Data-Driven Prediction of Glass Transition Temperature Using Molecular Structural Features

Sunny Kaushik[1 [0009−0002−2769−8878]], Rohit Mogli[1 [0009−0002−5239−2985]], Riddhika Mahalanabis[1 [0000−−0002−3586−7881]], and Balakrishnan Ashok[1 [0000−0002−6378−0150]]

Centre for Complex Systems and Soft Matter Physics, International Institute of Information Technology Bangalore (IIITB), 26/C Electronics City Phase 1, Bengaluru - 560100, INDIA.
bashok@iiitb.ac.in ; bashok1@gmail.com

Abstract. The accurate prediction of glass transition temperature (T_g) is crucial for materials design but often relies on melting point dependencies, limiting its applicability in inverse design problems. We present a data-driven approach leveraging machine learning (ML) and symbolic regression to predict T_g based solely on molecular structure. Using the BIMOG dataset, we extract key structural features—including molecular branching, computed from SMILES representations using RDKit and PySMILES, and atomic composition ratios (C, CH, O, etc.)—to enhance predictive accuracy. We apply multiple ML models, including Linear Regression, Random Forest, Gradient Boosting, XGBoost, and Extra Trees, achieving R^2 scores comparable to traditional approaches that depend on melting point data. Finally, we employ genetic programming for symbolic regression to derive an interpretable equation for T_g. Our results demonstrate that incorporating structural descriptors allows for accurate and generalizable T_g prediction without requiring melting point information, making this method well-suited for inverse materials design. This work highlights how computational approaches can improve the tractability of complex materials problems, aligning with the broader goal of integrating physics-based and data-driven methods for materials discovery.

Keywords: Glass Transition Temperature Prediction · Inverse Design of Materials · Machine Learning for Materials Design

1 Introduction

The glass transition temperature (T_g) is a critical property that governs the transition of materials from a rigid, glassy state to a viscoelastic state. This transition influences mechanical performance, thermal stability, and processing characteristics, making accurate T_g prediction essential for material design and engineering applications. However, experimental determination of T_g across diverse chemical compounds is resource-intensive, necessitating computational approaches for reliable prediction [1],[2].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

Previous studies by Alzghoul et al.[3] and Tao et al.[4] have demonstrated the potential of machine learning (ML) models such as Support Vector Machines and Random Forests for predicting T_g, for drug molecules and polymers. Tran et al.[5] have used ML and SMILES to predict T_g and other properties of polymers using various physical properties of the macromolecules. More recently, Armeli et al.[6] (2023) utilized the Extra Trees model to predict T_g for organic compounds, using the well-established, empirical Boyer-Beaman rule [7],[8] relating T_g to the melting temperature T_m through $T_g/T_m \approx 0.7$, which has been considered reliable across a large class of substances. However, the reliance of these ML models on such empirical relationships, introduces significant limitations. First, the accuracy of the Boyer-Beaman rule is heavily dependent on precise T_g measurements, which are often inconsistent and impractical for many compounds. For some organic molecules, T_m can reach extremely high values, making its determination infeasible under standard laboratory conditions. Furthermore, organic molecules, characterized by their complex, flexible structures and diverse functional groups, present formidable challenges for modeling. The intricate interactions during the melting process lead to thermodynamic calculations that are highly sensitive to slight variations in molecular interactions and entropic contributions [9]. This complexity underscores the need for alternative, feature-based approaches to enhance predictive reliability and reduce dependence on T_m. Galeazzo et al. [10], on the other hand, demonstrated the importance of molecular descriptors such as molecular mass and atomic composition in determining T_g. Their work highlights the potential of feature engineering to improve T_g predictions by incorporating structural and compositional properties directly, bypassing the reliance on melting temperature.

In this work, we propose a feature-engineered ML framework that eliminates reliance on T_m by leveraging intrinsic molecular properties. We introduce branching as a key structural descriptor, extracted from SMILES representations [12] using RDKit [13], and demonstrate its effectiveness alongside comparative atomic ratios such as carbon to hydroxyl (C:OH), double bond equivalent to carbon (DBE:C), methene to carbon (CH:C), and molecular mass. Our approach improves prediction accuracy and enhances interpretability through genetic programming, deriving an analytical equation for T_g directly from molecular features. By aligning with the broader goal of making complex real-life systems tractable through computational science, our method contributes to the development of data-driven materials design frameworks that integrate ML with structural insights.

In section 2 we describe the methodology for branching calculation. In section 3 we explore the impact of additional molecular ratios; and propose a novel approach to predict T_g independent of T_m by integrating the combined effects of molecular mass and branching. We then present an interpretable equation for T_g derived via symbolic regression.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

2 Methodology

Dataset and Algorithms The dataset used in this study is sourced from the Bielefeld Molecular Organic Glasses (BIMOG) database [11], a comprehensive collection of experimental glass transition temperature data. It includes a diverse range of chemical compounds, each annotated with detailed molecular descriptors essential for accurate T_g prediction. In our methodology, we applied machine learning algorithms to the BIMOG dataset, incorporating additional molecular features. Specifically, we derived a branching feature from the SMILES descriptor [14] and included the number of functional groups as an input variable. The models evaluated includes Linear Regression, Random Forest, Gradient Boosting, XGBoost, and Extra Trees. The dataset was split into a training set (90%) and a testing set (10%) to assess model performance.

To quantify predictive accuracy, we used the R^2 score (the coefficient of determination), to measure how well the model explains the variance in T_g. This is defined as: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where y_i is the actual T_g, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values. An R^2 score of 1 indicates perfect predictive accuracy, while a score of 0 suggests the model provides no explanatory power. Higher R^2 values signify better model performance.

Calculation of Branching Molecular branching is quantified by counting atoms with degree > 2 in the SMILES [12]-based molecular graph, where atoms are nodes and bonds are edges. Atoms with higher degree (e.g., junction carbons) signal branch points, while terminal and linear atoms have degrees 1 and 2, respectively. This is efficiently computed using `RDKit` [13] and `PySMILES` [14]. Chiral centers — bonded to four distinct groups — further add configurational complexity and contribute to the branching factor. Figure 1a shows chiral centers in Nystose ($C_{24}H_{42}O_{21}$); Figure 1b displays the T_g distribution.

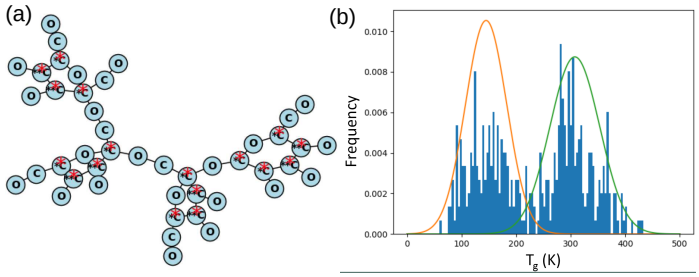

Fig. 1: (a) Nystose ($C_{24}H_{42}O_{21}$) molecule representation with chiral centers (b) Distribution of T_g values in the BIMOG[11] dataset

Deriving an Interpretable Equation for (T_g) Prediction We now outline our approach to deriving a transparent and interpretable equation for predicting T_g, combining exploratory visualization and symbolic regression to refine

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

the model. Our methodology consists of two stages: (1) Visualization-Based Exploratory Analysis to identify key functional relationships. (2) Symbolic Regression with PySR [15] to derive and optimize an explicit mathematical expression.

Stage 1: Visualization-Based Exploratory Analysis In the first stage, we examined the relationships between T_g and various molecular descriptors by plotting T_g as a function of each feature independently. This helped identify possible functional dependencies and trends, providing initial insights into how each descriptor contributes to T_g independent of others.

Stage 2: Symbolic Regression Using PySR While visualization provides valuable insights, it does not fully capture the complex interactions among features. To address this, we employed symbolic regression using PySR [15], which applies evolutionary algorithms to discover interpretable mathematical expressions that best fit the data. The PySR framework operates in two loops. In the inner loop, a population of candidate equations undergoes mutation, crossover, simplification, and optimization, guided by an objective function such as maximizing the R^2 score. The outer loop employs an island-based framework, which evolves equations independently within isolated populations, with periodic migrations fostering diversity and preventing premature convergence. This dual-loop approach balances exploration (diverse equation search) and exploitation (optimizing promising candidates), ultimately yielding a robust and interpretable equation for predicting T_g.

3 Results and Discussions

Incorporating Functional Group Ratios Galeazzo et al. [10] proposed an equation for T_g based on the molar mass and atomic oxygen-to-carbon (O/C) ratio of organic compounds of the following form:

$$T_g = A + B \cdot M + C \cdot M^2 + D \cdot (O/C) + E \cdot M \cdot (O/C) \tag{1}$$

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 DBE: C, C: OH, and CH: C, also played a crucial role in characterizing the organic compounds under investigation, as can be seen in Fig2(a). By incorporating these ratios into our model, we observed some enhancement in the accuracy and predictive power of our results.

The comparative analysis presented in Fig. 2(b), 2(c) and 2(d)underscores the significant contribution of these additional ratios in T_g prediction. Here, the three cases demonstrated are (1) when only the O:C ratio is considered (2) when all ratios except O:C are considered; and (3) when all ratios are considered, respectively.

Importance of Branching as a feature Feature importance analysis (Fig.3) highlights the strong influence of branching on T_g, with a correlation of 0.8. Fig.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

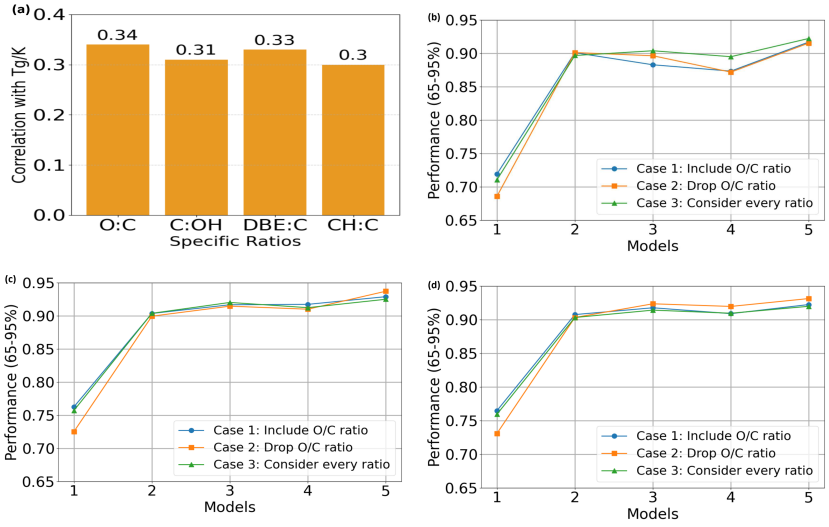

Fig. 2: (a) Correlation with different ratios; (b) Model performance with Molecular Mass; (c) Model performance with Branching; (d) Model performance with both Molecular Mass and Branching. Models are encoded as: 1-Linear Regression, 2-Random Forest, 3-Gradient Boosting, 4-XGBoost, 5-Extra Trees.

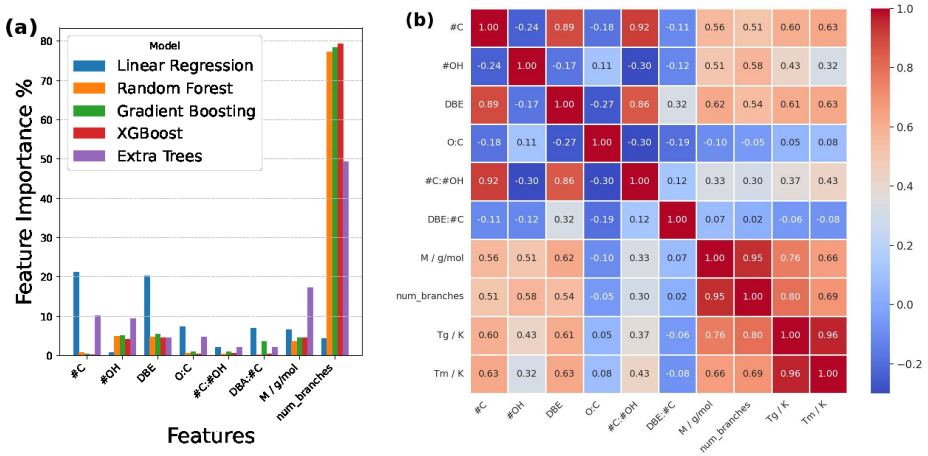

Fig. 3: (a) Feature importance plot (b) Correlation heat map

2(b) shows the model performances considering only molecular mass, which is also a strong predictor of T_g, having a correlation coefficient of 0.76. In Fig. 2(c), we remove the information about molecular mass and consider only the branching number. Interestingly, we observe that T_g is predicted with an accuracy of 92-93% in this case. Thus it can be said that branching is also a strong predictor of T_g. In Fig. 2(d), both molecular mass and branching have been

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

considered together. In this case, even though the overall performance remains almost consistent, the performance is enhanced even when the O/C ratio is not considered.

ML Models	All Ratios Considered			Only O:C Considered			All Ratios Except O:C Considered		
	I	II	III	I	II	III	I	II	III
Random Forest	0.9724	0.9030	0.0694	0.9722	0.9073	0.0649	0.9740	0.9030	0.0710
Gradient Boosting	0.9803	0.9139	0.0664	0.9783	0.9174	0.0609	0.9793	0.9233	0.0560
XGBoost	0.9681	0.9093	0.0588	0.9696	0.9087	0.0609	0.9611	0.9194	0.0417
Extra Trees	0.9735	0.9196	0.0539	0.9748	0.9222	0.0526	0.9741	0.9311	0.0430

Table 1: Performance of ML models in predicting T_g under Different Ratio Considerations. Columns I shows R^2 when T_m is included, column II shows the same when T_m is replaced by branching, thus only containing structural descriptors. Column III shows the difference between the R^2 scores.

In Fig. 2(b) - 2(d), it is significant that even though the Boyer-Beaman 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%. Table 1 shows the R^2 scores for a T_g prediction based on melting point (T_m) and one based on branching, without considering T_m. We see that although, in general, models including T_m are slightly better in performance, branching based models are close enough, with the mean difference in R^2 score ranging from 0.05 to 0.06. This observation is significant because it shows that a T_m dependency is not required for T_g prediction. Dependencies on T_m limit a model's usefulness for inverse materials design, where the goal is to identify new molecular structures with target T_g values. Since T_m is often unknown for novel materials, relying on it introduces an additional prediction or experimental step, increasing uncertainty. Moreover, while T_g and T_m are correlated, their relationship is system-dependent and not universally predictive. A model dependent on T_m does not provide a direct structure-property relationship, making it less effective for guiding molecular modifications to achieve a desired T_g. Recursive dependencies between predicted T_m and T_g can also propagate errors, reducing reliability. By directly predicting T_g from molecular features like branching and molecular composition, our approach overcomes these limitations, making it better suited for inverse design applications.

Equation for T_g predicted using Symbolic Regression The exploratory analysis revealed key relationships affecting T_g. Branch number dependency was polynomial in nature, capturing an upward trend with an accuracy of 82%. Molecular mass showed a logarithmic relationship with T_g, with a 78% accuracy, indicating a slow increase in T_g as molecular mass grew. The atomic ratios ($C : OH$ and $O : C$) were modeled using linear equations, achieving an accuracy of 78%. Starting with these forms after several iterations of evolutionary refinement, the final equation was generated (Equation 2), incorporating both linear and non-linear relationships between the molecular descriptors and T_g, providing a

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

more accurate and interpretable model. We obtain the relation

$$T_g = A + B\ln(M) + CM^2 + D(\text{branching})^{2/3} + EM(O/C) + F(C/OH) + G(DBE/C) + H(CH/C), \quad (2)$$

where DBE stands for double bond equivalent. $\frac{O}{C}$, $\frac{DBE}{C}$, $\frac{C}{OH}$ and $\frac{CH}{C}$ are the comparative ratios for carbon, oxygen, methene and hydroxyl groups. M is molecular mass and branching is the number of branches the molecular structure has. A, B, C, D, E, F, G and H are constants determined for each molecule. The initial equation generated using PySR did not include the logarithmic function, which was derived from the exploratory analysis and that improved the accuracy from 78.2% to 80.32%. The symbolic regression approach confirmed the primary influence of molecular mass and branching on T_g. The ratios were found to have a secondary impact, contributing to the refinement of the model without overshadowing the primary factors.

Limitations The model's accuracy was limited to ∼80% due to the relatively small dataset (400 chemical compunds). Increasing model complexity (e.g., deeper decision trees) improved accuracy to ∼85% but led to overfitting, reducing generalizability. The dataset exhibited a bimodal distribution with peaks at T_g = 144.90 K and 307.47 K (Figure 1b) based on a Gaussian Mixture Model fit. Consequently, predictions in these regions achieved ∼85–90% accuracy. Intermolecular interactions and dynamic properties like relaxation times, fragility etc. that would be related to T_g are also ignored. While Tran et al.[5] have used various descriptors to obtain T_g for polymers, our model uses a smaller descriptor set. Our method would not suffice for polymers where inter-chain interactions, segmental mobility, and molecular weight effects dominate T_g behaviour. The classical Flory-Fox expression [16] that works well for high molecular weight polymers would not work for the small molecular compounds we are considering. The symbolic regression process favors compact, interpretable equations, which may trade off accuracy for simplicity, underfitting complex interactions between features, especially in cases where multiple factors jointly influence Tg.

4 Conclusion

In this work, we have developed a machine learning-driven approach for predicting the glass transition temperature (T_g) of molecular organic glasses using structural features, eliminating the need for melting point dependencies. By extracting branching information from SMILES representations and incorporating atomic composition ratios alongside molecular mass, we achieved improved T_g predictions using multiple machine learning models, including Random Forest, Gradient Boosting, and XGBoost. Our results demonstrated that these structural features provide predictive power comparable to traditional models relying on melting point data. Furthermore, we applied genetic programming for symbolic regression, deriving an interpretable equation for T_g based purely on molecular descriptors. This study highlights the importance of leveraging molecular structure features for T_g prediction, enabling a more direct and reliable

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20

approach to the design of new organic glass materials. By avoiding recursive dependencies on melting point, our method facilitates the discovery of novel molecular organic glasses with tailored thermal properties. Future work could explore extending this approach to larger and more diverse datasets, incorporating additional molecular descriptors, and integrating it with computational materials screening frameworks.

Acknowledgments. Support received from the Machine Intelligence & Robotics Centre (MINRO), IIITB for a 2024 Summer Research Internship for SK, is acknowledged.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Berthier, L., Biroli, G.: Theoretical perspective on the glass transition and amorphous materials. Rev. Mod. Phys. **83** (2), 587–645 (2011)
2. Dudowicz, J., Freed, K.F., Douglas, J.F.: The glass transition temperature of polymer melts. J. Phys. Chem. B **109** 45, 21285–21292 (2005)
3. Alzghoul, A., Alhalaweh, A., Mahlin, D., Bergstrom, C.A.: Experimental and computational prediction of glass transition temperature of drugs. J. Chem. Inf. Model. **54** (12), 3396–3403 (2014)
4. Tao, L., Varshney, V., Li, Y.: Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. J. Chem. Inf. Model. **61**, 5395-5413 (2021)
5. Doan Tran H et al.: Machine-learning predictions of polymer properties with Polymer Genome. J. Appl. Phys. **128**, 171104 (2020)
6. Armeli, G., Peters, J.H., Koop, T.: Machine-learning-based prediction of the glass transition temperature of organic compounds using experimental data. ACS Omega **8** (13), 12298–12309 (2023)
7. Beaman, R.G.: Relation between (apparent) second-order transition temperature and melting point. J. Polym. Sci. **9** (5), 470-472 (1952)
8. Boyer, R.F.: Relationship of first-to second-order transition temperatures for crystalline high polymers. J. Appl. Phys. **25** (7), 825-829 (1954)
9. Preiss, U.P., Beichel, W., Erle, A.M., Paulechka, Y.U., Krossing, I.: Is universal, simple melting point prediction possible? ChemPhysChem **12** (16), 2959–2972 (2011)
10. Galeazzo, T., Shiraiwa, M.: Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings. Environ. Sci.: Atmos. **2** (3), 362-374 (2022)
11. Bielefeld University : Bimog database (2024), http://tgml.chemie.uni-bielefeld.de/BIMOG, accessed: 2024-09-06
12. Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. **28** (1), 31–36 (1988)
13. Bento, A.P.et al: An open source chemical structure curation pipeline using RDKit. J. cheminformatics **12**, 51 (2020)
14. Kroon, P.: Pysmiles: Smiles parsing in python. https://github.com/pckroon/ pysmiles (2024), accessed: 2024-09-06
15. Cranmer, M.: Interpretable machine learning for science with pysr and symbolic regression. jl. arXiv preprint arXiv:2305.01582 (2023)
16. Fox, T.G., Flory, P.J.: Second-order transition temperatures and related properties of polystyrene, J. Appl. Phys., **21** (6): 581–591 (1950)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_20