# Leveraging Data-Driven strategy for Accelerating the Discovery of Polyesters with Targeted Glass Transition Temperatures

Xiaoying He[1], Mengxian Yu[1], Jian-Peng Han[2], Jie Jiang[3], Qingzhu Jia[1], Qiang Wang[1], Zheng-Hong Luo[2], Fangyou Yan[1], and Yin-Ning Zhou[2]

[1]Tianjin University of Science and Technology
[2]Shanghai Jiao Tong University
[3]East China University of Science and Technology

October 29, 2023

**Abstract**

To overcome the limitations of empirical synthesis and expedite the discovery of new polymers, this work aims to develop a data-driven strategy for profoundly aiding in the design and screening of novel polyester materials. Initially, we collected 695 polyesters with their associated glass transition temperatures (Tgs) to develop a quantitative structure-property relationship (QSPR) model. The model underwent rigorous validation (external validation, internal validation, Y-random and application domain analysis) to demonstrate its robust predictive capabilities and high stability. Subsequently, by employing an in-silico retrosynthesis strategy, over 95000 virtual polyesters were designed, largely expanding the available space for polyester materials. External assessments highlight the good extrapolation ability of the QSPR model. Furthermore, we experimentally synthesized diverse virtual polyesters with Tgs covering a sufficient large temperature range. It is believed that this data-driven approach can drive future product development of polymer industry.

## Introduction

Synthetic polymers are indispensable in our daily life.[1, 2] Polyesters, in particular biodegradable polyesters, are widely used in automotive parts, medical apparatus, packaging products, electronic devices, and other fields owing to their good thermomechanical properties and biocompatibility.[3-5] Polyesters are generally consisting of ester containing repeating units produced by esterification reaction between diacids and diols.[6, 7]Thus, combination of different functional diacids with various diols can yield an enormous space of polyester materials. As a result, it becomes a non-trivial task to design and synthesize polyesters with targeted properties.

Glass transition temperature ($T_g$) of polymer governs the dynamic state of polymer chains, and further affects the performance and application domains. For example, the high-$T_g$ polyester with a rigid ring structure improves the thermal stability of polyester materials, aiming to provide bio-based polymers for the plastic consumer market.[8-10] In addition, aliphatic polyesters with low $T_g$ have been studied as environmentally friendly pressure-sensitive adhesives because of their low cost and potential biodegradability.[11, 12] $T_g$ is therefore an essential indicator for determining the properties of polymers.

Given the large polymer design space, it is difficult, time-consuming, and ineffective to screen polymers with targeted properties (e.g., specific $T_g$) through experimental procedures.[13-15] To enable rapid polymer molecular design and high-throughput screening of ideal products prior to laboratory synthesis and analysis, data-driven alternatives,[16-23] such as the quantitative structure-property relationship (QSPR) modeling[24-28]and machine learning (ML) approaches[29-33] have been successfully used to predict the properties for diverse polymers.

In this regard, Wang et al. used successfully trained machine learning to predict the gas permeability of more than 11,000 homopolymers and found that the upper bound of $CO_2/CH_4$separation was exceeded by synthesizing two promising polymeric membranes.[34] Recently, Tao et al. first studied the performance of 79 different models by combining polymer representation, feature engineering and ML algorithms.[35, 36] They then designed millions of hypothetical polyimides by polycondensation of existing dianhydrides and diamines or diisocyanates, and built an ML model to predict a diversity of their properties and verified the predictive ability of the ML through molecular dynamics simulations. Finally, a new polyimide with excellent thermal stability was successfully synthesized experimentally. By trained machine learning models, Wang and Jiang have screened nearly 30,000 hypothetical polymers with fractional free volume (FFV) $> 0.2$, enabling the design of high FFV polymers. [37] Wang et al. present a method for designing high-temperature polymer dielectrics by combining tailored structural units, and the design method is justified by analyzing ML predictions and experimental results.[38]Chen et al. developed an ML model that accurately predicts different frequency-dependent dielectric constants( ), and subsequently utilized the model to successfully design ten polymers with the desired and$T_g$ for application in the capacitor and microelectronics fields.[39] Meanwhile, Lin et al. proposed using a material genome approach design and screening of new heat-resistant resin materials by define gene and extracting key features of properties.[40-42] This strategy was subsequently used in the design of various high-performance polymers. These advances highlight the innovative potential of data-driven approaches in the design of polymers. However, rational design of polyester materials is still less explored. Therefore, it is promising to use data-driven methods to predict the target properties of polyesters and to design new polyesters to complement the existing library.

Herein, we report a data-driven strategy to enable the evaluation of the relationship between molecular structure and $T_g$of polyesters and further guide the design of novel polyesters with specific $T_g$s. The workflow is illustrated in**Figure 1** . First, an multiple linear regression (MLR)-based QSPR model is developed by employing ring repeating unit (RRU)[43, 44] to uniquely represent polyesters and norm descriptors for feature engineering. The predictability, robustness, and chance correlation of the model are evaluated by internal validation, external validation, and$Y$ -randomized analysis, respectively. We then construct a virtual library by designing over 95000 hypothetical polyesters by in-silico retrosynthesis. Later on, the$T_g$ prediction is performed by using the well-trained QSPR model. Ultimately, several polyesters with specific$T_g$ are synthesized and characterized to validate the data-driven polymer design strategy. This work puts the QSPR modeling approach a further step forward by expanding the application scope from properties prediction to model-based design of polymers.

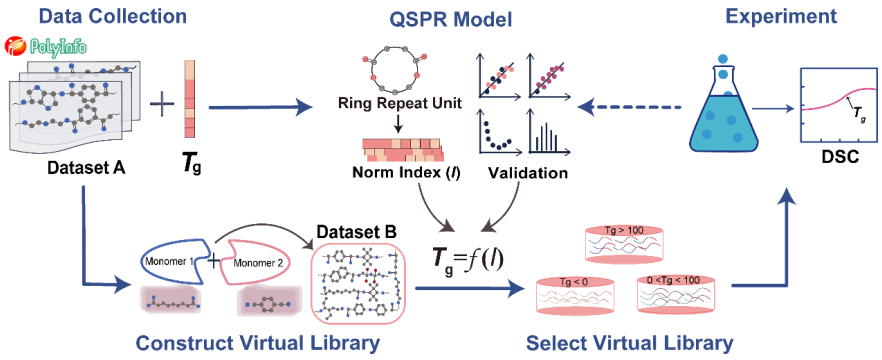


**Figure 1.** The workflow of this work

## Methods

### 2.1 Data Collection and Pretreatment

An available dataset, Dataset A, is established by collecting experimental $T_g$ values of polyesters from the polymer database,*PolyInfo* [45]. The data selection criteria include: number average molar mass ($M_n$ ) >

6000 g mol$^{-1}$, weight average molar mass ($M_w$) > 10000 g mol$^{-1}$, and $T_g$ measured by differential scanning calorimetry (DSC). For polyesters with multiple $T_g$ values from different sources, a median is chosen. The final dataset includes 695 polyesters with $T_g$ values ranging from -78 to 345 (Figure S1). The dataset is randomly divided into a training set and a testing set by the ratio of 4:1.

### 2.2 Modeling process

The diagram of modeling process shown in **Figure 2** includes 4 steps.

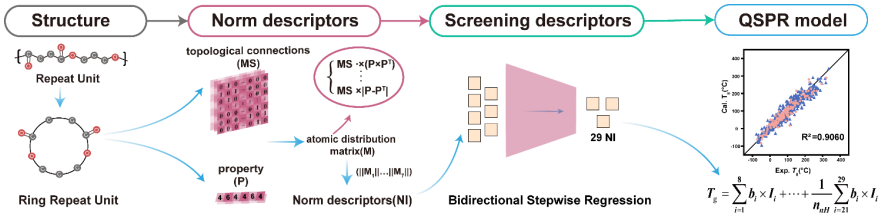


**Figure 2** The pipeline of QSPR model construction

(i) The first step is unique representation of the molecular structure, which is vital for reliable property prediction.[43, 44]Here, the structures of polyesters are approximated by RRU.

(ii) Subsequently, norm descriptors are adopted to characterize the RRU-based polymer structures, which contain details about the properties and topological connections of each atom. Among them, the topological connection relation is represented by step matrix ($MS$), showing the position relationship of each atom in a molecule. In this work, 10 basic step matrices (i.e., $MS_F$, $MS_A$, $MS_B$, $MS_C$, $MS_{AB}$, $MS_{ABC}$, $MS_{bon}$, $MS_{ABC\_aro}$, $MS_{ABC\_cyc}$, and $MS_{bon\_cyc}$) are derived according to the definitions by Eqs (S1)-(S10) in Supporting Information.

The property information of an atom refers to some basic properties of the atom (e.g. ionization energy, and number of outermost electrons), which are expressed in the form of a property matrix ($P$), as shown in Table S1. As such, the atomic distribution matrix ($M$) is generated by combining $MS$ and $P$ according to Eq. (1). Finally, the normal descriptors are obtained by using 7 norm indexes, whose formulas are expressed by Eqs. (S11)-(S17) in Supporting Information.

(iii) The dimensionality of the norm descriptors is then reduced by using bidirectional stepwise regression. Ultimately, 29 norm descriptors are screened out. Lastly, the QSPR model between chemical structures and targeted properties is developed through MLR.

### 2.3 Model validation and evaluation

The statistical parameters, including squared correlation coefficient ($R^2$), average absolute error (AAE), and correlation coefficient for leave-one-out cross validation ($Q^2_{\text{LOO-CV}}$) are employed to evaluate the goodness-of-fit of the model. The definitions of these statistical parameters are given in Eqs (2)-(4). Additionally, the testing set is used in the external validation to assess the external predictive abilities of the model. To evaluate the model's robustness, the internal validation results are offered. Application domain analysis and Y-randomized analysis are performed to confirm the reliability of the model and to exclude chance-related aspects from the modeling process.

Note: $T_{g\ i,exp}$ and $T_{g\ i,cal}$ represent experimental and calculated values of the polyesters, respectively. $n$ is the number of data points.$n_{train}$ is the number of data points in the training set.

### 2.4 Chemical space under exploration

In order to explore the chemical structure space and broaden the existing library of polyester at this stage, a larger polyester library is built by retrosynthesis. **Figure 3a-3c** illustrates the process of building the virtual polyester library. By assuming that the polyesters are formed by the polycondensation of a diol and a diacid,

the collected polyesters are split into secondary building blocks, namely, diols and diacids. Subsequently by de-duplicating the same structure, we obtain a total of 267 diacids and 358 diols, which are listed in Sheet S2 in the Supporting Information (data.xlsx). Ultimately, 95586 hypothetical polyesters are generated and their $T_g$s are predicted by the as-developed QSPR model (**Figure 3d** ), which are described as dataset B and given in Sheet S3 in the Supporting Information(data.xlsx). **Figure 3e** lists the structures of some hypothetical polyesters.

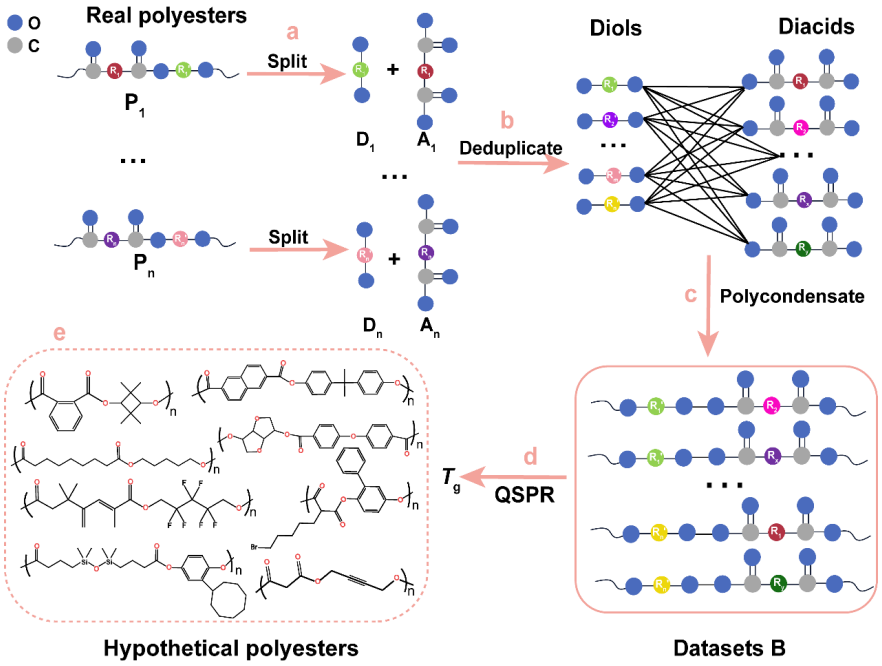


**Figure 3.** Design process of virtual polymers. Colors of $R_i$ represent different substructures. a. Splitting polyesters into diacids and diols, where $D_i$ and $A_i$ represent the diol and the diacid, respectively. b. De-duplicating the molecules with same structure that are generated during splitting. c. Exhaustive combination of diol and diacid into hypothetical polyesters. d. Prediction of $T_g$ for virtual polyesters by using the as-developed QSPR model. e. Examples of hypothetical polyesters.

### 2.5 Synthesis and characterization of polyesters

Synthesis of 10 selected polyesters from dataset B with diverse $T_g$s is carried out via a successive esterification and condensation two-stage polymerization process in a 250 mL three-neck flask equipped with mechanical overhead stirrer, vacuum-tight stirrer bearings and distillation columns. Detailed synthesis procedure for each polyester is shown in supporting information.

The $T_g$ values of final products are measured by differential scanning calorimetry (DSC) analysis. DSC was carried out on a Q2000 (TA Instruments, USA) with a temperature range of -70 to 200 . The data are collected from the second heating thermogram and the heating curves are presented in supporting information.

## 3.Results and Discussion

### 3.1 QSPR model and validation

A QSPR model was developed with 29 NI descriptors to quantify the relationship between the molecular structure characteristics and $T_g$s of 695 polyesters, as shown in Eq. (4). The norm descriptors ($I$ ) and the corresponding coefficients ($b$ ) in the model are listed in Table S2 of the Supplementary Information.

**Figure 4** shows the predictability of the QSPR model. With 556 data points used as the training set and 139 data points used as the testing set, the scatter plot of predicted and experimental $T_g$ values is shown in **Figure 4a** . It is obvious that the majority of the data points are distributed close to the diagonal line, which indicates that the model provides good prediction accuracy, with AAE $< 20$ and $R^2 > 0.90$. The calculated and experimental $T_g$s for the 695 polyesters are shown in the Sheet S1 of the Supporting Information(data.xlsx). Moreover, $R^2_{\text{training}} = 0.9054$ and $R^2_{\text{testing}} = 0.9077$ are significantly greater than 0.6, proving the good predictive performance of the model. Meanwhile, the two $R^2$ are very close, implying that the model has strong generalizability and is capable of well learning the relationship between the chemical structure of polyesters and their associated $T_g$.

n = 695; $R^2 = 0.9060$; $Q^2_{\text{LOO-CV}} = 0.8889$; AAE = 17.7197

where $n_A$ is number of atoms, $n_{nH}$ is number of non-hydrogen atoms; $MS_F$ are calculated with the polyester structures (H-suppressed); $b_i$ is the parameters and $I_i$ is the norm descriptors.

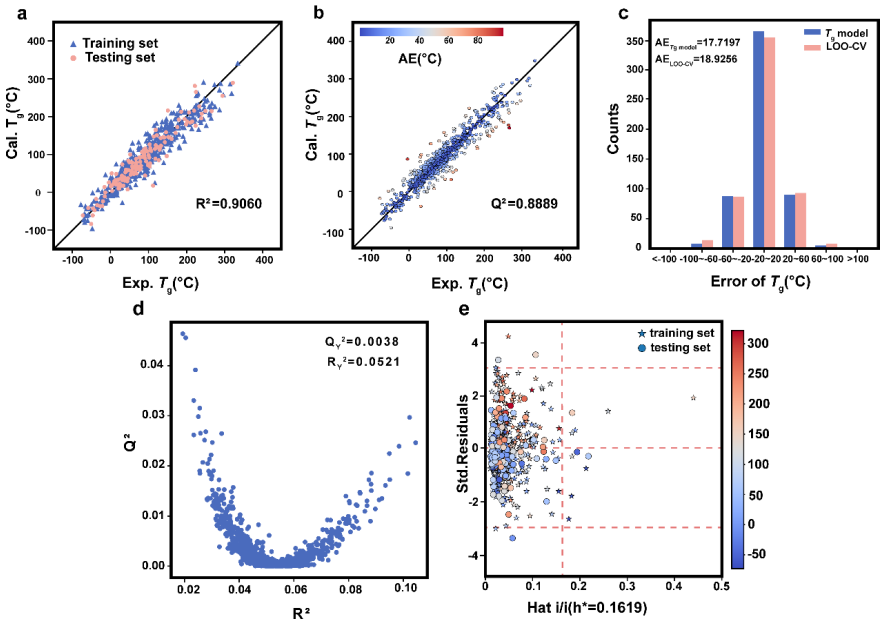


**Figure 4.** Results of the QSPR model. (a) plot of calculated *vs.* experimental $T_g$ of training set and testing set, (b) plot of internal validation via LOO-CV, (c) distribution of errors of the QSPR model and LOO-CV, (d) result of the 10,000 $Y$ -randomization tests and (e) William plot.

To further confirm the robustness of the developed model, the scatter plots of the experimental and LOO-CV estimated values, as well as the error distributions for the LOO-CV and the QSPR model, are shown in **Figure 4b** and **Figure 4c** , respectively. Specifically, $Q^2_{\text{LOO-CV}}$ is 0.8889 and greater than 0.5, showing that the model is robust and stable. Further, the absolute error (AE) distribution of LOO-CV is generally in good agreement with that of the QSPR model, with most polyesters having an error of $T_g$ within 20 .

**Table 1.** Statistical parameters of the QSPR model

| Methods | Variables | Samples | Values |
|---|---|---|---|
| LOO-CV | $Q^2_{\text{LOO-CV}}$ | 695 | 0.8889 |
| | $\text{AAE}_{\text{LOO-CV}}$ | 695 | 18.9256 |
| External validation | $R^2_{\text{training}}$ | 556 | 0.9054 |
| | $R^2_{\text{testing}}$ | 139 | 0.9077 |

| Methods | Variables | Samples | Values |
|---|---|---|---|
| | $AAE_{training}$ | 556 | 17.6530 |
| | $AAE_{testing}$ | 139 | 17.9863 |
| Overall data set | $R^2$ | 695 | 0.9060 |
| | AAE | 695 | 17.7197 |
| $Y$-random validation | | 695 | 0.0521 |
| | | 695 | 0.0038 |

Subsequently, 10,000 times of $Y$ -random validation was performed to assess the chance correlation. Plot of $R_Y^2$ versus $Q_Y^2$ is shown in Figure 3d. The average values of $R_Y^2$ and $Q_Y^2$ for the 10,000 $Y$ -random validation are 0.0521 and 0.0038, respectively, which are much lower than the $R^2$ and $Q^2_{\text{LOO-CV}}$ of the developed model. Therefore, the influence of randomness of the dataset itself on the instability of the QSPR model can be ruled out. William plot was used to visualize the developed model's application domain. Almost all the plots, as depicted in **Figure 4e** , are within the tolerance of three standard deviations of [-3, 3] and critical leverage level ($h^*$=0.1619). It therefore can be concluded that the QSPR model is reliable to predict $T_g$. The values of relevant statistical parameters of QSPR model are shown in **Table 1** .

### 3.2 Virtual Library of Designed Polyesters

Traditional polyesters are mainly made by the condensation of diacids and diols. In silico retrosynthesis route enables the generation of a virtual library of 95,586 polyesters, i.e., dataset B, based on 695 original polyesters. From the computer-aided design point of view, we assumed that once the ester group is correctly present the corresponding polyester is yielded. To illustrate the relationship between datasets A and B more clearly, we then visualized the two datasets separately in two-dimensional chemical space, as shown in **Figure 5a** . It shows that the chemical diversity of the two datasets is quite similar, while compared to dataset A (blue points), dataset B (pink points) clearly covers more possible chemical structures of polyesters. This result implies these virtual polyesters significantly expand the chemical space of the existing polyesters and effectively overcomes the issue of data scarcity. In silico design of these virtual polyesters helps us to explore new polyesters with desired properties from a wider region of chemical space.

However, it is unrealistic to obtain over 95000 polyesters by experimental synthesis. By using the as-developed QSPR model, $T_g$s of the hypothetical polyesters in dataset B were predicted. **Figure 5b** depicts the distributions of $T_g$s in datasets A and B. Similar trend in both datasets demonstrates that the hypothetical polyesters are almost consistent with the polyesters already presented in the database A. Therefore, a trustworthy virtual library for polyester has been successfully established, expanding the existing space for polyester materials and providing certain data support for the synthesis analysis of polyester materials.

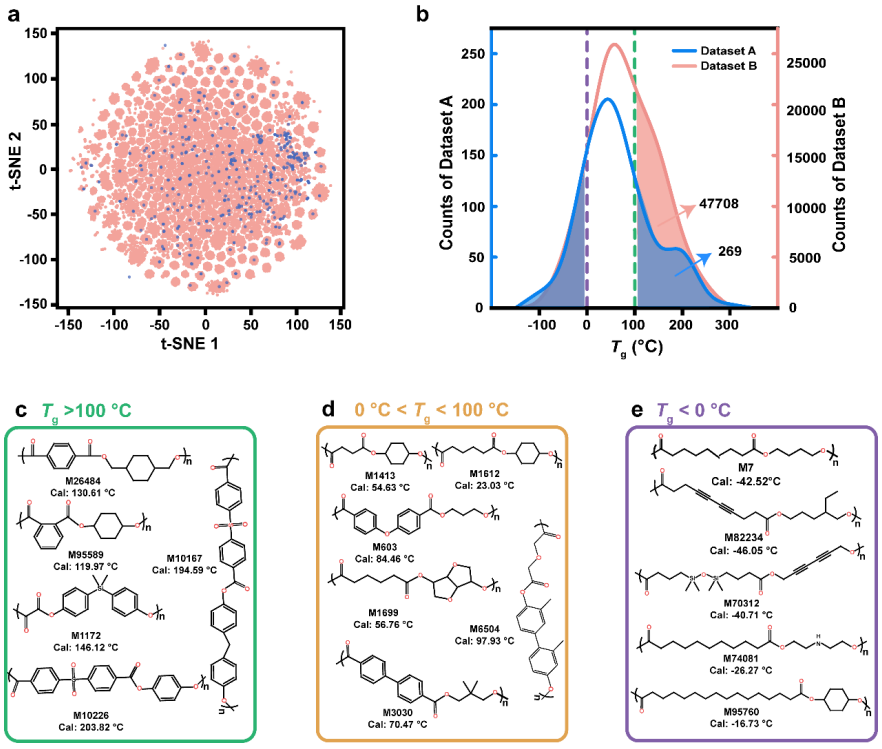


**Figure 5** . (a) Chemical space visualization of the dataset A (blue points) and the dataset B (pink points). (b) $T_g$ distribution of datasets A and B in red and blue, respectively. There are 269 and 47708 polyesters with $T_g$ higher than 100 in datasets A and B, respectively. Samples of polyesters in dataset B with different $T_g$ range show in (c) $T_g > 100$ , (d) $0 < T_g < 100$ and (e) $T_g < 0$ .

Polyesters with high $T_g$ have good heat resistance and thermal stability, thereby offering great potential in high temperature and harsh environments. Typically, we describe polyester materials as high-$T_g$ polyester if the $T_g > 100$ . As shown in **Figure 5b** , in dataset A, there are only 269 polyesters with $T_g$ higher than 100 , which may also include some polyesters that are not easy to synthesize experimentally. By comparison, there are 47708 virtual polyesters with $T_g$ higher than 100 in dataset B, which means that more than 170 times of potential candidates for high-$T_g$ polyester materials are explored. These screened high-$T_g$ polyesters provide a sound support for further synthesis of high-temperature resistant polyester materials. It can be found that the screened high-$T_g$ candidates all have ring groups, as shown in **Figure 5c** , such as benzene and alicyclic rings, and the presence of these ring groups increases the rigidity of the polyester chains, which leads to high $T_g$ values. Additionally, by screening the candidate polyesters with $0 < T_g < 100$ (**Figure 5d** ), compared with the candiates with $T_g > 100$,these have more aliphatic carbon chain and fewer rigid structures. And for $T_g < 0$ (**Figure 5e** ), almost all the candidates have longer aliphatic carbon chain structures. Lower $T_g$ is ascribed to the longer chains polyester molecules more flexible and facilitating an easier inter-chain segment movement. As shown in **Figure 5e** , the $T_g$ of M7 is lower than that of M95760, which is attributed to all aliphatic chain structure of M7, while the rigid hexatomic ring in M95760 trade-offs the chain flexibility.

### 3.3 Experimental validation

Lastly, 10 polyesters with different predicted $T_g$s were selected from the dataset B for experimental synthesis and characterization to verify the design rationality and further validate the predictability of the model. The DSC curves for determining $T_g$s are provided in the supporting information (Figure S2 with data summarized in Table S4).**Figure 6** visualizes that the experimentally determined $T_g$s are in good agreement with those

predicted ones. The AAE of these screened polyesters is 17.4045 , which is smaller than the model's AAE of 17.7197 . Also, the effect of various groups on $T_g$ were then compared. For instance, cyclic alkanes in diols contribute more to $T_g$than linear alkanes do for M4584, M4594, and M4642. For long chain acids, the $T_g$s of the polyesters decrease as the length of the chain increases (M1413, M1612 and M20). This is mostly caused by the increased chain flexibility by contrast to that of the cycloalkanes, which gives the polyesters lower $T_g$s. Additionally, it is found that para-phthalic acids (e.g., M436 and M26484) are more beneficial for improving $T_g$ when compared to ortho-phthalic acids.

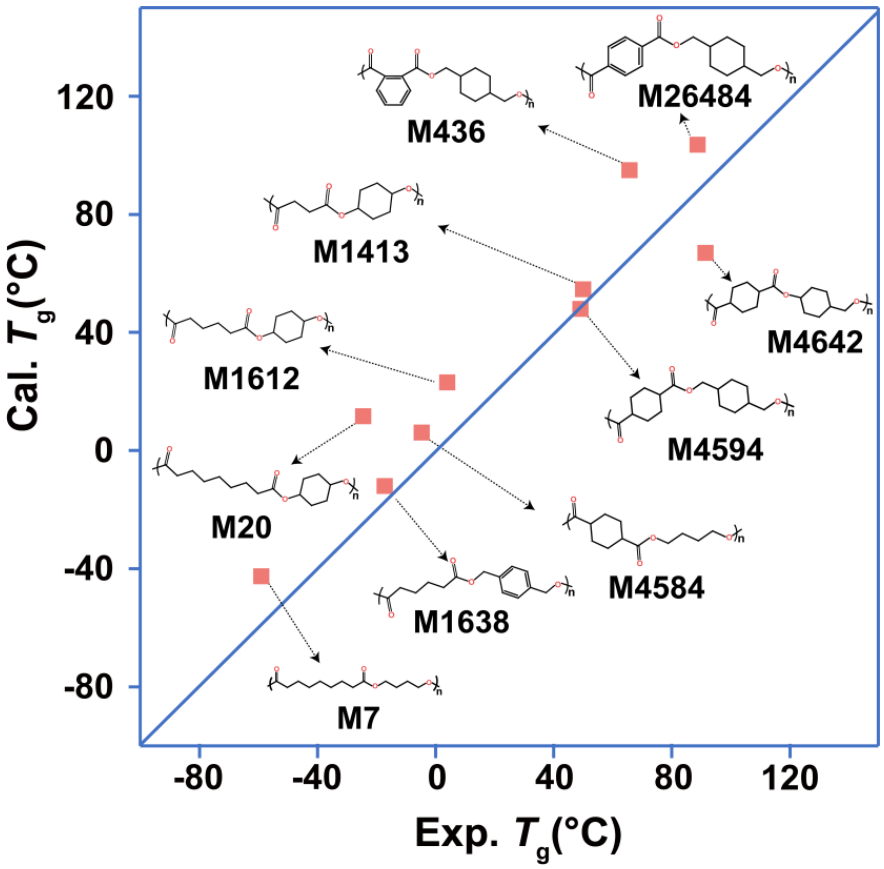


**Figure 6** . Comparison of $T_g$ of the selected 10 polymers between the experimental method and the theoretical calculation method.

## Conclusions

In this contribution, we have successfully developed the QSPR model for evaluating the chemical structure of 695 polyesters with respect to their $T_g$ through a series of rigorous validations. Specifically, $R^2 > 0.90$ and $Q^2_{\text{LOO-CV}} = 0.88$. Following this, a virtual library of nearly 100,000 polyesters has been built by in silico retrosynthesis, which greatly expands the available space for polyester materials. Their associated $T_g$s were predicted by the developed model as well. t-SNE shows significant chemical overlap with the known polyester database, i.e., dataset A, and the virtual library, i.e., dataset B, which demonstrates the rationality and feasibility of the design process.

Subsequently, 10 designed polyesters with different $T_g$s located in different temperature ranges were screened out for experimental synthesis. Good agreement between the experimental and predicted $T_g$s not only demonstrates the accurate prediction performance of the QSPR model, but also verifies the efficiency of

the design method. The rationality of the relationship between chemical structures and $T_g$s was analyzed accordingly.

The methodology presented and the results gained in this work offer the potential to accelerate the design of high-performance polyesters, and may drive future product development of polymer industry.

**Acknowledgments**

This work was financially supported by the National Natural Science Foundation of China (22222807 and 22278319).

**Supporting Information**

Additional results can be found online in the Supporting Information section.

**Data Availability and Reproducibility Statement**

The numerical data from Figures 3, 4, and 5 are tabulated in the Supplementary Material(data.xlsx) Sheet S2, Sheet S1 and Sheet S3. The code for the proposed model is available from the corresponding author upon reasonable request.

**References**

1. Geyer, R., Production, use, and fate of synthetic polymers, in*Plastic waste and recycling* . 2020, Elsevier. 13-32.

2. Haque, F.M., J.S. Ishibashi, C.A. Lidston, H. Shao, F.S. Bates, A.B. Chang, G.W. Coates, C.J. Cramer, P.J. Dauenhauer, and W.R. Dichtel, Defining the macromolecules of tomorrow through synergistic sustainable polymer research. *Chemical Reviews* . 2022;122(6):6322-6373.

3. Ikada, Y. and H. Tsuji, Biodegradable polyesters for medical and ecological applications. *Macromolecular Rapid Communications* . 2000;21(3):117-132.

4. Williams, C.K., Synthesis of functionalized biodegradable polyesters. *Chemical Society Reviews* . 2007;36(10):1573-1580.

5. Larrañaga, A. and E. Lizundia, A review on the thermomechanical properties and biodegradation behaviour of polyesters.*European Polymer Journal* . 2019;121:109296.

6. Pang, K., R. Kotek, and A. Tonelli, Review of conventional and novel polymerization processes for polyesters. *Progress in Polymer Science* . 2006;31(11):1009-1037.

7. Edlund, U. and A.-C. Albertsson, Polyesters based on diacid monomers. *Advanced Drug Delivery Reviews* . 2003;55(4):585-609.

8. Sanford, M.J., L. Pena Carrodeguas, N.J. Van Zee, A.W. Kleij, and G.W. Coates, Alternating copolymerization of propylene oxide and cyclohexene oxide with tricyclic anhydrides: access to partially renewable aliphatic polyesters with high glass transition temperatures. *Macromolecules* . 2016;49(17):6394-6400.

9. Mankar, S.V., M.N. Garcia Gonzalez, N. Warlin, N.G. Valsange, N. Rehnberg, S. Lundmark, P. Jannasch, and B. Zhang, Synthesis, life cycle assessment, and polymerization of a vanillin-based spirocyclic diol toward polyesters with increased glass-transition temperature.*ACS Sustainable Chemistry & Engineering* . 2019;7(23):19090-19103.

10. Pena Carrodeguas, L., C. Martín, and A.W. Kleij, Semiaromatic polyesters derived from renewable terpene oxides with high glass transitions. *Macromolecules* . 2017;50(14):5337-5345.

11. Ozturk, G.I., A.J. Pasquale, and T.E. Long, Melt synthesis and characterization of aliphatic low-Tg polyesters as pressure sensitive adhesives. *The Journal of Adhesion* . 2010;86(4):395-408.

12. Wang, X.-L., L. Chen, J.-N. Wu, T. Fu, and Y.-Z. Wang, Flame-retardant pressure-sensitive adhesives derived from epoxidized soybean oil and phosphorus-containing dicarboxylic acids.*ACS Sustainable Chemistry & Engineering* . 2017;5(4):3353-3361.

13. Zhang, Z., X. Han, W. Gong, K. Huang, J.h. Li, X. Chen, C. Lian, and H. Liu, Design and screening of zwitterionic polymer scaffolds for rapid underwater adhesion and long-term antifouling stability.*AIChE Journal* . 2023;69(8):e18084.

14. Audus, D. and J. De Pablo, Polymer informatics: opportunities and challenges. ACS Macro Letter 2017;6(10):1078-1082.

15. Chen, L., G. Pilania, R. Batra, T.D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports* . 2021;144:100595.

16. Kim, C., A. Chandrasekaran, T.D. Huan, D. Das, and R. Ramprasad, Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C* . 2018;122(31):17575-17585.

17. Mannodi-Kanakkithodi, A., A. Chandrasekaran, C. Kim, T.D. Huan, G. Pilania, V. Botu, and R. Ramprasad, Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond.*Materials Today* . 2018;21(7):785-796.

18. Patra, T.K., Data-driven methods for accelerating polymer design. *ACS Polymers Au* . 2021;2(1):8-26.

19. Andraju, N., G.W. Curtzwiler, Y. Ji, E. Kozliak, and P. Ranganathan, Machine-Learning-Based Predictions of Polymer and Postconsumer Recycled Polymer Properties: A Comprehensive Review. *ACS Applied Materials & Interfaces* . 2022;14(38):42771-42790.

20. Xu, P., H. Chen, M. Li, and W. Lu, New opportunity: machine learning for polymer materials design and discovery. *Advanced Theory and Simulations* . 2022;5(5):2100565.

21. Zhao, Y., R.J. Mulder, S. Houshyar, and T.C. Le, A review on the application of molecular descriptors and machine learning in polymer design. *Polymer Chemistry* . 2023; 14(29):3325-3346.

22. Martin, T.B. and D.J. Audus, Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polymers Au* . 2023; 3(3):239-258.

23. Gao, L., L. Wang, J. Lin, and L. Du, An Intelligent Manufacturing Platform of Polymers: Polymeric Material Genome Engineering.*Engineering* . 2023.

24. Yu, M., Y. Shi, X. Liu, Q. Jia, Q. Wang, Z.-H. Luo, F. Yan, and Y.-N. Zhou, Quantitative structure-property relationship (QSPR) framework assists in rapid mining of highly Thermostable polyimides. *Chemical Engineering Journal* . 2023;465:142768.

25. Schustik, S.A., F. Cravero, I. Ponzoni, and M.F. Diaz, Polymer informatics: Expert-in-the-loop in QSPR modeling of refractive index. *Computational Materials Science* . 2021;194:110460.

26. Wu, J.-Q., X.-Q. Gong, Q. Wang, F. Yan, and J.-J. Li, A QSPR study for predicting $\vartheta$ (LCST) and $\vartheta$ (UCST) in binary polymer solutions.*Chemical Engineering Science* . 2023;267:118326.

27. Khan, P. and K. Roy, QSPR modelling for investigation of different properties of aminoglycoside-derived polymers using 2D descriptors. *SAR and QSAR in Environmental Research* . 2021;32(7):595-614.

28. Rasulev, B., F. Jabeen, S. Stafslien, B.J. Chisholm, J. Bahr, M. Ossowski, and P. Boudjouk, Polymer coating materials and their fouling release activity: A cheminformatics approach to predict properties. *ACS Applied Materials & Interfaces* . 2017;9(2):1781-1792.

29. Pilania, G., C.N. Iverson, T. Lookman, and B.L. Marrone, Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. *Journal of Chemical Information and Modeling* . 2019;59(12):5013-5025.

30. Liang, Z., Z. Li, S. Zhou, Y. Sun, J. Yuan, and C. Zhang, Machine-learning exploration of polymer compatibility. *Cell Reports Physical Science* . 2022;3(6).

31. Alesadi, A., Z. Cao, Z. Li, S. Zhang, H. Zhao, X. Gu, and W. Xia, Machine learning prediction of glass transition temperature of conjugated polymers from chemical structure. *Cell Reports Physical Science* . 2022;3(6).

32. Wu, S., Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, and J. Shiomi, Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Computational Materials* . 2019;5(1):66.

33. Tao, L., G. Chen, and Y. Li, Machine learning discovery of high-temperature polymers. *Patterns* . 2021;2(4):100225.

34. Barnett, J.W., C.R. Bilchak, Y. Wang, B.C. Benicewicz, L.A. Murdock, T. Bereau, and S.K. Kumar, Designing exceptional gas-separation polymer membranes using machine learning. *Science Advances* . 2020;6(20):eaaz4301.

35. Tao, L., V. Varshney, and Y. Li, Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *Journal of Chemical Information and Modeling* . 2021;61(11):5395-5413.

36. Tao, L., J. He, N.E. Munyaneza, V. Varshney, W. Chen, G. Liu, and Y. Li, Discovery of multi-functional polyimides through high-throughput screening using explainable machine learning. *Chemical Engineering Journal* . 2023;465:142949.

37. Wang, M. and J. Jiang, Accelerating Discovery of High Fractional Free Volume Polymers from a Data-Driven Approach. *ACS Applied Materials & Interfaces* . 2022;14(27):31203-31215.

38. Wang, R., Y. Zhu, J. Fu, M. Yang, Z. Ran, J. Li, M. Li, J. Hu, J. He, and Q. Li, Designing tailored combinations of structural units in polymer dielectrics for high-temperature capacitive energy storage. *Nature Communications* . 2023;14(1):2406.

39. Chen, L., C. Kim, R. Batra, J.P. Lightstone, C. Wu, Z. Li, A.A. Deshmukh, Y. Wang, H.D. Tran, and P. Vashishta, Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Computational Materials* . 2020;6(1):61.

40. Zhang, S., S. Du, L. Wang, J. Lin, L. Du, X. Xu, and L. Gao, Design of silicon-containing arylacetylene resins aided by machine learning enhanced materials genome approach. *Chemical Engineering Journal* . 2022;448:137643.

41. Zhu, J., M. Chu, Z. Chen, L. Wang, J. Lin, and L. Du, Rational design of heat-resistant polymers with low curing energies by a materials genome approach. *Chemistry of Materials* . 2020;32(11):4527-4535.

42. Hu, Y., W. Zhao, L. Wang, J. Lin, and L. Du, Machine-learning-assisted design of highly tough thermosetting polymers. *ACS Applied Materials & Interfaces* . 2022;14(49):55004-55016.

43. Yu, M., Y. Shi, Q. Jia, Q. Wang, Z.-H. Luo, F. Yan, and Y.-N. Zhou, Ring Repeating Unit: An Upgraded Structure Representation of Linear Condensation Polymers for Property Prediction. *Journal of Chemical Information and Modeling* . 2023;63(4):1177-1187.

44. Antoniuk, E.R., P. Li, B. Kailkhura, and A.M. Hiszpanski, Representing Polymers as Periodic Graphs with Learned Descriptors for Accurate Polymer Property Predictions. *Journal of Chemical Information and Modeling* . 2022;62(22):5435-5445.

45. Polymer Database (PoLyInfo).*https://polymer.nims.go.jp/*(before 2021).