Interpretable Machine Learning Strategies for Accurate Prediction of Thermal Conductivity in Polymeric Systems

Chunbo Lin[1, *], Han Zheng[2]

[1.] College of Science, Zhejiang University of Technology, Zhejiang Province, China

[2.] College of Textile Science and Engineering (International Institute of Silk), Zhejiang Sci-Tech University, Zhejiang Province, China

* e-mail: 202103170412@zjut.edu.cn

Abstract. Polymers, integral to advancements in high-tech fields, necessitate the study of their thermal conductivity (TC) to enhance material attributes and energy efficiency. The TC of polymers obtained by molecular dynamics (MD) calculations and experimental measurements is slow, and it is difficult to screen polymers with specific TC in a wide range. Existing machine learning (ML) techniques for determining polymer TC suffer from the problems of too large feature space and cannot guarantee very high accuracy. In this work, we leverage TCs from accessible datasets to decode the Simplified Molecular Input Line Entry System (SMILES) of polymers into ten features of distinct physical significance. A novel evaluation model for polymer TC is formulated, employing four ML strategies. The Gradient Boosting Decision Tree (GBDT)-based model, a focal point of our design, achieved a prediction accuracy of R2=0.88 on a dataset containing 400 polymers. Furthermore, we used an interpretable ML approach to discover the significant contribution of quantitative estimate of drug-likeness and number of rotatable bonds features to TC, and analyzed the physical mechanisms involved. The ML method we developed provides a new idea for physical modeling of polymers, which is expected to be generalized and applied widely in constructing polymers with specific TCs and predicting all other properties of polymers.

Keywords: machine learning, thermal conductivity, polymer, interpretability analysis, the Gradient Boosting Decision Tree.

1. Introduction

Polymers play a crucial role in todays world, finding uses in advanced areas like implantable brain -computer interfaces, electronic chips, and wearable technologies.[1] Polymers characterized by elevated thermal conductivity (TC) are instrumental in augmenting the heat dissipation capacity of devices, thereby mitigating the potential adverse impacts of overheating on device functionality or user comfort[2]. Conversely, polymers exhibiting reduced thermal conductivity harness exceptional thermal insulation attributes, finding extensive utilization in thermal insulation applications, such as within construction sector walls and thermal management systems for electronic devices, aiming to diminish heat loss and enhance energy efficiency. The identification of polymers with specific thermal conductivities represents a noteworthy endeavor.

However, the current dominant approaches for screening polymers with specific thermal conductivities are molecular dynamics (MD) calculations[3] or experimental measurements[4]. Polymeric systems, characterized by their extensive size and substantial atomic count, render the application of MD methodologies for the calculation of TC inefficient. Experimentally ascertaining the TC of polymers necessitates intricate procedures, such as meticulous sample preparation and rigorous regulation of environmental variables, thereby imposing stringent demands on the precision of experimental methodologies.[5] Such methodologies demand substantial temporal investments, extending from days to weeks, to deduce the TC features of complex polymers, often yielding results with error margins that may be deemed unsatisfactory. To efficiently sift through a diverse array of materials for particular thermal conductivities, an urgent requirement emerges for a rapid and precise technique to forecast the TC of polymers.

In recent years, machine learning (ML) has witnessed a significant surge in its application, demonstrating remarkable success in achieving high levels of accuracy in forecasting outcomes such as carbon dioxide emissions[6] and properties of organic solar cells[7]. Previous studies[8] have leveraged a ML paradigm to estimate the TC of materials. They compiled a dataset comprising 469 amorphous polymers, converting the polymer-Simplified Molecular Input Line Entry System (p-SMILES) into 300-dimensional continuous value vectors, which yielded a prediction accuracy with a coefficient of determination (R^2) of 0.828. However, the complexity of the input features and the employment of 300-dimensional vectors, derived through linguistic processing devoid of physical significance, resulted in predictive performance that did not meet expectations.

We decode the p-SMILES of polymers into 10 features imbued with physical significance, thereby shrinking the feature space by a factor of 30 relative to the previous methodology[7]. Our approach entails the construction of a ML model predicated on Gradient Boosting Decision Trees (GBDT) for the estimation of polymers TC, culminating in an enhanced model accuracy with a coefficient of determination (R^2) of 0.93.

Furthermore, the features integrated into our model are intrinsically interpretable, laying the groundwork for interpretable analyses of the predictive framework. We have elucidated a series of characteristic polymer attributes, such as the number of rotatable bonds and the quantitative estimate of drug-likeness, that significantly influence the thermal conductivity of polymers. Furthermore, we have delineated the physical mechanisms underpinning the associations between these attributes and thermal conductivity. The ML methodology delineated in this study introduces a conceptual framework for the modeling of polymers, capable of predicting not merely the TC but encompassing all pertinent properties of the polymer under investigation.

2. Methods

We employed the third-party libraries *RadonPy*[9] and *RDKit*[11], sourced from GitHub, as our datasets. RadonPy comprises the Simplified Molecular Input Line Entry System (SMILES)[12] representations for 1077 polymers alongside TC data computed through MD methodologies. The SMILES notation for polymers typically encapsulates recurring monomeric units, contingent upon the structure and composition of monomers within the polymer. An exemplar is illustrated in Figure 1.

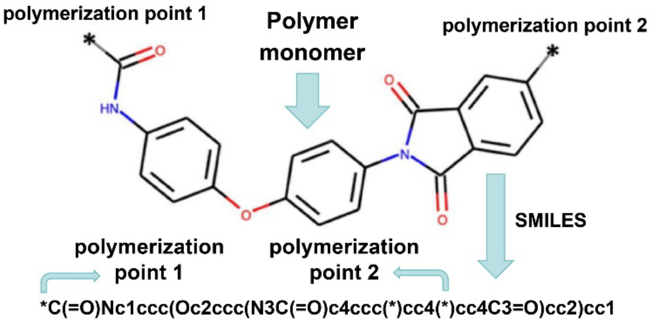

Figure 1. Illustration of a SMILES notation expressed as a string for depicting the molecular architecture of a polymer.

In this study, we opted to train our model using the initial 400 polymers from the dataset. The MolecularDescriptors module within the *RDKit* library facilitated the extraction of characteristic parameters imbued with physical significance for each polymer. These parameters were derived from the decoding of polymer SMILES into 10 eigenvectors (eg., Number of Rotatable Bonds[13]), thereby constituting a 10-dimensional feature space. The determination of TC in polymers conventionally necessitates weeks or even months of experimental measurements or alternatively, days of MD simulation calculations[14]. Consequently, to streamline the screening process for polymers possessing targeted TC, we opted to devise a regression model to delineate the relationship between SMILES representations and TC. This approach enables the accurate prediction of polymer TC within a significantly shorter timeframe.

We utilized *scikit-learn*[15], a *Python* library renowned for its capabilities in ML, to conduct training on Multi-layer Perceptron (MLP)[16], Random Forest (RF)[17], and GBDT[18] models. Additionally, for training eXtreme Gradient Boosting (XGBoost)[19] model, we employed the *Python* library *xgboost*[19]. Following rigorous experimentation, we identified the GBDT model as yielding the most favorable training results among the four models examined. Hyperparameter tuning facilitated the optimization of key parameters, with the Number of trees set to 300, Maximum depth of each tree set to 5, Minimum number of samples required to split a node set to 4, Minimum number of samples required at each leaf node set to 1, Learning rate set to 0.01, and Subsample ratio set to 0.9.

We operate under the assumption that the dataset obtained is accurate and that the TC computed through MD simulation represents the authentic TC of the polymer. Moreover, given that predictions of polymer TC are based solely on monomer information, it is presupposed that elements such as the degree of polymerization, temperature, and the spatial configuration of the polymer are considered ancillary influences on polymer TC.

3. Results and Discussion

In the present investigation, we transmute the SMILES notation into a ten-dimensional attribute sphere. The nomenclature for each feature within this multidimensional expanse, as delineated by the *RDKit* computational library, is cataloged in the inaugural column of Table 1. Additionally, Table 1 elucidates the physical significances and metrications of these features. For succinctness, the abbreviations denoting the physical properties of these polymers, as presented in the secondary column of Table 1, will henceforth represent these features. The selected features encompass: the molecular weights mean value (MWT), the Quantitative Estimate of Drug -likeness (QED), the molecules valence electron count (NVE), the computation of Balabans J metric (BBJ), the molecules total surface area (TPS), the tally of Hydrogen Bond Acceptors (NHA), the count of Rotatable Bonds (NRB), the Wildman-Crippen LogP valuation (MLP), the Wildman-Crippen MR valuation (MMR), and the enumeration of halogen elements (FHA).

Table 1. Ten features from SMILES for training applications

Feature (in RDKit)	Abbreviation	Physical meaning	Unit
MolWt	MWT	The average molecular weight of the molecule	amu
qed	QED	Quantitative Estimate of Drug-likeness	\
NumValenceElectrons	NVE	The number of valence electrons the molecule has	\
BalabanJ	BBJ	Calculate Balabans J value for a molecule	\
TPSA	TPS	The total surface area of a molecule	Å²
NumHAcceptors	NHA	Number of Hydrogen Bond Acceptors	\
NumRotatableBonds	NRB	Number of Rotatable Bonds	\
MolLogP	MLP	Wildman-Crippen LogP value	\
MolMR	MMR	Wildman-Crippen MR value	cm³/mol
fr_halogen	FHA	Number of halogens	\

To evaluate the comprehensive distribution of TC among the polymers encompassed in our dataset, Figure 2a was constructed. This figure, employing a kernel density estimation technique, delineates TC on the x-axis, with values spanning approximately from 0.06 to 0.7 W •m^{-1} •K^{-1}, while the y-axis quantifies the density of the numerical simulation. The prominence of the blue curve at any locus within Figure 2a signifies the aggregation of data points proximal to that specific value of TC. Manifesting a broadly symmetrical bell-shaped curve, and with the dataset affirming conformity to a normal distribution as evidenced by the Shapiro-Wilk test, it is posited that our training specimens are normally distributed.

To mitigate the inclusion of superfluous data ensuing from highly correlated features and to preclude inefficiencies during model training, Figure 2b was devised. This figure illustrates the Pearson correlation coefficients encapsulated within each square, quantifying the interrelation between pairs of features. The intensity of each squares hue signifies the correlation level between the corresponding features, with darker shades indicating higher correlation (dark red for positive, dark blue for negative) and lighter shades denoting weaker correlation. A pronounced correlation is notably observed between MWT and MMR, as well as NVE and MMR, each registering a coefficient of 0.98. Despite this, a deeper examination reveals that MMR, MWT, and NVE encapsulate distinct physical properties. To ensure no pertinent information is overlooked, we opted to retain both MWT and NVE within our feature set. Our analysis of the ten-dimensional feature space reveals a scarcity of highly correlated features, with the majority displaying negligible correlation. This indicates the selected features possess intrinsic value, underscoring the datasets overall rationality.

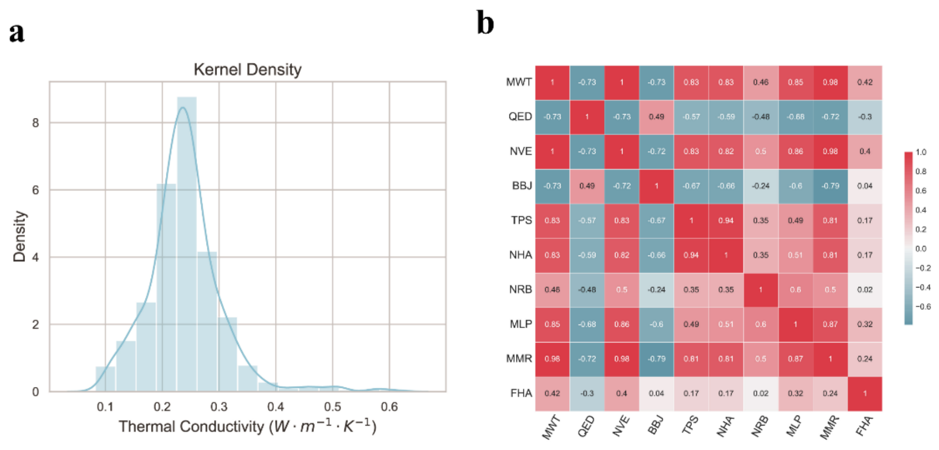

Figure 2. (a) Kernel density distribution of polymer TC within the dataset; (b) Heatmap depicting pearson correlation coefficients among ten distinct features.

We employed pairwise relationship plots to elucidate the bivariate relationships among features, as well as the distribution of individual features within the multivariate dataset, as depicted in Figure 3. For illustrative purposes, only the first 150 data points from the dataset were selected for plotting. The ten plots residing on the diagonal of this scatterplot matrix represent kernel density estimation plots for single features, delineating the distribution of each feature in isolation. Each blue point within Figure 3 symbolizes a polymer sample. Off-diagonal grids showcase small plots that elucidate the relationship between features labeled on the rows and those on the columns. For instance, the second plot in the first row elucidates the relationship between the MWT and QED features. This scatterplot matrix features a scatterplot in its lower left quadrant and a contour plot in the upper right, with contours illustrating the datas sparsity. The dataset contains a mi nimal number of anomalous samples, which were substituted with other normal samples from the dataset. At this juncture, the exploratory data analysis phase preceding ML modeling has been concluded.

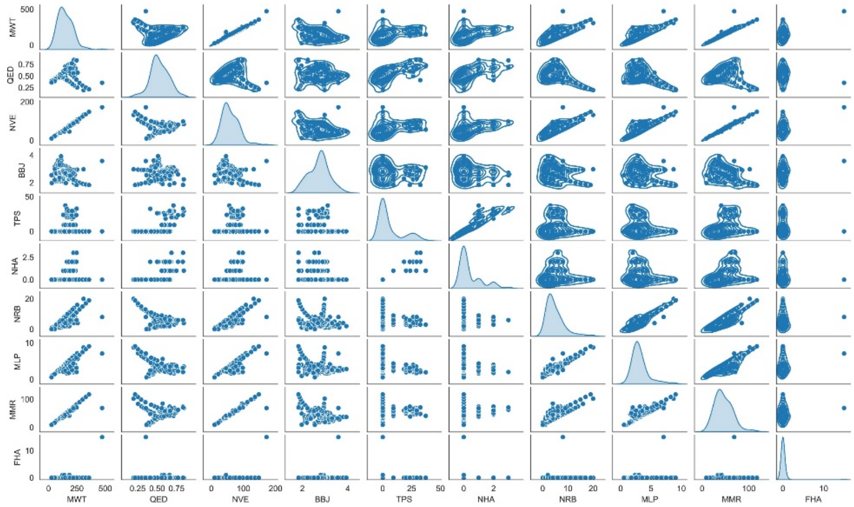

Figure 3. Pairwise correlation matrix with kernel density estimations for selected features.

We first normalized the dataset, which comprises feature spaces and TC of 400 polymers, yielding a novel dataset encompassing 4400 data points. Subsequently, this dataset was subjected to training employing four distinct ML models: MLP, RF, GBDT, and XGBoost. As delineated in the METHODS section, the training outcomes of these models are significantly influenced by their hyperparameters, which consequently affects the accuracy of TC predictions. We engaged in the selection of six hyperparameters for the GBDT model, subjecting it to training across a spectrum of hyperparameter values. These hyperparameters include the number of weak learners (*n_estimators*), the maximum depth of the tree (*max_depth*), the minimum number of samples required to split a node (*min_samples_split*), the minimum number of samples required at a leaf node (*min_samples_leaf*), the learning rate (*learning_rate*), and the subsample ratio of the training instance (*subsample*).

To mitigate variability in training outcomes attributable to differing data partitioning approaches, thereby enhancing the stability and reliability of our estimations, we implemented 10-fold cross-validation during the hyperparameter optimization process. Preliminary training sessions revealed a propensity for overfitting within the model. To augment the models generalization capability, introduce greater stochasticity, diminish its sensitivity to noise in the training data, and thus counteract overfitting, we judiciously decreased the number of trees and the maximum tree depth while increasing the minimum number of samples required for both node splitting and leaf nodes. Throughout the training phase, a grid search was employed to meticulously explore the hyperparameter space for all six hyperparameters, with the objective of identifying the most efficacious hyperparameter combination. For the MLP model, seven hyperparameters were optimized, whereas the RF models optimization involved four hyperparameters, and the XGBoost model was optimized across six hyperparameters. The outcomes of hyperparameter optimization for these models are systematically cataloged in Table 2.

Table 2. Hyperparameter optimization results for different models

MPL	Value	GBDT	Value	XGBoost	Value	RF	Value
hidden_layer_sizes	(50, 50)	n_estimators	300	n_estimators	100	n_estimators	150
alpha	0.005	max_depth	5	max_depth	5	max_depth	6
tol	0.0001	min_samples_split	4	min_child_weight	1	min_samples_split	2
max_iter	300	min_samples_	1	colsample_byt	0.7	min_samples_l	1

		leaf		ree		eaf	
learning_rate_init	0.01	learning_rate	0.01	learning_rate	0.05	\	\
momentum	0.9	subsample	0.9	subsample	0.7	\	\
validation_fraction	0.1	\	\	\	\	\	\

Figure 4 delineates the juxtaposition of the Predicted versus Ground Truth values across four ML models post hyperparameter optimization. The abscissa represents the true TC values of polymers, ascertained via MD simulations, while the ordinate corresponds to the models TC predictions. The dataset was partitioned into a test subset, depicted by orange dots (10%), and a training subset, illustrated by purple dots (90%). Each subplots header enumerates the metrics R², MAE, and RMSE, utilized to evaluate model precision. It was observed that the GBDT model markedly surpassed the MLP in predictive capability and marginally exceeded both the RF and XGBoost models. Specifically, the GBDT model achieved an R² of 0.93 on the training subset and an R² of 0.88 on the validation subset.

Prior research endeavors have similarly employed ML techniques for polymer TC prediction, utilizing a dataset of 469 polymers and decoding SMILES to a 300-dimensional feature space, yielding a prediction accuracy of R²=0.828. Our model demonstrates a 5.2% improvement in prediction accuracy over preceding research, as gauged by the R² metric on a validation set, despite utilizing a dataset of equivalent scale and a feature space reduced by a factor of 30. This achievement aligns with our projected expectations. These findings underscore the viability and promise of leveraging ML methodologies for predicting polymer TC, facilitating the identification of polymers with specific thermal conductivities, and even the discovery and creation of materials tailored for particular thermal applications.

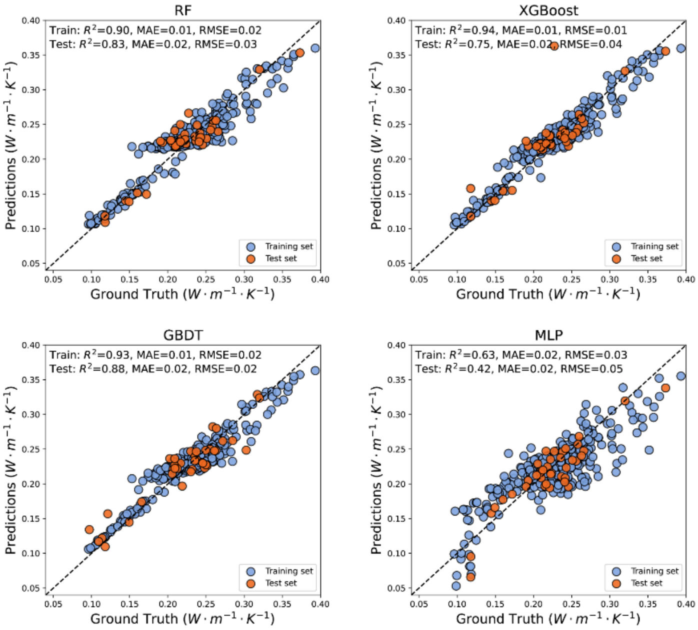

Figure. 4. Comparative pairwise plots of predicted versus ground truth TC, as calculated by MD, across training and test datasets for four models: MLP, RF, GBDT, and XGBoost, with evaluation metrics including R2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

Each feature extracted in our study is imbued with distinct physical significance, i.e., rendering our model inherently interpretable. To delineate the impact of these features on TC and identify those of paramount importance, we employed an interpretable ML framework to generate a representation of feature importance (refer to Figure 5). Employing Lundberg and Lees SHapley Additive ExPlanations (SH AP)[20], a methodology designed to furnish interpretations for individual predictions, allows us to leverage the game-theoretical foundation of Shapley values. Here, features exhibiting substantial absolute Shapley values are deemed crucial. Aiming for a global perspective on feature importance, the figures abscissa represents the mean of absolute Shapley values across features, while the ordinate lists the top 10 features, arranged in descending order of their SHAP importance. Within the context of the GBDT model, the QED emerges as the most influential feature, altering the average predicted absolute probability of TC by 0.5879 (0.5879 on the x-axis). Subsequently, the most significant features include MLP, MMR, NVE, and NRB, which have feature importance of 0.0875,0.0761,0.0712 and 0.0683, respectively. This further elucidates the validity of our method to retain features with distinct physical interpretations, such as NVE, despite their high correlations, in the preliminary phase of data analysis.

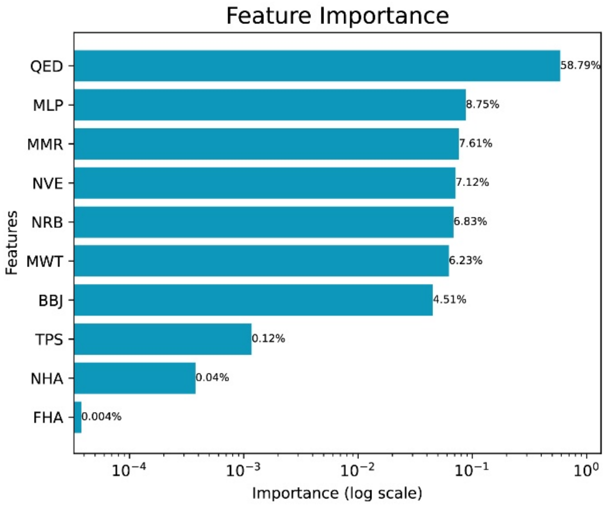

Figure 5. Bar chart of model feature importance contributions on a logarithmic scale. This figure displays the importance of SHAP features in predicting TC, as determined by the trained GBDT through SHAP feature importance analysis.

Upon determining the importance of various features, our investigation expanded to elucidate how these features—particularly the paramount ones—affect the TC. Our work aimed to decipher the physical underpinnings distinguishing the thermal conductivities of polymers characterized by disparate significant features. The SHAP summary plot (Figure 6) excellently encapsulates this objective by integrating feature importance with their effects. Each dot within the summary plot corresponds to a polymer sample associated with a specific feature. The y-axis enumerates the ten distinct features, whereas the x-axis quantifies the Shapley value attributable to a feature for a given sample, with the color gradient from red to blue denoting high to low values of the feature, respectively. Within the ambit of a single feature, dots sharing identical Shapley values converge along the x-axis, and such congruent points exhibit a vertical jitter towards the y-axis. This mechanism facilitates an understanding of the Shapley value distribution for each feature.

Our investigation delves into the influence of key features identified within our research on the TC of polymers, aiming to unearth the physical rationales underpinning these observations. Initially, the manifestation of a positive Shapley value corresponding to a sample exhibiting a low QED value suggests that diminutive values of this particular feature positively impact the models output. Consequently, polymers characterized by elevated QED values are inclined towards lower TC, whereas those with reduced QED values tend to demonstrate enhanced TC. To our knowledge, this constitutes the inaugural correlation of a polymers TC with its QED value, unveiling a potential inverse relationship between the two parameters.

Polymers characterized by elevated QED values frequently exhibit complex molecular architectures, which might encompass multiple ring structures, appendant chains, or functional groups, rendering these polymers molecularly akin to pharmaceutical entities.[21] It is postulated that such intricacy and the extent of branching could attenuate intermolecular forces, thereby diminishing the materials thermal energy conduction efficacy. Furtherm ore, an observed trend indicates that polymers with higher NRB values manifest enhanced TC, suggesting a positive correlation between NRB values and TC. This phenomenon is attributed to the premise that thermal energy transmission in polymers is contingent not solely on intermolecular interactions but also on the molecules translational, vibrational, and rotational degrees of freedom. A substantial NRB may denote increased intramolecular degrees of freedom for absorbing and redistributing thermal energy, alongside augmented molecular flexibility to foster a more ordered structure. Under certain conditions, these molecules possessing greater specific heat capacity and inherent energy have the potential to amplify thermal transport at the macroscopic level.

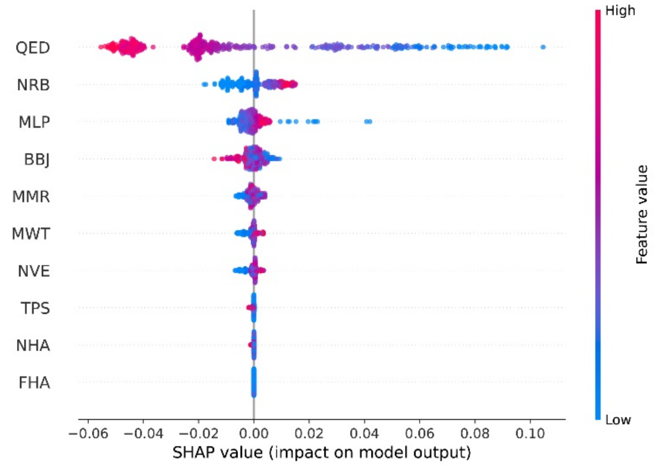

Figure 6. Average impact of model features on predictive outcomes: a SHAP summary visualization.

4. Conclusions

In summary, we have devised a model employing ML techniques that adeptly forecasts the TC of polymers characterized by known SMILES notations. This model leverages data pertaining to the physical features and TCs of 400 polymers in ref. 1 and ref. 2. For the first time, our methodology eschews traditional text-processing tactics in favor of interpreting the SMILES notations of polymers into ten physically significant features, thereby circumventing the generation of a high-dimensional, sparse vector. The model predicts the TC of polymers on the test set with an accuracy of R^2=0.88. Furthermore, through the lens of interpretable analysis, we have unearthed potential inverse relationships between the TC of polymers and their QED, alongside direct correlations with NRB. These correlations are elucidated from a physical standpoint, examining factors such as intermolecular forces and molecular freedom degrees. Our model excels in identifying the traits that predicate a polymer's TC by analyzing its monomeric units, thereby serving as a good pre-screening method in the quest for polymers of specified TCs on a grand scale. In addition, our work may provide some ideas for the design of polymers with specific TC in terms of physical properties. This ML method we designed to study the TC of polymers can also be applied in the study of other properties of polymers, which is highly generalizable and applicable.

References

[1] Wolf P D, Reichert W M. Indwelling neural implants: Strategies for contending with the in vivo environment[M]//Thermal Considerations for the Design of an Implanted Cortical Brain—Machine Interface (BMI). CRC Press, 2008.
[2] Candadai A A, Nadler E J, Burke J S, et al. Thermal and mechanical characterization of high performance polymer fabrics for applications in wearable devices[J]. Scientific Reports, 2021, 11(1): 8705.
[3] Müller-Plathe F. A simple nonequilibrium molecular dynamics method for calculating the thermal conductivity[J]. The Journal of chemical physics, 1997, 106(14): 6082-6085.
[4] Zhao D, Qian X, Gu X, et al. Measurement techniques for thermal conductivity and interfacial thermal conductance of bulk and thin film materials[J]. Journal of Electronic Packaging, 2016, 138(4): 040802.
[5] Henry A. Thermal transport in polymers[J]. Annual review of heat transfer, 2014, 17.
[6] Mardani A, Liao H, Nilashi M, et al. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques[J]. Journal of Cleaner Production, 2020, 275: 122942.
[7] Padula D, Simpson J D, Troisi A. Combining electronic and structural features in machine learning models to predict organic solar cells properties[J]. Materials Horizons, 2019, 6(2): 343-349.
[8] Ma R, Zhang H, Luo T. Exploring high thermal conductivity amorphous polymers using reinforcement learning[J]. ACS Applied Materials & Interfaces, 2022, 14(13): 15587-15598.
[9] Hayashi Y, Shiomi J, Morikawa J, et al. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics[J]. npj Computational Materials, 2022, 8(1): 222.
[10] Kusaba M, Hayashi Y, Liu C, et al. Representation of materials by kernel mean embedding[J]. Physical Review B, 2023, 108(13): 134107.

[11] Landrum G. Rdkit documentation[J]. Release, 2013, 1(1-79): 4.

[12] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. Journal of chemical information and computer sciences, 1988, 28(1): 31-36.

[13] Lipinski C A, Lombardo F, Dominy B W, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings[J]. Advanced drug delivery reviews, 1997, 23(1-3): 3-25.

[14] Ma R, Zhang H, Luo T. Exploring high thermal conductivity amorphous polymers using reinforcement learning[J]. ACS Applied Materials & Interfaces, 2022, 14(13): 15587-15598.

[15] Pedregosa F, Varoquaux G, Gramfort A, et al. *Scikit-learn*: Machine learning in *Python*[J]. the Journal of machine Learning research, 2011, 12: 2825-2830.

[16] Cybenko G. Approximation by superpositions of a sigmoidal function[J]. Mathematics of control, signals and systems, 1989, 2(4): 303-314.

[17] Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[18] Hastie T, Tibshirani R, Friedman J, et al. Boosting and additive trees[J]. The elements of statistical learning: data mining, inference, and prediction, 2009: 337-387.

[19] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[20] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30.

[21] Tian S, Wang J, Li Y, et al. The application of in silico drug-likeness predictions in pharmaceutical research[J]. Advanced drug delivery reviews, 2015, 86: 2-10.