ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

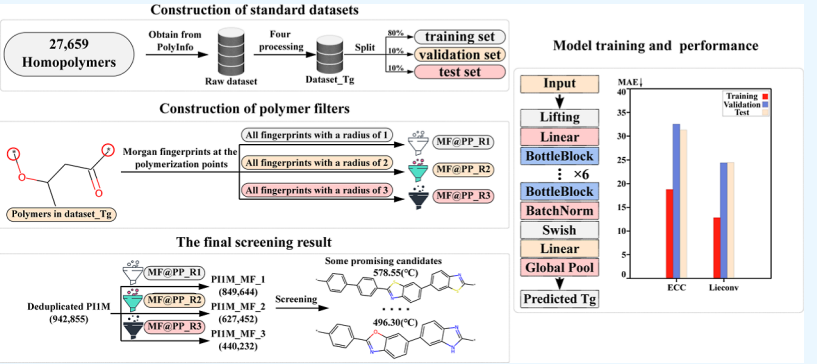
Large-Scale Glass-Transition Temperature Prediction with an Equivariant Neural Network for Screening Polymers

Zheng Long, Hongmei Lu,* and Zhimin Zhang*

Cite This: *ACS Omega* 2024, 9, 5452−5462

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The practically infinite chemical and morphological space of polymers makes them pervasive with applications in materials science but challenges the rational discovery of new materials with favorable properties. Polymer informatics aims to accelerate materials design through property prediction and large-scale virtual screening. In this study, a new method (Lieconv-Tg) has been developed to predict glass-transition temperature (Tg) values from repeating units of polymers based on Lieconv, which is equivariant with transformations from any specified Lie group. The introduction of equivariance allows the prediction of molecular properties from their 3D structures, independent of orientation and position. A total of 27,659 homopolymers with Tg values were collected from PolyInfo, and a standard data set containing 7166 polymers (named data set_Tg) was created for training a robust Lieconv-Tg model. Using the 3D coordinates as input, Lieconv-Tg performs better than Edge-Conditioned Convolution (ECC), and the mean absolute error (MAE) is significantly reduced by ∼6 from ∼30 to ∼24 on both the validation set and the test set, and the R^2 value for both the validation set and the test set can reach 0.90. Lieconv-Tg is thus used to screen promising candidates from a benchmark database named PI1M with 995,800 generated polymers. However, there are some implausible repeating units in PI1M. To get more reasonable candidates from PI1M, a new filtering method has been accomplished by utilizing Morgan fingerprints at the polymerization points (MF@PP) of repeating units in data set_Tg. The combination of a standard data set, Lieconv-Tg, and a more reasonable screening strategy provides new directions in materials design for polymers.

INTRODUCTION

Polymers have become ubiquitous in daily products and advanced technology components.[1−6] This versatility is due to the virtually infinite chemical and morphological space of polymers. A subtle change in the chemical structure of the polymers can significantly influence their properties. Based on intuition gained from previous experiments, the traditional trial-and-error process is difficult to traverse the entire chemical space of polymers and may miss some polymers with favorable properties. Some recent investigations into the rational design of polymers through property prediction and large-scale virtual screening suggest that polymer informatics[7−9] may provide methods for accelerated materials design.

For polymer properties, the glass-transition temperature (Tg) is the temperature at which the transitions occur between the rubbery and glassy states. Many important physical properties show drastic changes at Tg, such as the mechanical modulus, heat capacity, and dielectric constant. Tg thus plays a fundamental role in the properties and applications of polymers.[10−14]

With the advances in machine learning (ML), a surge in the exploitation of polymer informatics has emerged to predict polymer properties from the chemical structure of polymers in recent years,[7−9,15−47] including traditional ML and deep learning (DL) methods. Traditional ML algorithms, such as

Received: September 8, 2023
Revised: January 4, 2024
Accepted: January 11, 2024
Published: January 26, 2024

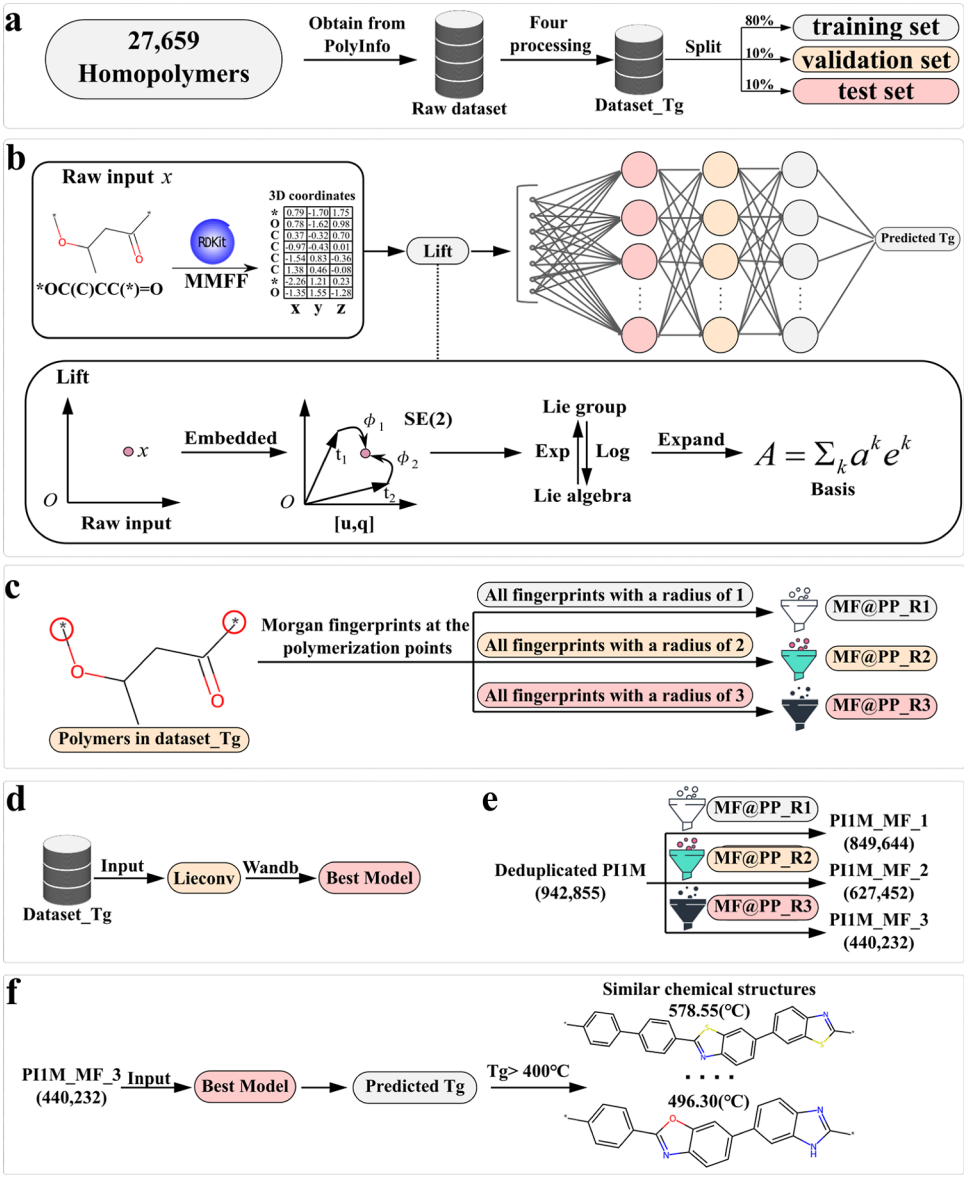


Figure 1. Overview of Lieconv-Tg. (a) Generation and split of data set_Tg. The raw data set containing 27,659 homopolymers with Tg values was collected from PolyInfo, and four processing steps created a standard data set containing 7170 polymers (named data set_Tg). The data set_Tg is divided into the training set, validation set, and test set in the ratio of 8:1:1. (b) General framework of Lieconv. Raw input x is the initial form of input data (spatial data represented as coordinates and values $\{(x_i f_i)\}_{i=1}^{N}$). Lift illustrates the lifted embeddings for the Lie group using SE(2) groups as an example in the form $[u,q]$, where $u \in G$ is an element of the group and $q \in \chi/G$ identifies the orbit. The elements of the Lie group are transformed into the corresponding Lie algebras via a logarithmic map and then are expanded to a basis to complete the subsequent calculations. The final basis was inputted into the Lieconv-Tg neural networks to predict the Tg values for polymers. (c) Generation of filters. Different filters were formed using different fingerprints with radii (from 1 to 3, MF@PP_R1, MF@PP_R2, MF@PP_R3). (d) Using Wandb to obtain the best Lieconv-based model. (e) Using different filters (MF@PP_R1, MF@PP_R2, and MF@PP_R3) to filter the deduplicated PI1M (PI1M_MF_1, PI1M_MF_2, and PI1M_MF_3). (f) When the predicted Tg value > 400 °C is used as a screening condition to filter PI1M_MF_3, the chemical structures of these polymers at the polymerization points have similar chemical structures.

support vector machine (SVM),[25,26,45] kernel ridge regression (KRR), Gaussian process regression (GPR), and random forest (RF), have been widely used to predict polymer properties. For instance, Varnek et al. used SVM to construct a unified model for predicting the Tg of both cross-linked and linear polymers of any type.[45] This model was built from 389 polymer data obtained from the literature (270 values for training and the remaining 119 for testing). The final performance on the testing set showed a root-mean-square error (RMSE) of 35.9 K. KRR has been used to train models for some high-throughput computed polymer properties (e.g., band gap and dielectric constant).[24] Ramprasad and co-workers, using three hierarchical levels of descriptors for feature representation of polymers, establish some GPR models in the polymer genome platform to predict various properties of polymers.[23] The best performance for the Tg task on the test set was[24] degrees. Pilania et al. used an RF model to predict the Tg of polyhydroxyalkanoate homopolymers and copolymers.[44] With 20 descriptors as a primary algorithm, the final model achieved an RMSE value of 11.12 K for the test set. The overly optimistic result may be due to the extremely small size of the database used for training (120 values) and testing

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

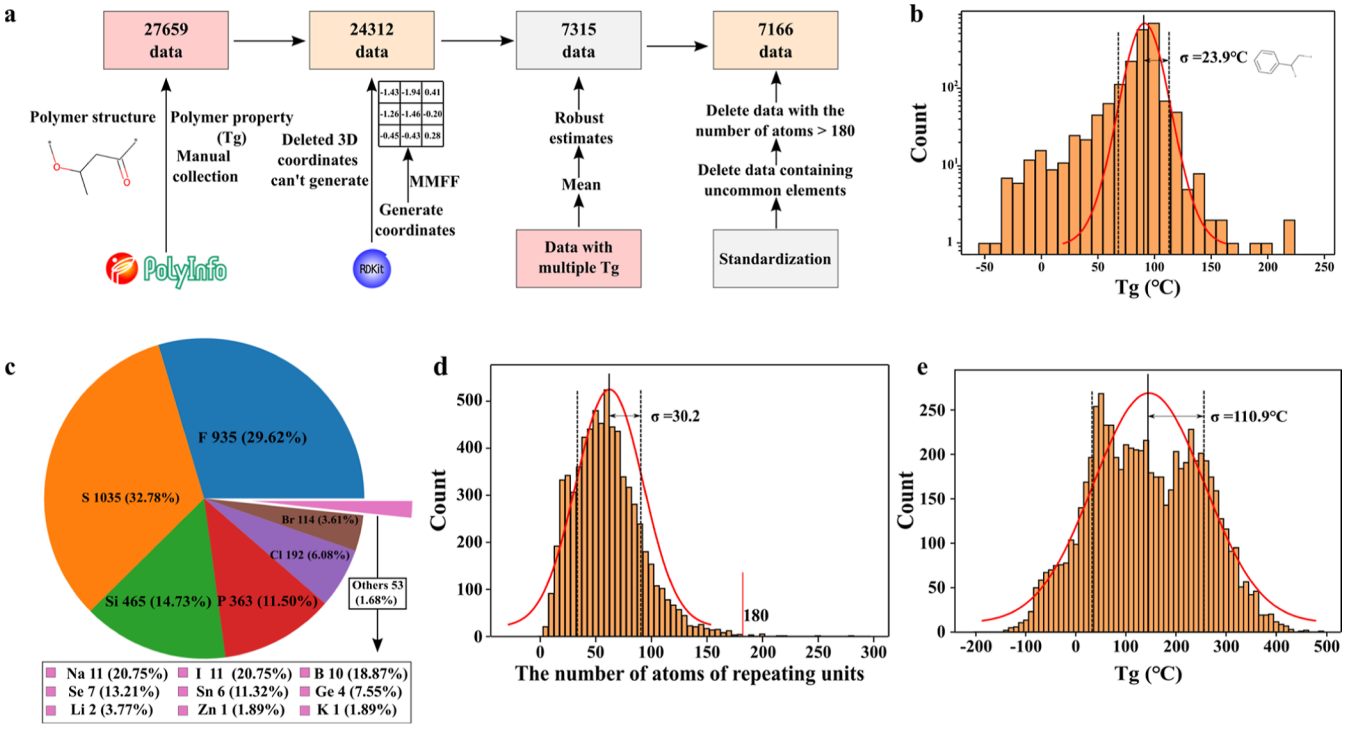


Figure 2. Process of generating data set_Tg. (a) 27,659 data obtained from PolyInfo were processed in a series of steps to produce the data set_Tg containing 7170 data. (b) Tg value distribution statistics for 2026 records for polystyrene. (c) After averaging the data for polymers with multiple Tg values, the frequency of occurrence of each element in the data set was counted (excluding C, H, O, and N). The elements in the other section are the elements that occur less frequently. (d) After deleting polymers containing elements with few occurrences, the distribution of the number of atoms contained in each polymer repeating unit. The repeating units with the number of atoms above 180 were removed from the raw data set. (e) Distribution of Tg values for the polymers in data set_Tg.

(13 values). Finally, Cheng et al. investigated different training methods to build a predictive ML model for polyimide (PI) Tg prediction. The model was based on a database of 225 PIs with 225 Tg values (165 values were used as the training set and the rest as the test set). The best performance of the ML model was around 20 K, but even the powerful GPR models have some problems. It is difficult for traditional ML algorithms to achieve better performance on large data sets because of the small capacity of these models.

With the growing polymer databases[48−51] and computational/experimental data sets in materials science, DL algorithms are increasingly utilized in polymer informatics. The learning capacity of deep learning is enhanced by sophisticated networks, such as convolutional neural networks (CNN), graph neural networks (GNN), recurrent neural networks (RNN), and transformers. They can capture subtle chemical information from the increasing availability of large amounts of data. For example, Miccio et al. converted the simplified molecular input line entry system (SMILES)[37] of 331 polymers into two-dimensional (2D) matrices (binary images). These binary images are used as inputs to predict the Tg values of polymers, and the relative errors of about 6% have been observed on the test set.[36] Das et al. proposed an RNN model based on the SMILES of polymers to train the Tg prediction model.[35] With the best hyperparameters, the model achieved a testing accuracy of 92%. Volgin et al. developed a graph convolutional neural network to predict the Tg of PI.[46] A data set with 214 polyimides combined with "transfer learning" to train the Tg prediction model. The result showed that the MAE of the best model was 22.5 °C, and the R^2 was 0.62. Aldeghi et al. used a graph representation of polymers and an associated GNN architecture to complete the prediction of polymer properties.[34] Tao et al. compared the impact of various ML methods (such as CNN, GNN, and RNN) and different feature representations on the performance of Tg models.[32,33] Amir et al. first converted the polymer data to tokens. Then, they used these tokens to train a transformer-based model and implement predictions for various properties of polymers.[39] Ramprasad and co-workers trained a large BERT model to predict a wide variety of properties of polymers.[41]

Like the translation equivariance of convolutional layers in CNN for images, more equivariances are also valuable for extracting subtle chemical information in polymers. This study employs an equivariant neural network (Lieconv) to predict Tg values from repeating units of polymers because of its powerful inductive bias and excellent performance (Lieconv-Tg). Lieconv is a new method that constructs a convolutional layer with translation and rotation equivariance from any specified Lie group.[31] It performs excellently in many domains, such as the image classification data set rotMNIST[52] and the molecular regression data set QM9.[53] Therefore, it has the potential to obtain a model with good performance, in terms of the Tg prediction for polymers. As shown in Figure 1a, a standard data set (data set_Tg) containing 7166 polymers was curated from 27,659 homopolymers with Tg values manually obtained from PolyInfo to build a highly stable and accurate Lieconv-Tg model and predict the Tg values of polymers.[51] Next, data set_Tg is divided into the training set, validation set, and test set in the ratio of 8:1:1. To increase the reliability of the virtual screening, the repeating units that contain elements not appearing in the data set_Tg are removed from the

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

PI1M[51] first (deduplicated PI1M 942,855). With fewer restrictions on the generation of polymer repeating units in PI1M, there are some chemically intuitively implausible repeating units. In order to obtain more rational candidate pools from the deduplicated PI1M, several filters are constructed by utilizing Morgan fingerprints[54] at the polymerization points (MF@PP) of repeating units in data set_Tg (see Figure 1c). The Sweep module of the Weights & Biases (Wandb)[55] and some additional processes are leveraged to obtain the optimal hyperparameters of the Lieconv-Tg method (see Figure 1d). Due to the differences in the data sets, it is difficult to compare Lieconv-Tg with previous methods fairly. Therefore, we constructed a model based on ECC[30] using the same 3D coordinates as input with optimized hyperparameters for comparison. Due to the excellent performance of the Lieconv-Tg model, it can be used to screen promising candidates from the filtered PI1M data sets (see Figure 1e, PI1M_MF_1, PI1M_MF_2, and PI1M_MF_3) which is constructed by using the MF@PP filters. When an extremely strict screening condition is set for polymers in PI1M_MF_3, the structure of these polymers is very similar (see Figure 1f). The further exploration of the repeating units in PI1M_MF_3 provides further evidence of what contributes to Tg and demonstrates the powerful capabilities of the Lieconv-Tg model. In order to screen out more polymers with heat resistance, the PolyAskln G[46] database of more than 6 million (6,726,935) polyimides was further screened using the Lieconv-Tg model. More details about the data set and Lieconv-Tg can be found in the Results and Discussion and Methods section.

METHODS

Data Set. All the data (27,659 homopolymers) containing Tg values have been hand-curated from one of the largest polymer databases, PolyInfo. They are homopolymers with experimental Tg values from the literature. These large amounts of reliable data are ideal for training deep neural network models. In general, the collected raw data set should be further processed before building the deep neural network models. Four critical steps were constructed to create the standardized data set (Figure 2a). Since the input of Lieconv requires arbitrary continuous (spatial) data represented as coordinates and values $\{(x_i f_i)\}_{i=1}^{N}$, the coordinates of polymer repeating units are generated by experimental torsion knowledge distance geometry (ETKDG)[56] in RDKit[57] cheminformatics package and optimized with Merck molecular force field (MMFF).[58] However, RDKit reports errors when the 3D coordinates of some repeating units. They could not obtain the optimized 3D coordinates as inputs to the model, so they were eliminated from the raw data set. Considering that the 3D coordinates generated from the force fields are the input in the model training, it is also of great importance to compare the coordinates of different force fields in RDKit, including the universal force field (UFF) and MMFF. The results are shown in Figure S1. The 3D coordinates, after optimization using the MMFF force field, can train the model better than that using the UFF force field. It is also important to note here that since the 3D coordinates are generated directly from SMILES, the model cannot distinguish between conformational isomers. Due to the fixing of the random seed in the coordinate generation process, the same SMILE input will result in the same optimized 3D coordinates. When the inputs to the model are the same, the same Tg predictions are obtained. Many homopolymer data have identical repeating units among the raw data set. For example, the polystyrene has 2026 records in the raw data set (Figure 2b illustrates the distribution of Tg values for these homopolymers), and it is impossible to distinguish these homopolymers with the collected information on them. For homopolymers with multiple Tg values, the average value is an approximation of the Tg value for that homopolymer. Ultimately, to further increase the stability of the raw data set, we applied additional restrictions to remove some outlier data (including polymers with "rare" atoms and the number of atoms in the repeating unit deviating far from common polymers) from the raw data set on which averaging had been performed. The frequency of elements in the raw data set is shown in Figure 2c (excluding C, H, O, and N since they are common elements in polymers). Some homopolymers containing elements that occurred less frequently (e.g., lithium, potassium, and zinc did not occur more than five times in all 7166 data sets) were removed from the original data set.

In some polymers, the number of atoms in the repeating unit deviated far from common polymers (the distribution of the number of atoms in the repeating unit is shown in Figure 2d), so polymers with the number of atoms in the repeating unit above 180 were removed from the raw data set in this step. The final data set contains 7166 homopolymers, which form data set_Tg composed of 10 usual elements (C, H, O, N, F, Si, P, S, Cl, and Br) and is sufficient to train robust and predictive Lieconv-Tg models for diverse polymers. In order to establish the models, the data set_Tg is divided into the training set, validation set, and test set in the ratio of 8:1:1. The training and other processes of the models are all on these three sets. It is important to note that the experimental Tg values used in the data set_Tg may be from different test methods: differential scanning calorimetry (DSC), dynamic mechanical thermal analysis (DMTA), thermal mechanical analysis (TMA), etc. The experimental Tg values obtained between the different methods may vary; therefore, it was also necessary to carry out independent modeling of these different data types (as shown in Figure S2).

In order to perform large-scale virtual screening for promising polymer materials, PI1M, a hypothetical polymer database generated by the RNN method with approximately 1 million polymers (995,800), forms a candidate pool of polymers. To increase the reliability of the virtual screening, the repeating units that contain elements not appearing in the data set_Tg are removed from the PI1M first (deduplicated PI1M 942,855). There are some implausible repeating units in PI1M because of only three simple restrictions when filtering the repeating units in the deduplicated PI1M database: (1) The repeating unit can complete the chemical validation in RDKit. (2) The number of '*' in the repeating unit is 2. (3) In a repeating unit, '*' is only attached to other atoms at one end. To obtain more promising candidates from the deduplicated PI1M and accelerate the design of polymer materials, further filtering of the deduplicated PI1M was performed using MF@PP. Chemically, polymers produced by the polymerization reactions should have unique chemical structures around the polymerization points. In terms of polymers produced by condensation polymerization, some characteristic groups such as the ether bond (-O−), ester bond (−OCO−), and amide bond (−NHCO−) are retained in their main chains. The polymers produced by free radical polymerization require the polymeric monomer to generate free radicals in the presence of an initiator. Therefore, π bonds are generally present in their

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

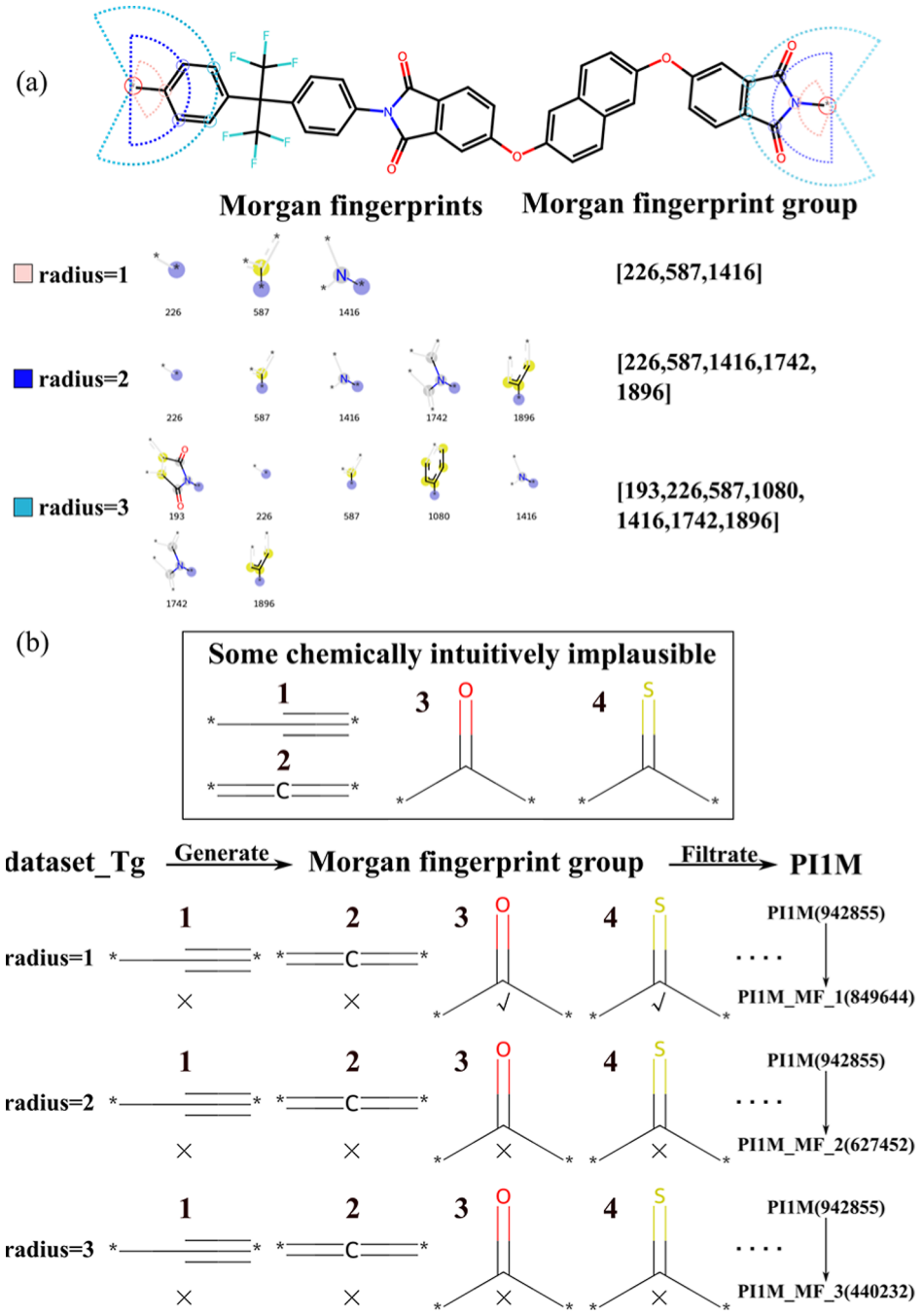


Figure 3. Generation of the Morgan fingerprint groups using Morgan fingerprints (radius from 1 to 3) and using them to filter the deduplicated PI1M. (a) Morgan fingerprints (radius from 1 to 3) at the polymerization points of repeating units in data set_Tg were collected and formed into corresponding Morgan fingerprint groups. The dashed areas of different colors represent the areas of chemical information that can be obtained for different radii of the Morgan fingerprint groups. (b) Different chemical spaces consisting of different sets of Morgan fingerprint groups (radius from 1 to 3) were used to filter the deduplicated PI1M and to generate the corresponding filtered data set (PI1M_MF_1, PI1M_MF_2, and PI1M_MF_3).

monomers such as alkenes, alkynes, carbonyl compounds, and some heterocyclic compounds. Overall, numerous repeating units that do not have a reasonable chemical structure in the deduplicated PI1M can be filtered out using MF@PP. The Morgan fingerprints at polymerization points were obtained from the repeating units in the data set_Tg, which are real homopolymers from the literature. All Morgan fingerprints at the two polymerization points of each repeating unit form a Morgan fingerprint group. Figure 3a shows the process of generating different Morgan fingerprint groups using the Morgan fingerprint radii from 1 to 3. By generating all the Morgan fingerprint groups with the same radius obtained from data set_Tg, a special chemical information space is formed to filter the deduplicated PI1M. Using different radii of the Morgan fingerprint groups corresponds to setting different degrees of restriction (as shown in Figure 3b, the larger the radius, the stricter the restriction). To further demonstrate the validity of this method, the distribution of the Synthetic Accessibility Score (SAscore)[59] for these data sets was analyzed. As shown in Figure S3, the larger the radius of the

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

fingerprint used for processing, the lower the average SAscore for the data set. It means that the portion of data with a higher SAscore is easily filtered out, which means that the ones that are more difficult to synthesize (relatively more implausible data) are more likely to be filtered out. The filtering results of the deduplicated PI1M using the Morgan fingerprint radii from 1 to 3 (PI1M_MF_1, PI1M_MF_2, and PI1M_MF_3) are uploaded to the GitHub repository of Lieconv-Tg. The three filtered data sets (especially PI1M_MF_2 and PI1M_MF_3) can be used to complete the large-scale virtual screening, which may accelerate the design of polymeric materials.

Lieconv. The key idea of Lieconv is to use specified Lie groups to construct a convolutional layer that is equivariant with the corresponding transformations. Thus, the crucial step of Lieconv is to correlate the polymer data with the Lie group and complete the relevant calculations. Consequently, the raw inputs, x_i in χ, are first transformed into Lie group elements u_i in G to accomplish group convolution called lift in the Lieconv methods. If χ is a homogeneous space of G, then every two elements in χ are associated by an element in G, and one can lift elements by simply selecting an origin o and defining $Lift(x) = \{u \in G/uo = x\}$: all elements in the group that map the origin to x. This measure allows the lifting of tuples of coordinates and features $\{(x_i f_i)\}_{i=1}^{N} \rightarrow \{(u_{ik} f_i)\}_{i=1,k=1}^{N,K}$, with up to K group elements for each input. To find all the elements $\{u \in G/uo = x\}$, one needs to find one element u_x and use the elements in the stabilizer of the origin $H = \{h \in G/ho = o\}$ to generate the rest with $Lift(x) = \{u_x h \text{ for } h \in H\}$. The Lie group space G is not always a vector space that can perform a group convolution. However, in the Lie algebra of G, the tangent space at the identity, $g = T_{id}G$, is a vector space and can be understood commonly as a space of infinitesimal transformations from the corresponding Lie group. So, the Lie group element u_i in G is further generalized to Lie algebra in g. In a vector space, one can easily expand elements in a basis $A = \Sigma_k a^k e^k$ and utilize the components for calculations. The logarithmic map log: $G \rightarrow g$ gives a mapping from the Lie group to the Lie algebra, converting infinitesimal transformations to corresponding Lie algebra, and an inverse mapping exp: $g \rightarrow G$ can be defined. In order to further expand the performance of this method, the quotient space $Q = \chi/G$ (consisting of the distinct orbits of G in χ) was considered. It is because the most general equivariant mappings will use this orbit information throughout the network. The space of elements should not be G but rather $G = \chi/G$, and $x \in \chi$ is lifted to the tuples (u,q) for $u \in G$ and $q \in Q$. The next step is to achieve group convolution. Adopting the convention of left equivariance, one can define a group convolution between two functions on the group, which generalizes the translation equivariance of convolution to other groups

Definition 1. *Let $k, f: G \rightarrow \mathcal{R}$, and $\mu(\cdot)$ be Haar measure on G. For any $u \in G$, the convolution of k and f on G at u is given by*

$$h(u) = (k^* f)(u) = \int_G k(\nu^{-1}u) f(\nu) d\mu(\nu). \quad (1)$$

The general framework of Lieconv is shown in Figure 1b. Algorithms 1 and 2 provide a concise overview of the lifting procedure and the new convolution layer, respectively. These two algorithms can be found on page 6 of this ref 31.

Architecture. The inputs of ECC and Lieconv-Tg require arbitrary continuous (spatial) data represented as coordinates and values of $\{(x_i f_i)\}_{i=1}^{N}$. Due to the differences in the network architecture of the models, the input 3D coordinates are also processed differently to get the final predicted values. More details of the ECC can be found in Figure S4. More information about Lieconv can be found in Figure 1b. The coordinates of repeating units can be obtained by ETKDG in RDKit. The obtained coordinates are further optimized with the MMFF. The optimized coordinates are then fed into the subsequent network through a Lift process, which transforms raw inputs x_i into group elements u_i. After a single linear layer, it is input into a module (six BottleBlocks and the Lieconv layer in this module complete the group convolution). Finally, after a BatchNorm layer, a Swish layer (activation function layer), a Linear layer, and a Global Pool layer, the output obtained is the predicted value of Tg. A more specific architecture for ECC is shown in Figure S5. A more detailed architecture of the Lieconv-Tg neural network is shown in Figure S6.

Software and Hardware. ECC was implemented using Spektral (version 1.2) and TensorFlow (2.5.0). Lieconv_Tg was implemented using Lieconv and PyTorch (1.12.1). RDKit (2022.9.1) was used to produce 3D coordinates of the molecule. Since Spektral is done using TensorFlow and Lieconv (0.1) is based on PyTorch, this work uses two different architectures. To mitigate this issue, we created the environment file for each framework. It is relatively easy to build the required environment according to the environment files. The Lieconv-Tg method and related files are available on our GitHub repository (https://github.com/LZ0221/Lieconv-Tg). All of the training and prediction methods for the model and the environment preparation are placed in this repository. The Jupyter notebook and virtual screening results are also uploaded to this repository. All models are trained on an RTX 3090 GPU.

RESULTS AND DISCUSSION

Optimization of Hyperparameters. The performances of the DL models are influenced by their hyperparameters. In order to explore the potential of the model, the Sweep module of Wandb and some additional procedures are utilized to obtain the optimal hyperparameters of all models Lieconv-Tg and ECC. The optimized hyperparameters of ECC and Lieconv-Tg and their meanings are listed in Table S1 and Table S2, respectively. For all three models, 40 trials on Wandb are carried out. Although 40 is less than 1 or 2 orders of magnitude of the entire hyperparameter combination space, Wandb can capture better-performing hyperparameters using a Bayesian search method. It utilizes Gaussian process[60] to model the relationship between the hyperparameters and the model metric and chooses hyperparameters to optimize the probability of improvement. As illustrated in Figures S7 and S8, most models eventually converge (at the right bottom of these figures), where the model has good performance with lower loss, demonstrating the reliability and efficiency of the Bayesian search method. Although the optimal model in the entire hyperparameter space may be missed, this choice balances the considerable cost (the vast hyperparameter space) and the slight performance loss. For the 40 trials, the epoch at which the trial terminates is 100, and the learning rate is fixed. As shown in Figure S9 (ECC) and Figure S10 (Lieconv-Tg), it is still difficult to distinguish these better-performing models.

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

The difference in their loss curves is slight, especially the loss in the more important validation set. In order to obtain a more robust model, five sets of potential hyperparameters (Tables S3 and S4) were initially selected from 40 trials. Two additional processes were performed to explore model performance: (1) The epoch of the model training is increased to 300 to ensure that the model capability can be thoroughly mined out. The performance of the ECC method reaches the peak before 100 epochs, while the performance of the Lieconv-Tg model is not thoroughly mined until it is trained up to 200 epochs (Figure S11). (2) A learning rate decay strategy was applied for all models to get better performance. The performance of five potential models based on ECC and Lieconv-Tg is illustrated in Figures S12 and S13 and Tables S5 and S6, respectively. Ultimately, the optimal model is obtained by comparing the MAE of the validation. The optimal models of ECC and Lieconv-Tg are trial 2 and trial 35, respectively. They have a relatively small MAE on the validation set. In addition, three different Lie groups were used to predict the Tg of polymers, which can be seen as a comparison of the performance of models with different equivariances. The Lie group T(3) has translation equivariance, and the Lie group SO(3) has rotation equivariance. The Lie group SE(3), which contains both translation and rotation equivariances, performs better (Figure 4). It also indicates that more equivariance provides a stronger inductive bias, which can lead to a better model.

Model Evaluation and Comparison. Using the sweep module of Wandb and some additional procedures, we eventually obtained two optimal models for the two types of DL methods. The optimal results of all models are shown in Figure 5 (trials 2 and 35 for ECC and Lieconv-Tg, respectively). The optimal Lieconv-Tg model can achieve MAEs of 12.92, 24.37, and 24.42 for the training, validation, and test sets. The optimal ECC model can achieve MAEs of 18.78, 32.53, and 31.29 for the training, validation, and test sets. The comparison shows that the MAEs of the Lieconv-Tg model on the validation and test sets are significantly better than the ECC models (using the same 3D coordinates as input). The MAE is reduced significantly by ∼6 from ∼30 to ∼24 on both the validation set and test set. The R^2 values for both the validation and test sets can reach 0.90. It demonstrates that the Lieconv-Tg model has a better performance that can reasonably screen promising candidates from a hypothetical database and provide a rational direction for the materials design of polymers.

Large-Scale Tg Value Prediction. Comparing the three optimal models for the two types of DL methods, the Lieconv-Tg model has excellent prediction and generalization ability and can be used to predict the Tg values of polymers and assist the virtual screening.

PI1M_MF_2 and PI1M_MF_3 form promising candidate pools to screen polymers with favorable properties. The use of PI1M_MF_1 as a promising candidate pool for virtual screening is not considered here, as the Morgan fingerprints of radius 1 contain too little chemical information. A loose restriction may allow inappropriate polymers to appear in the screening results. It is interesting to note that even this small chemical information (MF@PP_R1) can filter out tens of thousands of inappropriate polymers in the deduplicated PI1M. It further demonstrates that the chemical information at the polymerization points captures the subtle structure of polymers and can complete a reasonable screening of the polymers in the deduplicated PI1M. The data set_Tg, with

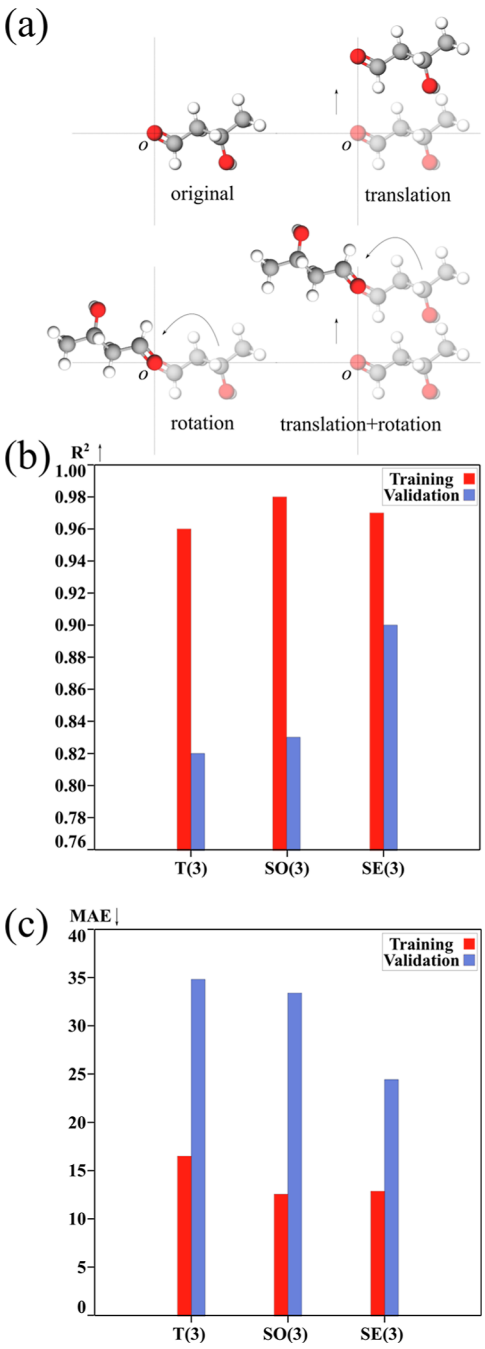


Figure 4. Comparisons of Lieconv-Tg with different Lie groups. (a) Difference between these three kinds of Lie groups. The Lie group T(3) has translation equivariance, and the Lie group SO(3) has rotation equivariance. The Lie group SE(3) contains both translation and rotation equivariances. (b) Performance (R^2) of the Lieconv model with different Lie groups in the training and validation sets. (c) Performance (MAE) of the Lieconv model with different Lie groups in the training and validation sets. The Lie group SE(3), which includes both translation and rotation equivariances, performs better (with a higher R^2 and a lower MAE in the validation set).

7166 real polymers, has 2511 polymers with Tg > 200 °C. These polymers have great potential to be used in harsh environments with high temperatures, but there is a need for more high Tg candidates to accelerate materials design for polymers. Through the Lieconv-Tg model, additional ∼49,435 new candidates are found in PI1M_MF_2 with Tg > 200 °C.

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

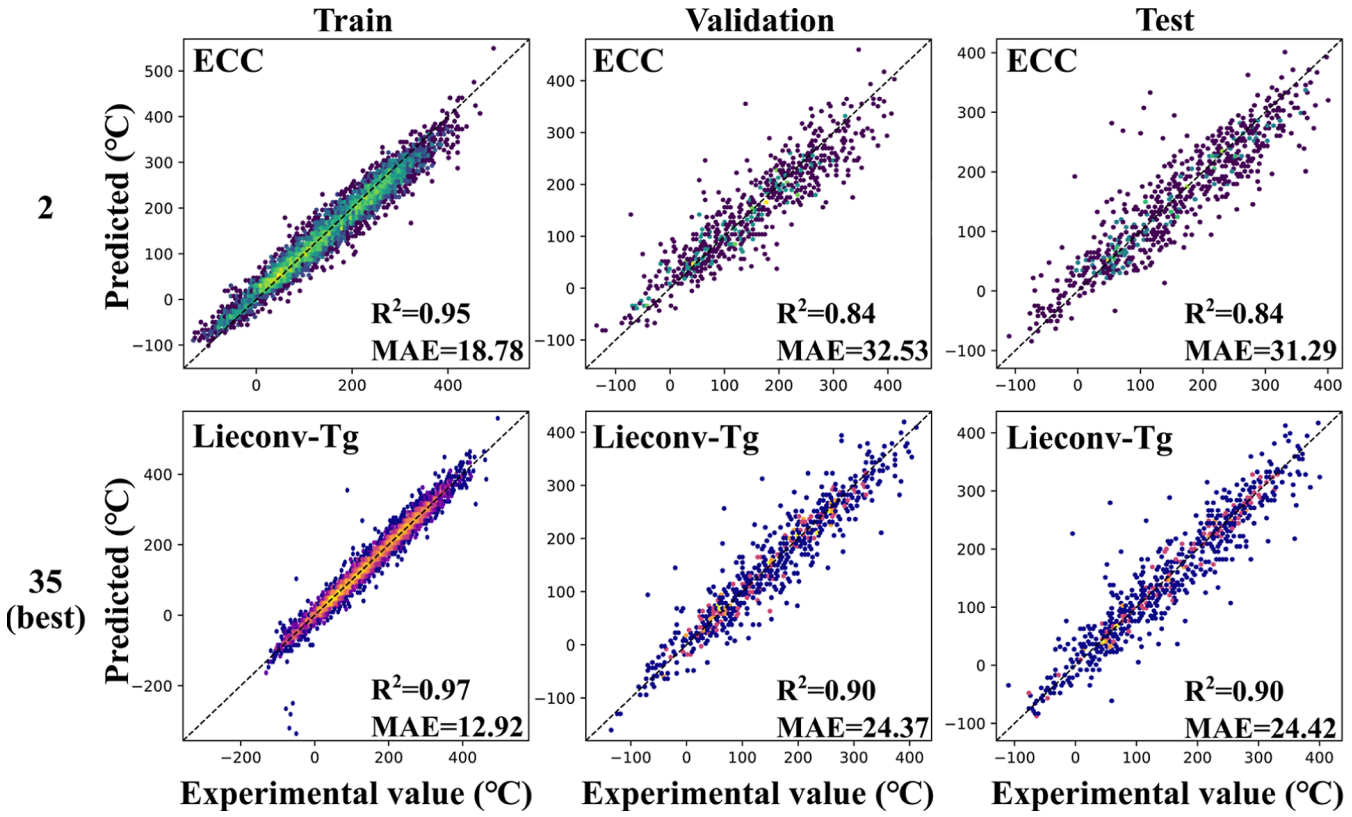


Figure 5. Performance of the Lieconv-Tg and baseline models (ECC) for predicting the glass-transition temperature (Tg). Each set of plots shows the experimental values against those predicted by the DL models for the training, validation, and test sets. The diagonal line is shown as a black dotted line. The scatter/density shows the predictions of each model for the training, validation, and test sets. The color intensity is proportional to the probability density, with brighter colors indicating areas of higher point density. The average coefficient of determination (R^2) and mean absolute error (MAE) across all two types of models are shown (from top to bottom are ECC and Lieconv-Tg).

Thus, this high-throughput screening finds ∼20 times more promising candidates for high-temperature polymers than the 2511 known high-temperature polymers in data set_Tg. If a harsher environment with a required Tg > 300 °C is considered, data set_Tg and PI1M_MF_2 have 615 and 3481 polymers that can potentially meet this requirement. Again, this high-throughput screening method identifies 5 times more promising candidates from PI1M_MF_2 than from data set_Tg. If PI1M_MF_3 is used, 37,951 candidates with Tg values > 200 °C and 2805 candidates with Tg > 300 °C can also be obtained. Finally, a Jupyter Notebook is uploaded to our GitHub repository to show how to filter polymers for specific requirements such as Tg > 400 °C, SAscore <3, Morgan fingerprint groups, and contained elements. In order to screen out more polymers with favorable properties, the PolyAskln *G*[46] database of more than 6 million (6,726,935) polyimides was further screened using the optimal model. Using the same processing steps as for the PI1M data set, we filtered the PolyAskln *G* data set. The results of the filtering were also put into the GitHub repository.

Further Exploration of PI1M_MF_3. When the predicted Tg value > 400 °C or < −110 °C is used as a screening condition to filter PI1M_MF_3, the chemical structures of these polymers at the polymerization points have intuitive patterns (as shown in Figures S14 and S15). Therefore, all Morgan fingerprints with a radius of 3 were collected for these repeating units, their Morgan fingerprint occurrences were counted, and these processes were used to mine the potential chemical information from them. For the Morgan fingerprints of the polymers with Tg > 400 °C in PI1M_MF_3, as shown in Figure 6a, the three most frequently occurring substructures (587, 1873, and 1380) indicate the presence of a ring structure. In particular, the 587 substructure exhibits a polymerization point associated with the ring structure in the main chain. The following six substructures are all large ring systems in the main chain throughout the polymer structure. In general, when the main chain contains more aromatic or aromatic heterocyclic rings, the proportion of single bonds that can rotate is relatively reduced. As the size of the chain segments becomes large, the rigidity of the molecular chain increases, and the Tg is higher. A continuous ring structure in the main chain further increases the Tg. For the Morgan fingerprints of polymers with Tg < −110 °C in PI1M_MF_3, as shown in Figure 6b, the structures of the polymers are primarily based on the oxy-silica system. Polymers with a saturated single bond in the main chain have a lower Tg due to the ease of internal rotation of the single bond, larger molecular chain flexibility, and smaller chain segment sizes. The above analysis further demonstrates the high rationality and powerful capabilities of the Lieconv-Tg model. All Morgan fingerprints of radius 3 and their occurrence frequencies of the polymers for Tg > 400 °C and Tg < −110 °C are shown in Figure S16 and Figure S17.

CONCLUSIONS

With the development of polymer informatics and the growing amount of available data in databases, data-driven deep learning approaches show great potential in polymer science. In this study, we utilized the data obtained from Polyinfo to

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

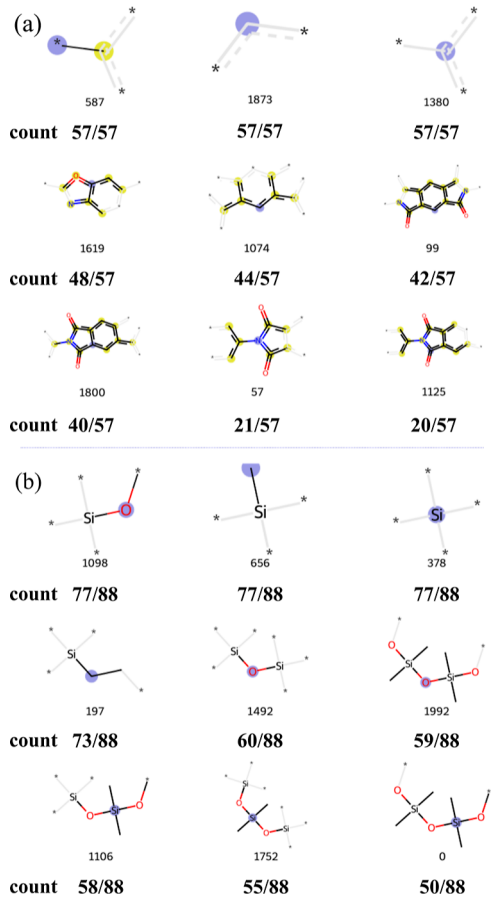


Figure 6. Morgan fingerprints with high occurrence frequencies in polymers with Tg > 400 °C and Tg < −110 °C in the PI1M_MF_3, the count represents the number of occurrences/total number of repeating units. (a) High frequency of Morgan fingerprints in repeating units with Tg > 400 °C all indicate the presence of a ring structure in the main chain of the polymer, which makes the Tg higher. (b) High frequency of Morgan fingerprints in repeating units with Tg < −110 °C all indicate the presence of the oxy-silica system in the main chain of the polymer, which makes the Tg lower.

build a standard data set (data set_Tg) containing 7166 data and optimized the hyperparameters of the models by Wandb and some additional procedures. The Lieconv-Tg model with strong equivariances shows the best performance in predicting the Tg values for polymers compared to ECC using the same 3D coordinates as input. The results show that the Lieconv-Tg model can achieve R^2 of 0.97, 0.90, and 0.90 for the training, validation, and test sets, respectively. The MAEs of the model on the training, validation, and test sets are 12.92, 24.37, and 24.42, respectively. Lieconv is significantly better than the results obtained by the ECC based on the same 3D coordinates as min input (MAE is reduced significantly by ∼6 from ∼30 to ∼24 on both the validation set and test set). This performance of MAE for Tg prediction around 20 seems to be a good enough result for any NN, but they usually get such good results only on small experimental data sets (<1000). Lieconv was trained on data set_Tg (a standard larger experimental data set with a wide variety of polymers), so Lieconv still performs better and has stronger generalization capabilities. The Tg values of ∼1 million generated polymers in PI1M were predicted by the Lieconv-Tg method. In order to select the more plausible and promising polymers from the deduplicated PI1M, a new approach is proposed to filter the deduplicated PI1M based on MF@PP and obtain a better result. Promising polymers are obtained many times more than in data set_Tg, whether PI1M_MF_2 or PI1M_MF_3 is used. The Tg values and SAscores can filter and rank candidate polymers, helping researchers to find polymers with high Tg values that are easy to synthesize quickly. Further analysis of the prediction results for PI1M_MF_3 confirms the power of the Lieconv-Tg method. The Lieconv-Tg method and these large-scale reasonably promising candidates will provide new directions in the materials design of polymers. Although data set_Tg is already relatively large, it is still too small compared to the entire chemical space of polymers. With the rising learning capacity for the model, more training data and ways to make the most of these data are required. Using unsupervised pretraining and supervised fine-tuning strategies, transfer learning can make full use of unlabeled data to train models with better performance and is a promising research direction for polymer property prediction.

■ ASSOCIATED CONTENT

Data Availability Statement

The Lieconv-Tg method and related files are available on our GitHub repository (https://github.com/LZ0221/Lieconv-Tg). All the training and prediction methods for the model and the environment preparation are placed in this repository. The Jupyter notebook and virtual screening results are also uploaded to this repository.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c06843.

Data set of Lieconv-Tg (data set_Tg), feature representation of data set_Tg, architecture of the models, optimization of hyperparameters, and further exploration of PI1M_MF_3 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Zhimin Zhang − *College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR, China*; orcid.org/0000-0002-4167-4234; Email: zmzhang@csu.edu.cn

Hongmei Lu − *College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR, China*; Email: hongmeilu@csu.edu.cn

Author

Zheng Long − *College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c06843

Author Contributions

Zheng Long and Zhimin Zhang conceptualized and planned the research. Zheng Long collected the data, wrote the code, trained the models, completed the virtual screening, and analyzed the screening result. Zheng Long and Zhimin Zhang interpreted the results and wrote the manuscript. Hongmei Lu and Zhimin Zhang supervised the work.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (22373117 and 22273120) and Exxon Mobil Asia Pacific Research and Development Company Ltd (A4011260). We acknowledge the High Performance Computing Center of Central South University for support. The studies meet with the approval of the university review board. We are grateful to all employees of this institute for their encouragement and support of this research.

■ REFERENCES

(1) Wong, C. *Polymers for electronic & photonic application*; Elsevier, 2013.
(2) Peacock, A. J.; Calhoun, A. *Polymer chemistry: Properties and application*; Carl Hanser Verlag GmbH Co KG, 2012.
(3) Haque, F. M.; Grayson, S. M. The synthesis, properties and potential applications of cyclic polymers. *Nat. Chem.* **2020**, *12* (5), 433−444.
(4) Hiemenz, P. C.; Lodge, T. P. *Polymer chemistry*; CRC Press, 2007.
(5) Huan, T. D.; Boggs, S.; Teyssedre, G.; Laurent, C.; Cakmak, M.; Kumar, S.; Ramprasad, R. Advanced polymeric dielectrics for high energy density applications. *Prog. Mater. Sci.* **2016**, *83*, 236−269.
(6) Liechty, W. B.; Kryscio, D. R.; Slaughter, B. V.; Peppas, N. A. Polymers for Drug Delivery Systems. *Annu. Rev. Chem. Biomol. Eng.* **2010**, *1* (1), 149−173.
(7) Mueller, T.; Kusne, A. G.; Ramprasad, R. Machine learning in materials science: Recent progress and emerging applications. *Rev. Comput. Chem.* **2016**, *29*, 186−273.
(8) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **2017**, *3* (1), 54.
(9) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: Current status and critical next steps. *Mater. Sci. Eng.: R: Rep.* **2021**, *144*, 100595.
(10) Hergenrother, P. M. The use, design, synthesis, and properties of high performance/high temperature polymers: an overview. *High Perform. Polym.* **2003**, *15* (1), 3−45.
(11) Zhou, H.; Xue, C.; Weis, P.; Suzuki, Y.; Huang, S.; Koynov, K.; Auernhammer, G. K.; Berger, R.; Butt, H.-J.; Wu, S. Photoswitching of glass transition temperatures of azobenzene-containing polymers induces reversible solid-to-liquid transitions. *Nat. Chem.* **2017**, *9* (2), 145−151.
(12) Van Krevelen, D. W.; Te Nijenhuis, K. *Properties of polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*; Elsevier, 2009.
(13) Meyer, J. Glass transition temperature as a guide to selection of polymers suitable for PTC materials. *Polym. Eng. Sci.* **1973**, *13* (6), 462−468.
(14) Müller, C. On the glass transition of polymer semiconductors and its impact on polymer solar cell stability. *Chem. Mater.* **2015**, *27* (8), 2740−2754.
(15) Kim, C.; Pilania, G.; Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **2016**, *28* (5), 1304−1311.
(16) Pankajakshan, P.; Sanyal, S.; de Noord, O. E.; Bhattacharya, I.; Bhattacharyya, A.; Waghmare, U. Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights. *Chem. Mater.* **2017**, *29* (10), 4190−4201.
(17) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28* (20), 7324−7331.
(18) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22* (12), 3762−3767.
(19) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21* (7), 785−796.
(20) Kim, C.; Chandrasekaran, A.; Jha, A.; Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Commun.* **2019**, *9* (3), 860−866.
(21) Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **2019**, *5* (1), 66.
(22) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110* (10), 5714−5789.
(23) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122* (31), 17575−17585.
(24) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3* (1), 2810.
(25) Yu, X. Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers. *Fibers Polym.* **2010**, *11* (5), 757−766.
(26) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889−2919.
(27) Cassar, D. R.; de Carvalho, A. C. P. L. F.; Zanotto, E. D. Predicting glass transition temperatures using neural networks. *Acta Mater.* **2018**, *159*, 249−256.
(28) Liu, W.; Cao, C. Artificial neural network prediction of glass transition temperature of polymers. *Colloid Polym. Sci.* **2009**, *287* (7), 811−818.
(29) Zhang, Z.; Friedrich, K. Artificial neural networks applied to polymer composites: a review. *Compos. Sci. Technol.* **2003**, *63* (14), 2029−2044.
(30) Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017; pp 3693−3702.
(31) Finzi, M.; Stanton, S.; Izmailov, P.; Wilson, A. G. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*; PMLR: 2020; pp 3165−3176.
(32) Tao, L.; Varshney, V.; Li, Y. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *J. Chem. Inf. Model.* **2021**, *61* (11), 5395−5413.
(33) Tao, L.; Chen, G.; Li, Y. Machine learning discovery of high-temperature polymers. *Patterns* **2021**, *2* (4), 100225.
(34) Aldeghi, M.; Coley, C. W. A graph representation of molecular ensembles for polymer property prediction. *Chem. Sci.* **2022**, *13* (35), 10486−10498.
(35) Goswami, S.; Ghosh, R.; Neog, A.; Das, B. Deep learning based approach for prediction of glass transition temperature in polymers. *Mater. Today: Proc.* **2021**, *46*, 5838−5843.
(36) Miccio, L. A.; Schwartz, G. A. From chemical structure to quantitative polymer properties prediction through convolutional neural networks. *Polymer* **2020**, *193*, 122341.
(37) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(38) Park, J.; Shim, Y.; Lee, F.; Rammohan, A.; Goyal, S.; Shim, M.; Jeong, C.; Kim, D. S. Prediction and Interpretation of Polymer Properties Using the Graph Convolutional Network. *ACS Polym. Au* **2022**, *2* (4), 213−222.

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462

(39) Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a Transformer-based language model for polymer property predictions. *npj Comput. Mater.* **2023**, *9* (1), 64.
(40) Magar, R.; Wang, Y.; Barati Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Comput. Mater.* **2022**, *8* (1), 231.
(41) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14* (1), 4099.
(42) St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **2019**, *150*(23)..
(43) Wen, C.; Liu, B.; Wolfgang, J.; Long, T. E.; Odle, R.; Cheng, S. Determination of glass transition temperature of polyimides from atomistic molecular dynamics simulations and machine-learning algorithms. *J. Polym. Sci.* **2020**, *58* (11), 1521−1534.
(44) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59* (12), 5013−5025.
(45) Higuchi, C.; Horvath, D.; Marcou, G.; Yoshizawa, K.; Varnek, A. Prediction of the Glass-Transition Temperatures of Linear Homo/Heteropolymers and Cross-Linked Epoxy Resins. *ACS Appl. Polym. Mater.* **2019**, *1* (6), 1430−1442.
(46) Volgin, I. V.; Batyr, P. A.; Matseevich, A. V.; Dobrovskiy, A. Y.; Andreeva, M. V.; Nazarychev, V. M.; Larin, S. V.; Goikhman, M. Y.; Vizilter, Y. V.; Askadskii, A. A.; Lyulin, S. V. Machine Learning with Enormous "Synthetic" Data Sets: Predicting Glass Transition Temperature of Polyimides Using Graph Convolutional Neural Networks. *ACS Omega* **2022**, *7* (48), 43678−43691.
(47) Yu, M.; Shi, Y.; Liu, X.; Jia, Q.; Wang, Q.; Luo, Z.-H.; Yan, F.; Zhou, Y.-N. Quantitative structure-property relationship (QSPR) framework assists in rapid mining of highly Thermostable polyimides. *Chem. Eng. J.* **2023**, *465*, 142768.
(48) Ellis, B.; Smith, R.; CRC Press, 2008.Polymers: a property database
(49) CAMPUS. https://www.campusplastics.com.
(50) CROW. http://www.polymerdatabase.com.
(51) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M., PoLyInfo: Polymer Database for Polymeric Materials Design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 2011; pp 22−29.
(52) Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; Bengio, Y., An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, 2007; pp 473−480.
(53) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131* (25), 8732−8733.
(54) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(55) Weights & Biases. https://wandb.ai/site.
(56) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562−2574.
(57) Landrum, G. e. a. RDKit: Open-source cheminformatics. http://www.rdkit.org.
(58) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 490−519.
(59) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8−11.
(60) Williams, C.; Rasmussen, C. Gaussian processes for regression; *Advances in neural information processing systems*, 1995, p 8.

https://doi.org/10.1021/acsomega.3c06843
ACS Omega 2024, 9, 5452−5462