Your Paper's Title Starts Here: Please Center

Use Helvetica (Arial) 14

FULL First Author[1a] FULL Second Author[2b]
Others[3c]
[1]Full address of first author including country
[2]Full address of second author including country
[3]List all distinct addresses in the same way
[a]email@example.com, [b]email@example.com, [c]email@example.com

Abstract

This document explains and demonstrates how to prepare your camera-ready manuscript for Trans Tech Publications. The text area for your manuscript must be 17 cm wide and 25 cm high (6.7 and 9.8 inches resp.). Do not place any text outside this area. Use good quality white paper of approximately 21 x 29 cm or 8 x 11 inches (please do not change the document setting from A4 to letter). Your manuscript will be reduced by approximately 20% by the publisher. Please keep this in mind when designing your figures and tables etc.

Keywords: Keyword1, Keyword2, Keyword3.

1 Introduction

Glass transition temperature (Tg) is a critical property that determines the behavior of materials as they transition from a rigid glassy state to a more flexible state. Accurate prediction of Tg is essential for a wide range of applications, from material science to atmospheric studies. Despite its importance, the experimental determination of Tg for a vast array of chemical compounds remains limited due to resource-intensive processes. As the temperature surpasses Tg, the material undergoes a significant change in its physical properties. The increased thermal energy allows for enhanced molecular mobility, resulting in the material becoming more flexible and exhibiting viscoelastic properties. Although the material does not flow like a liquid, it loses its brittle nature and adopts a rubbery, elastomeric state. This transition has profound implications across various domains, including the mechanical performance, thermal stability, and processing characteristics of polymers. In physics, understanding Tg is essential for predicting material behavior under different thermal conditions, influencing domains such as polymer physics, materials science, and thermomechanics.

Recent advancements in machine learning provide promising alternatives for Tg prediction. Studies by Alzghoul et al.[1] (2014) and Tao et al.[2] (2019) have demonstrated the potential of support vector machines and random forest models for Tg prediction, albeit on limited datasets. Armeli et al.[3] (2023) have utilized the Extra trees model for the prediction of glass transition temperatures of the organic compounds.

Armeli et al.[3] leveraged the well-established Boyer-Beaman[4,5] rule, which asserts a proportional relationship between the glass transition temperature (Tg) and the melting temperature (Tm) of organic compounds. This rule, with a Tg/Tm ratio approximating 0.7, has long been considered reliable across various substances. However, a significant challenge remains in accurately determining the melting point temperature of organic compounds. Organic molecules, characterized by their complex, flexible structures and diverse functional groups, present a formidable challenge in modeling due to intricate interactions during the melting process. This complexity often results in thermodynamic calculations that are highly sensitive to slight variations in molecular interactions and entropic contributions.

In our study, we tackled this challenge by introducing the concept of branching within organic compounds, a novel approach that diverges significantly from traditional reliance on melting point temperature and molecular mass. Notably, we have developed a unique algorithm, grounded in the RDKit[6] library but enhanced to account for chiral atoms, which allows for a precise calculation of branching. By focusing on the branching feature alone—without considering the molecular mass—we achieved a surprising outcome: the prediction accuracy for the glass transition temperature reached an impressive 93%. This demonstrates that the branching feature not only compensates for the exclusion of the melting point temperature but also provides a robust alternative that aligns closely with the empirical observations suggested by the Boyer-Beaman[4,5] rule.

Our results indicate that by prioritizing branching, we have successfully circumvented the traditional difficulties associated with predicting the melting point temperature. This breakthrough not only validates our approach but also suggests a new direction for future research in the field. The substantial performance of our model, even in the absence of melting point data, highlights the potential of branching as a critical determinant in the thermal properties of organic compounds, offering a novel contribution to the existing body of knowledge. This advancement, when compared with the work of Armeli et al[3]., underscores the efficacy of branching as a predictor and marks a significant step forward in the prediction of glass transition temperatures. Shiraiwa et al[9]. established the O: C ratio in their equation of glass transition temperature(T_g). However, our research demonstrates that other molecular ratios specifically C: OH, DBE: C, and CH are equally influential in predicting T_g. Through a comprehensive comparative analysis, we found that these additional ratios provide comparable predictions, challenging the traditional reliance on the O: C ratio alone. The results, illustrated in Figure 4, show that results were comparable after considering other ratios.

2 Materials and Methods

The dataset utilized in this study is sourced from the Bielefeld Molecular Organic Glasses (BIMOG) database[8], which is a comprehensive collection of experimental data on glass transition temperatures. The BIMOG database[8] was esestablished to support research in material science by providing reliable and accessible data on the glass transition temperatures of various organic compounds. The dataset includes a wide range of chemical compounds, each with a detailed molecular descriptors that are crucial for accurate Tg prediction.

The methodologies involved the application of Machine Learning algorithms applied to the BIMOG dataset[8] with an additional feature of branching derived from the SMILE descriptor value using the newly developed algorithm to find the branching which even considers the chiral atoms present in the compound. The Machine Learning Algorithms applied where the functional groups along with the addition of branching feature were considered for training the model, where the best-suited algorithms are Linear Regression, Random Forest, Gradient Boosting, XGBoost, and Extra Trees, to predict Tg values after dividing the dataset into training set which constitutes 90% and testing dataset which constitutes 10% of the dataset.

For the evaluation of the model, we have utilized the coefficient of determination, commonly known as the R^2 score, which is a statistical measure that explains the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of goodness-of-fit and typically ranges from 0 to 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values. The R^2 score can be interpreted as the percentage of the response variable variation that is explained by the model. An R^2 score of 1 indicates that the model perfectly explains the variability of the response data, whereas an R^2 score of 0 indicates that the model does not explain any of the variability in the response data. Higher R^2 values indicate better model performance.

2.1 Calculation of Branching using SMILE

We constructed a graph from the SMILES[10] formula, representing atoms as nodes and bonds as edges. Inspired by RDKit[6] and PySMILES[7], we implemented a method to calculate the number of branches in a molecule. This involves identifying the atom corresponding to a node in the graph and counting the number of edges connected to it, with each edge representing a bond with another atom. A branching atom is typically defined as an atom with a degree greater than 2. Atoms with a degree of 1 are usually terminal atoms, like hydrogen in simple organic molecules. Atoms in linear chains, like alkane carbons, typically have a degree of 2. Atoms with a degree greater than 2 indicate branching points where multiple chains or groups are connected. The branching count is the total number of atoms with a degree greater than 2.

2.2 Identifications of chiral center in a molecule

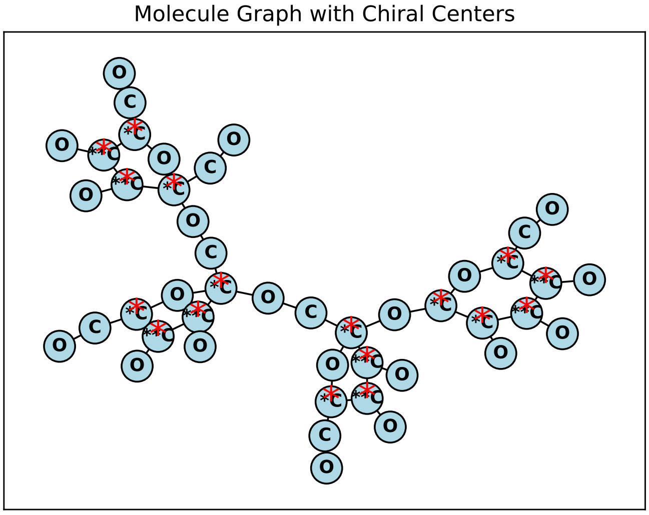


Figure 1: nystose (C24H42O21) molecule representation with chiral centers

The process involves iterating through a SMILES[10] string and identifying different tokens, such as atoms, bond types, branches, and chiral centers. Chiral centers are recognized by detecting the presence of '@' or '@@' in the SMILES[10] string, and these tokens are marked as chiral in the code. The function then checks each node in the molecular graph to see if it represents a chiral center by examining each node's 'element' attribute. If the attribute starts with '@' or '@@', it indicates chirality. For each node identified as a chiral center, the function calculates its position in the graph and places a red asterisk in the graphical representation. Figure 1 shows a graphical representation of Nystose (C24H42O21), with the chiral centers marked by a red asterisk above the corresponding atoms.

2.3 Machine Learning Algorithms Applied

In this study, we employ Random Forest, Gradient Boosting, and XGBoost models to predict the glass transition temperature (Tg). These models are based on ensemble learning techniques that combine multiple models to improve prediction accuracy. Random Forest utilizes the bagging (Bootstrap Aggregating) technique, where multiple decision trees are trained on different random subsets of the data, and their predictions are averaged to reduce variance and prevent overfitting. On the other hand, Gradient Boosting and XGBoost are based on boosting, a technique where models are built sequentially, with each new model focusing on correcting the errors made by the previous ones. Boosting enhances model performance by reducing bias and improving accuracy, making it particularly effective for complex datasets.

2.3.1 Decision Trees based Algorithms

A decision tree is a flowchart-like structure used for decision-making and predictive modeling. It consists of nodes representing decisions or tests on features, branches representing the outcomes of those decisions, and leaf nodes representing final predictions or outcomes. The decision tree algorithm works by selecting the best feature to split the data at each node, based on criteria like Gini impurity or information gain. The data is then split into subsets, and this process is repeated recursively for each subset until a stopping condition is met, such as a maximum tree depth or a minimum number of samples per leaf.

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting. Each tree in the forest is trained on a random subset of the data with replacement, known as bootstrap sampling, and a random subset of features is considered for each split. After training, the predictions of all the trees are combined, typically by averaging for regression tasks or majority voting for classification tasks, to make the final prediction.

Extra Trees, or Extremely Randomized Trees, is similar to Random Forest but introduces more randomness in the tree-building process. Instead of selecting the best split based on a criterion, Extra Trees randomly select split points for each feature. Additionally, Extra Trees uses the entire dataset to train each tree instead of bootstrap samples. This increased randomness can improve model diversity and reduce overfitting.

2.3.2 Linear Regression

Linear Regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The relationship is modeled as a linear equation where the dependent variable is expressed as a sum of the weighted independent variables plus a bias term. The goal is to find the weights (coefficients) that minimize the difference between the actual and predicted values, typically measured by the sum of squared residuals.

2.3.3 Gradient-Based Ensemble Methods

Gradient Boosting is an ensemble technique that builds models sequentially, where each new model corrects the errors of the previous models. It starts with an initial model, often the mean of the target values for regression tasks. The algorithm then computes the residuals, or errors, of this model and trains a new decision tree to predict these residuals. The predictions from this tree are added to the previous model, weighted by a learning rate, to improve the overall prediction accuracy. This process is repeated for a specified number of iterations.

Extreme Gradient Boosting (XGBoost) is an optimized version of Gradient Boosting that includes several enhancements to improve speed and performance. XGBoost adds regularization terms to the loss function to prevent overfitting, efficiently handles sparse data, and utilizes parallel computing to speed up training. Like Gradient Boosting, XGBoost builds trees sequentially, with each tree improving the residual errors of the previous trees, but with additional optimizations for better performance.

3 Results and Discussions

3.1 Redfining Shiraiwa's equation

The Shiraiwa et al[9] model is an advanced theoretical approach for predicting the glass transition temperature (T_g) of organic compounds, particularly those found in atmospheric aerosols. Developed by Shiraiwa et al.[9], this model integrates molecular properties with environmental parameters to provide a comprehensive understanding of the T_g of organic mixtures. It is based on the molar mass and atomic oxygen-to-carbon (O/C) ratio of organic compounds:

$$T_g = A + B \cdot M + C \cdot M^2 + D \cdot \frac{O}{C} + E \cdot M \cdot \frac{O}{C}$$

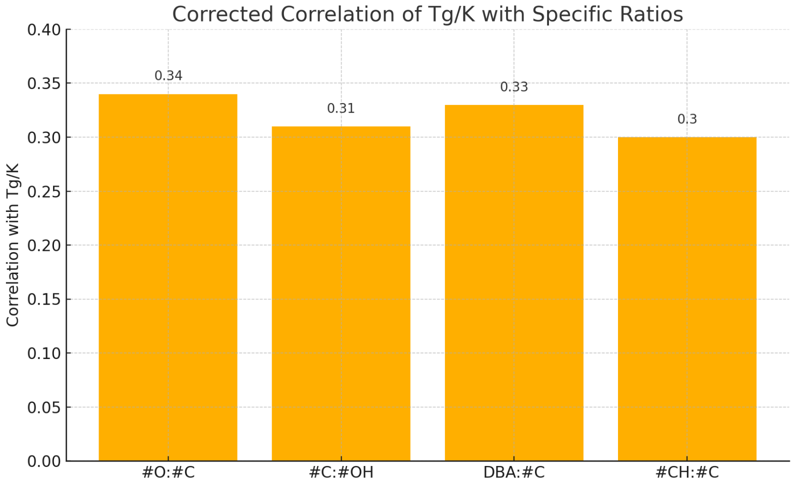
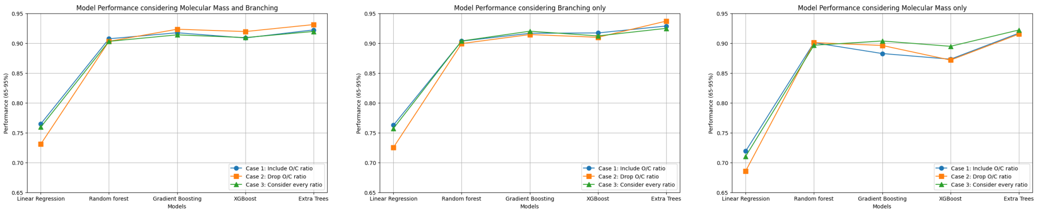


Figure 2: Comparison of Ratios of different entities.

In our preliminary analysis, we initially focused solely on the O: C ratio, aligning with existing literature. However, through iterative experimentation, it became evident that additional atomic ratios, such as DBA: C, C: OH, and CH: C, also played a crucial role in characterizing the organic compounds under investigation. By incorporating these ratios into our model, we observed a significant enhancement in the accuracy and predictive power of our results. Figure 3 illustrates the proximity of these ratios to the O: C ratio, suggesting their collective importance in defining the molecular composition. The comparative analysis presented in Figure 4 underscores the substantial improvement achieved by considering these additional ratios.

3.2 Replacement of Melting point temperature and Molecular Mass with Branching

The strong correlation between mass and temperature is well-established which can be attributed to the Boyer-Beaman[4,5] rule, a well-known empirical rela-



(a) Molecular Mass and Branching (b) Branching only (c) Molecular Mass only

Figure 3: Model Performance Considering Different Features

tionship, that posits a direct correlation between the glass transition temperature (T_g) and the melting point temperature (T_m) of organic compounds, expressed as $T_g/T_m = 0.7$.

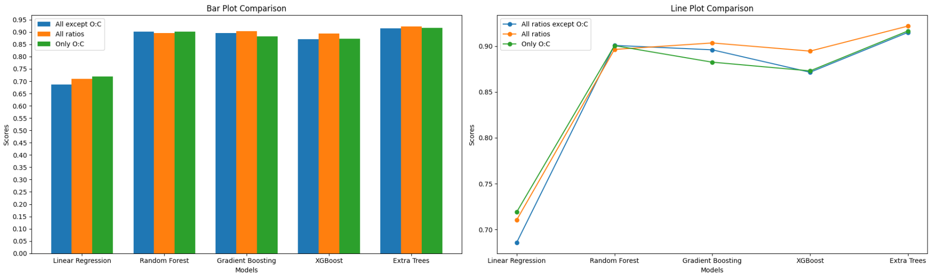


Figure 4: Considering Molecular Mass only (not branching and melting point temperature)

The Boyer-Beaman[4,5] rule, introduced by Robert F. Boyer and Ronald G. Beaman, offers a predictive model for estimating the glass transition temperature (T_g) based on the molecular structure and composition of organic compounds. This rule provides a straightforward method to approximate T_g, a crucial property for understanding the thermal behavior and processing characteristics of polymeric materials. Specifically, it estimates T_g as a proportion of the melting temperature (T_m):

$$T_g = g \cdot T_m$$

where g is a constant, approximately equal to 0.7. In our study, we have developed a methodology to quantify branching and have replaced the traditional dependency of T_g on T_m and Molecular Mass with a model that incorporates branching. This approach has proven to be comparably effective, as demonstrated by the results shown in Figure 5, where substituting the melting temperature with branching yielded results consistent with the original Boyer-Beaman model.

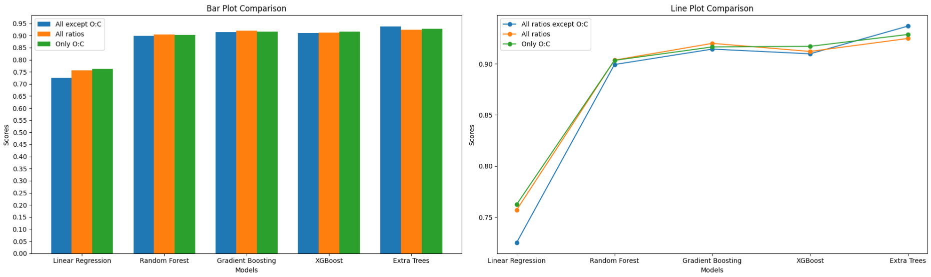


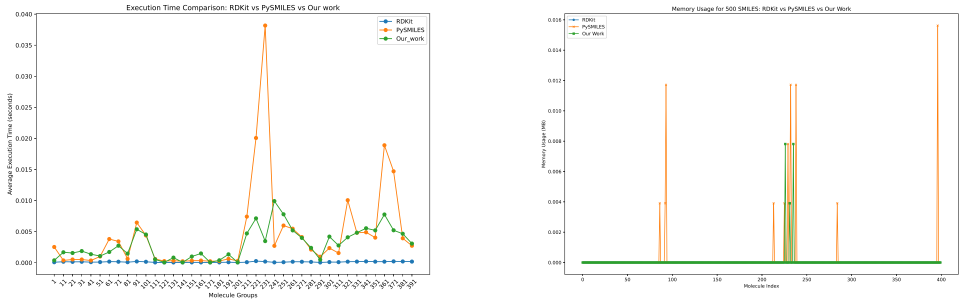
Figure 5: Melting Point Temperature and Molecular Mass replaced with Branching

In Figures 5 and 6, a significant observation is that the Boyer-Beaman[4,5] relationship between melting point temperature and glass transition temperature is not explicitly considered, yet the prediction of glass transition temperature is achieved with an accuracy of 92-93%. Notably, in the analysis, molecular mass—a key factor known for its strong correlation with temperature, evidenced by a correlation coefficient of 0.76 with the glass transition temperature—was also excluded.

Figure 5 presents the accuracies obtained when only molecular mass is considered, without including branching or melting point temperature. This figure highlights the critical role of molecular mass in predicting the glass transition temperature, as it surpasses other factors directly correlated with the glass transition temperature.

Conversely, in Figure 6, both molecular mass and melting point temperature are replaced by branching. Surprisingly, this substitution results in superior performance across all three scenarios: first, when all ratios except O: C are considered; second, when all ratios are considered; and third, when only the O: C ratio is considered. These findings underscore the significance of branching in predicting glass transition temperature, as it not only compensates for but also outperforms the combined effects of melting point temperature and molecular mass in this context.

3.3 Memory Efficiency and Execution Time Analysis



(a) Execution Time Comparison

(b) Memory Usage Comparison

Figure 6: Comparison of Execution Time and Memory Usage

To compare the execution time of the three libraries—RDKit[6], PySMILES[7], and our method—we will evaluate how each converts SMILES[10] formulas into their respective molecular representations. RDKit[6] has a predefined function that transforms SMILES[10] formulas into molecular objects. Similarly, PySMILES[7] includes a function that converts SMILES[10] formulas into a graph data structure using tokenization. Our method also features a function that graphically converts SMILES[10] formulas, similar to PySMILES[7].

To measure execution time, we will iterate through all the SMILES[10] formulas in the dataset, calling the corresponding functions for each library and using the time library to calculate the execution time. For memory usage, we will use the memory profile library to measure how much memory each function consumes during execution.

References

[1] A. Alzghoul, A. Alhalaweh, D. Mahlin, and C.A.S. Bergstrom: Experimental and computational prediction of glass transition temperature of drugs, J. Chem. Inf. Model., Vol. 54, No. 12 (2014), pp. 3396–3403. ACS Publications.

[2] J. Yang, L. Tao, J. He, J.R. McCutcheon, and Y. Li: Machine learning enables interpretable discovery of innovative polymers for gas separation membranes, Sci. Adv., Vol. 8, No. 29 (2022), Article eabn9545. American Association for the Advancement of Science.

[3] G. Armeli, J.-H. Peters, and T. Koop: Machine-learning-based prediction of the glass transition temperature of organic compounds using experimental data, ACS Omega, Vol. 8, No. 13 (2023), pp. 12298–12309. ACS Publications.

[4] R.F. Boyer: Relationship of first-to second-order transition temperatures for crystalline high polymers, J. Appl. Phys., Vol. 25, No. 7 (1954), pp. 825–829. AIP Publishing.

[5] R.G. Beaman: Relation between (apparent) second-order transition temperature and melting point, J. Polym. Sci., Vol. 9, No. 5 (1952), pp. 470–472. Wiley Online Library.

[6] A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij, and A. R. Leach: Open-Source Cheminformatics Software RDKit, Journal of Cheminformatics, Vol. 12 (2020), pp. 1–16. Springer.

[7] PySMILES, GitHub repository. Available: ”https://github.com/pckroon/pysmiles”.

[8] BIMOG Database, Bielefeld University:tgml.chemie.uni-bielefeld.de/BIMOG database.

[9] T. Galeazzo, and M. Shiraiwa: Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings, Environmental Science: Atmospheres, Vol. 2, No. 3 (2022), pp. 362–374. Royal Society of Chemistry.

[10] Simplified Molecular Input Line Entry System, Wikipedia, The Free Encyclopedia. Available: ”https://en.wikipedia.org/wiki/SimplifiedMolecularInputLineEntrySystem”.