Efficient Prediction of Polymer Glass Transition Temperatures through Machine Learning Methods

Xianghe Meng[1,a]

[1]School of Physics and Technology, Inner Mongolia University, Inner Mongolia 010021, China.

[a]0211120356@mail.imu.edu.cn

Abstract. The glass transition temperature (T_g) plays a crucial role in defining polymer properties. Despite the widespread use of machine learning for material design and property prediction, there are still challenges concerning the interpretability and model performance when predicting T_g. In this study, Simplified Molecular Input Line Entry System strings are utilised to encode the polymer structure, which are then transformed into molecular descriptors for analytical training and prediction of T_g using Artificial Neural Network and Random Forest models. Meticulous hyperparameter tuning of the Random Forest model was performed, resulting in reasonable T_g predictions. This methodology forges a connection between polymer structure and T_g, opening up new avenues for research in the field of polymers.

Keywords: Glass Transition Temperature; Polymer; Machine Learning; Molecular Descriptor.

1. Introduction

Polymer is a kind of compound produced by the polymerization of several identical or different monomer molecules, which is widely used in many fields. In recent years, high performance polymers have found widespread application in specialty coatings, which can impart excellent high temperature resistance, electrical properties, abrasion resistance, non-stick, anti-corrosion and anti-rust properties to the coating. Among the various properties of polymers, the glass transition temperature (T_g) holds particular significance [1]. The T_g refers to the temperature corresponding to the transition from a glassy state to an elastic state occurs. The glass transition is an intrinsic feature of amorphous polymers and represents the macroscopic manifestation of polymer motion. When amorphous polymer nanostructures are cooled below their T_g after nanofabrication, they may be trapped in a solid-like glassy state because of the abrupt loss of local molecular mobility. Below the T_g, structural relaxation may occur in polymer nanostructures, which can significantly affect the performance and lifespan of the polymer devices or temporary templates. As the temperature approaches T_g, structural relaxation accelerates. Therefore, T_g is a crucial parameter for ensuring the long-term stability of polymer nanostructures [2]. The determination of the T_g holds great significance for exploring polymer material properties making it a significant area of research in polymer physics.

With the advancement of T_g research, multiple methodologies are available to test this phenomenon. Based on the alterations in physical properties at the T_g, various procedures have been developed to quantify T_g: techniques involving alterations in volume, procedures utilising changes in thermodynamic features, and procedures utilising changes in mechanical properties [3,4,5]. However, experimental measurement of T_g poses significant difficulty. For instance, polymer materials may undergo multiple phase transitions during heating, and measuring virgin materials proves even more challenging [6]. Firstly, it is challenging to determine the T_g of materials with minimal thermal effects using Differential Scanning Calorimetry [7,8]. Additionally, the thermal history influences the measurement, and removing it results in higher costs. Similarly, Static Mechanical-Thermo Analysis is also affected by the thermal history of the sample and necessitates specific sample preparation [9]. Therefore, it is crucial to devise a simple, user-friendly, universally applicable, and highly precise for determining the T_g of polymers.

In recent years, the field of artificial intelligence, represented by machine learning, has progressed swiftly, due to the rise in computing power from both CPU and GPU. Machine learning has now become a widely utilised approach for designing materials and predicting properties, resulting in a

significant reduction in experimental cycles and cost. Karl Niendorf and Bart Raeymeakers utilised Artificial Neural Network (ANN) and Random Forest (RF) models to predict the electrical conductivity of composites and achieved more accurate results [10]. Currently, several studies have attempted to use machine learning to predict the T_g of polymers by representing the polymer structure as a list of several elements, which are used as inputs to a machine learning model. However, this approach does not provide a clear understanding of the physical principles, and its performance and model interpretability are limited [11]. In this study, we encode polymer structures as Simplified Molecular Input Line Entry System (SMILES) strings and then generate several molecular descriptors to construct a dataset for machine learning and then use the resulting dataset for machine learning training. We adjust the hyperparameters of the ANN and RF models by employing the grid-search method. The tuned RF model predicts the T_g with the r^2, MAE, and Root mean square error (RMSE) of 0.98, 5.95 K, 7.42K, respectively. Additionally, we employed Shapley Additive exPlanations (SHAP) analysis to examine the impact of eigenquantities on T_g and discovered that T_g is lowered as the polymer's rotatable bonds count rises, the topologically polar surface area decreases, the topological indices become more intricate, and the Wildman-Crippen (MR) values decrease. This analysis contributes to our comprehension of T_g. The method proposed in this paper establishes a direct correlation between the structure and attributes of materials, making a significant pioneering contribution to polymer research, and opening up new avenues to explore in the wider field.

2. Results and discussion

SMILES strings consist of a series of characters comprising atoms, chemical bond, rings, and other molecular structural information. The key benefit of SMILES is its simplicity and legibility, making it perfectly suitable for computer chemistry, molecular processing libraries, and chemical information systems. SMILES enables the storage and transfer of molecular structure data in a concise and manageable format, thereby aiding its utilisation in chemical research and computational chemistry. However, SMILES strings are limited to describing the connectivity of molecules and cannot provide an accurate portrayal of their structure and properties. In contrast, molecular descriptors are capable of capturing a more comprehensive and precise picture of a molecule's features. This feature supports the enhancement of machine learning models' accuracy. Therefore, this paper utilises machine learning to predict the T_g of polymers by converting SMILES strings into molecular descriptors.

Molecular descriptors play a significant role in presenting molecular features for cheminformatics. These descriptors are the outcome of a mathematical and logical process that converts chemical information encoded in symbolic representation of a molecule into meaningful numerical data or standardized experimental outcomes [12]. This paper filters seven molecular descriptors from the RDKit library's 207 descriptors based on their likelihood of association with the T_g [13]. The selected descriptors encompass the molecule's average molecular weight (MWT), the number of valence electrons (VE), the Balaban's J value (BJ), the molecule's topological polar surface area (TPSA), the number of rotatable bonds (RB), the Wildman-Crippen LogP (MLP), and Wildman-Crippen (MR) values.

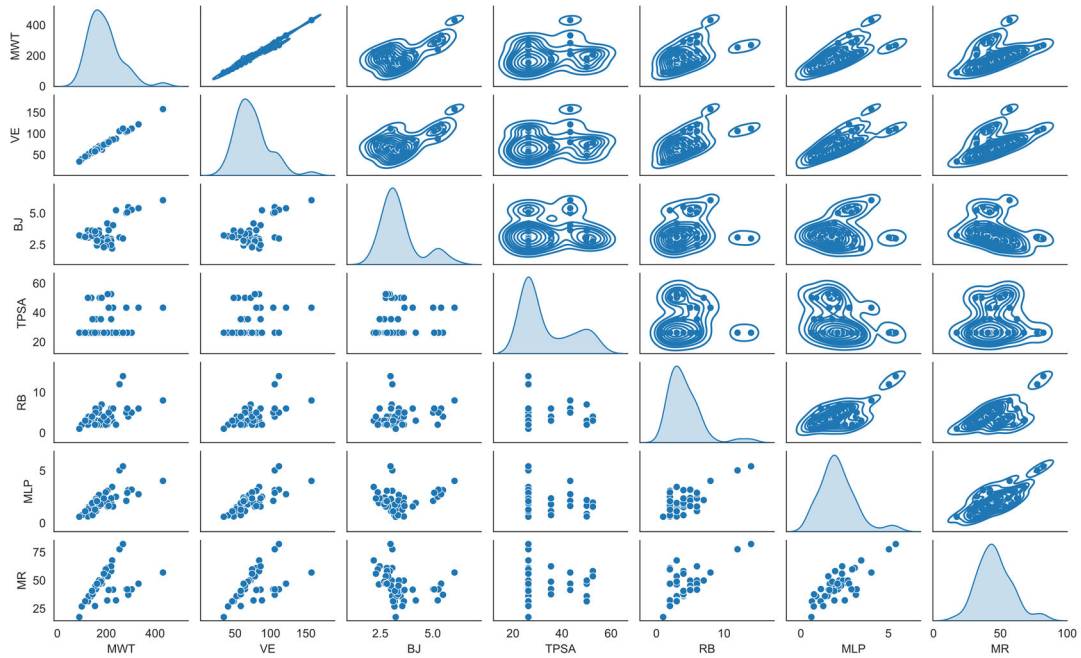

Fig.1 | Interrelationships between features in the data set.

Plots with identical horizontal and vertical coordinates on the diagonal indicate their respective distributions. As depicted in Fig.1, the plotted graph of MWT on the horizontal axis against VE on the vertical axis demonstrates a linear relationship accompanied by a high correlation. Conversely, a low correlation between MWT and BJ is evident from the second plotted graph. Furthermore, the data set appears relatively uniform in its distribution. The diagram in the top right section of this visual demonstrates the addition of contour lines to the lower left section of the locality. This results in the folding of the visual along its diagonal. The density of the data can be observed; denser areas have more sample points and sparser areas have fewer. On the diagonal, the individual data sets approximate a normal distribution. The results presented above indicate the feasibility of the dataset we have created.

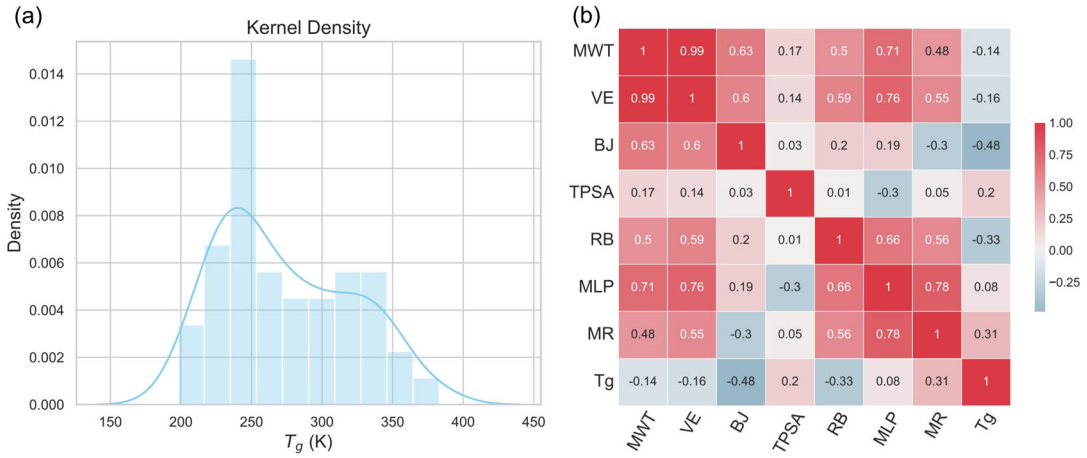

Fig. 2 | Analysis: Interrelationships between features in the data set.

Figure 2a displays the sample density distribution concerning the T_g, and Figure 2b presents the correlation of the seven features and the target quantity. The T_g dataset is primarily focused on the range of 200 K to 340 K, and it follows a normal distribution overall. This suggests that our dataset's creation is sound. As depicted in Fig. 2b, an analysis of feature importance was conducted, whereby a lighter colour signifies a lower correlation between the associated features. The majority of features illustrated in the figure exhibit a lighter hue, indicating a generally low correlation between them. In machine learning, it is paramount to minimize the correlation between individual features. Two features with high correlation are essentially one valid feature, and redundant information can hinder

the efficiency of machine learning. To produce reliable results following training, it is essential to have minimal correlation between each feature in the machine learning method.

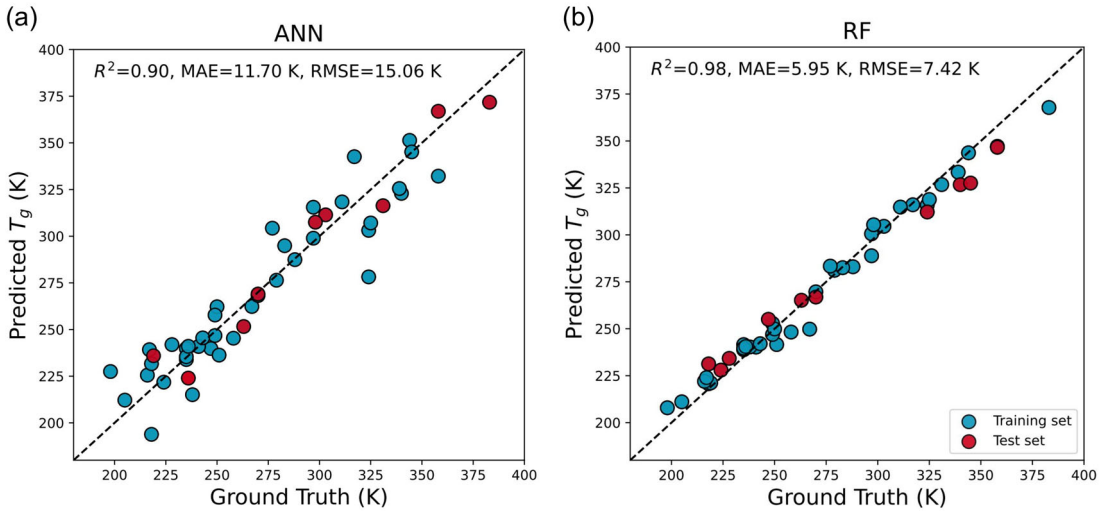

Fig. 3 | Training and prediction results for 96 sample points.

The training set is represented by the blue points, while the red points represent the test set. The left figure displays the training outcome of ANN, while the one to its right shows the training outcome of RF. The vertical coordinate represents the predicted value, while the horizontal coordinate indicates the actual one. A closer proximity to the diagonal line indicates better predictive capability. Based on an analysis of the points' distribution within these two graphs and various indicators, we conclude that the training effect of RF is superior to that of ANN, reaching 0.98. Based on our findings, it is likely that the suggested model efficiently and accurately predicts the T_g of polymers. Moreover, future investigations can be pursued to advance the model's performance by employing further feature parameters and large-scale data sampling. Furthermore, these projections can be utilised in the engineering of polymer substances to gain a more profound understanding and recommendations for the progression and implementation of such materials.

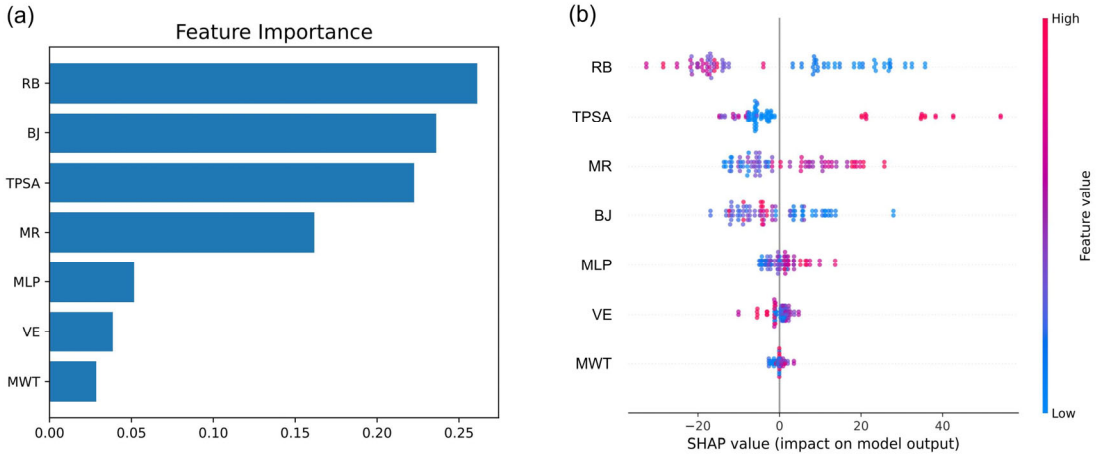

Fig. 4 | Importance of features extracted from RF and SHAP of each feature.

The ranking of each feature quantity's degree of influence on T_g is shown in Fig 4a, where RB, BJ, TPSA, and MR are the key factors that have an impact on T_g. Figure 4b is an analysis of the directionality of the effect of the amount of features on T_g by means of SHAP analysis, sorted from top to bottom by the importance of each feature, with the horizontal axis being the SHAP value of the model, and each point in the graph representing a sample, with the colour of the point representing the magnitude of the feature value, with the closer the red colour indicating the larger the value of the feature, and the closer the blue colour indicating the smaller the value of the feature[14]. A positive

SHAP value signifies a favourable contribution to the model's forecast of the T_g, whereas a negative SHAP value represents an unfavourable contribution to the same forecast. According to the graph, RB exerts the strongest impact on the T_g. As RB increases, the polymer's T_g declines, which implies a negative correlation between RB and the prediction of the T_g. The trend of TPSA is contrary to that of RB. As TPSA rises, the T_g increases, and vice versa.

The results of the above interpretability analyses are consistent with our understanding of physics：a higher number of rotatable bonds implies greater freedom of movement for the polymer chains. This increased freedom allows the chains to overcome the energy barrier associated with the glass transition more easily, leading to a lower T_g. A reduced topological surface polarity area restricts the movement of molecular polymer chains through space, resulting in fewer degrees of freedom. Consequently, the kinetic behaviour of these chains becomes constrained, reducing the rate of molecular movement. When the polymer has a small topologically polar surface area, the motion of its molecular chains decelerates, leading to a decrease in the rate of rearrangement and recombination of the chains, thereby resulting in a reduced the T_g. The greater complexity of the polymer results in heightened molecular interactions that impede the movement of the molecular chains, causing a reduction in their degree of freedom. Furthermore, the intricate structure may pose challenges in modifying the spatial configuration of the polymer molecular chains, leading to an inability to achieve efficient transformation and rearrangement, ultimately resulting in a lower the T_g of the polymer. Higher MR values suggest that the polymer molecule's structure contains stronger molecular polarity or more polar functional groups. These groups or polarity result in improved interactions between molecules, making it harder for the molecular chains to move when in the solid state. Consequently, higher temperatures are necessary to impart enough dynamism in the molecules to transform them into the glassy state. These findings have the potential to enhance our comprehension of the features of polymers.

3. Conclusion

This paper provides an overview of the significance and challenges associated with measuring the T_g of polymers. We then present advancements in predicting T_g through the application of machine learning techniques. Lastly, it details the creation of a dataset for machine learning training by encoding the polymer structure as a SMILES string and transforming it into various molecular descriptors. The ANN and RF models' hyperparameters were optimized through grid search hyperparameter tuning. The RF model achieved a predictive power of 0.98, with a MAE of 5.95K and a RMSE of 7.42K for T_g. This methodology establishes a clear connection between material structure and properties, paving the way for new research avenues in polymer-related fields

4. Methods

This paper employs the RDkit package to convert SMILES encoding into various features. Machine learning modeling and hyperparameter optimization are conducted using the Scikit-learn library. SHAP package is utilized for interpretability analysis [13,14]. All code implementations are based on the Python programming language. Matplotlib library is utilized for data visualization.

References

[1] D. McKechine, C. Jordan, D. Wadkin-Snaith and J. Karen. Glass Transition Temperature of a Polymer Thin Film: Statistical and Fitting Uncertainties. Polymer, 122433 (2020).

[2] H.X. Wang, T.X. Chang, X.H. Li, W.D. Zhang, Z.J. Hu and M.J. Alain. Scaled down glass transition temperature in confined polymer nanofibers. Nanoscale 8, 14950-14955, (2016).

[3] S. Kim, M. Lee, H. Chang Kim, K. YongJoo, L. Won Bo and W. You-Yeon. Determination of Glass Transition Temperatures in Bulk and Micellar Nanoconfined Polymers Using Fluorescent Molecular Rotors as Probes for Changes in Free Volume. Macromolecules, 3c00968 (2023).

[4] X.D. Xia, L. Jackie, J.J. Zhang and W. George J. Uncovering the Glass-Transition Temperature and Temperature-Dependent Storage Modulus of Graphene-Polymer Nanocomposites through Irreversible Thermodynamic Processes. International Journal of Engineering Science, 103411 (2020).

[5] W. Sun, V.P. Anastasios and K. Thomas. Effect of Thermal Lag on Glass Transition Temperature of Polymers Measured by DMA. International Journal of Adhesion and Adhesives 52, 31-39 (2014).

[6] T. Nguyen, and M. Bavarian. A Machine Learning Framework for Predicting the Glass Transition Temperature of Homopolymers. Industrial & Engineering Chemistry Research 61, 12690-12698 (2022).

[7] O. Moussa, A.P. Vassilopoulos,and T. Keller. Experimental DSC-Based Method to Determine Glass Transition Temperature During Curing of Structural Adhesives. Construction and Building Materials 28, 263-268,(2012).

[8] P. Liu, L. Yu, H.S. Liu, L. Chen and L. Li. Glass Transition Temperature of Starch Studied by a Gigh-Speed DSC. Carbohydrate Polymers 77, 250-253(2009).

[9] P. M. Khandare, J. W. Zondlo, and A. S. Pavlovic. The Measurement of the Glass Transition Temperature of Mesophase Pitches Using a Thermomechanical Device. Carbon,00202-2 (1996).

[10] K. Niendorf, and B. Raeymaekers.Using Supervised Machine Learning Methods to Predict Microfiber Alignment and Electrical Conductivity of Polymer Matrix Composite Materials Fabricated with Ultrasound Directed Self-Assembly and Stereolithography. Computational Materials Science 206, 111233 (2022).

[11] L. A. Miccio, and G. A.Schwartz. From Chemical Structure to Quantitative Polymer Properties Prediction through Convolutional Neural Networks. Polymer 193, 122341 (2020).

[12] H. Moriwaki, Y. S. Tian, N. Kawashita and T. Takagi.Mordred: a Molecular Descriptor Calculator. J Cheminform 10, 4,(2018).

[13] G. Landrum. RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling. Greg Landrum 8, 31 (2013).

[14] M. Vega García and J. L. Aznarte. Shapley Additive Explanations for NO2 Forecasting. Ecological Informatics 56, 101039 (2020).