(05), 953–963.

Zeng, M., Kumar, J. N., Zeng, Z., Savitha, R., Chandrasekhar, V. R., & Hippalgaonkar, K. (2018). Graph convolutional neural networks for polymers property prediction. *arXiv preprint arXiv:1811.06231*.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 558–575.

Zhang, Z., Liu, Q., & Wu, D. (2022). Predicting stress–strain curves using transfer learning: Knowledge transfer across polymer composites. *Materials & Design*, *218*, 110700.

Ziegler, T. (1991). Approximate density functional theory as a practical tool in molecular energetics and dynamics. *Chemical Reviews*, *91*(5), 651–667.