# Project “MSE803”: Data-based Atom-level Explanation of Polymer Properties

Yuhao Liu, PhD student

Dept. of Mechanical Engineering, University of Wisconsin-Madison

**Date:** 12/15/2023

*MSE 803: Special Topics in Material Science*

## Chapter



## Figures List

# Data-based Atom-level Explanation of Polymer Properties

Yuhao Liu

**ABSTRACT:** This research delves into the macro-scale characteristics of polymers by integrating atomic-scale analysis with advanced machine learning techniques. Polymers, known for their complex properties due to the varying structures and configurations of their constituent monomers, present a challenge for standard machine learning approaches, which often lack the capability to adequately represent the intricate details of polymer chains. Addressing this gap, our study introduces a novel method that combines atom-level insights with single-task machine learning to enhance the understanding of polymer behaviors. This approach, centering on the concept of atomic similarity within polymers, aims to enhance the clarity and precision of machine learning predictions, particularly in relation to the mechanical properties of polymers, exemplified by assessing Young's modulus. Our findings reveal the limitations of conventional machine learning explain approaches.

**KEYWORDS:** *machine learning, molecular dynamics simulations, atom-level explanation*

## INTRODUCTION

Polymer is a kind of wide used material. We can see it everywhere in our life, living, clothing, and producing etc. The reason is located in its easy processing and strong design ability, which means we can design the polymer structure by arranging different basic structure. The arrangement and even the number of those basic structure may play a key role on its mechanical property. This is also the beauty it has. Based on this, some subjects have been derived from here, such as polymer chemistry and polymer physics. Although many polymers have been designed successfully and entered our life. But there are still many secrets behind those material, how those basic structures affect its final mechanical property is not well understood. Therefore, it is important and meaningful to build a link or relation between its atom level structure and its macro-scale property. In this project, we use the polymer of Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl) to see how its molecular unit affects its Young's modulus. It is very hard or nearly impossible to explore every polymer about how their structure affects its final mechanical property, because there are more than one billion kinds of polymer have been discovered, not to say every one may own different properties. We only provide an example here to see how we can build this kind of relationship.

Machine learning has become a powerful tool used in scientific research. Its development has been driven the need to process large volumes of data and make intelligent decisions faster than humans can. Machine learning emerged from the desire to create machines that could mimic human learning and problem-solving abilities. Its development has been propelled by technological advancements, the availability of data,

and a wide range of applications that demand adaptive, efficient problem-solving capabilities. With the contribution from the researcher in Material Science, machine learning has been widely used in the subject of Material Science, such as the predictive modeling of material properties, material design and discovery, and structural analysis and characterization etc. In this study, SMILE has been used to represent the polymer structure, which is simple representation for computer and human to understand. But it can not include the topology information within here, such as cross-linking, branching, etc. Therefore, it is difficult to understand how those topological structure affects their property when we only use SMILE to represent its molecular. To address this issue, an atom-level explanation of polymer properties method based on the atom-level explanation of polymer similarity will be used.

Although there are many researches have shown the powerful prediction ability of ML method, it is still necessary to use some other method to validate the predicted result of ML method. Molecular Dynamic simulation is based on Newton's motion law, using potential function of atom structure to calculate the relevant polymer property. In this project, we used this method to make sure our prediction is reasonable.

**MODEL SETTING**

PoLyInfo, the most extensive database, holds details of more than 18,000 polymers, featuring 12,854 homopolymers along with their chemical structures and nearly 100 varieties of properties. This specific homopolymer collection is ideal for neural network training in our research. Within this collection, 923 homopolymers have documented Young's modulus values. By utilizing these noted property values and the associated monomer structures, machine learning models can be developed to establish a correlation between polymer composition and properties.

The intact process is shown in Figure 1, starting with the converting molecular structures into the SMILES format. Once in SMILES format, these structures are then fed into pre-trained machine learning (ML) models. The initial predicted values are obtained from these models. The process continues by iteratively removing individual atoms from the molecular structure. After each removal, the modified structure is converted back into the SMILES format. The new SMILES strings are input into the ML models again to obtain new predicted values. Finally, the impact value of each atom on the overall structure is calculated, concluding the cycle.

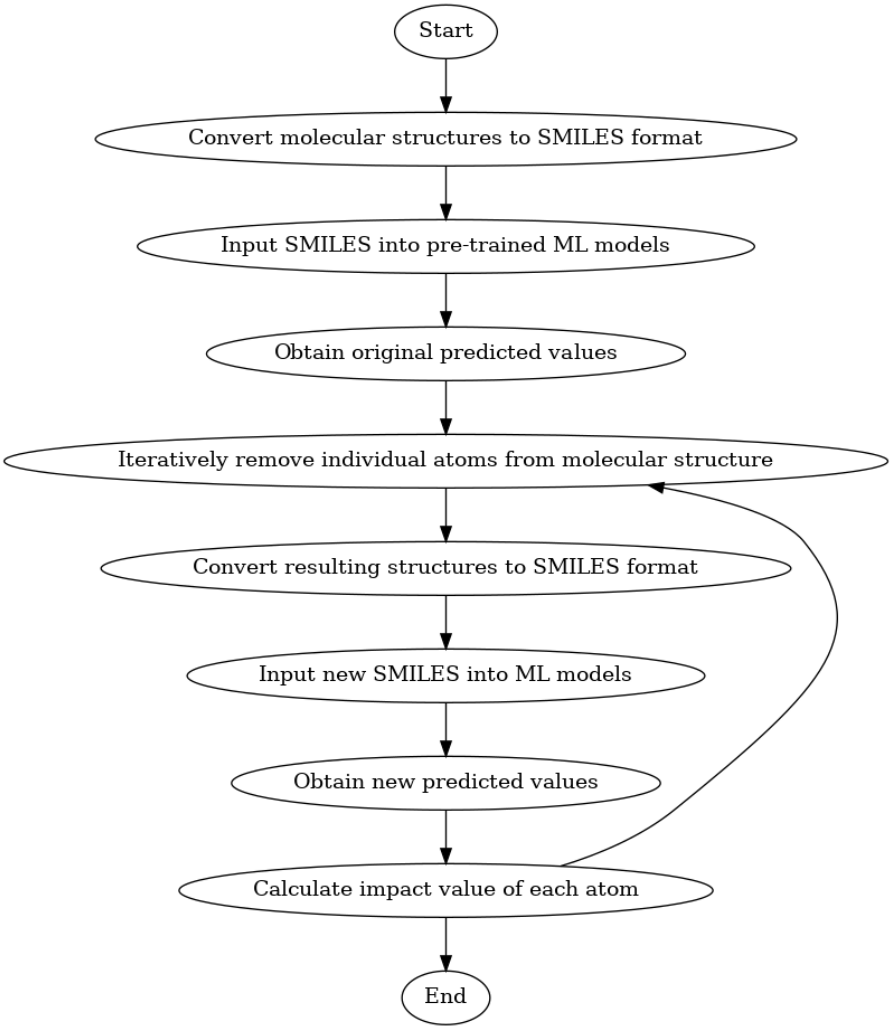


Figure 1. The flow chart of the machine learning

**RESULTS AND DISCUSSIONS**

In this study, Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl) has been used to see how molecular unit affects its Young's modulus. The molecular structure is shown in Figure 2. The reason using this polymer is because there are limited units we can observe in the structure (only 17 units).

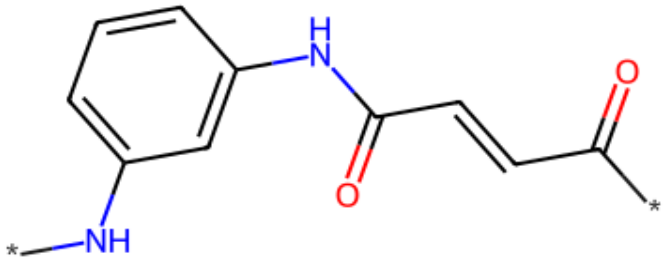


Figure 2. The structure of Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl).

As shown in Figure 3(a), the map shows the molecule with a network of interactions, where different colors represent the impact of removing certain parts of the molecule on the prediction value. Green indicates areas where removal decreases the prediction

value, suggesting these areas have a positive effect on the property being predicted. Gray shows no change in prediction value upon removal, indicating these parts do not significantly influence the predicted property. Pink areas represent a negative effect; removing these parts increases the prediction value, meaning they are likely to be inhibitory or destabilizing features for the property in question. In Figure 3(b), this graph presents the predicted Young's modulus values when each atom is individually removed from the molecule. The x-axis lists the atoms in sequential order as they appear in the molecular structure, while the y-axis shows the predicted Young's modulus in GPa. The line graph exhibits fluctuations, with certain atoms having a more pronounced effect on the predicted value than others. This suggests that some atoms (atom 6, 10, 12) play a more significant role in the molecular property being modeled. The dashed red line indicates a predicted value using impact structure, which serves as a reference to understand how the removal of each atom affects the Young's modulus relative to this baseline. In summary, the figure demonstrates how a single-task machine learning model can offer insights into the contribution of individual atoms to a molecule's mechanical property (in this case, the Young's modulus).

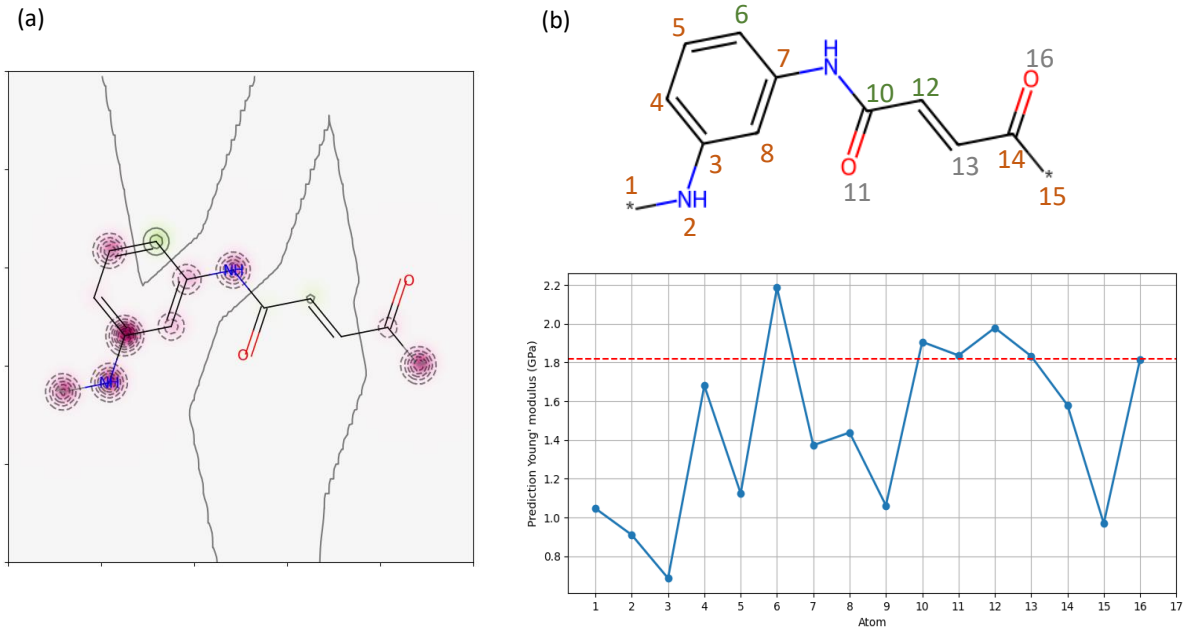


Figure 3. The atom-level explanation obtained using a single-task ML model. (a) Impacts map of molecule. Color scheme: removing bits decreases prediction value (i.e. positive effect) (green), no change in prediction value (gray), removing bits increases prediction value (i.e. negative effect) (pink). (b) The specific predicted values after removing each individual atom.

The Figure 4 provides a graphical representation of a computational process involving molecular dynamics (MD) simulations, which consists of several steps. First, building a polymer model using Materials Studio. Second, relaxation of configuration. In this step, it involves the use of LAMMPS for the relaxation process. Finally, conducting tensile experiments using LAMMPS. We can see the final calculated Young's modulus is 1.65 GPa, which is close to the predicted value using ML method. We can see there

is some difference we can see between both methods. It is still acceptable. Figure 5 gives the train R^2 and test R^2, which shows a high prediction accuracy. Based on both validations, it means our ML model is reasonable.

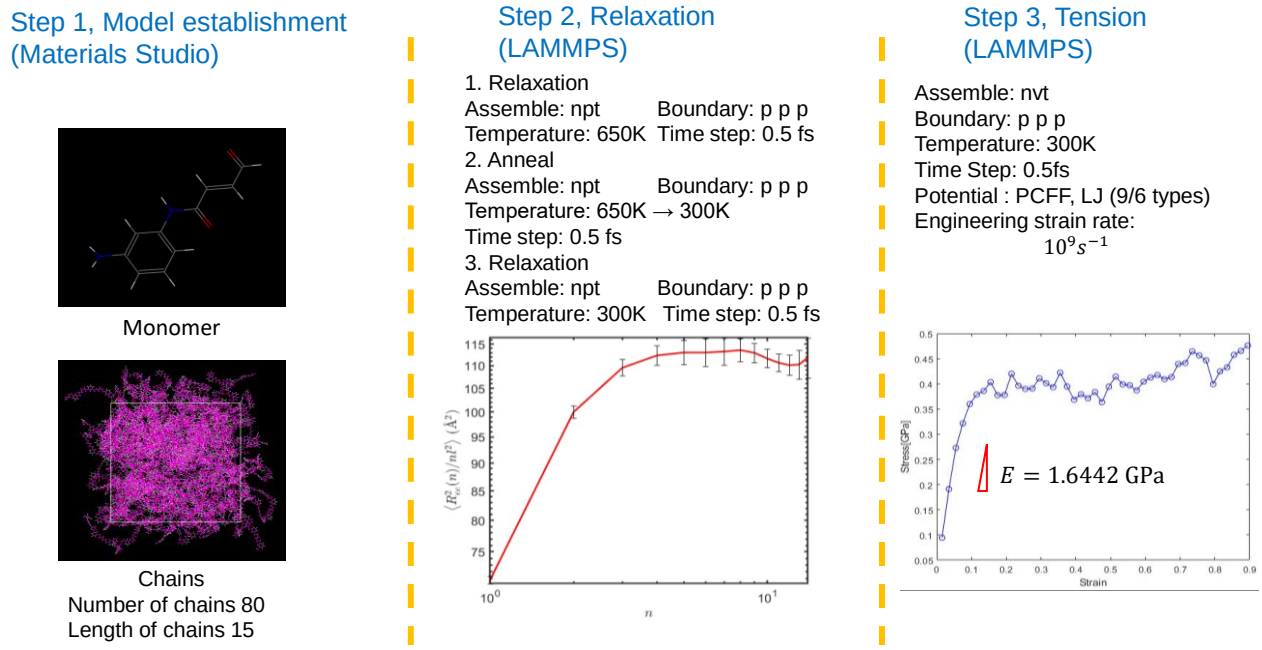


Figure 4. The computational process of MD simulation involves several steps. (a) Building a polymer model using Materials Studio. (b) Relaxation of configuration of the homopolymer. (c) Conducting Tensile Experiments using LAMMPS.

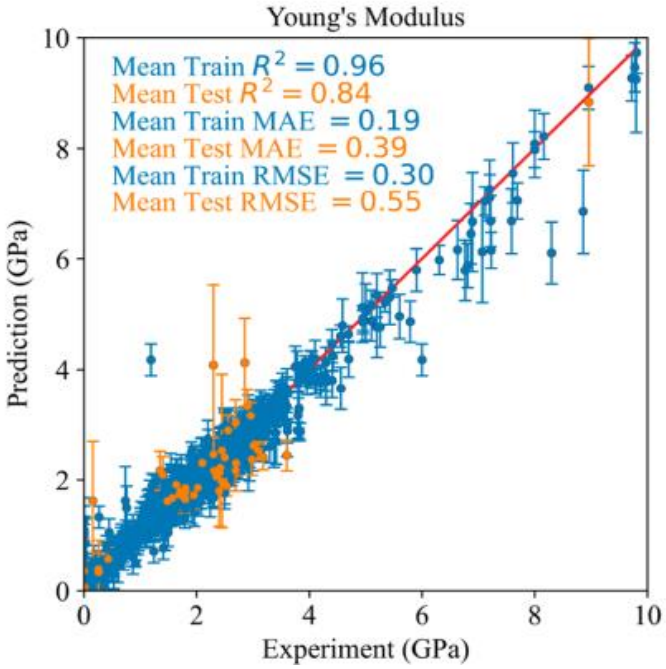


Figure 5. The training R^2 and testing R^2 of Single-task Model.

In this study, SHapley Additive exPlanations (SHAP) values are also calculated based on single-task ML model determine the impact of substructures of Poly(imino-1,3-

phenyleneiminobut-2-ene-1,4-dioyl). Shapley Additive exPlanations (SHAP) is a method used in the field of machine learning to explain the output of a model. It's based on the concept of Shapley values from cooperative game theory. SHAP values provide a way to understand the contribution of each feature in a dataset to the prediction made by a machine learning model. In Figure 6, it shows the SHAP values calculated based on single-task ML model. From the important substructures shown in Figure 6, it can be observed that the substructures have relatively small scales, with a small number of atoms and bonds. By analyzing the molecular graph, it can be observed that both nitrogen atoms contribute to improving the predictive values. This observation aligns with the atom-level explanation obtained using a single-task ML, which also indicated the positive effect of both nitrogen atoms. Similarly, the two carbon atoms with double bonds have a negative impact on the prediction, which is consistent with the atom-level explanation obtained using a single-task ML. In the SHAP Value results, the five carbon atoms in the benzene ring exhibit a positive effect. However, this finding contradicts the atom-level explanation obtained using a single-task ML. Additionally, SHAP Values can sometimes present contradictory outcomes, as the same atom may belong to multiple different substructures simultaneously. This contrasts with the atom-level explanation, where such situations do not occur. Both oxygen atoms do not show a significant impact on the predictive values in both SHAP Values and the atom-level explanation.

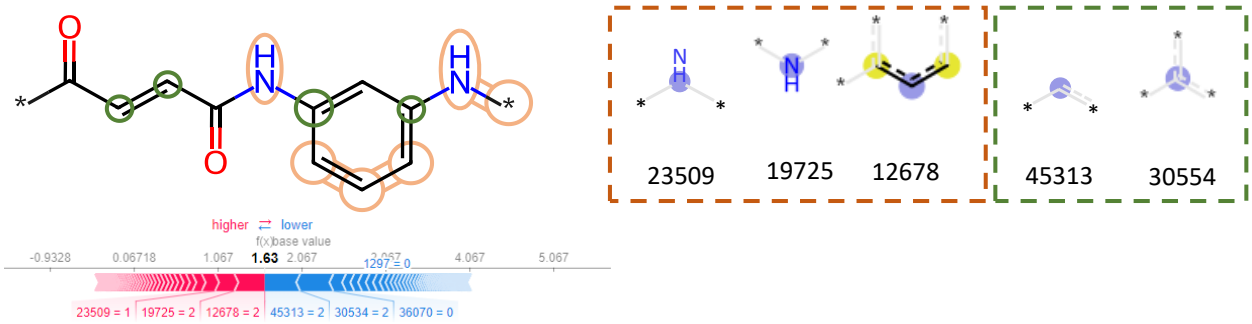


Figure 6. The SHAP values derived from a single-task machine learning model. This figure ranks crucial substructures in order of significance, where each dot symbolizes the influence of a specific sample from the training dataset. Additionally, it highlights the most influential substructures linked to Young's modulus, as determined by SHAP values. This is exemplified in the individual SHAP value diagram for Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl), where red and blue arrows denote positive and negative impacts of substructures, respectively. In cases where a substructure's feature value is "0", signifying its absence in a molecule, the length of the arrow still accurately reflects its importance.

## CONCLUSIONS

In this study, we explore the role of molecular unit on its Young's modulus, and build the relationship between its molecular unit and its final property. We find single-task ML model has better prediction for the elastic modulus, compared with MD simulation,

only 8% predicted error can be observed. Besides that, the SHAP values, calculated using single-task machine learning models, provide insightful analysis into the impact of substructures of Poly(imino-1,3-phenyleneiminobut-2-ene-1,4-dioyl). There are some findings that are not always consistent with the atom-level explanation obtained from the single-task ML model.