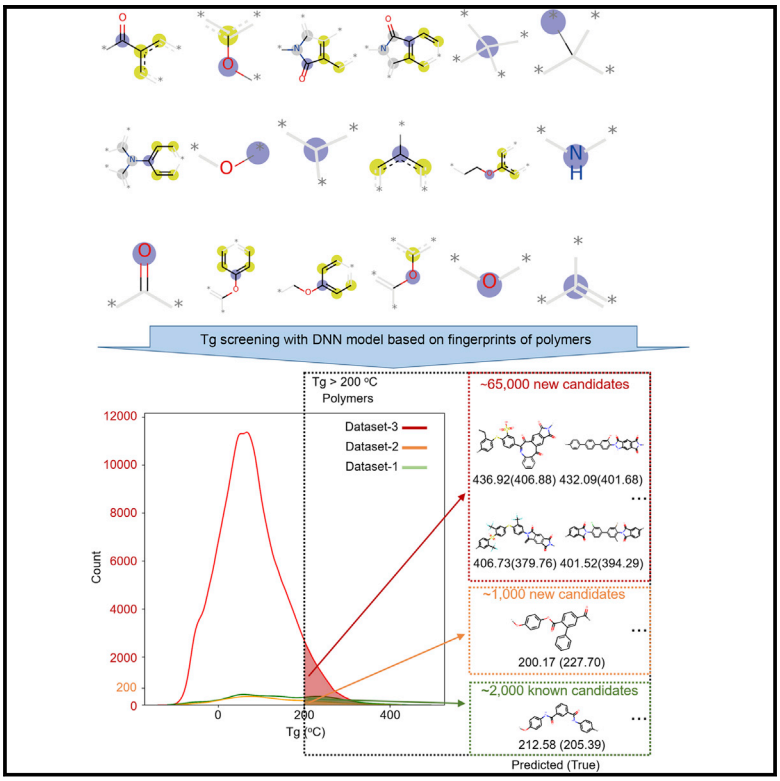
Patterns

Machine learning discovery of high-temperature polymers

Graphical abstract



Highlights

- Large datasets for polymer's glass transition temperature are collected
- Transferability of ML models depends on feature representations
- Molecular dynamics models and experimental results validate the formulated ML model
- Extensive promising candidates for high-temperature polymers are screened by ML model

Authors

Lei Tao, Guang Chen, Ying Li

Correspondence

yingli@engr.uconn.edu

In brief

Polymers with outstanding high-temperature properties have been identified as promising materials for aerospace, electronics, and automotive applications. However, the current design and development of high-temperature polymers has been an experimentally driven and trial-and-error process guided by experience, intuition, and conceptual insights. Therefore, we formulate a machine learning model that can quantitatively predict the glass transition temperature of a polymer from its chemical structure, such that more promising high-temperature polymers can be efficiently filtered out through high-throughput screening.

Tao et al., 2021, Patterns 2, 100225
April 9, 2021
https://doi.org/10.1016/j.patter.2021.100225

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Machine learning discovery of high-temperature polymers

Lei Tao,[1,3] Guang Chen,[1,3] and Ying Li[1,2,4,*]
[1]Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA
[2]Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, CT 06269, USA
[3]These authors contributed equally
[4]Lead contact
*Correspondence: yingli@engr.uconn.edu
https://doi.org/10.1016/j.patter.2021.100225

THE BIGGER PICTURE The design and development of high-temperature polymers has been an experimentally driven and trial-and-error process guided by experience, intuition, and conceptual insights. However, such an Edisonian approach is often costly, slow, biased toward certain chemical space domains, and limited to relatively small-scale studies, which may easily miss promising compounds. To overcome this challenge, we formulate a data-driven machine learning (ML) approach, integrated with high-fidelity molecular dynamics simulations, for quantitatively predicting the glass transition temperature of a polymer from its chemical structure and rapid screening of promising candidates for high-temperature polymers. Our work demonstrates that ML is a powerful method for the prediction and rapid screening of high-temperature polymers, particularly with growing large sets of experimental and computational data for polymeric materials.

Proof-of-concept Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

To formulate a machine learning (ML) model to establish the polymer's structure-property correlation for glass transition temperature T_g, we collect a diverse set of nearly 13,000 real homopolymers from the largest polymer database, PoLyInfo. We train the deep neural network (DNN) model with 6,923 experimental T_g values using Morgan fingerprint representations of chemical structures for these polymers. Interestingly, the trained DNN model can reasonably predict the unknown T_g values of polymers with distinct molecular structures, in comparison with molecular dynamics simulations and experimental results. With the validated transferability and generalization ability, the ML model is utilized for high-throughput screening of nearly one million hypothetical polymers. We identify more than 65,000 promising candidates with $T_g > 200^{\circ}C$, which is 30 times more than existing known high-temperature polymers (~2,000 from PoLyInfo). The discovery of this large number of promising candidates will be of significant interest in the development and design of high-temperature polymers.

INTRODUCTION

Lightweight and high-strength polymers with outstanding high-temperature properties have been identified as promising materials for aerospace, electronics, and automotive applications.[1–3] These high-temperature polymers are expected to have long-term durability at high temperatures, high thermal decomposition temperatures, or high glass transition temperature T_g. For example, polytetrafluoroethylene is a synthetic fluoropolymer of tetrafluoroethylene with a maximum service temperature >260°C, which has been widely used for non-stick coatings and insulations.[4] The other successful high-temperature polymers are perfluoroalkoxy alkanes, polyether ether ketone (PEEK), and fluorinated ethylene propylene. The high-temperature properties of these polymers are realized through the heteroatoms in the polymer chain of thermoplastics.[5–7] However, the molecular engineering and design of hydrocarbon polymers and other polymers with high-temperature properties remain to be explored. The current design and development of high-temperature polymers have been an experimentally driven and

trial-and-error process guided by experience, intuition, and conceptual insights. For example, different experimental strategies have been developed to synthesize high-temperature hydrocarbon polymers, such as (1) enhancement of the tacticity of the polymer chains,[8,9] (2) introduction of bulky pendant groups into the side chain,[10–12] and (3) incorporation of cyclic structures into the backbone chain.[13–15] Nevertheless, this Edisonian approach is often costly, slow, biased toward certain chemical space domains, and limited to relatively small-scale studies, which may easily miss promising compounds.[16] Thus, a robust and reliable high-throughput screening method is essential for the discovery and design of high-temperature polymers.[17]

For high-temperature polymers, a critical property is the T_g,[10,13,14] which determines the polymer's phase transition between a rubbery state and a glassy state, yielding orders of magnitude difference in elastic modulus.[18] Until now, T_g is well known to be related to many factors, including molecular weight,[19] chain stiffness,[20] side groups,[21] additives,[22] regularity.[23] Considering these aspects, researchers have proposed theoretical correlations between the chemical structure and the T_g of polymers. These empirical methods are built upon the assumption that the chemical groups in the repeating units of the polymer chain contribute to the T_g additively with different weighting factors.[24–26] For example, Van Krevelen and Te Nijenhuis[18] and Hoftyzer and colleagues[26] have proposed the "Molar Glass Transition Function," based on nearly 600 experimental T_g values of polymers, with different group contributions and structural corrections to T_g. This approach provides an effective way for molecular interpretation of T_g. However, this additive method is only applicable to the polymers containing previously investigated chemical structures.[18] Later, Dudowicz et al.[27] formulated an analytic theory to estimate T_g of polymer melts as a function of the relative rigidities of the chain backbone and side groups, monomer structure, polymer mass, and pressure, based on the generalized Lindemann criteria. This analytical theory can explain the general trends in the variation of T_g related to the microstructure of the polymer, e.g., influences of side-chain length, and relative rigidities between side groups and chain's backbone. Nevertheless, it cannot be used to directly predict the T_g of the polymer based on its chemical structure. Very recently, Xie et al.[28] established a relationship between T_g and molecular structure of 32 conjugated polymers with a single adjustable parameter ζ. ζ is an effective mobility value, determined by assigned atomic mobility for the repeating unit of conjugated polymers. The experimental results confirm that ζ is strongly correlated to the T_g of conjugated polymers, although they differ drastically in aromatic backbone and alkyl side-chain chemistry. Yet, quantitatively predicting a polymer's T_g from its chemical structure remains a significant challenge. We still lack a universal model that connect a polymer's T_g to its repeating unit and molecular structure.

With advancements in molecular simulation and high-performance computing, all-atom molecular dynamics (MD) simulations can reasonably predict a polymer's T_g,[29] despite the limitations of computational cost, cooling rate, and uncertainty.[30–33] Nevertheless, it is not feasible to use these expensive MD simulations to explore the vast chemical space of polymers, defined by the almost infinite combinations of their chemical elements and molecular structures. With the growing amount of polymer database,[16,30–33] data-driven methods are emerging to build correlations between chemical structure and the T_g of polymers, including quantitative structure-property relationships (QSPR) method[34–36] and machine learning (ML).[37–39] For the QSPR method, a large array of molecular descriptors are extracted from the polymer's repeating unit, which applies to any chemical structure.[40] For example, Katritzky et al. have extracted more than 400 constitutional, topological, geometrical, and quantum chemical descriptors for the repeating unit of the polymer.[40] Subsequently, a multi-step linear regression analysis is adopted to train these descriptors, leading to a good match between predicted and experimental $T_g$values for 88 homopolymers. Wu et al.[41] encoded a descriptor vector of seven different fingerprints, such as standard, extended, hybridization, maccs. And their Bayesian linear model reported an R value of 0.916 for T_g prediction. Liu and Cao[42] have adopted the artificial neural network to predict the T_g for 113 polyacrylates and polystyrenes, as a function of four molecular descriptors: the molecular average polarizability, the energy of the highest occupied molecular orbital, the total thermal energy, and the total entropy. Later, Cai et al.[43] have combined a support vector regression with particle swarm optimization, using six quantum chemical descriptors as inputs, to predict T_g values for 32 methacrylate polymers. However, the QSPR method suffers two major drawbacks: (1) it is expensive to quantify a large array of molecular descriptors, such as quantum chemical descriptors, which require the time-consuming density-functional theory calculations; (2) the QSPR method might generate many parameters that are challenging to physically interpret, such as topological bond connectivity and Kier shape index.[40]

Considering these aspects, several ML models have been established to predict a polymer's T_g directly from its chemical structure. For instance, Ramprasad and co-workers[37–39,44] utilized three hierarchical levels of descriptors, including atomic level, QSPR, and morphological descriptors, for feature representation of polymers. They fitted their datasets of 451–1,321 polymers with the Gaussian process regression model in the polymer genome platform.[38,45–48] When using 1,321 polymers for training, their ML model reported a root-mean-square error of 27 K and R^2 of 0.92.[39] In addition to molecular descriptors as feature representation, ML models, such as convolutional neural networks (CNNs) with image-based input, have also been examined. For example, Miccio et al.[49,50] converted the Simplified Molecular Input Line Entry System (SMILES) notations of 331 polymers into a two-dimensional (2D) matrix (binary images) by the presence or absence of composing characters in the SMILES formulation. This approach can be used to predict the unknown T_g of polymers with average relative errors as low as 6%, particularly without time-consuming calculations of molecular descriptors. Table 1 summarizes the database, feature representation, models, and prediction metrics from these theoretical, QSPR and ML studies.

Despite these extensive studies, we are still facing several significant challenges in creating ML models to directly predict a polymer's T_g based on its chemical structure.[16] Firstly, most of these data-driven models are built upon a small dataset of polymer T_g values with less than 1,000 data points, focusing on a certain category of polymers, such as polyacrylates and polystyrenes. It is very difficult to generalize these models for other

Table 1. Summary of theoretical, QSPR, and machine learning (ML) models investigated in the literature

Database	Features	Model	R^2	Ref.
600	chemical groups	group contributions approach	N/A[a]	18
32	an effective mobility value	single adjustable parameter	N/A[b]	28
113	quantum chemical descriptors	artificial neural networks	0.955[c]	42
37	quantum chemical descriptors	support vector regression	0.97	43
251	Descriptors	computational neural networks	0.96	51
389	descriptors	support vector regression	0.78	52
133	descriptors	random forest	N/A[d]	53
88	descriptors	multi-layer perceptron neural network	0.96	54
77	descriptors	support vector machine (SVM)	0.92	55
54	descriptors	artificial neural network	0.91	56
52	descriptors	artificial neural network	0.978[e]	57
451	hierarchy fingerprint	Gaussian process regression	0.94	38
751	hierarchy fingerprint	Gaussian process regression	0.87	37
1,321	hierarchy fingerprint	Gaussian process regression	0.92	39
5,917	combined fingerprint	Bayesian linear model	0.916[f]	41
331	SMILES-based binary images	convolutional neural network	N/A[g]	49
234	SMILES-based binary images	fully connected neural networks	N/A[h]	50
6,923 + 5,690 + 1 million	descriptors Morgan fingerprint SMILES-based binary images	lasso regression deep neural network convolutional neural network	0.80 0.85 0.87	this work

N/A, not applicable.
[a]About 80% of the calculated T_g values differed less than 20 K from the experimental values.
[b]Only root-mean-square error of 13°C was reported for all 32 alkylated conjugated polymers.
[c]R = 0.955 was reported for the prediction set.
[d]Only root-mean-square error of 4.76 K was reported for the test set of the model.
[e]R = 0. 978 was reported for the test set.
[f]R = 0. 916 was reported for the test set.
[g]The model performance was evaluated by relative error of 3%–8%.
[h]The model performance was evaluated by average relative errors of ~3%.

classes of polymers due to the limited range of chemical space. Secondly, it is challenging to choose appropriate feature representations to describe the chemical structures of polymers. Molecular descriptors, fingerprints, and images have been adopted to represent the chemical structures of polymers. It is not clear which feature representation is the most appropriate, leading to a predictive ML model for exploring a large chemical space of polymers. Finally, it is not straightforward to associate ML predictions on a polymer's T_g with physically meaningful quantities. Since most ML models are highly nonlinear with complicated architectures, it is difficult to pinpoint a specific set of physical quantities or chemical groups that are important in the prediction and design of a polymer's T_g.

To overcome the above challenges, we manually collected about 13,000 homopolymers structures from the largest polymer database, PoLyInfo.[58] Copolymers that are formed by two types of monomers are not collected here as the effect of their different components on T_g requires extra consideration,[59,60] and polymer composites are not included either when their T_g is affected by polymers interplaying with nanomaterials.[61,62] Focusing on homopolymers allows us to put our focus mainly on revealing the correlation of a polymer's chemical structure and its T_g. Among the around 13,000 homopolymers, 6,923 experimental T_g values are available, which form dataset-1, as shown in Figure 1. The remaining 5,690 polymers without reported T_g values form dataset-2. Also, a benchmark database, named PI1M[63], with nearly one million hypothetical polymers generated by a recurrent neural network (RNN) model, is taken as dataset-3, while the corresponding T_g values are unknown. Note that dataset-3 covers a similar chemical space as dataset-1 and dataset-2 because the RNN models are also trained on the PoLyInfo database, but significantly populate regions where PolyInfo data are sparse.[63] Such a large and diverse dataset allows us to develop four representative ML models based on dataset-1, namely Lasso_Descriptor, Lasso_Fingerprint, DNN_Fingerprint, and CNN_Image, by using the molecular descriptors, Morgan fingerprints, or images as inputs, and Lasso (least absolute shrinkage and selection operator), DNN (deep neural network) or CNN as the ML models. The predictivity and transferability of these ML models are tested on dataset-2 with distinct chemical substructures (Figure 1), in comparison with MD simulations and experimental results. Interestingly, our study reveals that the DNN_Fingerprint model can reasonably predict the T_g values of polymers from dataset-2, as the Morgan fingerprinting method[64] can take into account the chemical connectivity and appearance of different substructures of a polymer's repeating unit. More importantly, we use these ML models to identify key molecular descriptors and chemical substructures that can significantly

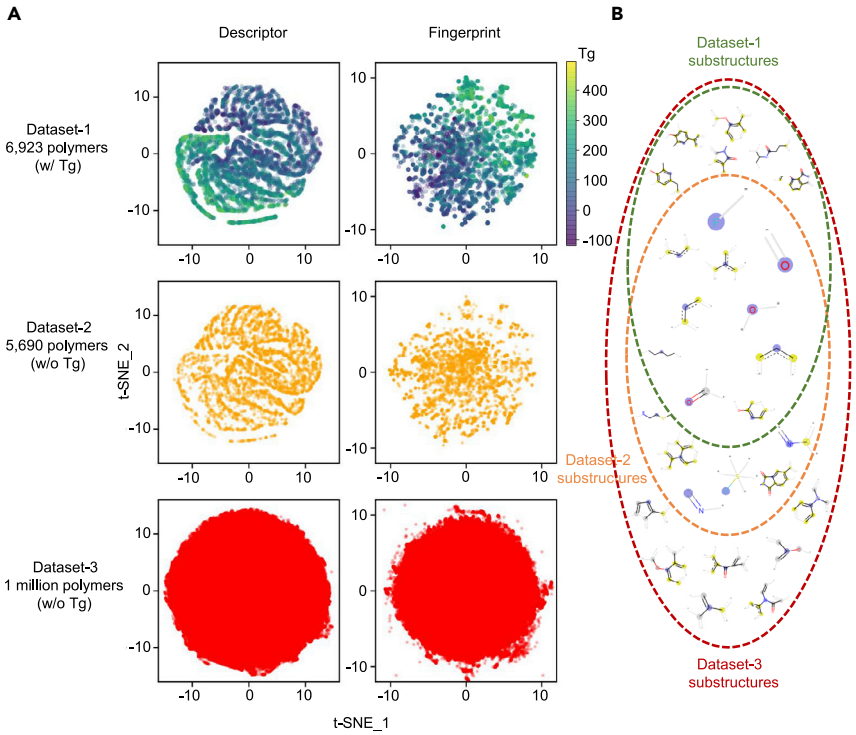


Figure 1. Chemical space visualization of dataset-1, dataset-2, and dataset-3

(A) 2D visualization based on descriptors and fingerprints using the t-SNE algorithm. Dataset-1 has reported T_g values, and each data point is colored based on the corresponding T_g value. Dataset-2 and dataset-3 do not have reported T_g values, colored with yellow and red, respectively.

(B) Set diagram showing representative substructures in dataset-1 (green circle), dataset-2 (yellow circle), and dataset-3 (red circle) based on Morgan fingerprint. Some substructures are common for all datasets, while some others are unique to certain datasets.

affect the polymer's T_g, providing physical insights into the prediction and design of the T_g for polymeric materials. We further examine the chemical functional groups of high-/low-T_g polymers and their common characteristics through Checkmol.[65] We also identify strong correlations between these common functional groups with the key chemical substructures revealed by our ML models. Eventually, we apply the validated DNN_Fingerprint model for rapid screening of one million hypothetical polymers in PI1M (dataset-3), and identify more than 65,000 promising candidates for high-temperature polymers with T_g > 200°C. We then use MD simulations to validate the predicted T_g values of the top four high-temperature polymers, which are previously unexplored and have not been tested to date. Thus, our study demonstrates that ML is a powerful method for the prediction and rapid screening of high-temperature polymers, particularly with growing large sets of experimental and computational data for polymeric materials. The key molecular descriptors and chemical substructures informed by ML models, combined with identified chemical functional groups, are important design motifs for the molecular engineering of high-temperature polymers.

RESULTS AND DISCUSSION

Dataset, feature representation, and chemical space

To formulate robust and predictable ML models for diverse polymers, we need to consider a larger dataset in contrast to previous studies (cf. Table 1). Dataset-1 contains 6,923 polymers from the largest polymer database, PoLyInfo,[58] as listed in Table 2. They are real polymers with experimentally measured $T_g$values reported in literature. Thus, it is ideal to use dataset-1 as a labeled dataset for ML model training. For experimentally measured T_g values, they depend on conditions, such as the cooling or heating rate, or even curing process and moisture content, thus there cannot be an exact value for T_g.[66–69] Although there are variations in experimental measurements, the reported T_g with a common experiment practice can be considered characteristic only of the polymer and not of the measuring method.[70] If measurement conditions are so extreme that the obtained T_g is not a proper representative of the real value, such records will mislead all analysis, including ML model training.

A total of 5,690 real polymers of dataset-2 were collected from the same data source as dataset-1, but their T_g values were not previously reported. Dataset-3 is based on an ML-generated database PI1M[63] with approximately one million hypothetical polymers. Note that PI1M is enumerated using a generative ML model, RNN, based on PolyInfo (dataset-1 plus dataset-2). These three datasets are regarded as similar to each other in terms of chemical space.[63] The collected three datasets in Table 2 are more than one order of magnitude of most datasets from the kinds of literature in Table 1, making up a broader range of chemical space involving various categories of polymers. The challenge of having ML models that can be generalized to all categories of polymers then becomes straightforward to address with the collected large datasets.

All polymers' chemical formulas and structures are represented by the SMILES notation,[71] which is a line notation for describing the structure of chemical species using short ASCII strings. For example, "*C(C*)C" represents the repeating unit for "poly(prop-1-ene)." It is worth noting that a special symbol "*" is used to indicate the polymerization points for the repeating unit. From the same molecular block, such as "CCC," the polymerization positions in *C(C*)C take into account the bonding information between repeating units, and determine the spatial structure of the polymer chain. The chemical species contained in these three datasets include C, O, N, Cl, F, Br, I, S, Si, B, P, Sn, Fe, Na, Li, Ge, Se, K, Co, Ni, Ca, Cd, Pb, Zn, and Te.

One challenge when creating ML models for evaluation of a polymer's T_g is choosing appropriate feature representation to describe the chemical structures being studied. Representation

Table 2. Comparison of three datasets

Dataset	No. of polymers	T_g (°C)	Source
Dataset-1	6,923	−118∼495	real polymers from PoLyInfo[58]
Dataset-2	5,690	unknown	real polymers from PoLyInfo[58]
Dataset-3	1 million	unknown	hypothetical polymers from PI1M[63]

options include descriptors, fingerprints, molecular graph, molecular embedding, quantum chemical quantities, images, etc. The effect of using different representations on T_g estimation has been demonstrated through systematic representation evaluation[72] or separate model development.[37–39,42,43,50–57] In addition, the development of new representations remains critical for the development of high-performance ML models. To carry out a thorough study considering different types of representations, we explore three types of feature representation based on the SMILES notation of each polymer: molecular descriptors, Morgan fingerprints, and images, as presented in Figure 2. In terms of molecular descriptors, the feature-generating engine alvaDesc[73] supports the calculation of about 5,305 descriptors within 32 categories, ranging from constitutional indices and ring descriptors to chirality descriptors.[73,74] The ensemble of descriptors represents the physical and chemical characteristics of polymers/molecules being studied, which have been widely adopted in the QSPR and ML models (Table 1). Thus, these molecular descriptors can provide physical information regarding charges, topological indices, functional groups, etc., of polymers. Among these 5,305 descriptors, 3,579 descriptors are all available for real polymers in dataset-1 and dataset-2. However, not all 3,579 descriptors are available to the one million hypothetical polymers in dataset-3. Around 5% of hypothetical polymers in dataset-3 cannot be processed using the alvaDesc. But it does not affect too much the chemical space visualization based on molecular descriptors for dataset-3. We should emphasize that the alvaDesc cannot process the * symbol in the SMILES notation and, thus, it misses the chemical connectivity of the repeating units.

In addition to molecular descriptors, we also choose the fingerprinting method (extended connectivity fingerprinting [ECFP])[64] to numerically represent the chemical connectivity in a repeating unit of the polymer. Specifically, the fingerprinting method has a significant advantage over the traditional group contribution and molecular descriptor methods, where all the possible build blocks and molecule descriptors have to be defined *a priori* and remain static. However, the fingerprinting method is more dynamic, and it can evolve to include new chemical structures and connectivities.[64] Essentially, to derive the ECFP of the repeating unit, we need to: (1) assign each atom with an identifier, (2) update each atom's identifiers based on its neighbors, (3) remove duplicates, and (4) fold list of identifiers into a 2,048-bit vector (a Morgan fingerprint). In this case, we transform each polymer's SMILES notation into a binary "fingerprint," by using the Daylight-like fingerprinting algorithm as implemented in RDKit[75] with radius 3 and 2,048 bits. Note that radius 3 is large enough to identify/encode large fragments of the chemical structure, with more than 45,000 distinct substructures detected from all datasets. Such a topological-based approach analyzes the various substructures of a molecule within a certain number of chemical bonds (here it is 3), and then hashes each substructure into a 2,048-bit vector, as shown in Figure 2. If the 45,000 distinct substructures are hashed into 2,048 buckets, collisions are inevitable. Then, the 1/0 (on/off) bit of a bucket does not indicate the occurrence of a specific substructure but represents the occurrence of several substructures. Besides, the number of occurrences for a substructure is not recorded through these buckets. To avoid the drawbacks of using buckets, we directly record each substructure and its number of occurrences. This dictionary of substructures is further used for the training of our ML models. We should emphasize that our fingerprinting method is different from previous studies using the ECFP and Morgan fingerprinting,[41,76,77] as we need to consider the number of occurrences for certain substructures in the training of ML models, to be discussed in the following section.

Based on the SMILES notation of polymers, we further define an ordered list of SMILES characters as a dictionary [“c”, “n”, “o”, “C”, “N”, “F”, “ = “, “O”, “(‘, ’)”, “*”, “[‘, ’]”, “1”, “2”, “3”, “#”, “Cl”, “/”, “S”, “Br”]. This dictionary creates a binary column for each character, with which one-hot encoding

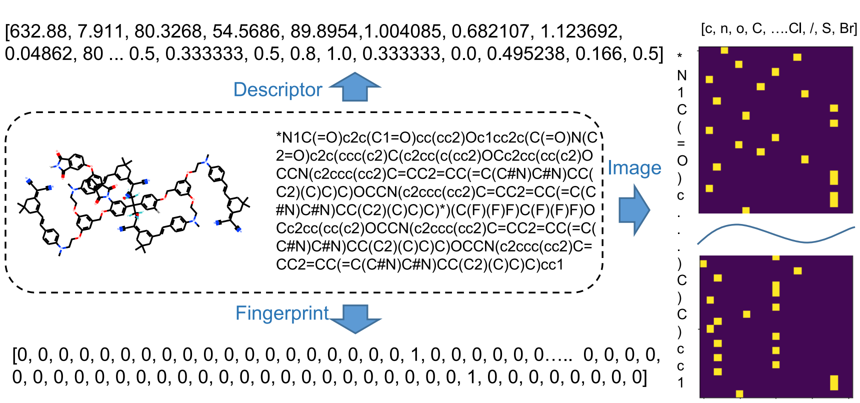


Figure 2. Three types of feature representation calculated based on the polymer's SMILES notation for ML models: molecular descriptor, Morgan fingerprint, and image

Table 3. Four ML models trained on dataset-1

Name	ML model	Features	R^2 (train/test)
Lasso_Descriptor	Lasso regression model	3,579 descriptors	0.80/0.71
Lasso_Fingerprint	Lasso regression model	2,048 fingerprints	0.74/0.73
DNN_Fingerprint	deep neural network	2,048 fingerprints	0.85/0.83
CNN_Image	convolutional neural network	310 × 21 binary images	0.87/0.80

algorithm[78] transforms each polymer's SMILES into a sparse matrix (a 2D binary image in Figure 2). The dimensions of all images are 21 (the number of characters in the dictionary) × 310 (the length of the longest SMILES code in the dataset). The key points of the one-hot encoding algorithm are: (1) defining a reasonable dictionary is the premise of a good model; (2) simple polymers (represented by a short SMILES code) return much sparser matrices than complex polymers (represented by a long SMILES code). Obviously, any change of dataset could lead to changes in the dictionary and corresponding images, significantly influencing the performance of a CNN model.

In view of the molecular descriptors and Morgan fingerprints, similarities between different datasets can be compared from their chemical space. To better visualize this space, the high-dimensional chemical spaces are reduced to a low-dimensional representation. By t-distributed stochastic neighbor embedding,[79] the chemical spaces can be shown in 2D plots as shown in Figure 1A. The top row of Figure 1A is for dataset-1, whose T_g values are marked with a color bar. The middle and bottom rows are for dataset-2 and dataset-3, respectively. We can see that, on both descriptor and fingerprint space, dataset-1 and dataset-2 distribute randomly on similar regions. The random distribution suggests that dataset-1 and dataset-2 are across similar chemical spaces. Dataset-3 is also found filling up a similar chemical space but significantly populate regions where PoLyInfo data (dataset-1 plus dataset-2) are sparse. Although Figure 1A shows similarities between dataset-1, dataset-2, and dataset-3, disparities still exist. For example, using Morgan fingerprints, we show some substructures of these polymers in dataset-1, dataset-2, and dataset-3 (Figure 1B). Besides the shared substructures enclosed in the overlapped area of the circles, all three datasets have their own unique substructures. As ML models are trained based on dataset-1, when they encounter a polymer in other datasets with new substructures, it is difficult to make an accurate prediction. Compared with the performance on dataset-1, whether the ML model can be well transferred to new dataset-2 and dataset-3 is more worthy of concern. ML models with good transferability and generalization ability are of significant importance for the discovery and design of high-temperature polymers.

ML models for the chemistry-T_g relation of polymers

Four ML models trained on dataset-1 (listed in Table 3) involve the Lasso model, the DNN model, and the CNN model. Lasso is a least-squares regression model with a shrinkage penalty, through which it performs variable selection by forcing the coefficients of trivial variables to become zero. Thus, the variables that are strongly associated with the output are identified in a variable selection process. DNN consists of connected units called nodes or neurons. Each node receives signals and triggers a process function to output new signals. Several nodes are grouped into layers and constructed into a complicated network architecture, which is processed between the input and output layers. DNN is capable of learning complex relationships between input and output. CNN is distinguished from DNN by its superior performance on image input. The convolutional layers with filters or kernels are the core building blocks of CNN. The optimized weights and biases in convolutional layers can identify the presence of various features in the input, showing an advanced performance, particularly in image processing. Although the ML algorithms are applicable for various kinds of problems, such as video recognition, image analysis, or natural language processing, their suitability and reliability are actually highly domain dependent. For the task of estimating a polymer's T_g based on structure features, ML models require a proper feature representation that depicts polymer physics and chemistry to the greatest extent.

Here, descriptors or fingerprints are used as the input features for Lasso regression models or DNN models. They have clear chemical or physical meanings for an organic molecule, but the time-consuming calculation is usually required considering a very large database of polymers. When representing polymers from the perspective of 2D images, the input is much easier to calculate.[49,50] Therefore, a CNN model using images is also investigated for comparison. Through these ML models, we aim to discover critical physical and chemical features affecting T_g, and to establish a reliable model for T_g screening of high-temperature polymers. Lasso regression is suitable for feature selection, while DNN and CNN models are more powerful to establish a correlation between chemical structure and T_g of polymers.[76]

The performances of these four ML models are illustrated by parity plots in Figures 3A–3D (see the supplemental experimental procedures Figures S1–S3 for model training details). Based on dataset-1, they all show good performances. The best one is the CNN_Image model, which produces an $R^2$of 0.87/0.80 for training/test sets. It indicates that, although there is no explicit physical meaning in the image representation, the CNN model is still able to establish a correlation between the image input and the physical property T_g of polymers. The DNN and Lasso models also lead to high R^2 values of 0.74–0.87. Their performances are satisfactory, considering the large chemical diversity of 6,923 polymers involved in dataset-1.

To examine the transferability of ML models on new polymers, these four ML models are applied to dataset-2 to predict their $T_g$values. The prediction accuracy of ML models is further validated with MD simulations (see Figure S4 and Table S2 for the MD simulation details and results). Twenty polymers are randomly selected from dataset-2. Their MD-simulated T_g and ML-predicted T_g values are compared in Figure 3E. Four ML models show different prediction performances on these

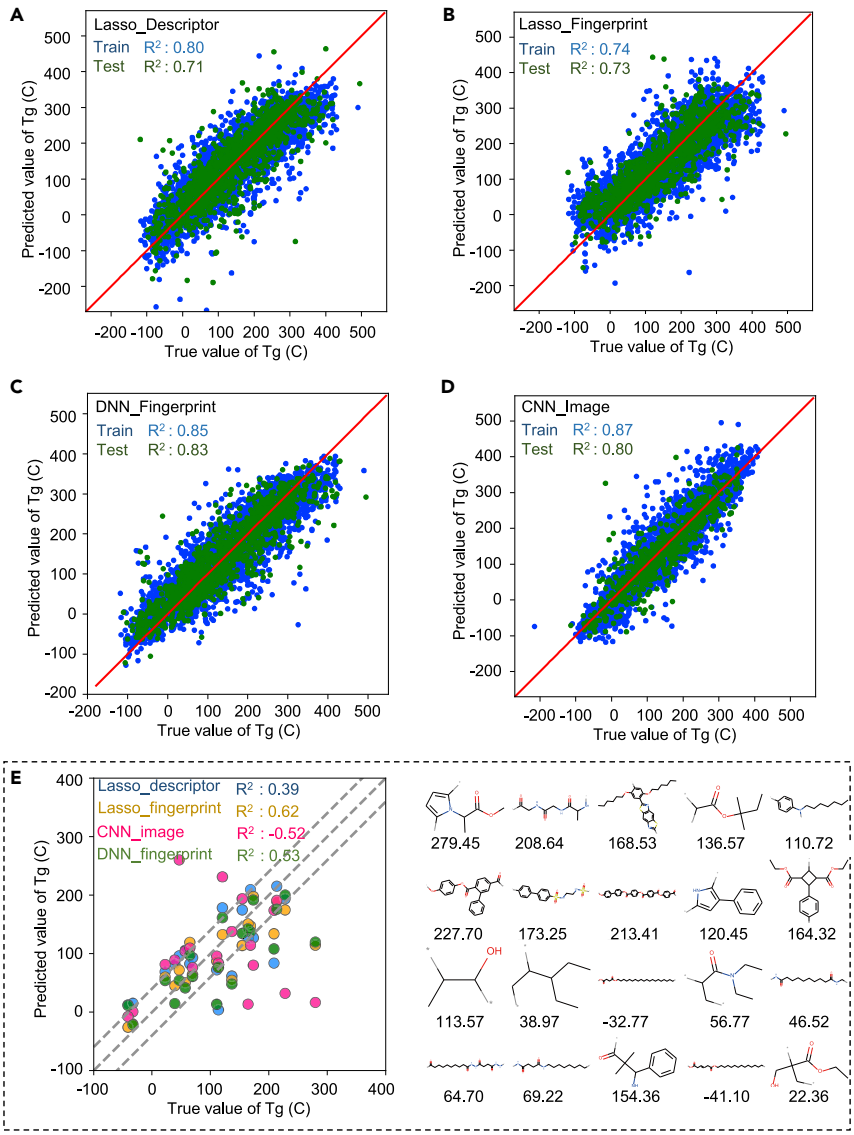


Figure 3. Performance of four ML models

(A) The Lasso regression model using descriptors as input features (Lasso_Descriptor model).

(B) The Lasso regression model using fingerprints as input features (Lasso_Fingerprint model).

(C) The DNN model using fingerprints as input features (DNN_Fingerprint model).

(D) The CNN model using images as input features (CNN_Image model).

(E) The comparison between the MD-simulated T_g and the ML-predicted T_g on 20 polymers randomly selected from dataset-2. Three dashed lines are a unity line and lines with a mean absolute error of 40°C. The chemical structure of these 20 polymers is followed by their MD-simulated T_g value.

polymers of dataset-2 (see Table S3 in the supplemental experimental procedures). The performances of CNN_Image model and Lasso_Descriptor model degrade remarkably to R^2 of −0.52 and 0.39, respectively, indicating poor transferability from dataset-1 to dataset-2. These two previously well-trained ML models on dataset-1 are found to be no longer accurate when giving a new and different dataset. Due to their worse generalization capabilities, the CNN_Image model and Lasso_Descriptor model are not considered for high-temperature polymer screening in the following sections.

On the contrary, the Lasso_Fingerprint and the DNN_Fingerprint models demonstrate good performance on these randomly selected polymers, with $R^2$of 0.63 and 0.53, respectively. Their small changes of R^2 from dataset-1 to dataset-2 suggest good transferability. Although with a little degradation, the prediction performances are still satisfactory considering: (1) dataset-2 is not exactly the same as dataset-1 in terms of substructures (cf. Figure 1), and (2) uncertainties may exist as the reference T_g values obtained by MD simulations can be higher than the true values due to the high cooling rate.[31,32,80,81] To avoid the uncertainties from MD simulations, validation using experimental results is more preferred. Thus, a newly reported experimental dataset is further utilized to verify the transferability of these two ML models. The experimental dataset contains 32 semiflexible (mostly conjugated) polymers[28] that are new to our ML models. These 32 polymers differ drastically in the aromatic backbone and alkyl side-chain chemistry (Table S4 in the supplemental experimental procedures), serving as an ideal experimental dataset to test our ML models. The predictions of the Lasso_Fingerprint model and the DNN_Fingerprint model lead to R^2 values of 0.20 and 0.68 (see Figure S5 in the supplemental experimental procedures for detailed results). Thus, the performance of the Lasso_Fingerprint model is found to be degrading on this new experimental dataset. According to these results, we find that the DNN_Fingerprint model has a consistent performance on different datasets with excellent transferability through the validations by MD simulations and experimental results. Also, Morgan fingerprints are identified to be more appropriate as feature representation for the ML model of polymer T_g in comparison with molecular descriptors and images.

As mentioned above, both molecular descriptors and images are representations of all the possible building blocks of a polymer's repeating unit, which must be defined *a priori* and remain static. However, Morgan fingerprints are an inherent more dynamic representation, as they can evolve to include new chemical substructures once encountered. Also, according to the previous theoretical models on T_g values of polymers,[18] we know that the number of occurrences for these substructures also plays an important role. Therefore, our Morgan fingerprints explicitly consider more than 45,000 distinct substructures and their frequency of occurrence, which allows us to study the effects of various substructures and their linkages on polymer T_g

Table 4. The top 10 physical descriptors and their absolute weight ratio from the Lasso model

Name	Description	Block	Ratio
AVS_B(i)	average vertex sum from Burden matrix weighted by ionization potential	2D matrix-based descriptors	0.0684
NssCH2	number of atoms of type ssCH2	atom-type E-state indices	0.0272
F02[C-N]	frequency of C–NA topological distance 2	2D atom pairs	0.0181
nHM	number of heavy atoms	constitutional indices	0.0145
BIC2	bond information content index (neighborhood symmetry of 2-order)	information indices	0.0138
NsCH3	number of atoms of type sCH3	atom-type E-state indices	0.0137
B03[F-F]	presence/absence of F–F at topological distance 3	2D atom pairs	0.0120
nCq	number of total quaternary C(sp3)	functional group counts	0.0113
nCrs	number of ring secondary C(sp3)	functional group counts	0.0098
C-006	CH2RX	atom-centered fragments	0.0097

values. Combined with the powerful and transferable DNN model,[82] the DNN_Fingerprint model trained from dataset-1 demonstrates the best performance on dataset-2 and a new experimental dataset of 32 conjugated polymers. We should emphasize that, if we only derive the Morgan fingerprints by hashing all the substructures into 2,048-bits, without considering their number of occurrences, the trained DNN model cannot reasonably predict the T_g values of these 32 conjugated polymers (see Figure S6 the supplemental experimental procedures for detailed results). Extensive studies using molecular descriptors, fingerprints, or images alone (Table 1) lead to well-trained ML models that are applicable for a certain category of polymers, but how well these models are suitable to predict other polymers is not getting much attention. Here, we demonstrate an appropriate feature representation through large dataset training, MD simulations, and experimental dataset verification, particularly from a perspective of the model's good transferability and generalization. The Morgan fingerprints with their number of occurrences are found most suitable in terms of T_g prediction, due to the encoded information of substructures and polymerization.

Machine learns physical rules for polymer T_g values

One of the challenges in using ML models for property predictions of organic molecules and polymers is correlating these predictions with meaningful physical quantities.[16,83] This is the major driving force of current research activities in interpretable artificial intelligence and ML methods.[84–86] Although our DNN_Fingerprint model demonstrates the best predictivity and transferability, it uses the fingerprinting representation of polymers, leading to the difficulty of pinpointing a specific set of physical quantities that are important in the prediction of a polymer's T_g. On the contrary, the performances of Lasso_Descriptor and Lasso_Fingerprint models are not as ideal as DNN_models, but they are still useful to establish reasonable correlations between a polymer's chemical structure and T_g with $R^2 > 0.7$ (cf. Figure 3). Furthermore, the Lasso method has an advantage for feature selection and extraction.[76,87] By applying L1-norm regularization on the weights, unimportant features are shrunk, and only important features are left. The feature importance is directly indicated by the obtained weight for each feature.[87,88]

Focusing on molecular descriptors, the Lasso_Descriptor model finds 444 descriptors having non-zero weights. More than 50% of the total absolute weight is contributed by 61 features. These features are considered important in determining T_g. The top 10 physical descriptors are listed in Table 4 (see the full list in Table S1 of the supplemental experimental procedures). Descriptors, such as "frequency of C-N at topological distance 2," "number of heavy atoms," "number of total quaternary C(sp3)," etc., are revealed to be principle features associated with the T_g of polymers. These structural and chemical parameters are expected to be the essential constituents of polymers in terms of T_g.

Several topological descriptors, such as F02[C-N] and B03[F-F], appear in the discovered top features as they encode the spatial relationship of the polymer backbone, such as the molecular size and free volume. Using topological descriptors alone is considered to be enough for a $T_g$prediction model when dealing with a very limited dataset of 251 polymers.[44] However, our Lasso_Descriptor model, dealing with a larger dataset, indicates the same level of importance as other factors, such as the functional group counts. Eleven functional groups (see the full list in Table S1 of the supplemental experimental procedures), such as "number of ring secondary C(sp3)," "number of hydroxyl groups," and "number of primary amines (aromatic)" are identified key factors affecting the T_g of polymers. They demonstrate no less significance than topological descriptors, and some critical functional groups are found to be good indicators to identify high-T_g or low-$T_g$polymers as shown later.

Focusing on Morgan fingerprints, the Lasso_Fingerprint model examines local substructures in a similar way. Among the 124 most common substructures found in dataset-1, 85 substructures have non-zero weights, and 18 substructures contribute more than 50% of the total absolute weight. These 18 substructures with the highest absolute weight are presented in Figure 4. These substructures also provide us physical insights into the T_g of polymers, including the importance of aromatic compounds[89] (substructures 16406, 24417, 17135, 17618, 11337, 11881, and 4916) and functional groups containing oxygen and nitrogen atoms (substructures 16406, 17748, 426, 24417, 770, 11337, 23586, 11881, 4916, 7305, and 24993), which indicates the positive influence of hydrogen bonds on T_g.[90] Also, some of these

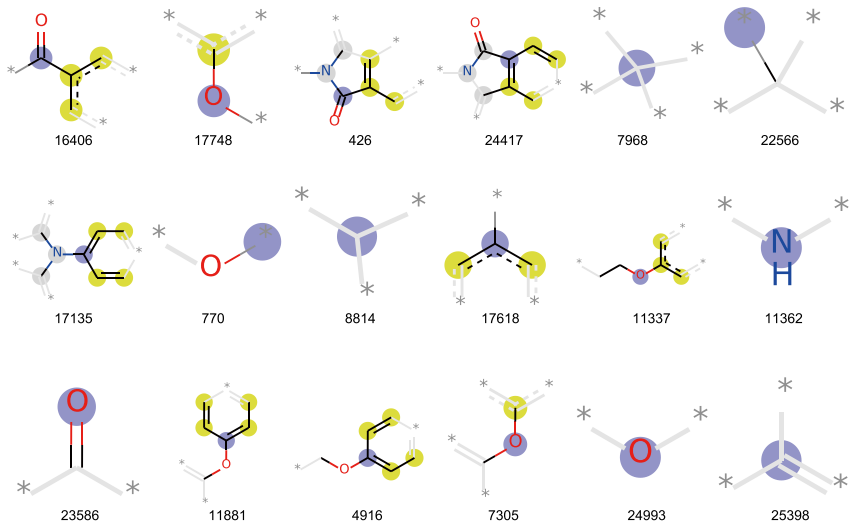


Figure 4. Substructures with the highest absolute weight based on Morgan fingerprint and Lasso ML model

The central atom of the substructures is highlighted in blue. Aromatic atoms are highlighted in yellow. Connectivity of Atoms is highlighted in light gray.

substructures are highly related to the important physical descriptors shown in Table 4, providing cross-validations between these two ML models.

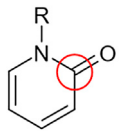
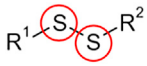
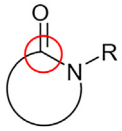
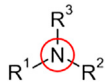
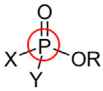
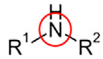
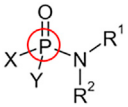
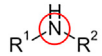
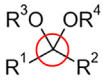
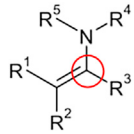
Besides the physical insights revealed by the Lasso regression models, critical functional groups can also be identified for their contributions to polymer T_g values as a posteriori analysis. Here, we can examine the polymers with high/low T_g values and their common characteristics (functional groups), and thereby gain insights into what physical quantities are important for enhancing/reducing their T_g values. We process all the polymers in dataset-1 through the Checkmol[60] package, and identify the functional groups only occurring in high-T_g (>200°C) and low-T_g (<50°C) polymers. These functional groups are listed in Table 5, where each functional group's key atom is highlighted in the red circle. For high-T_g polymers, we find that the functional groups, such as oxohetarene, lactam, amine, and enamine, play critical roles in their high-temperature property. In contrast, the functional groups, such as disulfide, phosphoric acid, and acetal, are only shown in the low-T_g polymers. These observations are consistent with the key substructures discovered from the fingerprint (Figure 4). For example, the substructures with oxygen "O" atom are revealed to be highly correlated to a polymer's T_g, and the most exclusive functional groups also involve the oxygen O atom in either high-T_g or low-T_g polymers, highlighting its important contribution to the T_g. Therefore, it is evident that the ML models indeed capture the critical features affecting a polymer's T_g.

These key features not only provide physical insights into understanding how the molecular structures influence a polymer's T_g, but also are design motifs that are important in the inverse molecular design of high-temperature polymers. For instance, the generative ML models, such as variational autoencoders (VAE)[91,92] and generative adversarial networks (GAN),[93,94] when integrated with reinforcement learning (RL),[95,96] can take into account the importance of these physical and chemical features. Such a strategy of combining the predictive ML model and generative ML model has been utilized in the inverse molecular design of small-drug-like molecules and organic molecules.[97,98] Successful examples include the chemical VAE,[99] ReLeaSE (reinforcement learning for structural evolution),[100] and ORGANIC (objective-reinforced generative adversarial network for inverse-design chemistry).[101] The generative ML model serves as an agent in generating molecules, while the predictive model acts as an external world to monitor the generation action taken by the agent. According to the feedback, either a reward or penalty can be assigned. Through training, the agent or the generative model learns to make good sequences of decisions in molecular generation toward a maximum reward. Therefore, our predictive ML model demonstrates its potential to be integrated with an inverse molecular design framework for high-temperature polymers or polymers with tailored T_g values.

High-throughput screening of high-temperature polymers

Since the DNN_Fingerprint model demonstrates the best transferability from dataset-1 to dataset-2 and to a new experimental dataset (32 conjugated polymers), we adopt this ML model for high-throughput screening to identify promising candidates for high-temperature polymers. Dataset-1, with 6,923 real polymers, has nearly 2,000 polymers with T_g larger than 200°C, as shown in Figure 5. These polymers have the great potential to be used in a harsh environment with high temperatures, but more candidates are still desired as many of these 2,000 polymers might not be easily synthesized and processed.[1] Dataset-2 and dataset-3, with 5,690 real polymers and one million hypothetical polymers, respectively, form a promising candidate pool for the screening of high-T_g polymers. Here, we aim to identify the polymers with T_g values larger than 200°C, because the T_g for high-temperature PEEK polymer is about 143°C.[102] Almost all predicted T_g values for dataset-2 and dataset-3 remain in the same range of dataset-1 (−118°C to 495°C), as shown in Figure 5. Excitingly, the population of potential promising candidates has been significantly increased. For example, dataset-1 has about 2,000 known polymers with $T_g \geq$ 200°C. Through our DNN_Fingerprint model, we find an additional 1,000 and 65,000 new candidates in dataset-2 and dataset-3 with $T_g \geq$ 200°C, respectively. Thus, through this high-throughput screening, we find 30 times more promising candidates for high-temperature polymers, in comparison with the 2,000 known high-temperature polymers in dataset-1. If we consider a harsher environment with required $T_g \geq$ 300°C (comparable with melting temperature of lead, 328°C), dataset-1, dataset-2, and dataset-3 have 309, 249, and 3,567 polymers, respectively, that can potentially satisfy this requirement.[1] Again, our high-throughput screening method identifies 11 times more promising candidates from dataset-2 and dataset-3 compared

Table 5. Important functional groups recognized using the Checkmol package

Within low-T_g polymers (<50°C)		Within high-T_g polymers (>200°C)	
Orthocarboxylic acid derivative	 R = H, alkyl, aryl X = OH, alkoxy, aryloxy, (substituted) amino, etc.	Oxohetarene	 R = H, alkyl, aryl
Disulfide	 R^1 = alkyl, aryl R^2 = alkyl, aryl	Lactam	 R = H, alkyl, aryl
Phosphoric acid derivative	 X, Y, Z = O, N, Hal residue	Tertiary arom_amine	 R^1 = aryl R^2 = aryl R^3 = aryl
Phosphoric acid ester	 R = alkyl, aryl X, Y = any O, N, Hal residue	Secondary aromatic amine	 R^1 = aryl R^2 = aryl
Phosphoric acid amide	 R^1, R^2 = H, alkyl, aryl X, Y = any O, N, Hal residue	Secondary mixed amine (aryl alkyl)	 R^1 = alkyl R^2 = aryl
Acetal	 R^1 = H, alkyl, aryl R^2 = H, alkyl, aryl R^3 = alkyl, aryl R^4 = alkyl, aryl	Enamine	 R^1 = H, acyl, alkyl, aryl R^2 = H, acyl, alkyl, aryl R^3 = H, acyl, alkyl, aryl R^4 = H, acyl, alkyl, aryl R^5 = H, acyl, alkyl, aryl

with dataset-1. The ML high-throughput screening for high-temperature polymers overcomes the challenges from theoretical analysis or MD simulations. Theoretical equations derived using small groups of polymers have difficulties in handling polymers of different categories, and are therefore not applicable to all data points of the vast chemical space. MD simulations, although capable of computing T_g vaues of various kinds of polymers, are restricted by the computational cost considering the vast amount of candidates to be screened. However, our high-throughput screening method processes the one million hypothetical polymers efficiently with proven reliability for T_g estimation.

We then focus our attention on the top four high-temperature polymers, with ML-predicted T_g > 400°C. These four polymers are unknown and hypothetical, although they share similar chemical structures as the other known high-temperature polymers, e.g., aromatic rings, sulfone groups, oxygen linkages, and amine groups. Each of these groups is highlighted during our analysis of the ML models as being related to the high-temperature properties of polymers (Figure 4; Table 5). Without making any assumptions or premises for the ML model, it is observed that the structures of the screened top four high-temperature polymers well follow the general rule controlling the T_g of polymers. The backbone structure with rigid benzene rings contributes to the stiffness of the chain, which is known to play a major role in determining the T_g of a polymer.[50,103,104] Also, there are no long alkyl chains that lead to lower glass transition.[105] Although the similar sulfur-containing polyimides, such as poly[(2,8-dimethyl-5,5-dioxodibenzothiophene-3,7-diamine)-alt-(biphenyl-3,3':4,4′-tetracarboxylic dianhydride)] (polymer ID: P130369 in PoLyInfo), have been tested with T_g values as high as 490°C,[106] the T_g values of these hypothetical polymers have not yet been reported. We take advantage of MD simulations to build all-atom molecular models for these hypothetical polymers and predict their T_g values (more details are given in the supplemental experimental procedures). As shown in Figure 5, our physics-based MD simulations confirm that these hypothetical polymers indeed have ultra-high T_g values. Furthermore, we find that the MD-predicted and ML-predicted T_g values are in relatively good agreement with each other (within the error of the prediction), indicating that the ML model could be

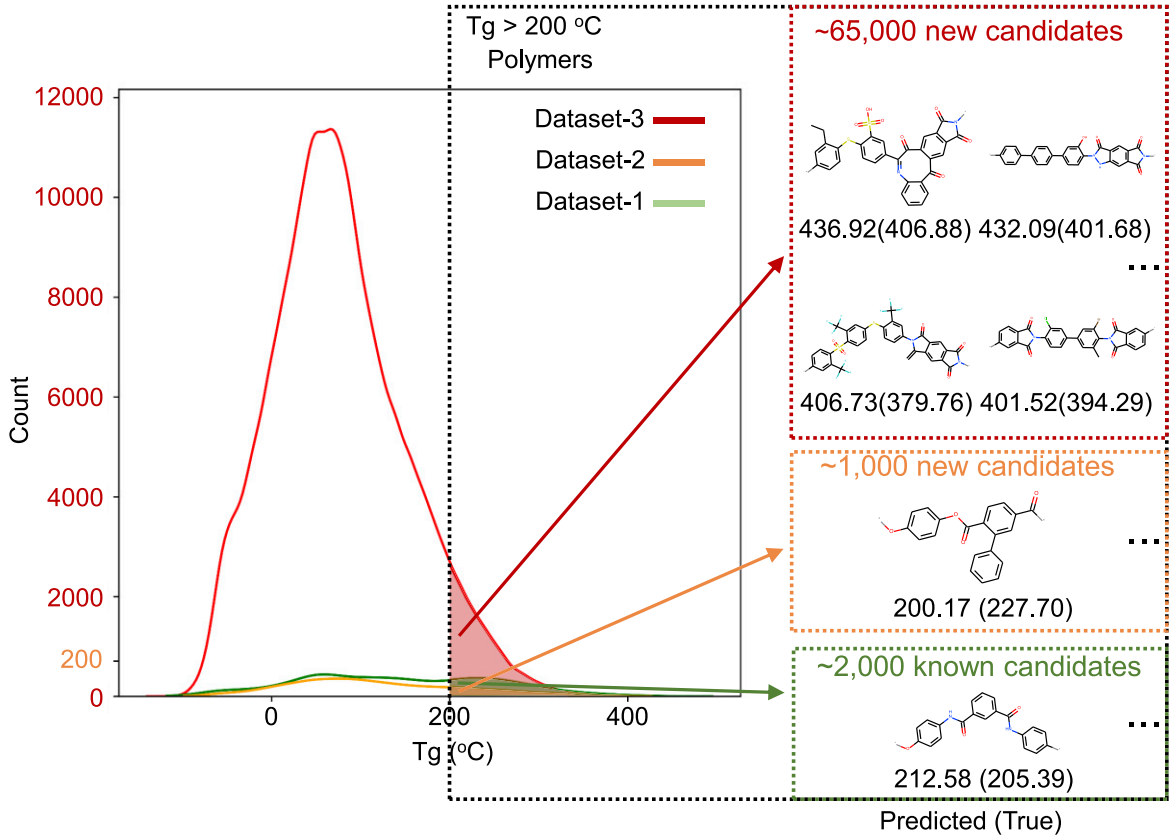


Figure 5. High-throughput screening of high T_g polymers with the DNN_Fingerprint model

The T_g distribution of the dataset-1, dataset-2, and dataset-3 are plotted in green, yellow, and red, respectively. The polymer samples on the right are following by their predicated T_g and true T_g values. For the sample in dataset-1 (green box), true T_g is the collected experimental value. For the samples in dataset-2 (yellow box) and dataset-3 (red box), true T_g is the MD-simulated value. More than 1,000 real polymers and 65,000 hypothetical polymers were discovered with T_g > 200°C.

used as a predictive tool for screening of previously unexplored chemical spaces for high-temperature polymers.

The key substructures (Figure 4) and functional groups (Table 5) related to the high-T_g polymers are revealed based on dataset-1. Their important roles are further confirmed on the identified high-T_g polymers with ML-predicted T_g > 200°C from dataset-2 and dataset-3. The key substructures of high-T_g polymers in dataset-1 (2,268 polymers), dataset-2 (1,155 polymers), and dataset-3 (65,283 polymers) are compared in Figure 6A (more details are given in Table S5 of the supplemental experimental procedures). For example, the substructure "16406" (a center carbon connected to aromatic compounds and oxygen) is recognized with percentages of 15.04%, 16.54%, and 27.55% of high-$T_g$polymers in dataset-1, dataset-2, and dataset-3, respectively. This indicates that the contributions of this substructure to the high-$T_g$polymers are similar across these different datasets. As mentioned above, one of the most important contributions comes from substructure "23586"—a single oxygen side chain, which consists of 53.40%, 53.16%, and 76.05% high-T_g polymers in dataset-1, dataset-2, and dataset-3, respectively. Overall, most of these 18 key substructures' contributions in different datasets are quite similar. Their comparable influences also explain the good transferability of the ML model based on the Morgan fingerprints. The frequency of occurrence is also an important aspect because of the probability of a substructure emerging during the inverse molecular design of high-T_g polymers. In terms of the functional groups, the six key functional groups exclusive to high-T_g polymers are compared in Figure 6B in a similar manner (also see Table S6 for detailed results). Interestingly, the six recognized functional groups are special ones only found in a few high-T_g polymers. For instance, the secondary aromatic amine functional group is identified in about 0.13% of the high-T_g polymers in dataset-1, while 3.32% of the high-T_g polymers in dataset-3 are found to have this functional group. Although training dataset-1 shows a quite negligible 0.13% of this functional group, its importance is successfully captured by the ML model using Morgan fingerprints and then demonstrated in dataset-3. In addition, we generally observe that polymers containing amine groups, oxygen along the backbone, and/or nitrogen rings, demonstrate high-temperature properties.[1] In short, our ML models for the chemistry-T_g relation of polymers seems to pinpoint meaningful physical-chemistry insights that can be used to enhance high-temperature performance and may be further utilized in the

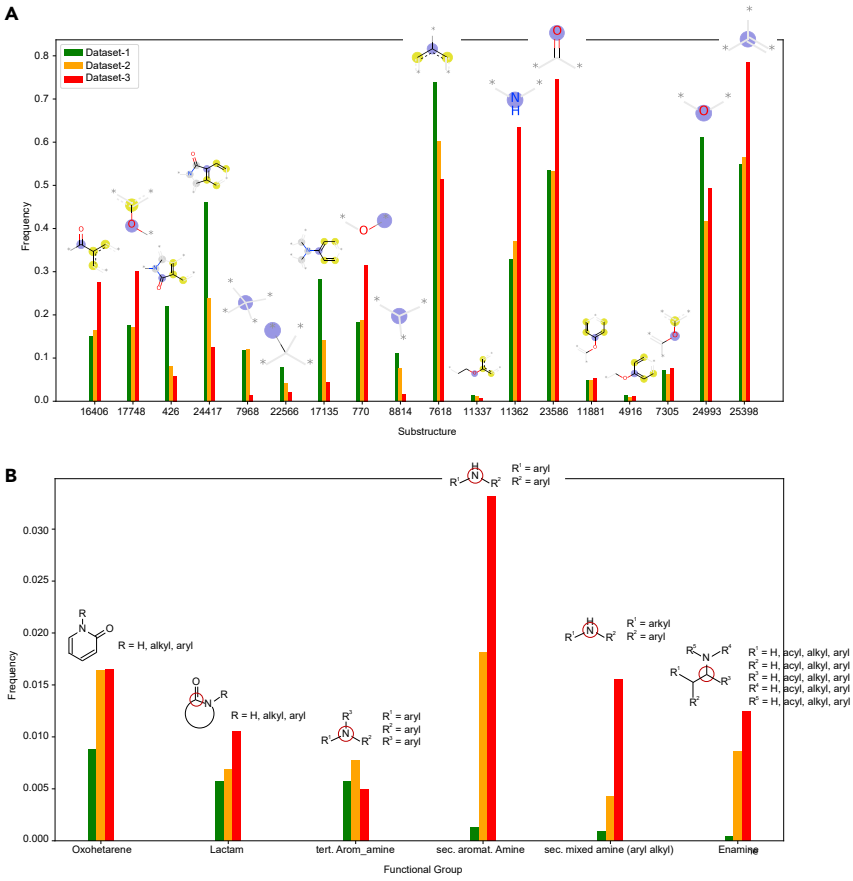


Figure 6. Comparison of key substructures and functional groups in high-T_g(>200°C) polymers

(A) Comparison of the 18 substructures recognized in Figure 4.

(B) Comparison of the six high-T_g-related functional groups recognized in Table 5.

inverse molecular design of high-T_g polymers that have not been experimentally studied.

Concluding remarks

Quantitatively predicting a polymer's T_g from its chemical structure is a significant challenge in material science and engineering, chemistry, and polymer science fields. Here, we use an ML-based approach to correlate a polymer's chemical structure with its T_g, taking advantage of a large and diverse dataset collected from PoLyInfo. The transferability and generalization ability of ML models are particularly focused and demonstrated by utilizing a large dataset of different categories of polymers. We consider three different feature representations of polymer's repeating unit, such as molecular descriptors, Morgan fingerprints, and images, and three different ML models, e.g., Lasso, DNN, and CNN. All of these ML models demonstrate comparable performances in training and testing on the experimentally available dataset-1. However, only the DNN_Fingerprint model exhibits the best transferability to dataset-2 with distinct substructures from dataset-1. We find that this excellent transferability is attributed to the dynamic representation of Morgan fingerprints, as they can evolve to include new substructures encountered. Furthermore, our Morgan fingerprints take into account the chemical connectivity between neighboring repeating units and the frequency of occurrence of different substructures, which play important roles in determining a polymer's T_g. Although Morgan fingerprints ignore all high-order polymer descriptors, e.g., stereoregularity, polarity, and chain length, the DNN_Fingerprint model gives satisfactory predictions on the T_g values of unknown polymers from dataset-2 and dataset-3. As we have discussed, choosing the appropriate feature representation for polymeric materials remains an open question in the ML field, which is also highly dependent on the specific application.[16,17,48,83]

Our ML approaches are designed with the specific goal to quickly predict a polymer's T_g from an extremely large set of known (dataset-2) and hypothetical (dataset-3) polymers. Such a high-throughput screening allows us to perform posterior correlations between high-T_g polymers with common functional groups and chemical substructures. These observations allow us to quantify physical quantities that are important in determining a polymer's T_g. For instance, our Lasso regression models reveal principal T_g-related features, including 61 molecular descriptors and 18 chemical substructures. Also, the functional groups exclusive to high-T_g (>200°C) or low-T_g (<50°C) polymers are further identified, which can cross-validate our Lasso regression models. It allows us to determine which chemical elements and molecular structures are worth experimental studies in molecular engineering and design of high-temperature polymers, leading to a molecular understanding of a polymer's T_g. With the DNN_Fingerprint model for high-throughput screening of nearly one million hypothetical polymers, we find more than 65,000 promising candidates with T_g > 200°C, which is 30 times more than existing known high-temperature polymers (~2,000 from dataset-1). The discovery of this large number of promising candidates will be of significant interest in the development and design of high-temperature polymers. The same task is very difficult to accomplish by screening with either theoretical equations or MD simulation due to their limitations in dealing with such large and diverse datasets. In summary, our study demonstrates that ML is a powerful method for the prediction and rapid screening of high-temperature polymers, particularly with growing large sets of experimental and computational data for polymeric materials. The key molecular descriptors and chemical substructures informed by ML models, combined with identified chemical functional groups, are important design motifs for the molecular engineering of high-temperature or high-performance polymers in an inverse materials design task.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Ying Li is the lead contact of this study and can be reached by e-mail: yingli@engr.uconn.edu.

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data and code are available at https://github.com/figotj/Polymer_Tg_.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100225.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Air Force Office of Scientific Research through the Air Force's Young Investigator Research Program (FA9550-20-1-0183; program manager: Dr. Ming-Jen Pan) and the National Science Foundation (CMMI-1934829). Y.L. would like to give thanks for the support from 3M's Non-Tenured Faculty Award. This research also benefited in part from the computational resources and staff contributions provided by the Booth Engineering Center for Advanced Technology (BECAT) at UConn. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Department of Defense. The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (Frontera project and the National Science Foundation award 1818253) for providing HPC resources that have contributed to the research results reported within this paper.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.L.; methodology, L.T., G.C., and Y.L.; software, L.T. and G.C.; validation, L.T.; formal analysis, L.T., G.C., and Y.L.; investigation, L.T.; resources, Y.L.; data curation, L.T.; writing—original draft, L.T.; writing—review & editing, L.T., G.C., and Y.L.; visualization, L.T.; supervision, Y.L.; funding acquisition, Y.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 15, 2020
Revised: January 21, 2021
Accepted: March 2, 2021
Published: April 9, 2021

REFERENCES

1. Hergenrother, P.M. (2003). The use, design, synthesis, and properties of high performance/high temperature polymers: an overview. High Perform. Polym. *15*, 3–45.
2. Meador, M.A. (1998). Recent advances in the development of processable high-temperature polymers. Annu. Rev. Mater. Sci. *28*, 599–630.
3. Mittal, K.L. (2005). Polyimides and Other High Temperature Polymers: Synthesis, Characterization and Applications, *Vol. 3* (CRC Press).
4. Sperati, C.A., and Starkweather, H.W. (1961). Fluorine-containing polymers. II. Polytetrafluoroethylene. In Fortschritte Der Hochpolymeren-Forschung (Springer), pp. 465–495.
5. Petrie, E. (2012). Extreme high temperature thermoplastics: gateway to the future or the same old trail. Pop. Plast. Packag, *57*, 30–43.
6. Imai, Y. (1995). Synthesis of novel organic-soluble high-temperature aromatic polymers. High Perform. Polym. *7*, 337–345.
7. Li, Q., Chen, L., Gadinski, M.R., Zhang, S., Zhang, G., Li, H.U., Iagodkine, E., Haque, A., Chen, L.-Q., and Jackson, T.N. (2015). Flexible high-temperature dielectric materials from polymer nanocomposites. Nature *523*, 576–579.
8. Kaminsky, W., Rabe, O., Schauwienold, A.-M., Schupfner, G., Hanss, J., and Kopf, J. (1995). Crystal structure and propene polymerization characteristics of bridged zirconocene catalysts. J. Organomet. Chem. *497*, 181–193.
9. McLain, S.J., Feldman, J., McCord, E.F., Gardner, K.H., Teasley, M.F., Coughlin, E.B., Sweetman, K.J., Johnson, L.K., and Brookhart, M. (1998). Addition polymerization of cyclopentene with nickel and palladium catalysts. Macromolecules *31*, 6705–6707.
10. Kobayashi, S., Matsuzawa, T., Matsuoka, S.-i., Tajima, H., and Ishizone, T. (2006). Living anionic polymerizations of 4-(1-adamantyl) styrene and 3-(4-vinylphenyl)-1,1′-biadamantane. Macromolecules *39*, 5979–5986.
11. Fetters, L.J., and Morton, M. (1969). Synthesis and properties of block polymers. I. Poly-α-methylstyrene-polyisoprene-poly-α-methylstyrene. Macromolecules *2*, 453–458.
12. Kobayashi, S., Kataoka, H., Goseki, R., and Ishizone, T. (2018). Living anionic polymerization of 4-(1-adamantyl)-α-methylstyrene. Macromol. Chem. Phys. *219*, 1700450.
13. Wang, W., Schlegel, R., White, B.T., Williams, K., Voyloy, D., Steren, C.A., Goodwin, A., Coughlin, E.B., Gido, S., and Beiner, M. (2016). High temperature thermoplastic elastomers synthesized by living anionic polymerization in hydrocarbon solvent at room temperature. Macromolecules *49*, 2646–2655.
14. Nakahara, A., Satoh, K., and Kamigaito, M. (2012). Random copolymer of styrene and diene derivatives via anionic living polymerization followed by intramolecular Friedel–Crafts cyclization for high-performance thermoplastics. Polym. Chem. *3*, 190–197.
15. Cai, Y., Lu, J., Zuo, D., Li, S., Cui, D., Han, B., and Yang, W. (2018). Extremely high glass transition temperature hydrocarbon polymers prepared through cationic cyclization of highly 3,4-regulated poly(phenyl-1,3-butadiene). Macromol. Rapid Commun. *39*, 1800298.
16. Chen, G., Shen, Z., Iyer, A., Ghumman, U.F., Tang, S., Bi, J., Chen, W., and Li, Y. (2020). Machine-learning-assisted de novo design of organic molecules and polymers: opportunities and challenges. Polymer *12*, 163.
17. Batra, R., Song, L., and Ramprasad, R. (2020). Emerging materials intelligence ecosystems propelled by machine learning. Nat. Rev. Mater. 1–24.
18. Van Krevelen, D.W., and Te Nijenhuis, K. (2009). Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions (Elsevier).
19. Dalnoki-Veress, K., Forrest, J., Murray, C., Gigault, C., and Dutcher, J. (2001). Molecular weight dependence of reductions in the glass transition temperature of thin, freely standing polymer films. Phys. Rev. E *63*, 031801.
20. Privalko, V., and Lipatov, Y.S. (1974). Glass transition and chain flexibility of linear polymers. J. Macromol. Sci. Phys. *9*, 551–564.
21. Yi, L., Li, C., Huang, W., and Yan, D. (2014). Soluble aromatic polyimides with high glass transition temperature from benzidine containing tert-butyl groups. J. Polym. Res. *21*, 572.
22. Huang, Y.-J., and Horng, J.C. (1998). Effects of thermoplastic additives on mechanical properties and glass transition temperatures for styrene-crosslinked low-shrink polyester matrices. Polymer *39*, 3683–3695.
23. Hiemenz, P.C., and Lodge, T.P. (2007). Polymer Chemistry (CRC press).
24. Wiff, D., Altieri, M., and Goldfarb, I. (1985). Predicting glass transition temperatures of linear polymers, random copolymers, and cured reactive oligomers from chemical structure. J. Polym. Sci. Polym. Phys. Ed. *23*, 1165–1176.
25. Barton, J.M. (1970). Relation of glass transition temperature to molecular structure of addition copolymers. In Journal of Polymer Science Part C: Polymer Symposia (Wiley Online Library), pp. 573–597.

26. Weyland, H., Hoftyzer, P., and Van Krevelen, D. (1970). Prediction of the glass transition temperature of polymers. Polymer *11*, 79–87.

27. Dudowicz, J., Freed, K.F., and Douglas, J.F. (2005). The glass transition temperature of polymer melts. J. Phys. Chem. B *109*, 21285–21292.

28. Xie, R., Weisen, A.R., Lee, Y., Aplan, M.A., Fenton, A.M., Masucci, A.E., Kempe, F., Sommer, M., Pester, C.W., and Colby, R.H. (2020). Glass transition temperature from the chemical structure of conjugated polymers. Nat. Commun. *11*, 1–8.

29. Han, J., Gee, R.H., and Boyd, R.H. (1994). Glass transition temperatures of polymers from molecular dynamics simulations. Macromolecules *27*, 7781–7784.

30. Choi, J., Yu, S., Yang, S., and Cho, M. (2011). The glass transition and thermoelastic behavior of epoxy-based nanocomposites: a molecular dynamics study. Polymer *52*, 5197–5203.

31. Patrone, P.N., Dienstfrey, A., Browning, A.R., Tucker, S., and Christensen, S. (2016). Uncertainty quantification in molecular dynamics studies of the glass transition temperature. Polymer *87*, 246–259.

32. Buchholz, J., Paul, W., Varnik, F., and Binder, K. (2002). Cooling rate dependence of the glass transition temperature of polymer melts: molecular dynamics study. J. Chem. Phys. *117*, 7364–7372.

33. Sharma, P., Roy, S., and Karimi-Varzaneh, H.A. (2016). Validation of force fields of rubber through glass-transition temperature calculation by microsecond atomic-scale molecular dynamics simulation. J. Phys. Chem. B *120*, 1367–1379.

34. Katritzky, A.R., Sild, S., Lobanov, V., and Karelson, M. (1998). Quantitative structure–property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers. J. Chem. Inf. Comput. Sci. *38*, 300–304.

35. Schut, J., Bolikal, D., Khan, I., Pesnell, A., Rege, A., Rojas, R., Sheihet, L., Murthy, N., and Kohn, J. (2007). Glass transition temperature prediction of polymers through the mass-per-flexible-bond principle. Polymer *48*, 6115–6124.

36. Camelio, P., Cypcar, C.C., Lazzeri, V., and Waegell, B. (1997). A novel approach toward the prediction of the glass transition temperature: application of the EVM model, a designer QSPR equation for the prediction of acrylate and methacrylate polymers. J. Polym. Sci. A Polym. Chem. *35*, 2579–2590.

37. Jha, A., Chandrasekaran, A., Kim, C., and Ramprasad, R. (2019). Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. Model. Simul. Mater. Sci. Eng. *27*, 024002.

38. Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. (2018). Polymer genome: a data-powered polymer informatics platform for property predictions. J. Phys. Chem. C *122*, 17575–17585.

39. Ramprasad, M., and Kim, C. (2019). Assessing and improving machine learning model predictions of polymer glass transition temperatures. arXiv, preprint arXiv:1908.02398.

40. Katritzky, A.R., Kuanar, M., Slavov, S., Hall, C.D., Karelson, M., Kahn, I., and Dobchev, D.A. (2010). Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. Chem. Rev. *110*, 5714–5789.

41. Wu, S., Kondo, Y., Kakimoto, M.-a., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., and Shiomi, J. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. Npj Comput. Mater. *5*, 1–11.

42. Liu, W., and Cao, C. (2009). Artificial neural network prediction of glass transition temperature of polymers. Colloid. Polym. Sci. *287*, 811–818.

43. Pei, J.F., Cai, C.Z., Zhu, Y.M., and Yan, B. (2013). Modeling and predicting the glass transition temperature of polymethacrylates based on quantum chemical descriptors by using hybrid PSO-SVR. Macromol. Theory Simul. *22*, 52–60.

44. Kim, C., Chandrasekaran, A., Jha, A., and Ramprasad, R. (2019). Active-learning and materials design: the example of high glass transition temperature polymers. MRS Commun. *9*, 860–866.

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

46. Chandrasekaran, A., Kim, C., and Ramprasad, R. (2020). Polymer genome: a polymer informatics platform to accelerate polymer discovery. In Machine Learning Meets Quantum Physics (Springer), pp. 397–412.

47. Doan Tran, H., Kim, C., Chen, L., Chandrasekaran, A., Batra, R., Venkatram, S., Kamal, D., Lightstone, J.P., Gurnani, R., and Shetty, P. (2020). Machine-learning predictions of polymer properties with Polymer Genome. J. Appl. Phys. *128*, 171104.

48. Chen, L., Pilania, G., Batra, R., Huan, T.D., Kim, C., Kuenneth, C., and Ramprasad, R. (2020). Polymer informatics: current status and critical next steps. arXiv, preprint arXiv:2011.00508.

49. Miccio, L.A., and Schwartz, G.A. (2020). From chemical structure to quantitative polymer properties prediction through convolutional neural networks. Polymer, 122341.

50. Miccio, L.A., and Schwartz, G.A. (2020). Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks. Polymer *203*, 122786.

51. Mattioni, B.E., and Jurs, P.C. (2002). Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks. J. Chem. Inf. Comput. Sci. *42*, 232–240.

52. Higuchi, C., Horvath, D., Marcou, G., Yoshizawa, K., and Varnek, A. (2019). Prediction of the glass-transition temperatures of linear homo/heteropolymers and cross-linked epoxy resins. ACS Appl. Polym. Mater. *1*, 1430–1442.

53. Pilania, G., Iverson, C.N., Lookman, T., and Marrone, B.L. (2019). Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. J. Chem. Inf. Model. *59*, 5013–5025.

54. Palomba, D., Vazquez, G.E., and Díaz, M.F. (2012). Novel descriptors from main and side chains of high-molecular-weight polymers applied to prediction of glass transition temperatures. J. Mol. Graph. Model. *38*, 137–147.

55. Yu, X. (2010). Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers. Fibers Polym. *11*, 757–766.

56. Liu, W. (2010). Prediction of glass transition temperatures of aromatic heterocyclic polyimides using an ANN model. Polym. Eng. Sci. *50*, 1547–1557.

57. Ning, L. (2009). Artificial neural network prediction of glass transition temperature of fluorine-containing polybenzoxazoles. J. Mater. Sci. *44*, 3156–3164.

58. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., and Yamazaki, M. (2011). In PoLyInfo: Polymer database for polymeric materials design (International Conference on Emerging Intelligent Data and Web Technologies, IEEE), pp. 22–29.

59. Lee, J.-C., and Litt, M.H. (2000). Glass transition temperature-composition relationship of oxyethylene copolymers with chloromethyl/(ethylthio) methyl, chloromethyl/(ethylsulfinyl) methyl, or chloromethyl/(ethylsulfonyl) methyl side groups. Polym. J. *32*, 228–233.

60. Fox, T.G. (1956). Influence of diluent and of copolymer composition on the glass temperature of a polymer system. Bull. Am. Phys. Soc. *1*, 123.

61. Hadipeykani, M., Aghadavoudi, F., and Toghraie, D. (2020). A molecular dynamics simulation of the glass transition temperature and volumetric thermal expansion coefficient of thermoset polymer based epoxy nanocomposite reinforced by CNT: a statistical study. Phys. Stat. Mech. Appl. *546*, 123995.

62. Hadipeykani, M., Aghadavoudi, F., and Toghraie, D. (2019). Thermomechanical properties of the polymeric nanocomposite predicted by molecular dynamics. ADMT J. *12*, 25–32.

63. Ma, R., and Luo, T. (2020). PI1M: a benchmark database for polymer informatics. J. Chem. Inf. Model. *60*, 4684–4690.

64. Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754.

65. Haider, N. (2010). Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. Molecules *15*, 5079–5092.

66. Baur, E., Ruhrberg, K., and Woishnis, W. (2016). Chemical Resistance of Commodity Thermoplastics (William Andrew).

67. Simatos, D., Blond, G., Roudaut, G., Champion, D., Perez, J., and Faivre, A. (1996). Influence of heating and cooling rates on the glass transition temperature and the fragility parameter of sorbitol and fructose as measured by DSC. J. Therm. Anal. Calorim. *47*, 1419–1436.

68. McKenna, G.B. (2020). Looking at the glass transition: challenges of extreme time scales and other interesting problems. Rubber Chem. Technol. *93*, 79–120.

69. Biron, M. (2004). Detailed accounts of thermoset resins for moulding and composite matrices. In Thermosets and Composites, pp. 183–327.

70. Rudin, A., and Choi, P. (2012). The Elements of Polymer Science and Engineering (Academic Press).

71. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. *28*, 31–36.

72. Ma, R., Liu, Z., Zhang, Q., Liu, Z., and Luo, T. (2019). Evaluating polymer representations via quantifying structure–property relationships. J. Chem. Inf. Model. *59*, 3110–3119.

73. Mauri, A. (2020). alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints. In Ecotoxicological QSARs (Springer), pp. 801–820.

74. (2020). alvaDesc molecular descriptors. https://www.alvascience.com/alvadesc-descriptors/.

75. Landrum, G. (2013). RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling (Academic Press).

76. Chen, G., Shen, Z., and Li, Y. (2020). A machine-learning-assisted study of the permeability of small drug-like molecules across lipid membranes. Phys. Chem. Chem. Phys. *22*, 19687–19696.

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

78. Alkharusi, H. (2012). Categorical variables in regression analysis: a comparison of dummy and effect coding. Int. J. Educ. *4*, 202.

79. Maaten, L.V.D., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

80. Yu, K.Q., Li, Z.S., and Sun, J. (2001). Polymer structures and glass transition: a molecular dynamics simulation study. Macromol. Theory Simul. *10*, 624–633.

81. Mohammadi, M., and Davoodi, J. (2017). The glass transition temperature of PMMA: a molecular dynamics study and comparison of various determination methods. Eur. Polym. J. *91*, 121–133.

82. Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 3320–3328.

83. Sivaraman, G., Jackson, N., Sanchez-Lengeling, B., Vasquez-Mayagoitia, A., Aspuru-Guzik, A., Vishwanath, V., and de Pablo, J. (2020). A machine learning workflow for molecular analysis: application to melting points (Machine Learning: Science and Technology).

84. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv, preprint arXiv:1702.08608.

85. Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. *2*, 573–584.

86. Molnar, C. (2020). Interpretable Machine Learning (Lulu. com).

87. Fonti, V., and Belitser, E. (2017). Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics *30*, 1–25.

88. Muthukrishnan, R., and Rohini, R. (2016). In LASSO: A feature selection technique in predictive modeling for machine learning (IEEE International Conference on Advances in Computer Applications (ICACA), IEEE), pp. 18–20.

89. Naito, K., and Miura, A. (1993). Molecular design for nonpolymeric organic dye glasses with thermal stability: relations between thermodynamic parameters and amorphous properties. J. Phys. Chem. *97*, 6240–6248.

90. Painter, P.C., Graf, J.F., and Coleman, M.M. (1991). Effect of hydrogen bonding on the enthalpy of mixing and the composition dependence of the glass transition temperature in polymer blends. Macromolecules *24*, 5630–5638.

91. Kusner, M.J., Paige, B., and Hernández-Lobato, J.M. (2017). Grammar variational autoencoder. arXiv, preprint arXiv:1703.01925.

92. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In Advances in Neural Information Processing Systems, pp. 2352–2360.

93. Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv, preprint arXiv:1511.06434.

94. Goodfellow, I. (2016). NIPS 2016 tutorial: generative adversarial networks. arXiv, preprint arXiv:1701.00160.

95. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (MIT press).

96. Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement learning: a survey. J. Artif. Intell. Res. *4*, 237–285.

97. Sanchez-Lengeling, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: generative models for matter engineering. Science *361*, 360–365.

98. Elton, D.C., Boukouvalas, Z., Fuge, M.D., and Chung, P.W. (2019). Deep learning for molecular design—a review of the state of the art. Mol. Syst. Des. Eng. *4*, 828–849.

99. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. *4*, 268–276.

100. Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. Sci. Adv. *4*, eaap7885.

101. Sanchez-Lengeling, B., Outeiral, C., Guimaraes, G.L., and Aspuru-Guzik, A. (2017). Optimizing Distributions over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-Design Chemistry (ORGANIC).

102. Cebe, P., Chung, S.Y., and Hong, S.D. (1987). Effect of thermal history on mechanical properties of polyetheretherketone below the glass transition temperature. J. Appl. Polym. Sci. *33*, 487–503.

103. Fox, T.G., Jr., and Flory, P.J. (1950). Second-order transition temperatures and related properties of polystyrene. I. Influence of molecular weight. J. Appl. Phys. *21*, 581–591.

104. Gibbs, J.H., and DiMarzio, E.A. (1958). Nature of the glass transition and the glassy state. J. Chem. Phys. *28*, 373–383.

105. Jordan, E.F., Jr., Feldeisen, D.W., and Wrigley, A. (1971). Side-chain crystallinity. I. Heats of fusion and melting transitions on selected homopolymers having long side chains. J. Polym. Sci. A-1: Polym. Chem. *9*, 1835–1851.

106. Tanaka, K., Kita, H., Okamoto, K., Nakamura, A., and Kusuki, Y. (1989). Gas permeability and permselectivity in polyimides based on 3,3',4,4'-biphenyltetracarboxylic dianhydride. J. Membr. Sci. *47*, 203–215.

Patterns, Volume *2*

Supplemental information

Machine learning discovery of high-temperature polymers

Lei Tao, Guang Chen, and Ying Li

Supplemental Experimental Procedures

Table of Contents

S1. Feature representation of polymers.

The chemical species contained in three datasets include C, O, N, Cl, F, Br, I, S, Si, B, P, Sn, Fe, Na, Li, Ge, Se, K, Co, Ni, Ca, Cd, Pb, Zn, Te. One challenge when creating ML models for evaluation of polymer's T_g is choosing appropriate feature representation to describe the chemical structures being studies. Based on the SMILES notation of each polymer, three types of feature representation are derived: molecular descriptors, Morgan fingerprints, and images.

S1.1 Molecular Descriptors

In terms of molecular descriptors, the feature-generating engine alvaDesc[1] supports the calculation of about 5305 descriptors within 32 categories, ranging from constitutional indices, ring descriptors to chirality descriptors.[1, 2] The ensemble of descriptors represent the physical and chemical characteristics of polymers/molecules being studied, which have been widely adopted in the QSPR and ML models. Thus, these molecular descriptors can provide physical information regarding charges, topological indices, and functional groups, etc., of polymers. Among these 5305 descriptors, 3579 descriptors are valid for real polymers in dataset-1 and dataset-2. However, not all 3579 descriptors are applicable to the 1 million hypothetical polymers in dataset-3. Around 5% of hypothetical polymers in dataset-3 cannot be processed using the alvaDesc. But it doesn't affect too much the chemical space visualization based on molecular descriptors for dataset-3. We should emphasize that the alvaDesc cannot process "*" symbol in the SMILES notation, and thus, it misses the chemical connectivity of the repeating units.

S1.2 Morgan Fingerprints

Here we also test a fingerprinting method where the chemical connectivity in a polymer's repeating unit is represented numerically. Fingerprinting has a distinct advantage over traditional group contribution methods, where all of the possible building blocks must be defined *a priori* and remain static; fingerprinting methods are an inherently more dynamic representation because they can evolve to include materials as they are synthesized. Further, they take into account the chemical connectivity between the different units. We transformed each polymer's SMILES notation into a binary "fingerprint" using the Daylight-like fingerprinting algorithm as implemented in RDKit.[3] This topological-based approach analyzes the various fragments of a molecule containing a certain number of bonds and then hashes each fragment to produce a binary fingerprint that computationally represents the molecule. After a polymer's repeat unit was read into memory via a molfile, it was broken down into fragments containing between 1 and 7 units, and the structure was hashed into a fingerprint with 2048 bits of information to encode all of the possible connectivity pathways of the monomer. This process is repeated for each group in the molecule to generate the full fingerprint. Each bit was treated as a single feature in our model, which allows us to study the effects of various functional groups and their linkages on the polymer's glass transition temperature. This fingerprinting technique is the simplest representation of the polymer chemistry and structure that is sufficient to capture trends observed in

the experimental data.[4]

When using Morgan fingerprints through the tool RDKit,[3] radius 3 and 2048 bits are used for the calculations. Using radius 3 is able to encode large fragments of the molecular structure, with which more than 45,000 distinct substructures are detected from all datasets. If the 45,000 distinct substructures are hashed into 2048 buckets, collisions are inevitable. The 1/0 (on/off) bit of a bucket does not indicate the occurrence of a specific substructure, but indicate the occurrence of several substructures. Also, the number of occurrences for a substructure is not recorded through the buckets. To avoid the drawbacks of using buckets, we directly record each substructure and the number of occurrences. This dictionary of substructures is used for ML models. For chemical space visualization purposes, the focus is placed on the occurrence of certain substructures instead of the number of occurrences. Thus, 2048 on/off bits are sufficient and easy to process for visualization.

S1.3 Images

Based on the SMILES notation of polymers, we define an ordered list of SMILIES characters as a dictionary [‘c’, ‘n’, ‘o’, ‘C’, ‘N’, ‘F’, ‘=’, ‘O’, ‘(‘, ‘)’, ‘*’, ‘[‘, ‘]’, ‘1’, ‘2’, ‘3’, ‘#’, ‘Cl’, ‘/’, ‘S’, ‘Br’]. The dictionary creates a binary column for each character, with which one-hot encoding algorithm[5] transforms each polymer’s SMILES into a sparse matrix (2D binary images). The dimensions of all images are 21 (the number of characters in the dictionary) × 310 (the length of the longest SMILES code in the dataset). The key points of the one-hot encoding algorithm are: (1) defining a reasonable dictionary is the premise of a good model; (2) simple polymers (represented by a short SMILES code) return much sparser matrices than complex polymers (represented by a long SMILES code). Obviously, any change of dataset could lead to changes in the dictionary and changes of images, influencing the performance of a CNN model significantly.

S2. Machine Learning Models.

S2.1 Lasso model with descriptors or fingerprint

Lasso model is constructed with the Scikit-learn package.[6] The constant for L1 regularization is set as 0.1. Dataset-1 is randomly spitted into the train set (80%) and test set (20%). When using descriptors, 3579 descriptors are calculated through alvaDesc package.[1] When using fingerprints, 124 most frequent fingerprints detected by RDKit[3] package are used as inputs. The number of input substructures is a hyperparameter affecting the performance of ML model. With several rounds of hyperparameter optimization, the most frequent 124 substructures are determined as inputs for our ML models.

S2.2 DNN model with fingerprint

The DNN model is constructed with the TensorFlow library.[7] It contains one hidden layer with 8 neurons, and the rectifier activation function (ReLU) is used. Dataset-1 is randomly spitted into the train set (80%) and test set (20%). The batch size is 32, and 100 epochs are run during the training. The loss history during training is shown in **Figure S1**.

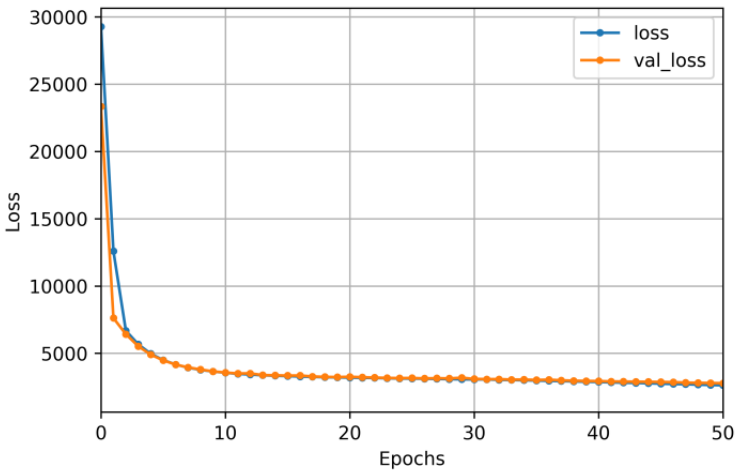


Figure S1. The loss history for training/validation dataset of the DNN model

S2.3 CNN model with images

The architecture of the CNN model consists of: (1) a first 2D convolution layer with 8 filters, ReLU activation function, and a kernel size of 10×10. (2) A second 2D convolution layer with 8 filters, ReLU activation function, and a kernel size of 4×4. (3) A max-pooling layer with pool size 2×2. (4) A dropout layer with a rate 0.3. ADAM optimizer[8] is used with a learning rate equals to 0.005. The training batch size is 16, and after 75 epochs, the final model is obtained. The dataset is split into a train set (80%) and a test set (20%). The loss history during training is shown in **Figure S2**.

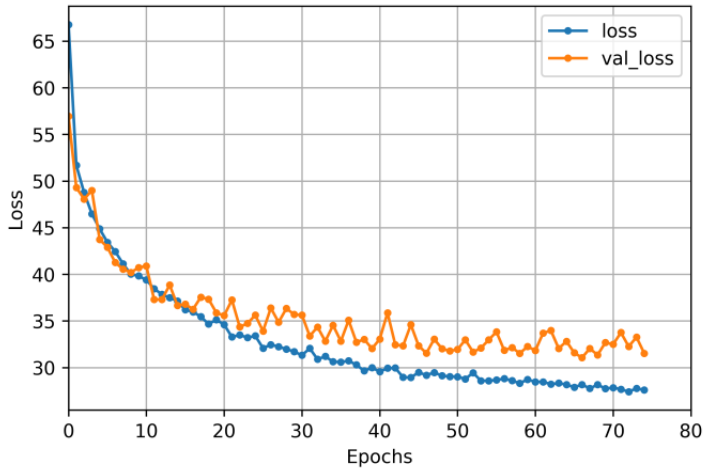


Figure S2. The loss history for training/validation dataset of the CNN model

S2.4 Distributions of the T_g values predicted by ML models.

The histograms of the T_g predicted by ML models for different datasets are displayed in **Figure S3**. For each dataset, different ML models produce a similar T_g distributions.

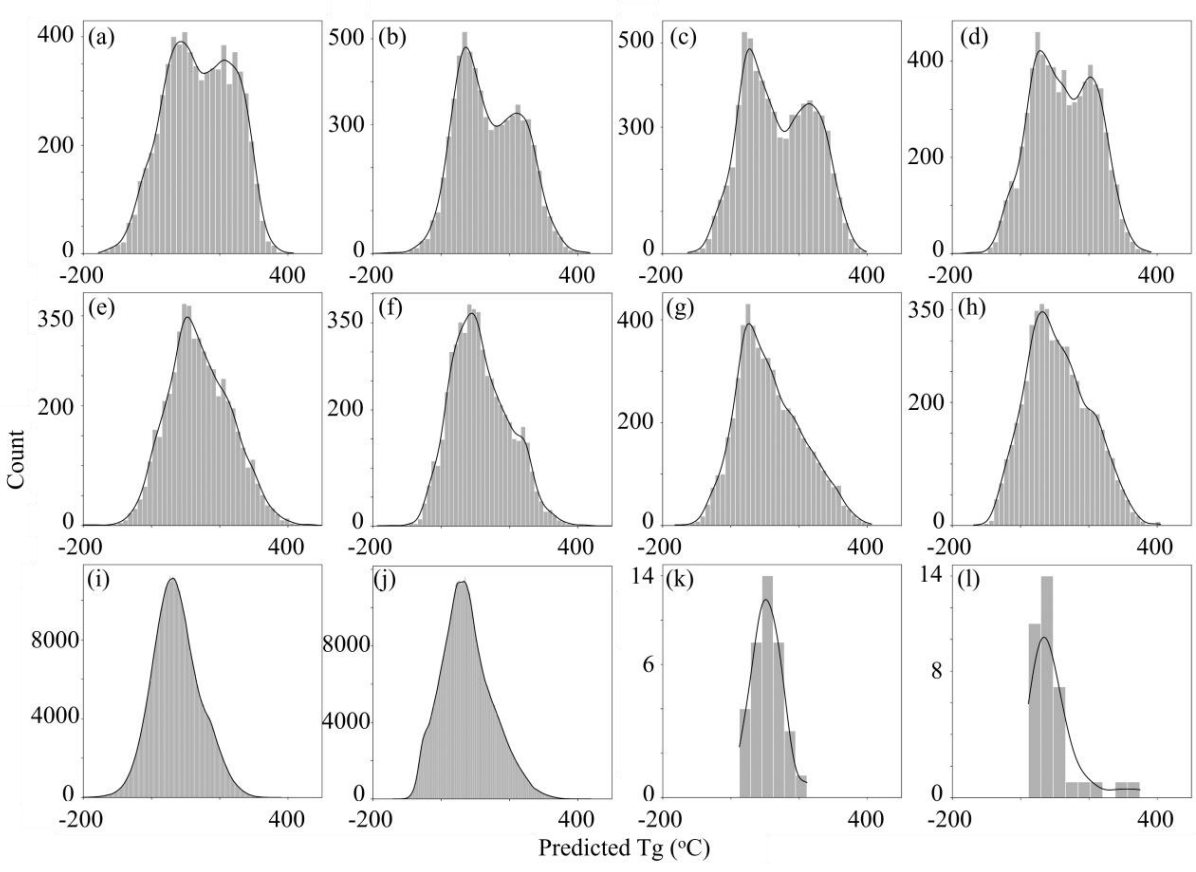


Figure S3. Histograms of the T_g predicted by ML models. (a) Lasso_Descriptor model on dataset-1. (b) Lasso_Fingerprint model on dataset-1. (c) DNN_Fingerprint model on dataset-1. (d) CNN_Image model on dataset-1. (e) Lasso_Descriptor model on dataset-2. (f) Lasso_Fingerprint model on dataset-2. (g) DNN_Fingerprint model on dataset-2. (h) CNN_Image model on dataset-2. (i) Lasso_Fingerprint model on dataset-3 (j) DNN_Fingerprint model on dataset-3. (k) Lasso_Fingerprint model on 32 conjugated polymers (l) DNN_Fingerprint model on 32 conjugated polymers.

S3. The 61 principal descriptors and their absolute weight ratio

Based on the Lasso_Descriptor model, 61 descriptors contribute more than 50% of the total absolute weight. The descriptors' detail[1, 2] and the corresponding weight ratio are list in **Table S1**.

Table S1. The 61 principal descriptors and their absolute weight ratio from the Lasso model.

No.	Name	Description	Block	ratio
836	AVS_B(i)	average vertex sum from Burden matrix weighted...	2D matrix-based descriptors	0.0684
3062	NssCH2	Number of atoms of type ssCH2	Atom-type E-state indices	0.0272
4206	F02[C-N]	Frequency of C - N at topological distance 2	2D Atom Pairs	0.0181
35	nHM	number of heavy atoms	Constitutional indices	0.0145
289	BIC2	Bond Information Content index (neighborhood s...	Information indices	0.0138
3060	NsCH3	Number of atoms of type sCH3	Atom-type E-state indices	0.0137
3553	B03[F-F]	Presence/absence of F - F at topological dista...	2D Atom Pairs	0.0120
2709	nCrs	number of ring secondary C(sp3)	Functional group counts	0.0113
2708	nCq	number of total quaternary C(sp3)	Functional group counts	0.0098
2864	C-006	CH2RX	Atom-centred fragments	0.0097
376	ChiA_X	average Randic-like index from chi matrix	2D matrix-based descriptors	0.0096
2765	nArNH2	number of primary amines (aromatic)	Functional group counts	0.0090
997	MATS2m	Moran autocorrelation of lag 2 weighted by mass	2D autocorrelations	0.0088
2842	nImidazoles	number of Imidazoles	Functional group counts	0.0087
2935	N-078	Ar-N=X / X-N=X	Atom-centred fragments	0.0084
1210	P_VSA_LogP_2	P_VSA-like on LogP, bin 2	P_VSA-like descriptors	0.0084
2711	nCrq	number of ring quaternary C(sp3)	Functional group counts	0.0083
2802	nS(=O)2	number of sulfones	Functional group counts	0.0080
4295	F03[N-N]	Frequency of N - N at topological distance 3	2D Atom Pairs	0.0079
2992	SaaN	Sum of aaN E-states	Atom-type E-state indices	0.0077
3378	B01[O-Si]	Presence/absence of O - Si at topological dist...	2D Atom Pairs	0.0077
1318	SpMaxA_EA(dm)	normalized leading eigenvalue from edge adjace...	Edge adjacency indices	0.0068
2973	P-120	C-P(X)2=X (phosphonate)	Atom-centred fragments	0.0066

3051	SssssSn	Sum of ssssSn E-states	Atom-type E-state indices	0.0066
1223	P_VSA_MR_7	P_VSA-like on Molar Refractivity, bin 7	P_VSA-like descriptors	0.0064
3425	B02[C-C]	Presence/absence of C - C at topological dista...	2D Atom Pairs	0.0063
2916	O-059	Al-O-Al	Atom-centred fragments	0.0062
2901	C-043	X--CR..X	Atom-centred fragments	0.0061
3448	B02[O-O]	Presence/absence of O - O at topological dista...	2D Atom Pairs	0.0060
1105	JGI4	mean topological charge index of order 4	2D autocorrelations	0.0060
3715	B05[F-X]	Presence/absence of F - X at topological dista...	2D Atom Pairs	0.0060
1296	Eta_D_epsiD	eta measure of hydrogen bond donor atoms	ETA indices	0.0059
286	CIC5	Complementary Information Content index (neigh...	Information indices	0.0059
2786	nROH	number of hydroxyl groups	Functional group counts	0.0059
217	X5A	average connectivity index of order 5	Connectivity indices	0.0059
1061	GATS2e	Geary autocorrelation of lag 2 weighted by San...	2D autocorrelations	0.0056
216	X4A	average connectivity index of order 4	Connectivity indices	0.0054
3362	B01[N-P]	Presence/absence of N - P at topological dista...	2D Atom Pairs	0.0053
4296	F03[N-O]	Frequency of N - O at topological distance 3	2D Atom Pairs	0.0053
3933	B08[S-Si]	Presence/absence of S - Si at topological dist...	2D Atom Pairs	0.0052
3266	CATS2D_00_NN	CATS2D Negative-Negative at lag 00	Pharmacophore descriptors	0.0051
1039	MATS4s	Moran autocorrelation of lag 4 weighted by I-s...	2D autocorrelations	0.0050
3251	CATS2D_05_PN	CATS2D Positive-Negative at lag 05	Pharmacophore descriptors	0.0050
2847	nThiazoles	number of Thiazoles	Functional group counts	0.0049
2710	nCrt	number of ring tertiary C(sp3)	Functional group counts	0.0048
1030	MATS3i	Moran autocorrelation of lag 3 weighted by ion...	2D autocorrelations	0.0047
3194	CATS2D_08_DL	CATS2D Donor-Lipophilic at lag 08	Pharmacophore descriptors	0.0047
2912	H-055	H attached to C0(sp3) with 4X attached to next C	Atom-centred fragments	0.0046
3167	CATS2D_01_DP	CATS2D Donor-Positive at lag 01	Pharmacophore descriptors	0.0046
3202	CATS2D_06_AA	CATS2D Acceptor-Acceptor at lag 06	Pharmacophore descriptors	0.0045
3482	B02[Cl-Cl]	Presence/absence of Cl - Cl at topological dis...	2D Atom Pairs	0.0044
2755	nArCO	number of ketones (aromatic)	Functional group counts	0.0043
384	SpDiam_X	spectral diameter from chi matrix	2D matrix-based descriptors	0.0043
4456	F05[N-Cl]	Frequency of N - Cl at topological distance 5	2D Atom Pairs	0.0043

3209	CATS2D_03_AP	CATS2D Acceptor-Positive at lag 03	Pharmacophore descriptors	0.0043
4090	B10[S-X]	Presence/absence of S - X at topological dista...	2D Atom Pairs	0.0042
2753	nArCHO	number of aldehydes (aromatic)	Functional group counts	0.0042
41	O%	percentage of O atoms	Constitutional indices	0.0041
2894	C-036	Al-CH=X	Atom-centred fragments	0.0040
2937	F-081	F attached to C1(sp3)	Atom-centred fragments	0.0039
4524	F06[C-Br]	Frequency of C - Br at topological distance 6	2D Atom Pairs	0.0039

S4. Molecular dynamics (MD) simulation details

Among the randomly selected 20 polymers from dataset-2, four fifth are linear polymers, and the rest are crosslinked polymers. Different model construction strategies are used. The polymer consistent force field (PCFF)[5] is used to define interatomic interactions for these molecular models. PCFF is a second-generation force field,[9-13] which has been parameterized against a wide range of experimental observables for organic compounds containing H, C, N, O, S, P, halogen atoms and ions. PCFF has a broad coverage of organic polymers, in calculations of cohesive energies, mechanical properties, compressibilities, heat capacities, elastic constants.

For linear polymers, a polymer chain is first built with 20 repeating units connected head-to-tail. After minimizing its energy, 60 polymer chains are used to construct a 3D-periodic amorphous cell, as a representative volume element (RVE). The configuration of the molecules is adjusted in a Monte Carlo fashion. Self-avoiding random walks in space are used to minimize close contacts between atoms[14], while ensuring a realistic distribution of torsion angles. A homogeneously packed cell is constructed as the linear polymer model.

The molecular simulation of the polymer crosslinking process is very challenging. In the past decade, significant efforts have been devoted to simulating the crosslinking reactions of thermosetting (e.g., epoxy) polymers through a series of singular bond changes.[2, 15-35] Overall, these crosslinking schemes can be classified into two categories[36]: single-step and multi-step crosslinking. In the single-step crosslinking process, the crosslinked network is formed by assigning crosslink bonds to nearby pairs of reactive atoms from a single (initial) coordinate snapshot.[37-39] Afterwards, MD simulations are performed to further reduce the distance between these bonded pairs. The selection of initial reactive pairs is determined by a Monte Carol (MC) method through partial path reversal moves. The MC method can effectively attempt multiple different bond partners to minimize the total length of all crosslink bonds. Such a path reversal move is equivalent to shuffling bond partners of reactive atoms before the crosslinking. Once the one-step crosslinking has been finished, the excess hydrogen atoms are removed, and atomic partial charges are assigned from the classical force field, e.g., OPLS (Optimized Potential for Liquid Simulations)[40], GAFF (general AMBER force field)[41], DREIDING[42] or PCFF[11]. In the multiple-step crosslinking process, a short cutoff distance is usually defined (about 3-4Å). Then, covalent bonds (elastic springs) are formed between reactive atoms within this cutoff distance. A short MD simulation is performed to relax the crosslinked network. Subsequently, the extra hydrogen atoms are removed with partial charges adjusted following the charge neutral principle and classical MD force field, with further MD relaxation. Then, the reactive cutoff distance is increased (about 0.5 or 1 Å) for the next round of crosslinking, until the curing degree is satisfied.[36] Essentially, the single-step and multi-step crosslinking methods use the same core polymerization scheme to identify the pairs of reactive atoms within a short distance, with MD simulations to equilibrate and relax the intermediate states during bond breaking and formation.[43-45]

For these crosslinked polymers, the multi-step crosslinking method is used. Reactive atoms are first assigned to monomers and crosslinkers. 400 of each are packed together within the 3D-periodic amorphous cell. Crosslinking steps include reaction radius updating, crosslinks creation, relaxation, and hydrogen adjustment, as aforementioned. These steps are repeated until 80 percent of reactive atoms on monomers have reacted.

After model construction, both linear polymer model and crosslinked polymer models contain ~20,000 atoms with the final cubic cell of dimension around 70 Å. Periodic boundary conditions are

applied in all directions. Using LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) package,[46] polymers are equilibrated first with a 21-step molecular dynamics equilibration protocol[43] shown in **Table S2**.

Table S2. 21-step equilibration scheme[43]

Step	Ensemble	Conditions	Duration (ps)
1	NVT	*T* max	50
2	NVT	*T* final	50
3	NPT	*T* final, 0.02 × *P* max	50
4	NVT	*T* max	50
5	NVT	*T* final	100
6	NPT	*T* final, 0.6 × *P* max	50
7	NVT	*T* max	50
8	NVT	*T* final	100
9	NPT	*T* final, *P* max	50
10	NVT	*T* max	50
11	NVT	*T* final	100
12	NPT	*T* final, 0.5 × *P* max	5
13	NVT	*T* max	5
14	NVT	*T* final	10
15	NPT	*T* final, 0.1 × *P* max	5
16	NVT	*T* max	5
17	NVT	*T* final	10
18	NPT	*T* final, 0.01 × *P* max	5
19	NVT	*T* max	5
20	NVT	*T* final	10
21	NPT	*T* final, *P* final	800

The simulation protocol contains NVT simulations, NPT simulations, compression, and decompression, etc., for polymers to achieve realistic final densities and configurations. The time step is 0.1 femtosecond throughout the whole simulation. The maximum pressure and temperature are 50,000 atmospheres and 1000 Kelvin. The temperature damping parameter and pressure damping parameter are 100 timesteps and 1000 timesteps, respectively. After 156 picoseconds equilibration, the model is simulated through a cooling process under the isobaric-isothermal ensemble (NPT) from >500 K to 100 K for 20 nanoseconds. The time step is 1 femtosecond, and the pressure is 1 atmosphere.

S5. T_g calculation from MD simulations

From the MD cooling process simulations, the specific volume vs. temperature curves are recorded for each model, as shown in **Figure S4**. Based on the curve, segments of the constant slope are for different phases (rubbery and glassy), which are used for the least-square line fit. The intersection of the two lines represents the T_g.[47-49] It is realized that the MD simulation is with a nanosecond time scale. Thus its cooling rate is much faster than in the experiments. Although the faster cooling rate results in a higher T_g, the MD simulated T_g is still proven to be close to the experimental value.[20, 50-52]

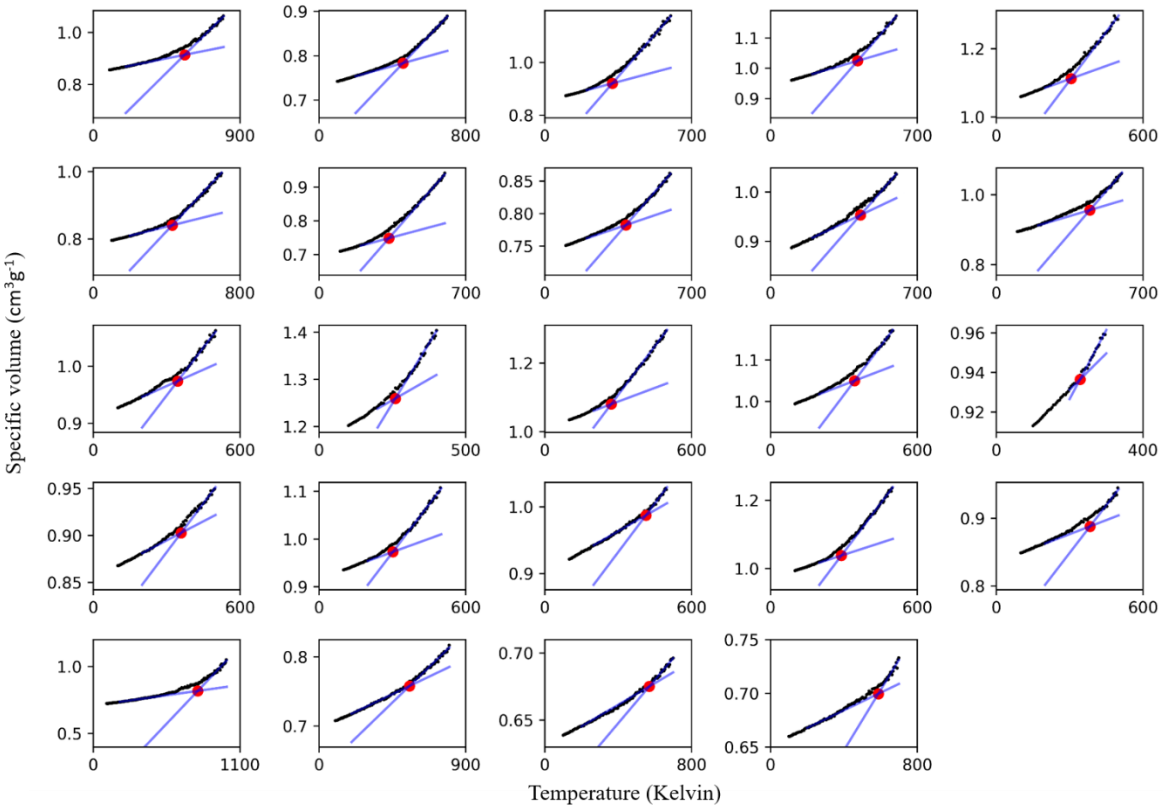
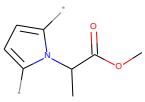
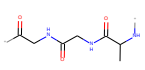
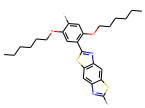
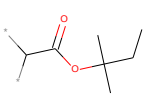


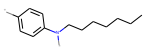
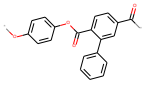
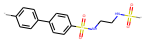
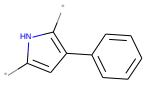
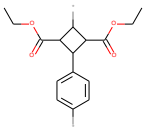
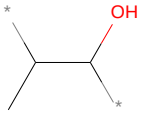
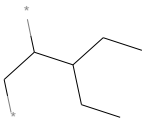
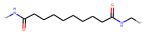
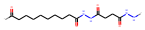
Figure S4. Plots of specific volume vs. temperature for the 20 randomly selected polymers from dataset-2 (the top 4 rows) and 4 polymers from dataset-3 (the bottom row).

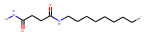
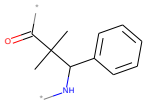
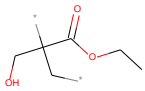
S6. ML predicted T_g v.s. MD simulated T_g for random selected 20 polymers in dataset-2

The T_g obtained from MD simulations and ML models are compared in **Table S3**. For the 20 polymers in dataset-2, four ML models are applied. The ones highlighted in red are crosslinked polymers, and the rest are linear polymers. Different MD simulation strategies are applied accordingly.

Table S3. T_g obtained for 20 polymers in dataset-2 (Celsius)

	MD	Lasso_ Fingerprint	Lasso_ Descriptor	CNN_ Image	DNN_ Fingerprint
	279.45	114.04	114.70	16.28	119.72
	208.64	133.84	83.58	173.79	107.85
	168.53	147.02	209.37	114.36	137.21
	136.57	54.27	61.25	137.65	47.85

	110.72	87.13	72.12	95.57	13.21
	227.70	174.66	193.64	31.86	200.17
	173.25	197.50	126.79	80.04	191.53
	213.41	189.22	215.08	191.01	175.68
	120.45	132.37	178.26	231.11	161.24
	164.32	150.69	125.34	13.36	142.24
	113.57	76.27	3.62	83.66	76.38
	38.97	44.91	82.33	88.83	14.50
	-32.77	-20.70	14.85	0.48	-22.02
	56.77	50.03	105.35	103.79	41.34
	46.52	71.86	96.03	260.06	54.20
	64.70	119.16	82.44	109.44	97.24

	69.22	64.40	92.89	76.03	60.70
	154.36	113.64	174.74	193.75	134.50
	-41.10	-26.13	12.22	-7.91	11.22
	22.36	59.85	68.87	81.24	54.04

S7. ML predicted T_g v.s. Experimental T_g for 32 polymers in a newly reported experimental study

Lasso_Fingerprint model and DNN_Fingerprint model are applied to the reported experimental dataset on 32 semiflexible (mostly conjugated) polymers that differ drastically in aromatic backbone and alkyl side chain chemistry.[53] The prediction performance is compared in **Figure S5**. The values ML predicted T_g and experimental values are listed in **Table S4**.

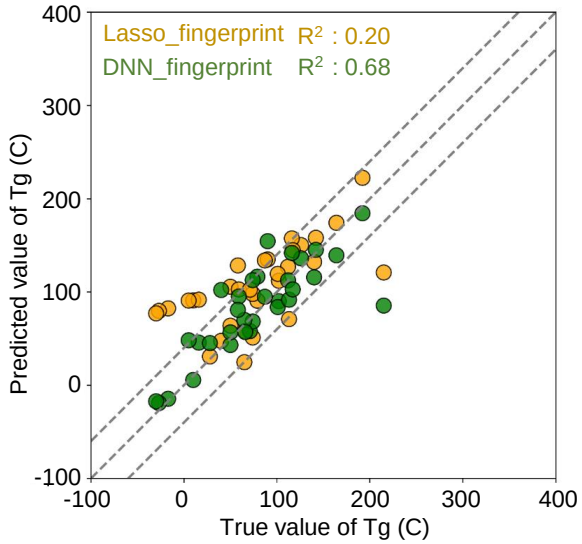
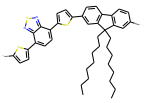
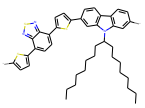
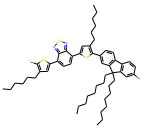
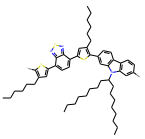
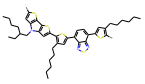
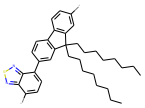
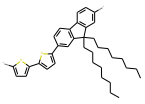
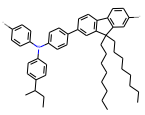
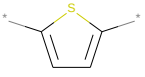
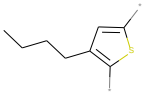
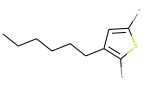
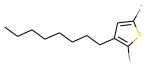
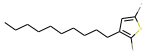
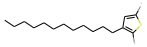
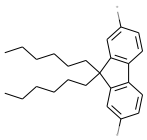
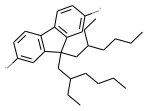
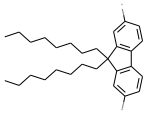
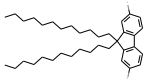
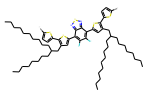
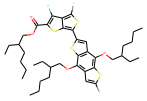
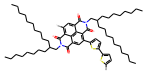
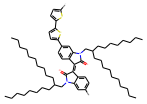
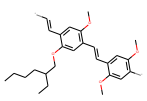
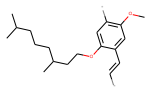
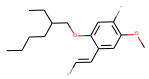
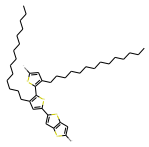
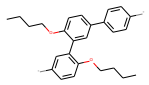
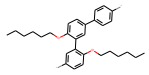
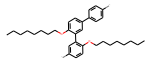
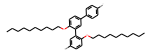
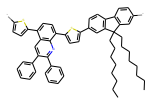
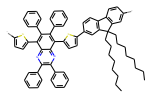


Figure S5. The comparison between the experimental T_g and the ML predicted T_g. Three dashed lines are a unity line and lines with a mean absolute error (MAE) of 40 °C.

Table S4. T_g of 32 polymers in a newly reported experimental dataset

Polymer	Experiment	Lasso_ Fingerprint	DNN_ Fingerprint
	126	150.36	136.11
	116	157.62	141.86
	79	90.65	116.50
	74	97.91	112.77
	90	134.99	154.62
	112	127.14	112.63
	102	112.19	90.16
	140	132.37	115.74
	215	120.91	85.48
	50	105.74	43.25
	10	91.06	5.64

	-17	82.56	-14.47
	-27	79.87	-19.01
	-30	77.18	-17.12
	101	119.56	83.95
	59	102.80	95.24
	71	102.55	58.10
	16	91.79	45.68
	28	30.89	45.29
	40	47.71	102.26
	74	51.23	68.50
	65	24.76	69.99
	113	70.95	91.76

	50	63.86	56.78
	66	58.00	57.03
	5	90.79	48.23
	164	174.27	139.40
	117	144.92	102.89
	87	133.91	94.95
	58	128.53	80.91
	142	158.13	145.25
	192	222.56	184.55

S8. With 1/0 (on/off) bits Morgan fingerprint, the ML models' prediction performance for the 20 polymers in dataset-2 and 32 polymers in a newly reported experimental study.

Based on the 20 polymers from dataset-2, the MD simulated T_g are compared with the ML prediction using 1/0 (on/off) bits Morgan fingerprint, as shown in **Figure S6a**. The performances of the ML models with 1/0 bits are comparable to that of the ML models considering the number of occurrences for substructures (**Figure 3e**).

Based on the 32 conjugated polymers from a newly reported experimental study, the experimental T_g values are compared with the ML predictions using 1/0 (on/off) bits Morgan fingerprint, as shown in **Figure S6b**. The performances of the ML models with 1/0 bits are worse than that of the ML models considering the number of occurrences for substructures (**Figure S5**). If not considering the number of occurrences for substructures, the model with 1/0 (on/off) bits Morgan fingerprint only leads to R^2 of -0.18, showing poor transferability. It is caused by the long side chain groups of these 32 conjugated

polymers, as presented in **Table S4**.

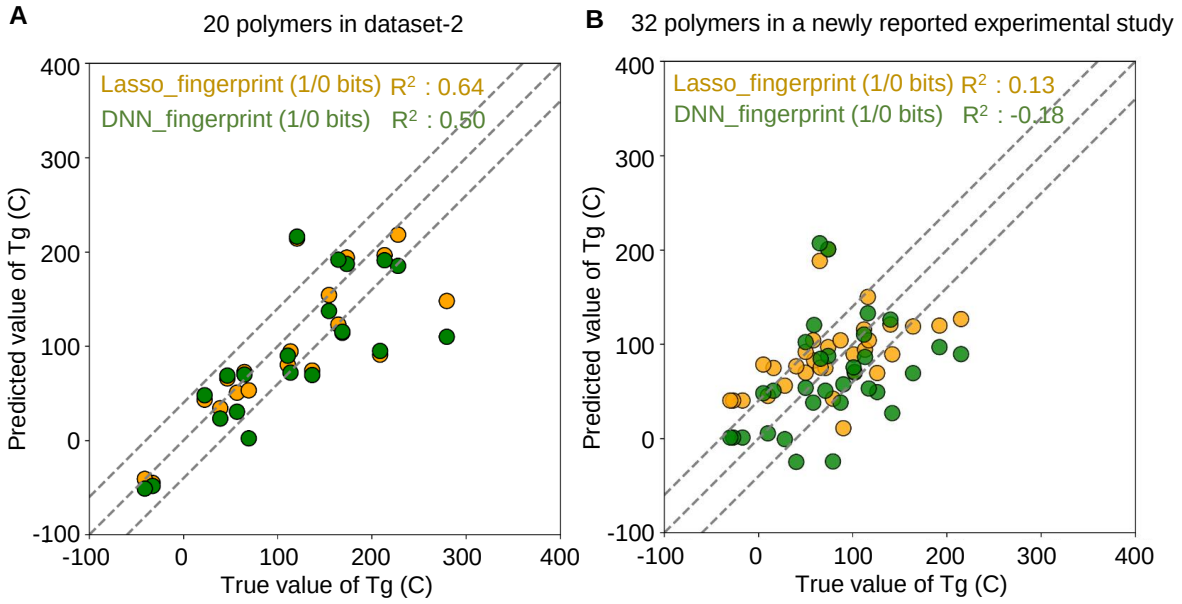


Figure S6. Comparison between the experimental T_g and the ML predicted T_g. (a) The ML predicted T_g v.s. MD simulated T_g for 20 polymers in dataset-2. (b) The ML predicted T_g v.s. Experimental T_g for 32 polymers in a newly reported experimental study. Three dashed lines are a unity line and lines with a mean absolute error (MAE) of 40 °C.

S9. Comparison of key substructures and functional groups in high-T_g (>200 °C) polymers from three datasets.

For 18 key substructures recognized in Figure 3, the percentage of high T_g polymers containing them are compared in **Table S5**. For the 6 high T_g related functional groups recognized in **Table 5**, the percentage of high T_g polymers containing them are compared in **Table S6**.

Table S5. Percentage of high-T_g polymers (>200 °C) containing key substructures

Substructure	Dataset-1	Dataset-2	Dataset-3
16406	15.04%	16.54%	27.55%
17748	17.59%	17.23%	29.99%
426	22.00%	8.05%	5.86%
24417	46.08%	23.90%	12.41%
7968	11.90%	12.12%	1.43%
22566	7.94%	4.16%	2.03%
17135	28.35%	14.20%	4.50%
770	18.39%	18.87%	31.54%
8814	11.11%	7.62%	1.65%
17618	74.87%	60.26%	51.47%
11337	1.50%	1.21%	0.65%
11362	32.94%	37.06%	63.53%
23586	53.40%	53.16%	76.05%
11881	4.85%	4.85%	5.39%
4916	1.41%	1.04%	1.09%
7305	7.10%	6.32%	7.55%

24993	64.73%	41.65%	49.41%
25398	54.81%	56.62%	79.72%

Table S6. Percentage of high-T_g polymers (>200 ^{o}C) containing exclusive functional groups

Functional group	dataset-1	dataset-2	dataset-3
Oxohetarene	0.88%	1.65%	1.65%
Lactam	0.57%	0.69%	1.05%
tert. arom_amine	0.57%	0.78%	0.50%
sec. aromat. Amine	0.13%	1.82%	3.32%
sec. mixed amine (aryl alkyl)	0.09%	0.43%	1.55%
Enamine	0.04%	0.87%	1.24%

References

1. Mauri, A. (2020). alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In Ecotoxicological QSARs, (Springer), pp. 801-820.
2. Kotelyanskii, M., Wagner, N. J., and Paulaitis, M. E. (1996). Building Large Amorphous Polymer Structures: Atomistic Simulation of Glassy Polystyrene. Macromolecules 29, 8497-8506. 10.1021/ma960071b
3. Landrum, G. (2013). RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Academic Press.
4. Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. Journal of chemical information and modeling 50, 742-754.
5. Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. International Journal of Education 4, 202.
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12, 2825-2830.
7. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., and Isard, M. In *Tensorflow: A system for large-scale machine learning*, 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), 2016; pp 265-283.
8. Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
9. Sun, H., Ren, P., and Fried, J. (1998). The COMPASS force field: parameterization and validation for phosphazenes. Computational and Theoretical Polymer Science 8, 229-246.
10. Bunte, S. W., and Sun, H. (2000). Molecular modeling of energetic materials: the parameterization and validation of nitrate esters in the COMPASS force field. The Journal of Physical Chemistry B 104, 2477-2489.
11. Sun, H. (1998). COMPASS: an ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds. The Journal of Physical Chemistry B 102, 7338-7364.
12. McQuaid, M. J., Sun, H., and Rigby, D. (2004). Development and validation of COMPASS force field parameters for molecules with aliphatic azide chains. Journal of computational chemistry 25, 61-71.
13. Kondratyuk, N. D., and Pisarev, V. V. (2019). Calculation of viscosities of branched alkanes from 0.1 to 1000 MPa by molecular dynamics methods using COMPASS force field. Fluid Phase Equilibria 498, 151-159.
14. Binder, K. (1995). Monte Carlo and molecular dynamics simulations in polymer science. (Oxford University Press).
15. Bandyopadhyay, A., and Odegard, G. (2012). Molecular modeling of crosslink distribution in epoxy polymers. Modelling and Simulation in Materials Science and Engineering 20, 045018.
16. Bandyopadhyay, A., Valavala, P. K., Clancy, T. C., Wise, K. E., and Odegard, G. M. (2011). Molecular modeling of crosslinked epoxy polymers: The effect of crosslink density on thermomechanical properties. Polymer 52, 2445-2452. 10.1016/j.polymer.2011.03.052

17. Clancy, T. C., Frankland, S. J. V., Hinkley, J. A., and Gates, T. S. (2009). Molecular modeling for calculation of mechanical properties of epoxies with moisture ingress. Polymer 50, 2736-2742. 10.1016/j.polymer.2009.04.021
18. Fan, H. B., and Yuen, M. M. F. (2007). Material properties of the cross-linked epoxy resin compound predicted by molecular dynamics simulation. Polymer 48, 2174-2178. 10.1016/j.polymer.2007.02.007
19. Li, C., Coons, E., and Strachan, A. (2014). Material property prediction of thermoset polymers by molecular dynamics simulations. Acta Mechanica 225, 1187-1196. 10.1007/s00707-013-1064-2
20. Li, C., and Strachan, A. (2011). Molecular dynamics predictions of thermal and mechanical properties of thermoset polymer EPON862/DETDA. Polymer 52, 2920-2928.
21. Lin, P.-H., and Khare, R. (2009). Molecular Simulation of Cross-Linked Epoxy and Epoxy−POSS Nanocomposite. Macromolecules 42, 4319-4327. 10.1021/ma9004007
22. Nouri, N., and Ziaei-Rad, S. (2011). A Molecular Dynamics Investigation on Mechanical Properties of Cross-Linked Polymer Networks. Macromolecules 44, 5481-5489. 10.1021/ma2005519
23. Odegard, G. M., Jensen, B. D., Gowtham, S., Wu, J., He, J., and Zhang, Z. (2014). Predicting mechanical response of crosslinked epoxy using ReaxFF. Chemical Physics Letters 591, 175-178. 10.1016/j.cplett.2013.11.036
24. Shenogina, N. B., Tsige, M., Patnaik, S. S., and Mukhopadhyay, S. M. (2013). Molecular modeling of elastic properties of thermosetting polymers using a dynamic deformation approach. Polymer 54, 3370-3376. 10.1016/j.polymer.2013.04.034
25. Shenogina, N. B., Tsige, M., Patnaik, S. S., and Mukhopadhyay, S. M. (2012). Molecular Modeling Approach to Prediction of Thermo-Mechanical Behavior of Thermoset Polymer Networks. Macromolecules 45, 5307-5315. 10.1021/ma3007587
26. Shokuhfar, A., and Arab, B. (2013). The effect of cross linking density on the mechanical properties and structure of the epoxy polymers: molecular dynamics simulation. Journal of Molecular Modeling 19, 3719-3731. 10.1007/s00894-013-1906-9
27. Varshney, V., Patnaik, S. S., Roy, A. K., and Farmer, B. L. (2008). A Molecular Dynamics Study of Epoxy-Based Networks: Cross-Linking Procedure and Prediction of Molecular and Material Properties. Macromolecules 41, 6837-6842. 10.1021/ma801153e
28. Yu, S., Yang, S., and Cho, M. (2009). Multi-scale modeling of cross-linked epoxy nanocomposites. Polymer 50, 945-952.
29. Abbott, L. J., and Colina, C. M. (2011). Atomistic structure generation and gas adsorption simulations of microporous polymer networks. Macromolecules 44, 4511-4519.
30. Soni, N. J., Lin, P.-H., and Khare, R. (2012). Effect of cross-linker length on the thermal and volumetric properties of cross-linked epoxy networks: A molecular simulation study. Polymer 53, 1015-1019. 10.1016/j.polymer.2011.12.051
31. Yang, S., and Qu, J. (2012). Computing thermomechanical properties of crosslinked epoxy by molecular dynamic simulations. Polymer 53, 4806-4817. 10.1016/j.polymer.2012.08.045
32. Demir, B., and Walsh, T. R. (2016). A robust and reproducible procedure for cross-linking thermoset polymers using molecular simulation. Soft Matter 12, 2453-2464. 10.1039/c5sm02788h
33. Li, C., and Strachan, A. (2010). Molecular simulations of crosslinking process of thermosetting polymers. Polymer 51, 6058-6070.
34. Liu, H., Li, M., Lu, Z.-Y., Zhang, Z.-G., Sun, C.-C., and Cui, T. (2011). Multiscale Simulation Study on the Curing Reaction and the Network Structure in a Typical Epoxy System. Macromolecules 44, 8650-8660. 10.1021/ma201390k
35. Gavrilov, A. A., Komarov, P. V., and Khalatur, P. G. (2014). Thermal Properties and Topology of Epoxy Networks: A Multiscale Simulation Methodology. Macromolecules 48, 206-212. 10.1021/ma502220k
36. Jang, C., Sirk, T. W., Andzelm, J. W., and Abrams, C. F. (2015). Comparison of crosslinking algorithms in molecular dynamics simulation of thermosetting polymers. Macromolecular Theory and Simulations 24, 260-270.
37. Sirk, T. W., Khare, K. S., Karim, M., Lenhart, J. L., Andzelm, J. W., McKenna, G. B., and Khare, R. (2013). High strain rate mechanical properties of a cross-linked epoxy across the glass transition. Polymer 54, 7048-7057.
38. Lin, P.-H., and Khare, R. (2009). Molecular simulation of cross-linked epoxy and epoxy− POSS nanocomposite. Macromolecules 42, 4319-4327.
39. Yarovsky, I., and Evans, E. (2002). Computer simulation of structure and properties of crosslinked polymers: application to epoxy resins. Polymer 43, 963-969.
40. Jorgensen, W. L., and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and

crambin. Journal of the American Chemical Society 110, 1657-1666.
41. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. Journal of computational chemistry 25, 1157-1174.
42. Mayo, S. L., Olafson, B. D., and Goddard, W. A. (1990). DREIDING: a generic force field for molecular simulations. Journal of Physical chemistry 94, 8897-8909.
43. Abbott, L. J., Hart, K. E., and Colina, C. M. (2013). Polymatic: a generalized simulated polymerization algorithm for amorphous polymers. Theoretical Chemistry Accounts 132, 1334.
44. Abbott, L. J., Hughes, J. E., and Colina, C. M. (2014). Virtual synthesis of thermally cross-linked copolymers from a novel implementation of polymatic. The Journal of Physical Chemistry B 118, 1916-1924.
45. Abbott, L., and Colina, C. (2013). Polymatic: A Simulated Polymerization Algorithm.
46. Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. Journal of computational physics 117, 1-19.
47. Rigby, D., and Roe, R. J. (1987). Molecular dynamics simulation of polymer liquid and glass. I. Glass transition. The Journal of chemical physics 87, 7285-7292.
48. Yu, K. q., Li, Z. s., and Sun, J. (2001). Polymer structures and glass transition: A molecular dynamics simulation study. Macromolecular theory and simulations 10, 624-633.
49. Hadipeykani, M., Aghadavoudi, F., and Toghraie, D. (2020). A molecular dynamics simulation of the glass transition temperature and volumetric thermal expansion coefficient of thermoset polymer based epoxy nanocomposite reinforced by CNT: A statistical study. Physica A: Statistical Mechanics and its Applications, 123995.
50. Buchholz, J., Paul, W., Varnik, F., and Binder, K. (2002). Cooling rate dependence of the glass transition temperature of polymer melts: Molecular dynamics study. The Journal of chemical physics 117, 7364-7372.
51. Li, C., Medvedev, G. A., Lee, E.-W., Kim, J., Caruthers, J. M., and Strachan, A. (2012). Molecular dynamics simulations and experimental studies of the thermomechanical response of an epoxy thermoset polymer. Polymer 53, 4222-4230.
52. Han, J., Gee, R. H., and Boyd, R. H. (1994). Glass transition temperatures of polymers from molecular dynamics simulations. Macromolecules 27, 7781-7784.
53. Xie, R., Weisen, A. R., Lee, Y., Aplan, M. A., Fenton, A. M., Masucci, A. E., Kempe, F., Sommer, M., Pester, C. W., and Colby, R. H. (2020). Glass transition temperature from the chemical structure of conjugated polymers. Nature communications 11, 1-8.