Heat diffusion coefficient study of polymers based on interpretable machine learning

Congyi Chen

School of Physical Science and Technology, Shanghaitech University, Shanghai 200100, China

chency2022@shanghaitech.edu.cn

Abstract. Polymers hold significant application value across various fields of modern society, with different application scenarios requiring specific thermal diffusivity coefficients. Finding polymer materials with targeted thermal diffusivities is crucial. However, due to the vast variety and complex structures of polymers, constructing a unified structured dataset for machine learning modeling is challenging. Although machine learning has shown great potential in materials science, it has rarely been applied to predict the heat diffusion coefficient of polymers. This paper constructs a dataset for predicting the thermal diffusion coefficient of polymers using a publicly available dataset by transforming the SMILES code of polymers into eight features with practical physical and chemical meanings. Using the Random Forest algorithm, training with 400 of these data and randomly selecting 200 of them for cross-validation, the accuracy of the test set reached 0.9. Additionally, through interpretability analysis, we found that the molecular weight of the polymer monomers, the TPSA (the polar surface area of the molecule), and the NRB (the number of rotatable bonds) are the main features affecting the polymer thermal diffusion coefficient. An increase in the TPSA and the NRB positively contributes to the thermal diffusivity, while an increase in molecular weight negatively contributes. Our study provides a new method for the prediction of polymer thermal diffusivity and creates a new paradigm for the study of polymer thermal diffusivity, promoting further development in this field.

Keywords: polymer, heat diffusion coefficient, machine learning.

1. Introduction

Polymer materials are used extensively in electronics[1], medical[2], aerospace[3], and other applications due to their lightweight, wear resistance, and ease of synthesis. Different fields and application scenarios require varying thermal diffusivity rates for polymer materials. Thermal diffusivity, also known as thermal conductivity, is a physical quantity that characterizes the speed at which an object warms, defined as the ratio of the amount of heat passing through a unit area in a unit of time to the temperature gradient over that area, measured in (m^2/s). Materials with high thermal diffusivity are valuable in the thermal management of electronic devices. Materials with excellent thermal diffusivity are important for advancing ultra-high frequency and high-power microelectronic devices[4]. Conversely, materials with very low thermal diffusivity serve as excellent thermal insulation materials, widely demanded in in aerospace insulation and other areas[5]. Thus, finding an efficient and accurate method to screen materials with specific thermal diffusivity is particularly urgent.

In recent years, machine learning methods based on big data have begun to play an increasingly important role in many fields, such as materials science and physics, achieving significant success in predicting and screening various material properties, including battery performance[6] and the search for potential high-performance battery materials[7]. Compared with traditional experimental methods, machine learning offers significant time and cost advantages in screening materials. The laser pulse method for determining the thermal diffusivity of a sample requires the sample to be synthesized beforehand, and the total measurement time can be up to ten hours, which is not favorable for large-scale screening of polymers with specific thermal diffusivity[8]. The machine learning model designed in this paper can analyze information from over 1000 polymers and predict and screen their thermal diffusivities within a few minutes. While computational simulation requires accurate modeling of the polymer material, machine learning eliminates the need to construct complex physical models and learns directly from existing data, avoiding errors that may result from simplifying the model for computation during simulation. Although machine learning has shown potential in predicting and screening material properties, it has rarely been applied to research in materials thermology, where the demand for materials with specific thermal diffusivities is high.

This paper designs a set of Random Forest machine learning models to predict the thermal diffusivity of polymers and analyzes the relationship between certain specific structures in polymer monomers and the thermal diffusivity of polymers. We first found 1077 polymer monomer SMILES versus thermal diffusivity in a public dataset. Subsequently, the SMILES were converted into eight features with specific physical meanings, and then a machine learning model was trained to learn the relationship between these features and thermal diffusivity with an accuracy of 0.90. Additionally, we found that three quantities, namely, the molecular weight, the polar surface area of the molecule, and the number of rotatable bonds, have the highest degree of influence on the thermal conductivity of polymers by means of interpretability analysis, and analyzed them using the SHAP method to determine the directionality of their influence on the thermal diffusion coefficient.

2. Methods

All code in this paper was implemented in Python. This study uses a publicly available dataset sourced from GitHub[9], which contains the SMILES codes of 1077 polymers and their corresponding heat diffusion coefficients. Since SMILES codes cannot be directly used as inputs to machine learning models, we used the rdkit.Chem.Descriptors module to encode SMILES into eight features with actual physical meaning. The features include the average molecular weight, polarity, hydrophobicity, rotational ability of chemical bonds, and some special functional groups.

*C*orrelation Calculation: Pearson correlation coefficients between features were calculated using pandas.DataFrame.corr[10].

Correlation Matrix Visualization: Correlation heatmaps were drawn using matplotlib[9] and seaborn[11].

Prior to model training, we normalized the feature data to eliminate differences between different measures. Normalization was performed using StandardScaler to ensure each feature had a mean of 0 and a standard deviation of 1. We chose the RandomForestRegressor model for training due to its ability to handle high-dimensional data and nonlinear relationships. To optimize model performance, we used the GridSearchCV method to adjust model hyperparameters. The specific parameters were as follows: number of trees (n_estimators): 100 to 700; maximum depth (max_depth): 3 to 10.

We randomly divided 1077 data into training and test sets, with 400 data used for training and 200 data used for cross-validation. We used the 20-fold cross-validation (ShuffleSplit) method to ensure the stability and reliability of the model.

Performance Metrics: The following metrics were used to assess the performance of the model: Mean Absolute Error (MAE), Mean Square Error (MSE), Coefficient of Determination (R^2).

In the test set, we assessed the predictive accuracy of the model by comparing the linear regression of the predicted values with the actual values. Feature importance analysis and SHAP (SHapley Additive exPlanations) value analysis were implemented through scikit-learn and shap libraries.

3. Results and Discussion

Initial processing of the dataset is crucial to ensure data quality and consistency, thereby improving model performance and accuracy. For the initial analysis, we plotted the kernel density distribution of the thermal diffusivity of the 1077 polymers in the dataset(Figure 1 below). The sample approximates a normal distribution with a maximum value of 2.27e-7 (m^2/s) and a minimum value of 2.96e-8 (m^2/s). There are no extreme values in the sample, with most polymers' thermal diffusivity centrally distributed in the range of 0.5-0.9e-7(m^2/s). Hence, the collected dataset is suitable for machine learning.

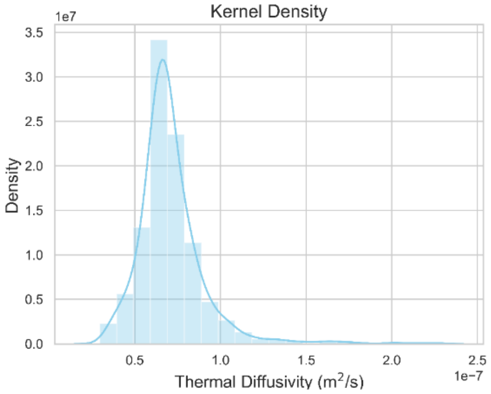

Figure 1. Kernel density distribution of polymer thermal diffusivity in the dataset

The initial dataset only contained the polymer monomer SMILES strings and their corresponding thermal diffusivity data. Since SMILES notation encodes molecular structure in a linear string format and does not represent the three-dimensional arrangement of atoms, nor reflect the interaction between functional groups, it is not suitable for direct input into machine learning models. Thus, we first encoded SMILES into eight quantifiable features with actual physical or chemical meanings.

Tabel 1. The abbreviations and meanings of the features.

qed	Quantitative estimation of drug similarity
MolWt	Average molecular weight of a molecule
TPSA	Polar surface area of a molecule
FractionCSP3(FCSP3)	Ratio of sp3 hybridized carbon atoms in a molecule
NumHDonors(NHD)	Number of hydrogen bond donors
NumRotatableBonds(NRB)	Number of rotatable bonds
MolLogP(MLP)	Degree of hydrophobicity of the molecule
fr_halogen(fr_h)	Number of halogens

These features cover the average molecular mass, polarity, hydrophobicity, rotational ability of chemical bonds, and some special functional groups. To avoid redundant features, which could negatively impact machine learning modeling by increasing training time, storage cost, risk of overfitting, and poor model interpretability, we calculated the correlation between each feature and the target quantity (thermal diffusivity) and plotted a correlation heatmap(Figure 2 below). The highest correlation among the feature values is 0.85, and the highest negative correlation is 0.73, indicating no strong correlation among the features. Therefore, the features can be used as a training set for predicting the thermal diffusivity of polymer monomers.

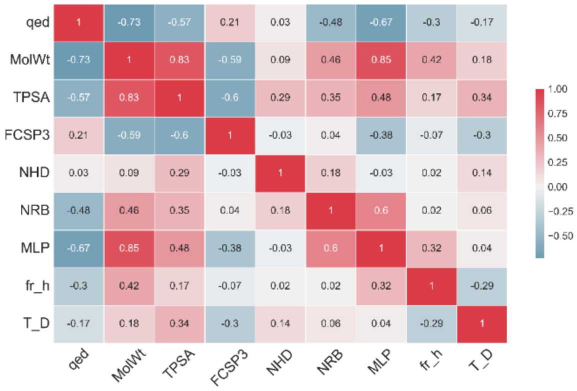

Figure 2. Heat map of correlation between individual features and target quantities

Selecting appropriate data from the public dataset for machine learning modeling is conducive to improving model training speed with minimal impact on prediction accuracy. We randomly selected 400 out of 1077 data for the training set and used grid search to generate random forest models with node layers ranging from 3 to 10 and the number of nodes per layer ranging from 100 to 700. The best-performing model was selected based on the MAE of the linear regression of the real data and predicted data. Then, 200 data were randomly selected from the remaining data, and the accuracy of the model prediction was verified using cross-validation. The R-squared of the test set is about 0.90, indicating good results. The test chart(Figure 3) is shown below.

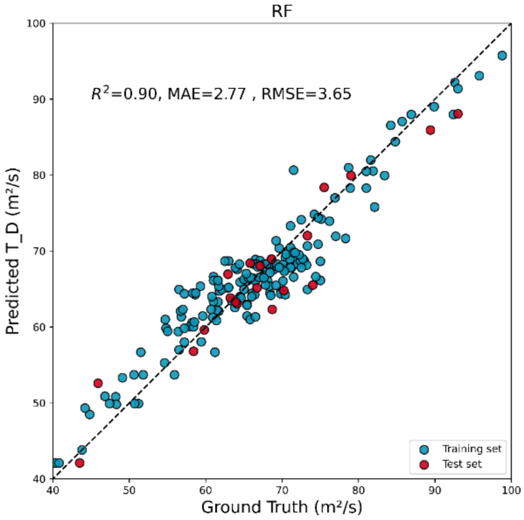

Figure 3. Cross-validation of Random Forest Model Performance for Predicting Polymer Thermal Diffusivity

Analyzing the importance of each feature in the prediction model, we found that the molecular weight has the highest influence on the thermal diffusivity of polymers with a feature importance of 0.266. The polar surface area of the molecule and the number of rotatable bonds contribute significantly to thermal diffusivity with feature importance of 0.209 and 0.178, respectively. Figure 4 (a) shows a visualization of the importance of each feature in the model.

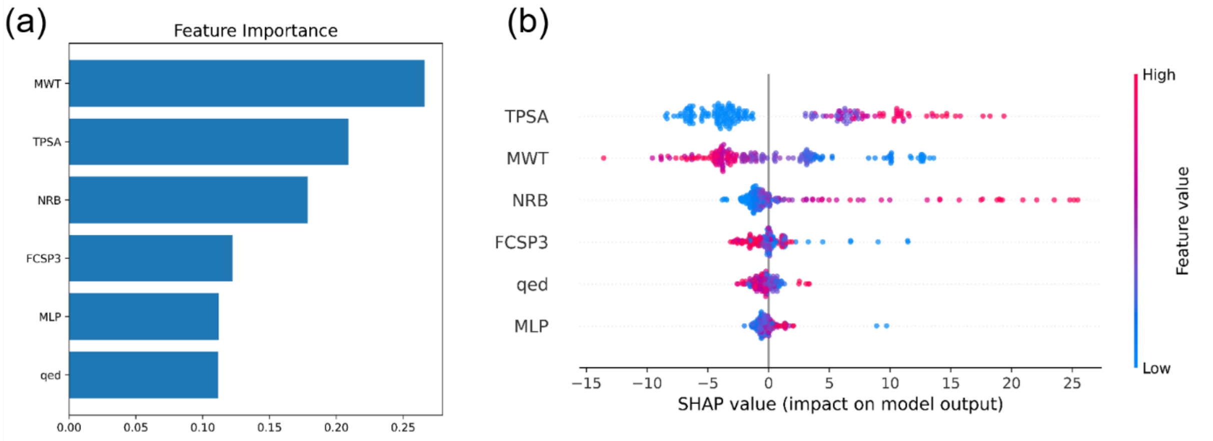

Figure 4. (a) Histogram of the importance of features in the model; (b) The SHAP figure

To reveal the directionality of feature values on the target quantities, we further analyzed the contribution of each feature to the thermal diffusivity of the polymer using SHAP values(Figure 4b). We found that smaller values of the polar surface area of the molecule correspond to lower thermal diffusivity, higher molecular weights result in lower thermal diffusivity, and fewer rotatable bonds lead to lower thermal diffusivity.

These findings can be explained by experimental results and theory. Polymers with lower molecular weight monomers tend to be structurally simpler, allowing greater freedom of molecular motion and more efficient heat transfer by thermal vibration compared to higher molecular weight monomers. Polymers with high molecular weight monomers tend to have more complex chains, which may collide and scatter duringchains' thermal vibration, resulting in lower heat transfer efficiency. TPSA describes the polar characteristics of the molecule and the physical quantity of molecular hydrophilicity. Polymers with higher TPSA values often exhibit hydrogen bonding and dipole interactions, leading to a more ordered structure that facilitates heat transfer in specific directions. Experiments have shown that highly polar polymers (e.g., polyamides, polyesters) have significantly higher thermal diffusivity than low-polarity polymers (e.g., polyolefins). A low number of rotatable bonds may result in a rigid molecular structure, restricting the efficiency of energy transfer within and between molecules, thus resulting in lower thermal diffusivity.

4. Conclusion

In this paper, we have used a machine learning approach to predict the heat diffusion coefficients of polymers, providing a new methodology for screening and designing polymer materials with specific heat diffusivities in both experimental and theoretical approaches. We constructed our dataset using publicly available datasets by encoding the SMILES in them as eight features with actual physical meaning. Subsequently, we used a random forest model with 400 data as a training set and cross-validated the data with 200 randomly selected data, achieving a test set accuracy of 0.9. We then performed an interpretability analysis of this highly accurate predictive model and found that three features, the relative molecular weight of the polymer monomers, the polar surface area of the molecules, and the number of rotatable bonds, have a large degree of influence on the heat diffusion coefficients. An increase in the polar surface area and the number of rotatable bonds of a molecule contributes positively to the thermal diffusivity, while an increase in the molecular weight contributes negatively to the thermal diffusivity. Our work opens up a new paradigm in the study of thermal diffusivity of polymers and provides a new methodology, which will help to advance the research in this field.

References

[1] S. M. Haque, J. A. Ardila-Rey, Y. Umar, A. A. Mas' ud, F. Muhammad-Sukki, B. H. Jume, H. Rahman and N. A. Bani, Energies 14 (10), 2758 (2021).
[2] H.-M. Huang, (MDPI, 2020), Vol. 12, pp. 2560.
[3] P. Balakrishnan, M. J. John, L. Pothen, M. Sreekala and S. Thomas, in Advanced composite materials for aerospace engineering (Elsevier, 2016), pp. 365-383.
[4] A. L. Moore and L. Shi, Materials today 17 (4), 163-174 (2014).
[5] M. Barrios and S. Van Sciver, Cryogenics 55, 12-19 (2013).
[6] Y. Zhang and M. Zhao, Energy Storage Materials 57, 346-359 (2023).
[7] L. Zhou, A. M. Yao, Y. Wu, Z. Hu, Y. Huang and Z. Hong, Advanced Theory and Simulations 4 (9), 2100196 (2021).
[8] W. Parker, R. Jenkins, C. Butler and G. Abbott, Journal of applied physics 32 (9), 1679-1684 (1961).
[9] J. D. Hunter, Computing in science & engineering 9 (03), 90-95 (2007).
[10] W. McKinney, presented at the SciPy, 2010 (unpublished).
[11] M. L. Waskom, Journal of Open Source Software 6 (60), 3021 (2021).