NeurIPS Open Polymer Prediction

Student Name	Student Number
Charles Chang	1006115525
Aakash Kanagala	1008091967
Kevin Zhu	1008451630

A proposal submitted in conformity with the requirements for CHE1147

Department of Chemical Engineering and Applied Chemistry
University of Toronto

Introduction

Predicting polymer properties using machine learning remains a central challenge in chemical and materials informatics. Accurate predictions can significantly reduce the time and cost of experimental testing and molecular dynamics (MD) simulations, accelerating the design of polymers with tailored functionalities for applications in packaging, electronics, coatings, and sustainable materials.

This study is based on the 2025 NeurIPS Open Polymer Prediction Challenge[1] dataset, which includes 7,973 polymer entries represented by SMILES notation and five key properties derived from MD simulations: glass transition temperature (Tg), fractional free volume (FFV), thermal conductivity (Tc), density, and radius of gyration (Rg). Tg determines thermal stability, FFV reflects porosity, Tc influences heat transfer, density relates to packing efficiency, and Rg represents molecular flexibility.

The project aims to explore the distributions and interrelationships of these properties and descriptors to guide preliminary feature selection and develop predictive models for Tg and Tc, which are key to thermal paste formulation.[2] Previous studies have achieved varying success: Casanola-Martín et al.[3] applied multiple linear regression to predict Tg across 900 homopolymers using a reduced set of 15 key descriptors, while Huang et al. introduced Uni-Poly, a multimodal framework combining SMILES, 2D, and 3D representations to outperform single-modality models.[4] Building on these efforts, this work seeks to identify feature–property relationships that enhance interpretability and prediction accuracy for polymer performance modeling.

Approach

Our initial proposed workflow involves three major sections. In the first stage, the data is loaded and preprocessed to remove empty and duplicate samples. Features are then extracted from SMILES molecular representation using RDKit, which is followed by data splitting. The next major stage is to identify the hyperparameters and to create a pipeline. This pipeline includes hyperparameters that are not number-based such as model type and scaling method. In addition, numerical hyperparameters such as learning rate are also included. Eventually, a satisfied model will be tested with a test set and if the result is promising, the final model will be trained based on both test and train data set to improve its generalization.

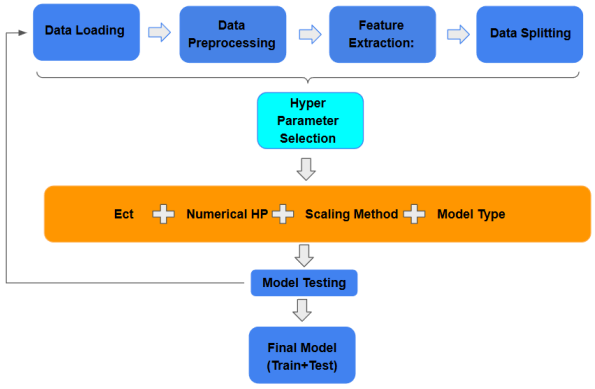

Figure 1: Flow chart for supervised learning model development.

Exploratory Data Analysis (EDA)

Distribution of Key Variables

There are four types of distributions observed across variables with one example discussed with respect to each distribution. The most frequent distribution is right-skewed distribution as seen in Tg, ranging from about –100 °C to 400 °C (Fig. 2). This indicates that while most polymers transition at lower temperatures, a few exhibit exceptional thermal stability. The other common distribution is bimodal represented by Tc where two distinct peaks around 0.2 and 0.35 W/m·K is observed(Fig. 2). This pattern suggests the presence of two polymer categories, possibly crystalline and amorphous materials, indicating that simple linear models may not adequately capture the complex structure–property relationships in the dataset. Discrete distribution is also observed in some RDkit extracted features (HBD, RingCount and Numaromatics Rings), which indicate their potential poor correlation with other target features. Lastly, narrow random distribution is also observed from FFV.

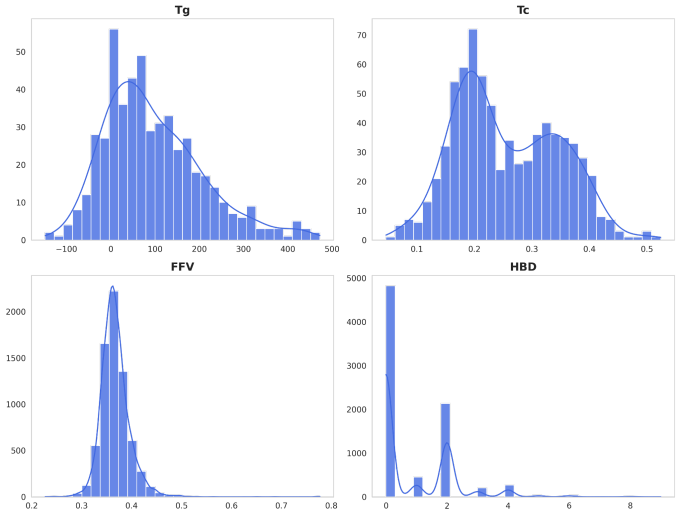

Figure 2: Polymer Property Distribution

Based on the above observation we propose stratified splitting and random splitting and it will be optimized as hyperparameters. In terms of the normalization method, Z score, MinMax scaling are proposed as hyperparameters since properties magnitudes vary significantly and Tg includes both negative and positive values, Z score and MinMax scaling preserve sign and relative magnitude are selected as tunable hyperparameters within the model pipeline. Some extreme Tg and Density values may represent simulation artifacts and are treated as outliers. Initial trends also suggest that density and fractional free volume may be inversely related, while polymers with larger radii of gyration could exhibit lower thermal conductivity; these hypotheses will be quantitatively validated in subsequent analysis. Based on the distribution trend, Kernel and XGBoost are proposed as linear models that would underestimate the skewed tail and overfit the dense region.

Trends and Insights

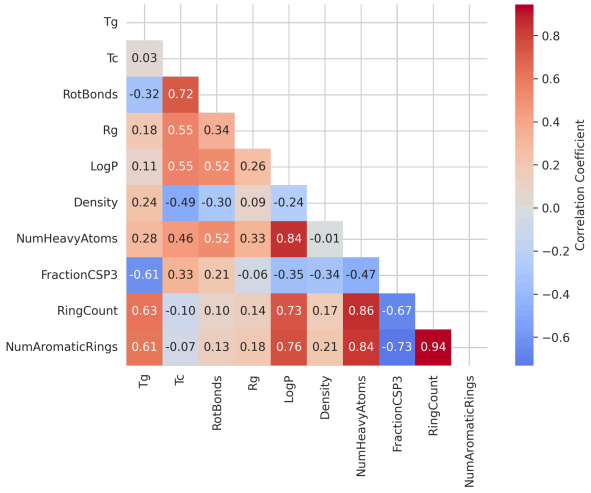

Table 1: Interactive Correlation of Tc and Tg

Tc	**Tg**
logP	FractionCSP3
Numheavyatoms	NumAromaticRings
R_g	RingCount
Rotabonds	-

Figure 3 : Selected Molecular Descriptor Correlation Heatmap

Figure 3 demonstrates the heatmap between 8 temporarily selected variables, and the highest interactive correlation amongst the features to targets are reported in Table 1. The 8 features are selected under two categories. First, features with a moderated individual correlation (>0.45, based on intuitive selection) with each target respectively. Second, features are kept if they have a high interactive correlation with either target reported from EDA analysis. The heatmap including all variables is presented in Appendix A. Wrapper feature selection will be performed when tuning the model to further optimize the model (by adding or removing features and evaluating the change in model performance). The high dimensional relationship justified our selection of Kernel and XGBoost as they capture the interactive effects.

Conclusion

The dataset presents several challenges for predictive modeling, primarily due to its limited number of data points, which restricts the model's capacity to generalize and underscores the need for careful testing. Predicting the glass transition temperature (Tg) adds complexity, as it is a bulk material property, whereas RDKit-derived descriptors capture only single-molecule characteristics. The imbalance and unequal spread among Tg and Tc values may bias predictions toward dominant ranges, while the nonlinear relationships between variables such as density and fractional free volume necessitate nonlinear feature engineering for improved physical interpretability. Additionally, potential high-temperature and high-density outliers must be verified to ensure they reflect realistic polymer behavior rather than simulation errors. To address these challenges, we propose a supervised learning workflow integrating molecular descriptors from RDKit with simulation-derived features, emphasizing rigorous preprocessing through outlier removal, missing-value handling, and adaptive scaling (Z-score and MinMax). Feature selection will be guided by correlation analysis, chemical interpretability, and wrapper methods to retain meaningful descriptors. Machine learning models including XGBoost and Kernel Ridge Regression will be optimized through cross-validation, with linear regression and instance-based models serving as baselines, and hyperparameter tuning applied to capture the nonlinear structure–property relationships effectively.

References

(1) *NeurIPS - Open Polymer Prediction 2025*. https://kaggle.com/neurips-open-polymer-prediction-2025 (accessed 2025-10-23).

(2) Wang, H.; Ihms, D. W.; Brandenburg, S. D.; Salvador, J. R. Thermal Conductivity of Thermal Interface Materials Evaluated By a Transient Plane Source Method. *J. Electron. Mater.* **2019**, *48* (7), 4697–4705. https://doi.org/10.1007/s11664-019-07244-0.

(3) Casanola-Martin, G. M.; Karuth, A.; Pham-The, H.; González-Díaz, H.; Webster, D. C.; Rasulev, B. Machine Learning Analysis of a Large Set of Homopolymers to Predict Glass Transition Temperatures. *Commun Chem* **2024**, *7* (1), 226. https://doi.org/10.1038/s42004-024-01305-0.

(4) Huang, Q.; Li, Y.; Zhu, L.; Zhao, Q.; Yu, W. Unified Multimodal Multidomain Polymer Representation for Property Prediction. *npj Comput Mater* **2025**, *11* (1), 153. https://doi.org/10.1038/s41524-025-01652-z.

Appendix

Appendix A - Figures

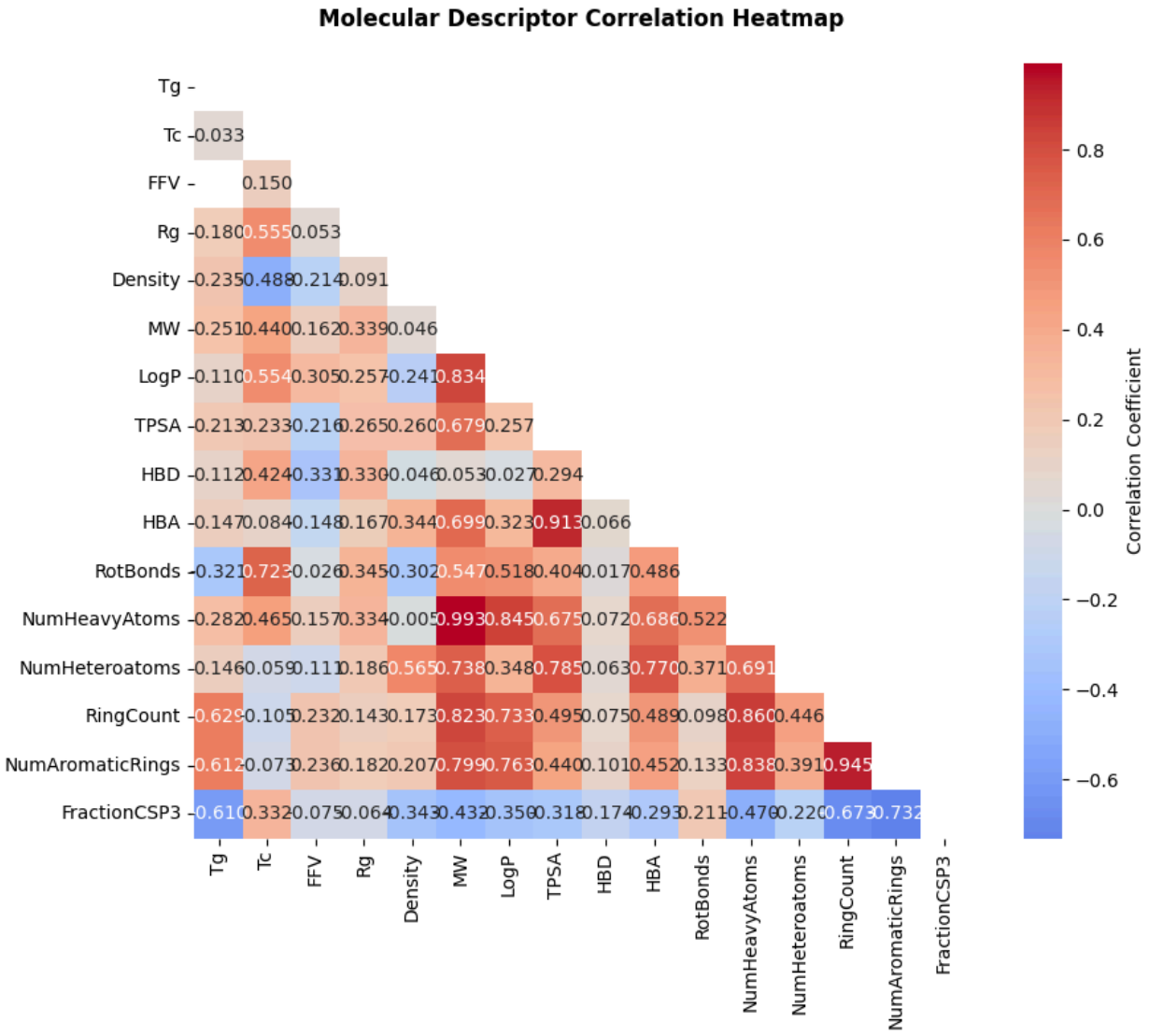

Fig 4: Molecular Correlation Heatmap including all variables

Appendix B - Tables

Table 2: Absolute Correlation of Tg and Tc with features and targets

S.No	Column	Absolute Correlation	Column	Absolute Correlation
0	Tc	1	Tg	1
1	RotBonds	0.722609	RingCount	0.628617
2	Rg	0.554758	NumAromaticRings	0.612252
3	LogP	0.554276	FractionCSP3	0.610485
4	Density	0.488476	RotBonds	0.320607
5	NumHeavyAtoms	0.464714	NumHeavyAtoms	0.281691
6	MW	0.439685	MW	0.250809
7	HBD	0.424403	Density	0.235291
8	FractionCSP3	0.331863	TPSA	0.212706
9	TPSA	0.233444	Rg	0.180072
10	FFV	0.149878	HBA	0.14733
11	RingCount	0.104799	NumHeteroatoms	0.146351
12	HBA	0.083799	HBD	0.111675
13	NumAromaticRings	0.073421	LogP	0.110265
14	NumHeteroatoms	0.058626	FFV	NaN

Table 3: Absolute Correlation of Tg and Tc with selected features and targets

	Tc	Absolute Correlation	Tg	Absolute Correlation
1	RotBonds	0.722609	RingCount	0.628617
2	Rg	0.554758	NumAromaticRings	0.612252
3	LogP	0.554276	FractionCSP3	0.610485
4	Density	0.488476	RotBonds	0.320607
5	NumHeavyAtoms	0.464714	NumHeavyAtoms	0.281691
6	FractionCSP3	0.331863	Density	0.235291
7	RingCount	0.104799	Rg	0.180072
8	NumAromaticRings	0.073421	LogP	0.110265