Fall 2024 MSE760 Final Project Report
University of Wisconsin-Madison

Data-Driven Atom-Level Explanation of Polymer Properties

Yuhao Liu, Jiahui Yang

December 13, 2024

Abstract

This study investigates the mechanical properties of polymers by linking atom-level structural features to macroscopic behavior through machine learning and molecular dynamics simulations. Polymers, with their highly versatile structures, pose challenges for conventional machine learning models due to their intricate molecular configurations. To address this, a single-task machine learning model was utilized to analyze the atomic contributions to the Young's modulus of Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl). The machine learning model's predictions were validated against molecular dynamics simulations, showing a prediction error of only 8%. Shapley Additive Explanations values were employed to further interpret the influence of substructures, highlighting both consistencies and discrepancies with atom-level explanations derived from the machine learning model. This approach provides an insightful framework for understanding and predicting the relationship between polymer structures and their properties, offering potential pathways for the design of advanced polymer materials.

Link to Final Project `git` repo:
https://github.com/jhyang13/UWMadison-MSE760-Final-Project

Contents

General information

1. Name: Jiahui Yang
2. Email: jyang753@wisc.edu
3. Home department: Mechanical Engineering
4. Status: PhD Student
5. Teammate: Yuhao Liu

Introduction

Polymers are an essential class of materials widely utilized in various industries due to their tunable mechanical, chemical, and physical properties. [1-3] The versatility of polymers arises from their molecular structures, which can be tailored by modifying the composition and arrangement of their monomeric units. [4-6] This adaptability enables their application in diverse fields such as construction, packaging, electronics, and medical devices. However, despite the significant progress in polymer synthesis and application, the relationship between the atomic-scale structure of polymers and their macroscopic properties, such as mechanical strength and elasticity, is not fully understood. [7-8] This knowledge gap presents a challenge in designing polymers with optimized properties for specific applications.

Machine learning (ML) has emerged as a powerful tool in materials science, offering a data-driven approach to modeling and predicting the properties of complex materials. [9-10] When combined with molecular dynamics (MD) simulations, ML enables the exploration of the intricate structure-property relationships in polymers. For instance, ML models can predict mechanical properties such as Young's modulus by leveraging atomistic features, while MD simulations provide a physics-based validation of these predictions. Recent studies have demonstrated the potential of explainable ML techniques, such as Shapley Additive Explanations (SHAP), to identify key substructures influencing polymer properties, offering insights that are critical for advanced polymer design.

Problem statement

Despite advancements in polymer research, understanding the connection between atomic-scale features and the macroscopic properties of polymers remains a significant challenge. Traditional methods rely on experimental approaches, which are time-consuming and costly, limiting their scalability to the vast chemical space of polymers. Additionally, simplified representations, such as SMILES, often fail to capture complex topological features like cross-linking and branching, further complicating efforts to model polymer behavior accurately. [11-12]

ML offers a promising alternative, but existing models often lack explainability and validation, undermining their utility in guiding polymer design. Moreover, while MD simulations provide high-fidelity property predictions, their computational cost restricts their application in large-scale polymer discovery. Integrating ML and MD approaches to develop efficient, scalable, and interpretable methods for linking polymer structures to properties is critical to overcoming these limitations. This integration would enable the discovery and optimization of high-performance polymers tailored for specific applications, addressing the growing demand for advanced materials in various industries.

Research Methods

This research integrates ML and MD simulations to investigate the relationship between polymer structures and their macroscopic properties, particularly Young's modulus. The methodology is divided into the following key steps:

- **Data Collection and Preparation**

Data was collected from the PoLyInfo database [13], which contains over 18,000 polymers, including 12,854 homopolymers with reported chemical structures and approximately 100 property types. Specifically, 923 polymers with documented Young's modulus values were selected for ML model training. The simplified molecular-input line-entry system (SMILES) [14] was used to represent polymer structures, ensuring compatibility with ML algorithms while maintaining interpretability.

- **Machine Learning Modeling**

Single-task ML models and multitask feedforward neural networks (FFNNs) were employed to predict mechanical properties, with a particular focus on Young's modulus. Morgan fingerprints were used for feature representation, capturing the presence and frequency of key molecular substructures. To enhance model interpretability, SHAP values were calculated, identifying the contributions of individual substructures to the predicted properties.

- **Molecular Dynamics Simulations**

MD simulations were conducted using the LAMMPS software package with a polymer consistent force field (PCFF). These simulations validated the predictions of ML models by calculating mechanical properties from stress-strain curves generated during tensile testing. MD models were built with realistic polymer configurations, equilibrated under high-temperature and high-pressure conditions to ensure physical relevance.

- **Validation and Analysis**

The performance of the ML models was evaluated by comparing predicted values with MD simulation results. Discrepancies between ML and MD were analyzed to refine the models and improve predictive accuracy. Additionally, insights from SHAP values were used to explore the structure-property relationships of polymers, highlighting critical substructures that influence mechanical properties.

The workflow depicted in Figure 1 involves the following steps:

- **Conversion to SMILES:** Molecular structures are represented using the SMILES format, enabling compatibility with ML algorithms.

- **Machine Learning Training:** A ML model is trained using the SMILES representation and corresponding Young's modulus values to establish a predictive relationship.

- **Iterative Atom Removal:** Atoms are systematically removed from the molecular structure, and predictions are recalculated to assess the impact of individual atomic contributions.

- **Substructure Contributions via SHAP:** SHAP values are computed to quantify the influence of substructures on the model's predictions, providing interpretable insights into structure-property relationships.

- **Atom-Level Explanation Mapping:** The changes in model predictions resulting from atom removal are mapped, highlighting the atomic-level contributions to the polymer's mechanical properties.

This workflow integrates data-driven modeling with interpretability, offering a detailed understanding of how molecular features influence macroscopic properties.

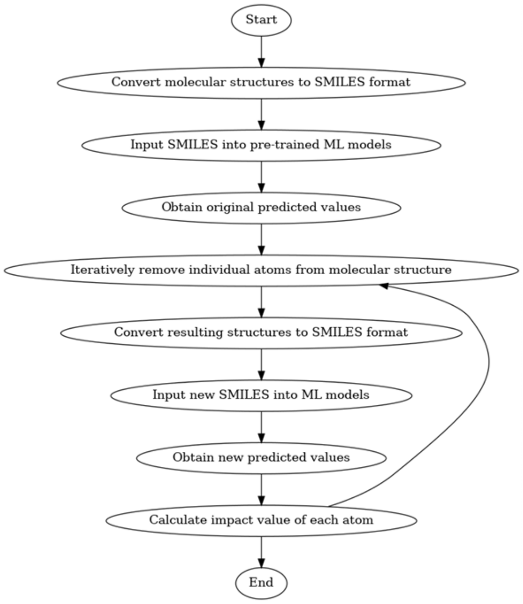
Start
Convert molecular structures to SMILES format
Input SMILES into pre-trained ML models
Obtain original predicted values
Iteratively remove individual atoms from molecular structure
Convert resulting structures to SMILES format
Input new SMILES into ML models
Obtain new predicted values
Calculate impact value of each atom
End

Figure 1. Workflow for Atom-Level Contribution Analysis Using ML.

Model Settings (Machine Learning Models)

- **Feature Representation**

Molecular structures were converted into SMILES format and further processed into Morgan fingerprints with a radius of 3. These fingerprints captured the presence and frequency of 129 key substructures relevant to mechanical properties.

- **Model Architecture**

Two single-task ML models were developed for Young's modulus predictions. The FFNN architecture consisted of 40 neurons in the hidden layer for modulus prediction, while the yield strength model employed four hidden layers with varying neurons (8, 8, 8, and 16). An ensemble approach averaged predictions from 12 models to enhance robustness.

- **Multitask Learning**

A multitask FFNN was implemented to predict multiple related properties simultaneously, leveraging shared information across tasks. This model contained three hidden layers with 128, 16, and 64 neurons, respectively.

Model Settings (Molecular Dynamics Simulations)

- **Polymer Modeling**

Homopolymers were constructed using Materials Studio, with each polymer chain containing approximately 2,000 atoms. These chains were assembled into amorphous configurations and equilibrated through annealing cycles to eliminate thermal stresses.

- **Simulation Parameters**

MD simulations were performed under periodic boundary conditions using the NPT ensemble for equilibration and the NEMD ensemble for tensile testing. Stress-strain curves were generated to extract Young's modulus, validating ML predictions.

- **Force Field and Interactions**

PCFF was used to model interatomic potentials, with nonbonded interactions calculated using Lennard-Jones (LJ) and Coulomb potentials. A cutoff distance of 12 Å was applied for truncating interactions.

- **Explainable Machine Learning (SHAP Values)**

SHAP values were computed to rank the importance of substructures within polymers, providing insights into their influence on mechanical properties. This analysis identified critical substructures that contribute positively or negatively to Young's modulus, guiding polymer design.

Figure 2 outlines the workflow for MD simulations of polymer mechanical properties:

- **Model Establishment (Materials Studio)**

Polymer chains are generated with 80 chains, each 15 units long. The monomer is assembled into a 3D structure, preparing the system for further simulations.

- **Relaxation (LAMMPS)**

The system undergoes three steps:

- Initial relaxation at 650 K using the NPT ensemble.
- Annealing, where the temperature is reduced from 650 K to 300 K.
- Final relaxation at 300 K to achieve equilibrium.

- **Tensile Testing (LAMMPS)**

Tensile simulations are performed at 300 K using the NVT ensemble, with a time step of 0.5 fs. The PCFF potential and Lennard-Jones (9/6) interactions model the system, and an engineering strain rate of $10^9 s^{-1}$ is applied.

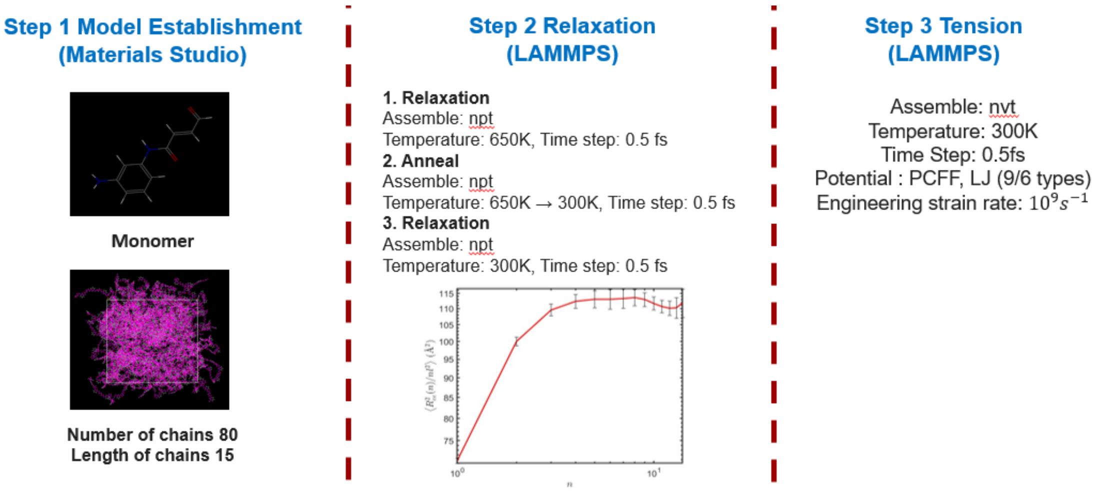

Figure 2. Workflow for MD Simulations: Model Establishment, Relaxation, and Tensile Testing.

Results & Discussion

This study integrates ML and MD simulations to investigate the mechanical properties of Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl) (Figure 3), with a focus on the contributions of individual atoms and molecular substructures to Young's modulus.

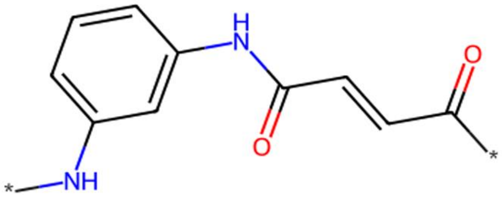

Figure 3. The molecular structure of Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl).

Key findings are summarized as follows:

- **Atom-Level Contributions to Young's Modulus**

The single-task ML model provided atom-level insights into how individual atoms influence the polymer's mechanical properties. As shown in Figure 4a, green regions in the molecular map represent atoms whose removal decreases the predicted Young's modulus, indicating a positive contribution to the property. Pink regions show a negative contribution, where removal increases the predicted modulus, while gray regions have a negligible impact. The graph in Figure 4b quantifies these effects, highlighting significant contributions from atoms 6, 10, and 12. These findings demonstrate the ability of ML to pinpoint critical atomic features within complex molecular systems.

- **Molecular Dynamics Simulations for Validation**

MD simulations, conducted in three steps (model building, relaxation, and tensile testing), confirmed the ML predictions. Figure 2 outlines the computational process: the polymer model was built in Materials Studio, relaxed under high temperature and pressure using LAMMPS, and subjected to tensile testing. The calculated Young's modulus of 1.65 GPa aligns closely with the ML-predicted value, with an 8% error margin. This agreement validates the robustness of the ML model in predicting mechanical properties.

➢ SHAP Analysis for Substructure Contributions

SHAP values were used to rank the contributions of substructures to Young's modulus. Figure 5 highlights significant substructures, with nitrogen atoms showing a positive impact on the predicted values, consistent with atom-level explanations. However, discrepancies were noted, such as benzene ring carbon showing positive contributions in SHAP analysis but negative effects in atom-level predictions. These differences likely stem from overlapping substructure definitions in SHAP analysis.

➢ Model Performance

The single-task ML model demonstrated strong predictive capabilities, achieving high R^2 values for both training and testing datasets, as shown in Figure 6. This confirms the model's ability to effectively correlate molecular structure with mechanical properties.

➢ Comparison of ML and SHAP Analysis

While both methods provided valuable insights, some inconsistencies were observed. For example, SHAP results attributed a positive impact to benzene ring carbons, while atom-level analysis identified a negative effect. These variations underscore the complementary nature of these techniques and the importance of combining them for a comprehensive understanding.

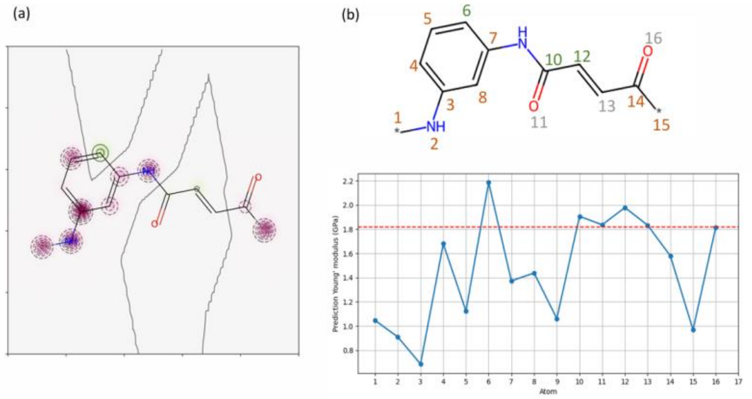

Figure 4. Atom-level explanation obtained using a single-task ML model. (a) Impact map of the molecule, with color coding: green indicates regions where removing bits decreases the prediction value (positive effect), gray indicates no change in prediction value, and pink indicates regions where removing bits increases the prediction value (negative effect). (b) Predicted values of Young's modulus after the removal of each individual atom.

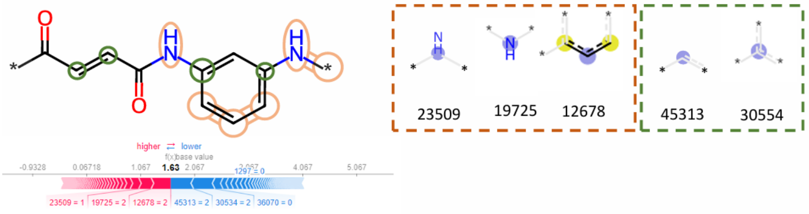

Figure 5. SHAP values derived from a single-task ML model. The figure ranks substructures by their significance, with each dot representing the influence of a sample from the training dataset. It highlights key substructures impacting Young's modulus, with red and blue arrows indicating positive and negative effects, respectively. Even when a substructure's feature value is "0" (absent in the molecule), the arrow length reflects its importance.

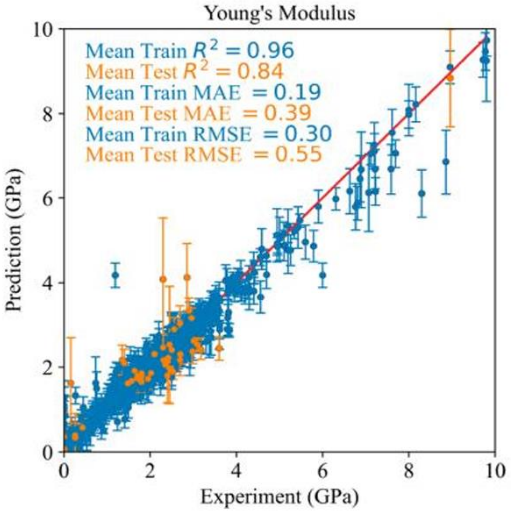

Figure 6. The training and testing R^2 values of the single-task ML model.

Conclusions & Future Work

This study integrates ML and MD simulations to predict polymer properties and understand structure-property relationships. Single-task ML models demonstrated high accuracy, while SHAP provided atom-level insights. Multi-task ML models further enhanced predictions by leveraging interrelated property correlations, improving generalizability. MD simulations validated the models, confirming their robustness and applicability for designing high-performance polymers for applications such as energy storage and flexible electronics.

Future work will expand ML models to predict additional properties and incorporate topological features like cross-linking and branching. Multi-task learning will facilitate the design of multifunctional materials by modeling complex inter-property relationships. Exploring larger chemical spaces with generative models and integrating uncertainty quantification will enhance reliability. Experimental validation will bridge computational insights with real-world applications, driving accelerated polymer discovery and material innovation.

Acknowledgments

I would like to acknowledge the assistance provided by OpenAI's ChatGPT in formulating parts of this report by reviewing grammar and refining sentences.

References

[1] National Research Council, et al. Polymer science and engineering: the shifting research frontiers. National Academies Press, 1994.
[2] Brinson, Hal F., and L. Catherine Brinson. "Polymer engineering science and viscoelasticity." An introduction 99 (2008): 157.

[3] Sawyer, Linda, David T. Grubb, and Gregory F. Meyers. Polymer microscopy. Springer Science & Business Media, 2008.

[4] Lutz, Jean-François, et al. "Sequence-controlled polymers." Science 341.6146 (2013): 1238149.

[5] Balasubramanian, Ganesh, et al. "Towards Improving the Efficiency of Organic Solar Cells by Coarse-Grained Atomistic Modeling of Processing Dependent Morphologies." Computing in Science & Engineering 23.3 (2021): 48-55.

[6] Hartmann, Laura, and H. G. Börner. "Precision Polymers: Monodisperse, Monomer‐Sequence‐Defined Segments to Target Future Demands of Polymers in Medicine." Advanced Materials 21.32‐33 (2009): 3425-3431.

[7] Monroe, Barbara, and David Oliviere, eds. Resilience in palliative care: achievement in adversity. OUP Oxford, 2007.

[8] Dusoe, Keith J., et al. "Ultrahigh elastic strain energy storage in metal-oxide-infiltrated patterned hybrid polymer nanocomposites." Nano letters 17.12 (2017): 7416-7423.

[9] Hill, Joanne, et al. "Materials science with large-scale data and informatics: Unlocking new opportunities." Mrs Bulletin 41.5 (2016): 399-409.

[10] Liu, Yue, et al. "Materials discovery and design using machine learning." Journal of Materiomics 3.3 (2017): 159-177.

[11] Pilania, Ghanshyam, et al. "Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers." Journal of Chemical Information and Modeling 59.12 (2019): 5013-5025.

[12] Deacy, Arron C., et al. "Sequence control from mixtures: switchable polymerization catalysis and future materials applications." Journal of the American Chemical Society 143.27 (2021): 10021-10040.

[13] Otsuka, Shingo, et al. "PoLyInfo: Polymer database for polymeric materials design." 2011 International Conference on Emerging Intelligent Data and Web Technologies. IEEE, 2011.

[14] Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." Journal of chemical information and computer sciences 28.1 (1988): 31-36.