

ARTICLE OPEN



Exploring high thermal conductivity polymers via interpretable machine learning with physical descriptors

Xiang Huang , Shengluo Ma¹, C. Y. Zhao , Hong Wang , and Shenghong Ju ^{1,2}✉

The efficient and economical exploitation of polymers with high thermal conductivity (TC) is essential to solve the issue of heat dissipation in organic devices. Currently, the experimental preparation of functional polymers with high TC remains a trial-and-error process due to the multi-degrees of freedom during the synthesis and characterization process. Polymer informatics equips machine learning (ML) as a powerful engine for the efficient design of polymers with desired properties. However, available polymer TC databases are rare, and establishing appropriate polymer representation is still challenging. In this work, we propose a high-throughput screening framework for polymer chains with high TC via interpretable ML and physical feature engineering. The hierarchical down-selection process stepwise optimizes the 320 initial physical descriptors to the final 20 dimensions and then assists the ML models to achieve a prediction accuracy R^2 over 0.80, which is superior to traditional graph descriptors. Further, we analyze the contribution of the individual descriptors to TC and derive the explicit equation for TC prediction using symbolic regression. The high TC polymer structures are mostly π -conjugated, whose overlapping p-orbitals enable easy maintenance of strong chain stiffness and large group velocities. Ultimately, we establish the connections between the individual chains and the amorphous state of polymers. Polymer chains with high TC have strong intra-chain interactions, and their corresponding amorphous systems are favorable for obtaining a large radius of gyration and causing enhanced thermal transport. The proposed data-driven framework should facilitate the theoretical and experimental design of polymers with desirable properties.

npj Computational Materials (2023)9:191; <https://doi.org/10.1038/s41524-023-01154-w>

INTRODUCTION

Polymers are extensively used in industry and daily life, owing to various advantages of chemical inertness, mechanical flexibility, and lightweight¹. As the organic electronics are becoming smaller while the power density keeps increasing, the thermal management and heat dissipation capability have attracted significant attention^{2,3}. However, conventional polymers are thermal insulators with reported thermal conductivity (TC) in the range from 0.1 to $0.5 \text{ W m}^{-1} \text{ K}^{-1}$, preventing the development of organic electronics⁴. Polymers with high TC are urgently demanded in organic energy storage and electronic devices to accommodate revolutionary innovations in organic electronics and optoelectronics⁵. The polymer morphology and topology were found to be closely related to TC⁶. Increasing the crystallite orientation and crystallinity can significantly reduce the phonon scattering and enhance the TC along the chain directions, which has been demonstrated by both experiments^{7–12} and theoretical simulations^{13–15}. A recent study has fabricated polyethylene (PE) films by disentanglement and alignment of amorphous chains with a metal-like TC of $62 \text{ W m}^{-1} \text{ K}^{-1}$, over two orders of magnitude greater than that of classical amorphous polymers⁷. Moreover, molecular dynamics (MD) simulations have suggested that individual crystalline PE chains have a very high or even divergent TC¹⁶. These findings provide opportunities for solving the heat dissipation problem of polymer devices.

Intra-chain atomic interactions are usually much stronger than inter-chain interactions in polymers, and enhancing the intra-chain thermal transport of polymers is essential to improve their TC. Experimental techniques such as micro-mechanical stretching^{7,8}, electrostatic spinning^{9,10}, and nanoscale templating^{11,12} are effective in improving the crystallinity of polymers and obtaining

more consistent chain orientation, resulting in an increase in the TC of amorphous polymers by 1–3 orders of magnitude. Strategies such as applying mechanical strains^{17,18}, parallel-linking of chains¹⁹, and modulation of dihedral energy²⁰ in MD simulations suggested that ordered chains and large radius of gyration (R_g) are favorable for high TC of polymers. According to Debye's theory, the TC of the polymer k can be expressed by the phonon group velocity v_g , mean free path l and volumetric heat capacity C_v , i.e., $k = v_g C_v l$. In general, the v_g and C_v of polymers mainly determined by the characteristics of the repeating units and the strength of the backbone bonding in the individual chain²¹. Thus, it is possible to realize the high TC polymer by adjusting the repeating unit of polymer chains, which also facilitates understanding the heat transport mechanisms along the chain for polymers with different hierarchical structures.

Despite the fact that the chain structure of polymers exhibits great influence on the thermal characteristic, the polymer library is quite large, with as many as 10^8 monomeric organic molecules known to exist in chemical space²². Current research on the TC of polymers is still an Edisonian process, guided by intuition or experience in a trial-and-error approach that is time-consuming and expensive²³. Most of the studies are conducted on simple structures such as PE^{5,7,16}, which makes it difficult to grasp the general rule of the factors affecting the TC of polymers and to discover polymer molecular structures with high TC in huge chemical space.

The field of polymer informatics²⁴, associated with the development of artificial intelligence and machine learning (ML) methods, attempts to utilize the data-driven centric method for physical property regulation or device development of organic materials to resolve the conflict between structural freedom and

¹China-UK Low Carbon College, Shanghai Jiao Tong University, Shanghai, China. ²Materials Genome Initiative Center, School of Material Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. ✉email: shenghong.ju@sjtu.edu.cn

efficiency/cost in the traditional trial-and-error approach. The research on polymer informatics has attracted extensive attention and succeeded in recent years^{25–27}, involving the prediction of organic optical^{28–30}, electrical^{31–33}, and thermal properties^{34–38}. Particularly, several efforts emerged in the search or design of structures with high TC as related to crystalline polymers³⁶, amorphous polymers^{37,38}, and copolymers³⁹. Most of these studies have employed graph descriptors³⁷ or polymer chemistry fragment statistics^{36,38,39} to describe monomer structures in informatics algorithms, also called fingerprints or representations. The graph descriptors generated rely on molecular/monomer graph information, formulated by knowledge domain feature engineering⁴⁰ or by attempting to form general descriptors⁴¹. Moreover, descriptors such as molecular access systems (MACCS)⁴² are obtained through statistics of different chemical fragments, and are closely related to molecular graphs. Subsequently, they are collectively referred to as graph descriptors. Fingerprints are required for the unique, complete, minimal representation of each candidate, and successful fingerprints are a challenging task⁴³. Besides, polymers are composed of many repeating units, which are more complex than organic small molecules and require accurate capture of information on monomer connection sites³². Graph descriptors have long been applied and validated in the development of drug-like small molecules⁴⁴, and the availability of open-source toolkits such as RDKit⁴⁵ and Mol2vec⁴⁰ has facilitated their accessibility, which is also one reason that graph descriptors are popular in polymer informatics. However, the graph descriptor is in the form of a string of numeric vectors. The completeness of the molecular structure determines the coupling association between the digits. Hence, the relationship between molecular monomers and material properties is difficult to grasp.

Exploring the ensemble of physically independent descriptors for the representation of molecular structures is important in qualitative structure-property relationship modeling and enables more intuitive guidelines for molecular structure evaluation⁴⁶. Feature engineering for the collection and reduction of physical descriptors are critical steps in determining effective capabilities in polymer informatics. The development of automatic, universal and efficient tools for the calculation of descriptors of organic molecules is of interest to researchers, which translates the chemical information encoded in the symbolic representation of molecules into useful numbers or some standardized experimental results⁴⁷. Several open-source and commercial software^{47–49} are available to calculate various types of molecular descriptors such as carbon atomic number, molecular weight, and Extended Topochemical Atom⁵⁰, which have been successfully applied in organic chemistry synthesis⁵¹, molecular antibacterial activity prediction⁵², and so on. In addition, the parameter conditions in experiments or simulations affect the molecular properties. For instance, the force-field-inspired descriptors such as types of bond, angle, and dihedral have been validated for the prediction of the specific heat of polymers, even if the datasets are from experiments³⁵. The dimensionality reduction of polymer features is another concern, as some descriptors may have little relevance to the target property, and a low-dimensional descriptor space is much easier to build up for the ML model⁵³. Feature extraction and selection are the dominant approaches to reduce the dimensionality of features. Feature extraction creates subsets from the original data space, such as principal component analysis (PCA), where the specific meaning of the new features obtained is difficult to understand⁵⁴. Feature selection retains the physical meaning of individual descriptors, while filters based on correlation evaluation have dependencies on mathematical models, like the Pearson and Spearman coefficients that consider the linear and monotonic relationships of the data, respectively⁵⁵. Further, the filter methods do not involve ML models, which may lead to the inapplicability of the gained features. The wrapper-based

feature selection techniques combine ML models to eliminate redundant features, including recursive feature elimination (RFE), sequential feature selection (SFS), and exhaustive feature selection⁵⁶. Testing different subsets of descriptors for informatics algorithms is the crucial feature of the wrapper approaches, and the key is the strategy of combining different descriptors. Typical RFE seeks to improve model performance by continuously reducing the low-impact features from the remaining features in iteratively constructed ML models, which refer to the ranking of feature weights assigned by models such as random forests⁵⁴. Thus, the RFE relies on the feature weight evaluation mechanism of the ML models.

Herein, focusing on the challenges of polymer monomer representation and feature selection, we propose an ML interpretable framework integrated with high-throughput MD simulations for the discovery of polymer structures with high TC, as illustrated in Fig. 1. It consists of four components: 1) polymer library construction; 2) MD simulation for the TC of polymers; 3) monomer feature representation and hierarchical down-selection; 4) ML models construction for TC prediction. The training data were collected from literature^{57,58}, and candidates from the databases of PoLyInfo⁵⁹ and PI1M⁶⁰ were applied for the virtual screening of high TC structures. All polymer monomers were identified by the SMILES (simplified molecular input line entry system) strings and formed one-dimensional polymer chains by replication. The TC of training datasets was calculated by MD simulations with the second generation of the general AMBER force field—GAFF2⁶¹. Inspired by drug-like molecular representation and molecular force fields, we obtained 320 physical descriptors by Mordred software⁴⁷ calculation and force field parameter file extraction, and retained 20 optimized descriptors by hierarchical down-selection. We then trained random forest (RF), extreme gradient boosting (XGBoost) tree-based models, and multilayer perceptron (MLP) neural network models separately to establish the relationship between the optimized descriptors and the TC of these benchmark polymer datasets. Further, we analyzed the feature importance of each optimized descriptor and extracted the chemical heuristic for high TC polymers design through SHAP analysis⁶². Using the trained ML models, 107 promising polymers with TC greater than $20.00 \text{ W m}^{-1} \text{ K}^{-1}$ were identified, which are served for symbolic regression to derive mathematical formulas for expressing the TC of promising polymers. Last, we discussed the thermal transport mechanisms of polymer chains and analyzed the intra-chain thermal transport linkages of polymers with different hierarchical structures. Overall, the proposed approach is beneficial for theoretical or experimental investigations of high TC polymers.

RESULTS

Distribution of polymer datasets in chemical space

Polymer data from literature^{57,58} were utilized as the benchmark database for training ML models, as well as PoLyInfo⁵⁹ and PI1M⁶⁰ databases were used for the virtual screening of polymer structures with high TC. The polymers are classified into 19 classes such as polyolefins, polyethers, and polyamides according to different elements and chemical functional groups⁶³. To validate the distribution of the selected 1735 benchmark data over the other two datasets, their chemical structures were visualized in 2D space by the uniform manifold approximation and projection⁶⁴, where the chemical structure of each monomer was transformed into the Morgan fingerprint⁴¹ of a 1024 vector with a radius of two atoms. It is observed that the polymer structures in the selected benchmark dataset (Fig. 2a) are well covered by the chemical space distribution of those in the PoLyInfo (Fig. 2b) and PI1M (Fig. 2c) databases. Note that the PI1M dataset was generated by a generative model of a recurrent neural network

trained with data from PoLyInfo, which fills the sparse region of the chemical space of the PoLyInfo dataset, but the distribution is consistent⁶⁰. Thus, the ML models trained with the selected data are well able to learn the chemical features of all candidates and can be effectively adopted for the virtual screening of polymer structures with high TC. In addition, we counted the distribution of polymer TC in the benchmark dataset in Supplementary Fig. 1, which has a wide range and most of the polymers have TC less than $10 \text{ W m}^{-1} \text{ K}^{-1}$, and only a few polymers have TC greater than $30 \text{ W m}^{-1} \text{ K}^{-1}$ (Insert in Supplementary Fig. 1a). The unbalanced data distribution makes the discovery of high TC polymer

structures a difficulty. To better improve the ML models generalization across the entire TC range, our learning problem was framed in logarithmic scale, i.e., $\log_2 \text{TC}$, as the target property for ML models⁶⁵.

Polymer descriptors hierarchical down-selection and ML Models Training

Polymer descriptors are hierarchically down-selected in three stages: removing features with low variance, primary filtering referred to different correlation coefficients, and final selection assisted with the ML model (shown in Supplementary Note 2). The

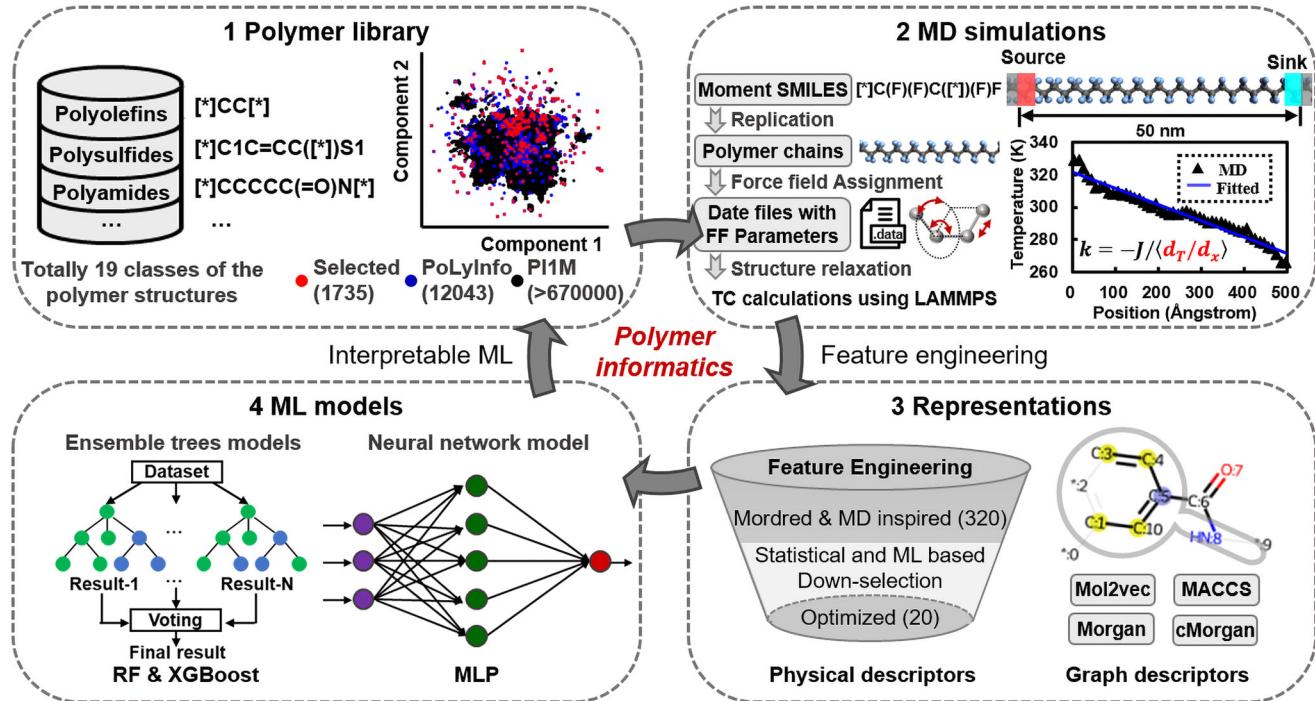


Fig. 1 Schematics of high-throughput screening of polymers with high TC via interpretable machine learning. which is implemented in four components: 1) polymer library construction, 2) MD simulation for the TC of polymers; 3) monomer feature representation and hierarchical down-selection; 4) ML model construction for TC prediction.

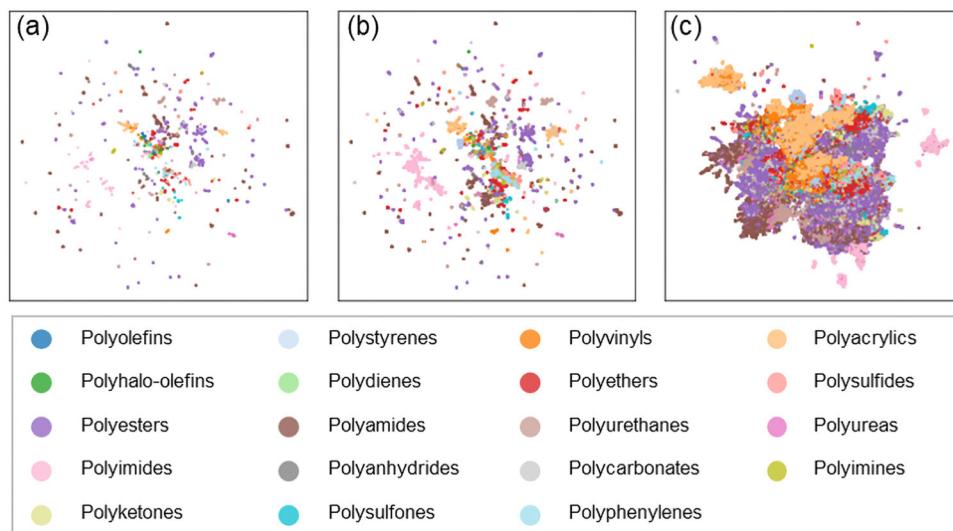


Fig. 2 Visualization of polymer data distribution in a 2D space by UMAP. a, b and c correspond to the selected, PoLyInfo, and PI1M datasets, respectively.

collected initial monomer physical descriptors are composed of 286 Mordred-based and 34 MD-inspired descriptors. The descriptors of MD-inspired and Mordred-based descriptors are listed in Supplementary Note 3. The removal of low variance descriptors is intended to eliminate descriptors with variance less than a specific threshold, whose contribution to the target property of all polymer data ($\log_2\text{TC}$ in this work) is considered to be nearly consistent. After the variance threshold was set to 0.10, the 264 descriptors were reserved for the next stage. We established the weight assignment mechanism based on the different correlation coefficients for further primary filtering of the descriptors, due to the various attentions of their mathematical models. The Pearson, Spearman, and Distance coefficients are used to evaluate linear, monotonic, and non-linear relationships between data, respectively, while the maximum information coefficient (MIC) reflects the association of two variables through information entropy, whether linear or nonlinear. The reliability of MIC depends on the data sample size and the value is reliable only with large datasets. The four metric coefficients of Pearson, Spearman, Distance, and MIC were incorporated and each was assigned a weighting factor of 0.25, and the thresholds were set to 0.05, 0.05, 0.153, and 0.132, respectively. The 53 descriptors with a cumulative weight value of 1 were retained through VAM. Random sequential feature selection (RFSF) combined with the RF model was then developed for optimized descriptors determination. Considering all possible combinations of descriptors for ML model training is time-consuming and expensive, so traditional SFS usually leads to sub-optimal solutions, where the recommended ensemble of optimized descriptors is not unique, and is influenced by the input order of the descriptors⁶⁶. Here, we disrupted the order of the input descriptors before each run, then combined them with 100 RF model training cycles, and acquired the final optimized descriptors based on a statistical approach. The threshold value depends on the occurrence times of the descriptors in 100 RF model training runs, and descriptors with a frequency larger than the threshold value were retained. We measured the performance of RF models trained by different descriptor ensembles with thresholds ranging from 0.39 to 0.32 in Supplementary Fig. 3, separately. By balancing the mean-square error (MSE) of ML model and the number of descriptors, 20 optimized descriptors were finally selected with a threshold of 0.34. The results of the optimized descriptors based on VAM and RSFS are shown in Fig. 3a, and their detailed descriptions are listed in Supplementary Note 4. Moreover, Fig. 3e exhibits the Pearson correlation matrices of the correlations among optimized descriptors (Other metrics, see Supplementary Fig. 2). It is found that most descriptors are positively correlated with each other and negatively correlated with TC. Only three descriptors are positive for TC, two of which are MD-inspired descriptors. For example, the descriptor MW_ratio reflects the ratio of the molecular weight of the mainchain to the molecular weight of the monomer, with values between 0 and 1. The MW_ratio of 1 indicates that the polymer is without side chains, which reduces the loss of heat flux along the chain and makes it possible to get large TC.

Figure 3b shows the results of the RF model trained with the optimized descriptors, with training and test R^2 of 0.87 and 0.84, respectively. To verify the extensibility of the optimized descriptors, XGBoost and MLP models were deployed for training (see Supplementary Fig. 4). The accuracy R^2 of the training and test sets for XGBoost is 0.95 and 0.87, and that for MLP are 0.81 and 0.88, respectively, which is comparable or even better than the RF model. Therefore, these three models are utilized in the subsequent discussion.

The prediction accuracy of ML models at different down-selection stages is illustrated in Fig. 3c (training and test data set prediction in Supplementary Fig. 5). The extra PCA with more than 95% variance was performed to compare with RFSF technology. According to the relationship between the number of principal

components and the cumulative variance in Supplementary Fig. 6, at least 19 components are required to exceed 95% variance. It is close to the number of sets of optimized descriptors. As seen in Fig. 3c, the tree-based models of the RF and XGBoost maintain relatively low MSE and high accuracy R^2 (See Supplementary Fig. 8) even with large descriptor dimensions because of their strong ability to prevent overfitting of the data. Moreover, the feature down-selection process is usually accompanied by the loss of information, which results in a decrease of model accuracy. However, the feature down-selection process also reduces the redundancy between data which suppresses the overfitting and improves the accuracy of the MLP model. Overall, the accuracy of all three models trained with the optimized descriptors from RFSF is higher than that of the models trained with the PCA-derived descriptors, which demonstrates the effectiveness of our approach.

The ML models with different graph descriptors were applied for comparison in Fig. 3d (training and test data set prediction in Supplementary Fig. 7). The Mol2vec⁴⁰ is an unsupervised ML approach to learn vector representations of molecular substructures, which requires a benchmark dataset for molecular structure training. Here, the pre-trained polymer embedding model was from elsewhere⁶⁰, which was created using the PolYInfo and PI1M datasets. The MACCS⁴² descriptor is the structural key-based descriptor with 166-bit keyset. The Morgan and Morgan count (cMorgan)⁴¹ descriptors are the extended connectivity fingerprints that capture molecular features relevant to molecular activity. The results in Fig. 3d and Supplementary Fig. 8 reflect the superiority of ML models trained with the optimized descriptors, no matter the models of RF, XGBoost, and MLP. The down-selection processes of physical descriptors examine individual/combined descriptors in relation to TC, while the graph descriptors aim to represent molecular/monomeric information as completely as possible. Whilst the elements or groups in the molecular graph have been indicated to correlate with the TC of polymer chains³⁶, it is more intuitive and effective to predict the $\log_2\text{TC}$ of polymer chains using the associated physical descriptors. But not absolute, which is also related to the parameters such as chain stiffness⁶⁷. We also evaluated ML models with a hybrid descriptor set composed of the optimized descriptors and one of the graph descriptors in Supplementary Fig. 9a and 9b. The performance of ML models trained with hybrid descriptors shows only a small improvement or even is comparable to that of trained with optimized descriptors, which reflects the fact that the optimized descriptors cover relatively complete information about the polymer structures. Furthermore, we applied the optimized descriptors or graph descriptors to train the directed message passing neural network (DMPNN) models in Chemprop⁶⁸, as shown in Supplementary Fig. 9c, d. Although the limited amount of data in the available benchmark dataset makes it difficult to output a high-performance DMPNN model, the performance of the optimized descriptors is the best compared to other descriptors. This illustrates the potential of optimized descriptors for applications in diverse and complex ML models.

Physical insights from an interpretable ML model

Figure 4 summarizes the effect of the features using SHAP, for the RF model trained on optimized descriptors. The SHAP approach attempts to address the unexplainable black-box challenge of ML algorithms by calculating the marginal contribution of features to the model output⁶². Hence, the features of each polymer structure in training data sets are assigned the SHAP values separately. As shown in Fig. 4a, the importance ranking of the optimized descriptors was referenced to the average SHAP value. Among the top 8 optimized descriptors, the number of MD-inspired and Mordred-based descriptors is equal, which reflects that the construction of the RF model is a joint contribution of these two

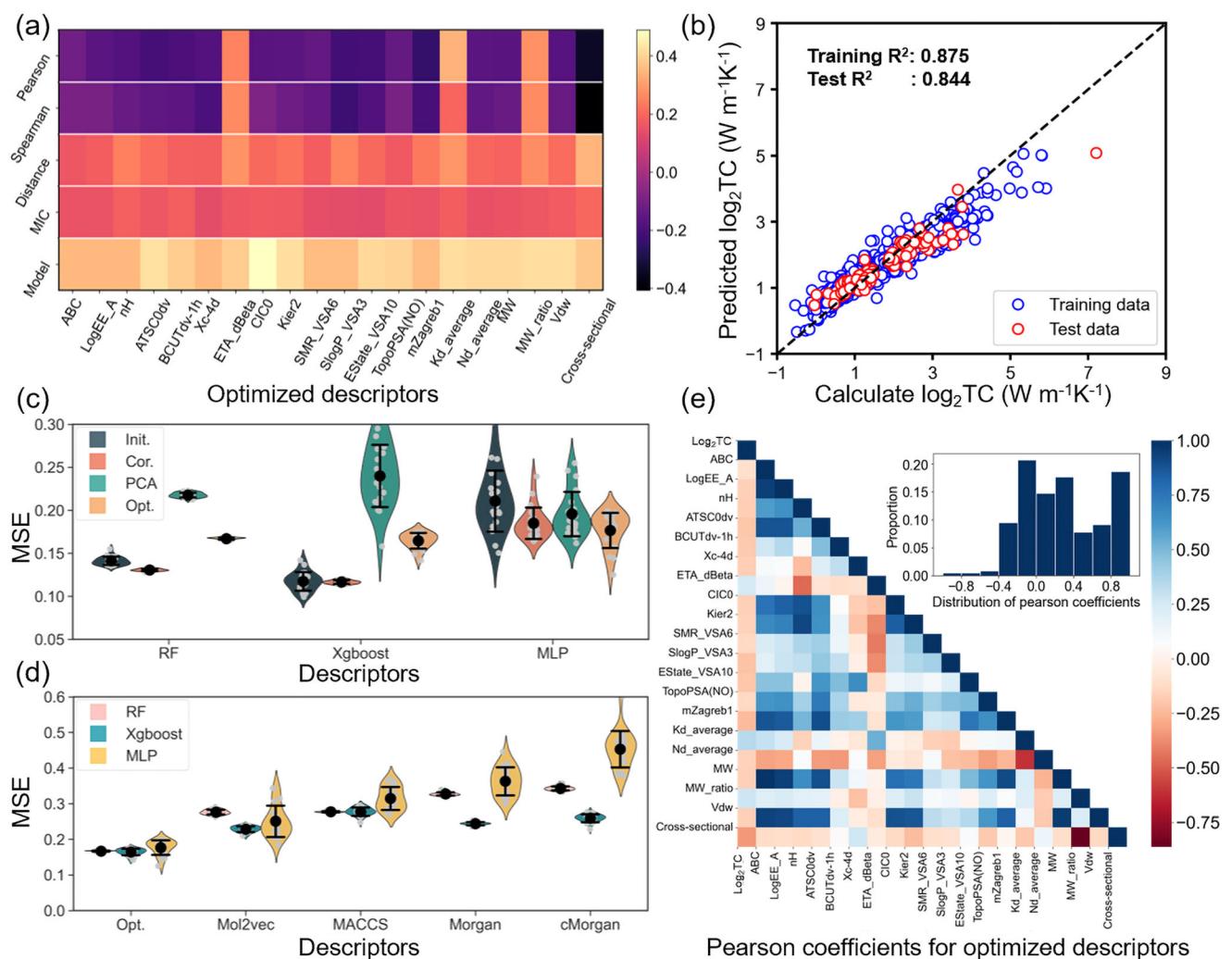


Fig. 3 Polymer descriptors down-selection and ML models training. **a** Optimized descriptors acquired by down-selection with four coefficients - Pearson, Spearman, Distance, and MIC coefficients - and RF model. **b** Accuracy of RF model based on optimized descriptors, where training R^2 is 0.875 and test R^2 is 0.844. **c** Mean-square error (MSE) of ML models at different down-selection processes, including initial (Init.), mathematical correlation (Cor.) coefficients screening, and RF model optimization (Opt.) stages. And, an additional PCA approach was applied to compare. **d** MSE of ML models with different polymer representation approaches. The violin plot represents the distribution of values, individual subsamples are shown in gray, and the mean and standard of MSE in black. **e** Pearson correlation matrices showing correlations among optimized descriptors and TC. The inset is the statistics of the Pearson coefficients distribution.

types of descriptors. The distribution of SHAP values for each descriptor is displayed in Fig. 4b, and the depth of shade of data points in the beeswarm plot represents the magnitude of TC of polymer structures in the training set. The distribution of SHAP values for the top-ranked features is relatively wide, and is monotonic about the feature values overall (Supplementary Fig. 10).

Here, we highlight the two MD-inspired descriptors of cross-sectional and Kd_average. The most important descriptor of cross-sectional indicates the effective cross-sectional area of the polymer chain, which is intuitive in relation to the TC. From Fig. 4c, the SHAP value for cross-sectional decreases monotonically with the descriptor. In 1-D polymer chains, the effective cross-sectional area relies on factors such as the complexity of the side chains and the chain orientation. Polymers with small cross-sectional areas facilitate the construction of centralized phonon transport channels along the backbone, and reduce the heat flux dissipated through the side chains. Thus, the TC is negatively related to the cross-sectional area, and polymers with high TC usually have a small cross-sectional area (Supplementary Fig. 11a). Moreover, the polymer chain structure is absent of disorder

compared to the amorphous structure, maintaining the symmetry of the crystal and reducing phonon scattering. However, the polymer chains may rotate and become disordered due to temperature and other effects, resulting in a rapid decrease in TC⁶⁹. The close correlation between the dihedral energy constant and polymer chain stiffness has been demonstrated, and the dihedral angle force constant Kd has been artificially increased in MD simulations to maintain PE chain stiffness and increase TC^{20,69}. The Kd_average is the average of all types of dihedral force constants from GAFF2 force field for polymer chain, which is roughly proportional to the corresponding SHAP value in Fig. 4d. Especially for polymer structures with great kd_average (>4 kcal mol⁻¹) usually have large SHAP values and TC (Supplementary Fig. 11b). Notably, the TC of polymer chains is influenced by multiple parameters and it is difficult to have the individual descriptor to determine its value. One example is that crystalline polynorbornene has been proven to be weakly sensitive to chain stiffness, even if increasing the dihedral angular force constant term in MD simulations⁶⁹. This confirms the significance of our proposed ML framework for predicting the $\log_2 TC$ of polymers.

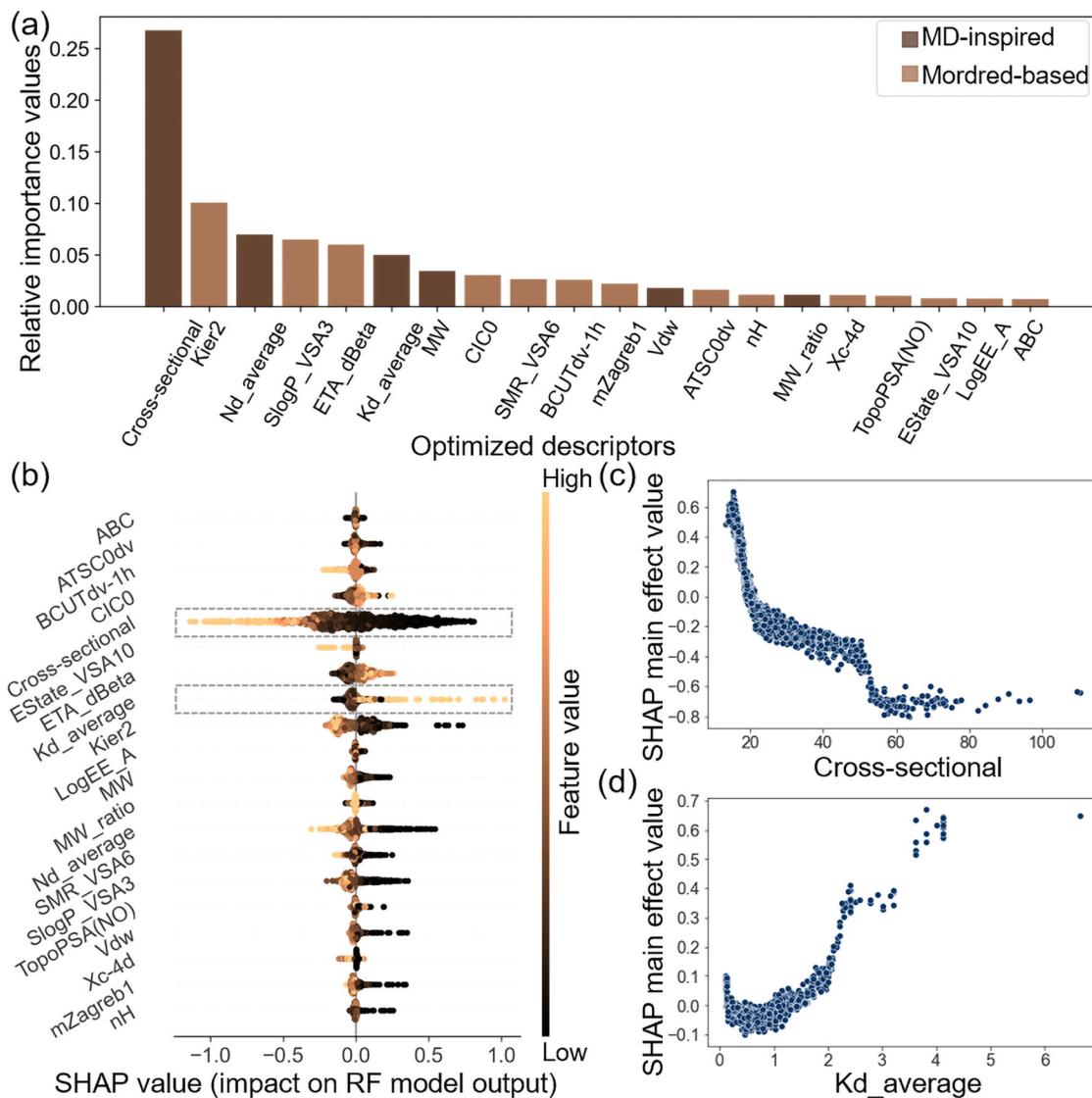


Fig. 4 Analysis of feature importance using SHAP on RF model trained by optimized descriptors. a Average SHAP values for 20 optimized descriptors. **b** Represent the SHAP values of each descriptor related to training data set polymers in a beeswarm diagram. **c, d** SHAP values for the Cross-section and Kd_average of the training data set polymers as a function of descriptor value. The cross-section is the effective cross-sectional area of the polymer chain, and the Kd_average is the average value of force constants of the dihedral angle from the GAFF2 force field.

Discovery of high TC polymers

The reliability of the optimized descriptors has been exhibited by the performance of the various trained ML models. Next, we applied these ML models to predict the \log_2 TC of polymer structures in the PoLyInfo and PI1M databases, in order to virtually screen promising polymers with high TC. The predicted polymer TC versus cross-sectional area from the ensemble of optimized descriptors combined with RF, XGBoost, and MLP are visualized in Fig. 5a–c, respectively. Where stars indicate polyethylene with \log_2 TC of 3.91, 4.66, and 5.30 predicted by RF, XGBoost, and MLP, respectively, and that calculated by MD simulation is 5.28. The dependence of TC on the cross-sectional area is evident here, as almost all of the predicted high \log_2 TC polymers have small cross-sectional areas. Moreover, since PI1M has the same chemical distribution space as PoLyInfo and fills the sparse area, which covers most of the \log_2 TC range of PoLyInfo and enriches the polymer structures in the high \log_2 TC region.

Comparing the results from different ML models, the tree-based models of RF and XGBoost predict the \log_2 TC of polymers in a

narrower space than that of the MLP. Though the excellent performance of the tree-based models in preventing overfitting, the extrapolation of the models is usually inadequate and the predictions are still limited to the range of \log_2 TC of the polymer structures in the training set. In contrast, the neural network model of MLP usually has better extrapolation capability, and is superior in finding small data such as high \log_2 TC polymer structures, despite the relatively low training accuracy R^2 of the model. This finding is similar to a previous study of predicting the permeability of gas separation membranes using ML²³. As well, previous work has revealed the length dependence of the TC of polymer chains. Within a certain length range, the diverging thermal conductivity k and chain length L can be fitted by $k \sim L^\beta$, where β indicates the relatively dominant phonon transport mechanism⁷⁰. Here, we considered polymer chains with TC greater than $20.00 \text{ W m}^{-1} \text{ K}^{-1}$ with an effective length of 50 nm as the outstanding polymers with high TC. Then, a balanced strategy to integrate the performance of three ML models was devised to recommend promising polymer structures for the

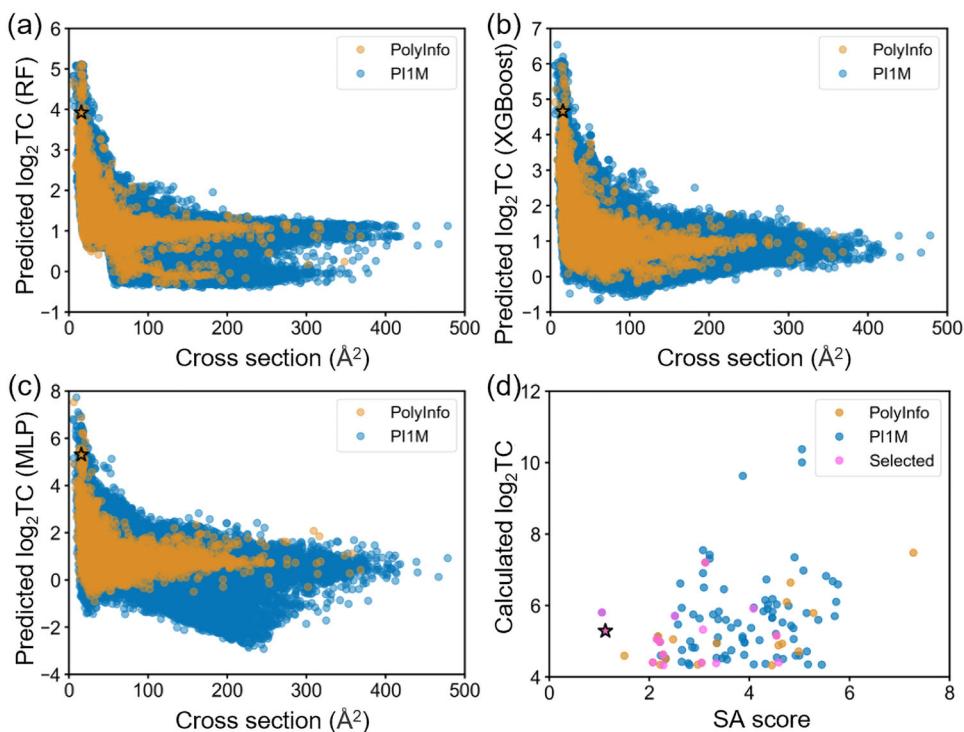


Fig. 5 Prediction of high TC polymers in PoLyInfo and PI1M databases using constructed ML models. **a, b** and **c** based on RF, XGBoost, and MLP models, respectively. **d** Synthetic accessibility (SA) score versus calculated log₂TC of screened high TC polymers (TC > 20.00 W m⁻¹ K⁻¹). The star indicates PE, and the TC in this work is 38.98 W m⁻¹ K⁻¹.

calculation of TC by MD simulations. We identified the polymer structures in the PoLyInfo dataset with RF, XGBoost, and MLP predictions of log₂TC up to 3.51, 3.50, and 4.33, and only the polymer structures with no less than 2 occurrences were picked for MD simulations. As a result, 24 polymer structures with high TC were discovered and verified. Similarly, we implemented this method to identify 84 high TC polymer structures in the PI1M database. After de-duplication, totally 107 high TC polymer structures were found in this work, and the Synthetic Accessibility (SA) scores were calculated as shown in Fig. 5d. The specific polymer structures can be seen in Supplementary Note 8. From Supplementary Fig. 12, we can see that most of the high TC polymers are simple linear or contain aromatic rings in the mainchain, which have small repeating unit lengths and no side chains. The SA score was initially utilized to estimate the synthetic accessibility of drug-like molecules based on molecular complexity and fragment contributions⁷¹, and was subsequently adopted for polymers^{37,38}. The SA score values ranged from 1 to 10, and synthesis is more difficult as the value increases. To take into account the effect of monomer linkages, polymer molecules with a polymerization degree of 6 were calculated for the SA score. Among them, 28 polymer structures with SA no more than 3.00, including polyethylene, polytetrafluoroethylene and poly(p-phenylene), and etc. Although it is currently difficult to fabricate each of these structures, we believe that more polymers like PE chains will be prepared for exploring the limits TC of polymers by combining advanced processes such as micromechanical stretching, electrostatic spinning, and nanoscale templating preparation in the near future^{5,7,16}.

Symbolic regression for TC prediction of promising polymers

Since the TC of polymer chains is influenced by complex multi-parameters, it is difficult to predict trends in TC values for different polymers from any single descriptor. Symbolic regression (SR) attempts to accelerate the discovery of materials with superior

properties by relating available descriptors through mathematical formulas to construct new combinatorial features⁷². SR does not require massive datasets, as long as a high consistency and accuracy⁷³. The 107 promising polymer structures (TC > 20.00 W m⁻¹ K⁻¹) with optimized descriptors were utilized for SR, where the ratio of training to test set was 3:1. The mathematical formula was acquired and selected using an efficient stepwise strategy with SR based on genetic programming (GPSR) as implemented in the gplearn code⁷⁴. The hyperparameters setup and the detailed formula determination process can be found in Supplementary Note 9. Pearson coefficients are first applied to filter optimized descriptors and create sub-descriptors, and an updated ensemble of 22 descriptors was obtained. The frequency of occurrence of optimized descriptors in 158 mathematical formulas (PC values ≥ 0.85 and complexity ≤ 10) is displayed in Fig. 6a, and the first eight descriptors were finally retained. It is worth emphasizing that the MD-inspired descriptors of cross-sectional area (cross-sectional) and dihedral force constants (Kd_average) appeared in each of the formulas. In Fig. 6b, we calculated the Pearson coefficients of the new set of descriptors with the TC, the results suggest these descriptors are closely associated with the TC. Subsequently, we reset the grid search hyperparameters in gplearn and used R^2 as the evaluation metric. Only formulas with high R^2 and low complexity (length of formula) are considered suitable for the prediction the log₂TC of polymer structures⁷⁵. Thus, 9073 mathematical formulas with complexity within 30 and R^2 over 0.6, which are characterized by complexity and accuracy R^2 via density plot in Fig. 6c. The four points of c, d, e, and f at Pareto front were identified by Latin hypercube sampling approach^{76,77}, and their corresponding formulas are expressed in Supplementary Table 8. The complexities of the four formulas are in the range of 20–30, and the fitting accuracies are all greater than 0.70. Moreover, the training accuracy is mostly positive to complexity. For example, the formula represented by point c with a complexity of 20 has a relatively low accuracy R^2 among the four points, but the fitting

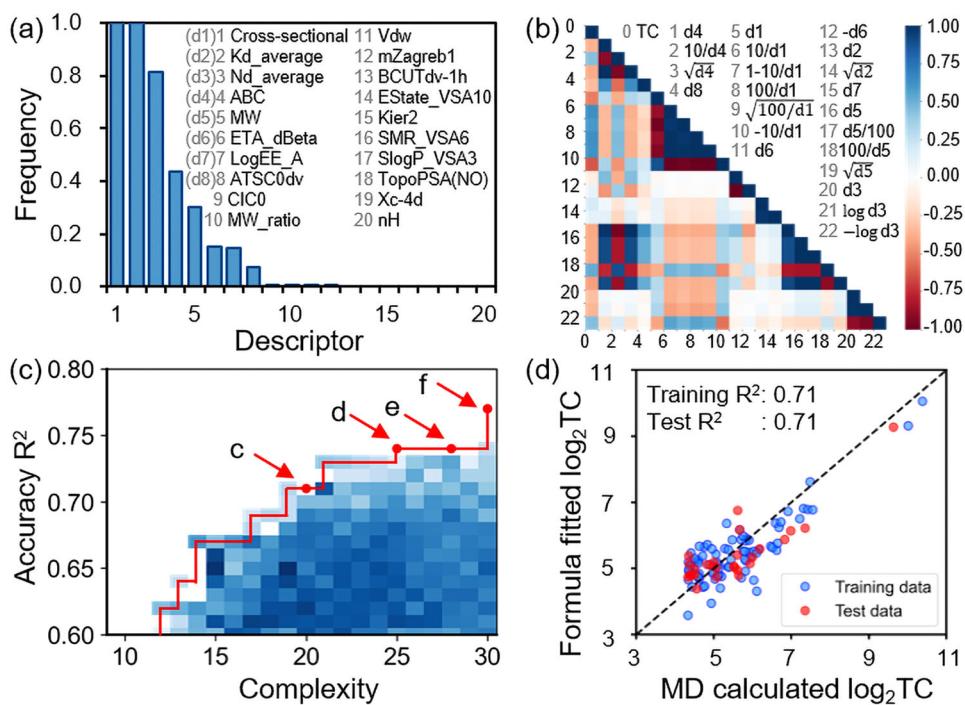


Fig. 6 GPSR for TC prediction of promising polymers. a Frequency of occurrence of optimized descriptors in 158 mathematical formulas (PC values ≥ 0.85 and complexity ≤ 10). **b** Pearson correlation matrices showing correlations among 22 descriptors and TC, where the descriptors d1–d8 correspond to descriptors 1 to 8 in (a). **c** Pareto front of accuracy R^2 vs. complexity of 9073 mathematical formulas shown via density plot. **d** MD calculated $\log_2\text{TC}$ vs. fitting results of the formula (point c) with a complexity of 20 and training accuracy R^2 of 0.71.

results are consistent with the MD calculated $\log_2\text{TC}$, as demonstrated in Fig. 6d. Meanwhile, all four identified formulas include the descriptors of the Cross-sectional, Kd_average and Nd_average, which verified that the TC of polymer chain is strongly correlated with the parameters such as cross-sectional area and dihedral stiffness. These formulas are meaningful in the initial rapid screening of high TC polymer chain structures.

Thermal transport mechanism of promising individual polymer chains

Taking into account factors such as TC and SA score, eight polymer structures (see Fig. 7a) were chosen for the analysis of phonon dispersion relations. Currently, polymer structures like $[\cdot]\text{C}=\text{Cl}\cdot$ and $[\cdot]\text{N}=\text{N}\cdot$ are challenging to be synthesized experimentally, but are contributing to our understanding of polymer thermal transport mechanisms. All of these polymer molecules are π -conjugated structures except for the PE and the Polytetrafluoroethylene (PTFE), which are simple linear structures. In π -conjugated polymer molecules, the overlap of p-orbitals has enhanced restraint in inhibiting chain rotation and forming the rigid backbone¹⁵. Figure 7b illustrates the phonon dispersion relations, which were obtained by phonon spectral energy density (Phonon-SED) analysis⁷⁸. The detailed description of the Phonon-SED approach can be found in the Method part. Since the acoustic modes are dominated by the thermal transport of heat carriers in polymer chains, phonon modes with frequencies below 25 THz are demonstrated. Moreover, the phonon group velocity v_g is approximated as the average of the slopes of all acoustic branches^{15,67}. The volumetric heat capacity C_v of each structure was evaluated from corresponding amorphous polymers, we constructed an amorphous system containing about 10,000 atoms according to the repeating units, respectively, and calculated the value of C_v after the equilibrium simulation at 300 K, more details can be found in Supplementary Note 10. So far, the phonon mean free path was derived from $l = k/v_g C_v$. Since the MD simulation

did not reach equilibrium for the model of $[\cdot]\text{N}=\text{N}\cdot$, we failed to obtain its C_v and l . The approximations of the above calculations allow the results to be rough, but it do help us to understand the underlying thermal conductivity mechanisms of these promising polymer structures by comparing the relative trends of the relevant parameters, as listed in Table 1.

The volumetric heat capacity of the eight polymer structures varies from 2.70 to 3.74 J cm⁻³ K⁻¹, which is not critical to the high TC of polymer chains²¹. As for the phonon group velocity, the six π -conjugated polymers have large values (more than 5900 m/s) due to overlapped p-orbital and delocalized electrons. Additionally, the small atomic mass enables a large phonon group velocity. The PTFE has a smaller phonon group velocity than that of PE due to the relatively larger mass of fluorine atoms compared to hydrogen atoms. The phonon mean free path provides valuable insights into phonon transport in the polymer chains. Overall, simple linear polymer chains easy to have long phonon mean free paths, especially for linear π -conjugated polymers such as $[\cdot]\text{C}=\text{Cl}\cdot$. These structures have large chain stiffness and few atoms except for the backbone, thereby having weak phonon-disorder scattering.

Thermal transport linkages between the various hierarchical polymer structures

To explore the thermal transport linkages between the different hierarchical structures of polymer chains and amorphous polymers, we selected 58 structures from 107 promising high TC polymer chains and calculated the TC of the corresponding amorphous polymers (ATC) using reverse non-equilibrium molecular dynamics (NEMD) simulations⁷⁹, as listed in Supplementary Table 4 and shown in Fig. 8a. Here, ATC was specifically defined as the TC of the amorphous polymer to distinguish it from that of polymer chains. Amorphous polymers normally have much lower TC than polymer chains due to their internal disordered chain entanglement, and thus polymers with ATC greater than

0.40 W m⁻¹ K⁻¹ can be considered to have outstanding thermal conductivity^{37,80}. Among the amorphous polymers simulated in this work, half of which have an ATC greater than 0.40 W m⁻¹ K⁻¹, while the equivalent percentage is only 2.3% in the reference (ref. ⁶³). In Fig. 8b, the radius of gyration (R_g) of amorphous polymers has a close positive correlation with ATC, and this work broadens the upper limit of R_g that in ref. ⁶³. Since polymer chains

with high TC are associated with strong atomic interactions and large chain stiffness, their corresponding amorphous structure is also conducive to maintaining large rigid chain segments.

According to different values of R_g , we selected six structures in Fig. 8c, Poly(p-phenylene), Poly(p-phenylenevinylene), Polyacetylene (PA), Poly[(E)-1-fluoroethene-1,2-diyl] (PEFD), PE and PTFE, to understand the thermal transport mechanism via energy flux decomposition analysis⁶³. The ATC of each amorphous polymer was quantified into six components of bond, angle, dihedral, convection, nonbonded and improper, where the nonbonded contribution contains pairwise and K-space contributions. From Fig. 8d, the intra-chain interactions of bonds, angles, and dihedrals dominate the ATC of amorphous polymers. Especially for π -conjugated polymers, the direct contribution of the dihedral term to the ATC is obvious. By comparing PA/PEFD pairs or PE/PTFE pairs, the system containing atoms with a large mass such as fluorine may inhibit the propagation of phonons and reduce the ATC. For a unified comprehension of the mechanism of the dihedral term on the contribution to the TC of different hierarchical structures, we investigated the role of chain orientation and chain rotation of polymers in Supplementary Note 10. Our results reveal that polymers with low dihedral energy are prone to poorly consistent chain orientation (Supplementary Fig. 16) and severe chain rotation (Supplementary Fig. 17), which are undesirable for heat flux transport in the intended direction. Furthermore, the TC of strained amorphous polymer or polymer chains is more sensitive to the reduction of dihedral energy rather than strain-free amorphous polymer, because it has a large original orientational order parameter.

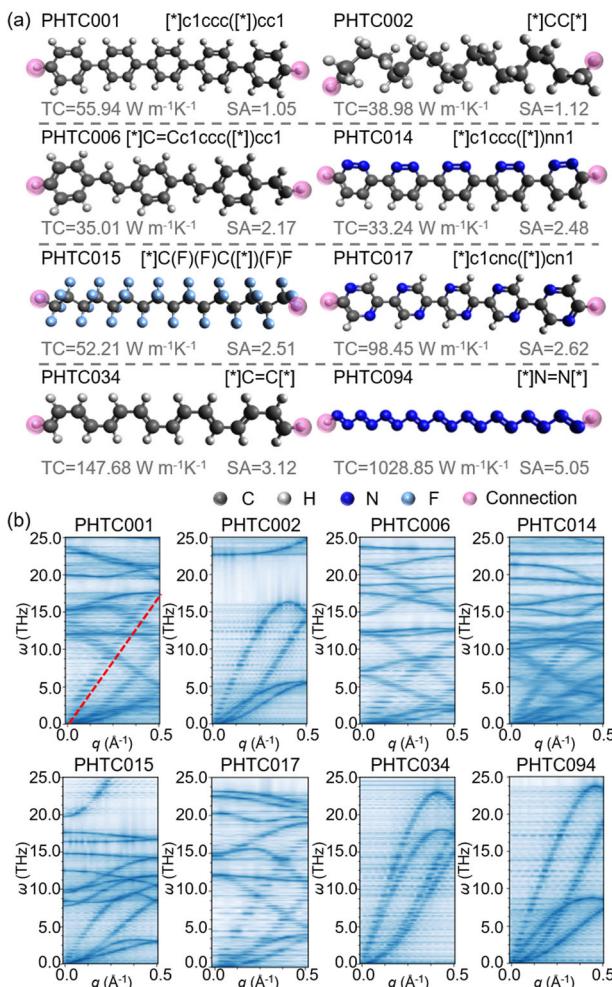


Fig. 7 Structure and phonon dispersion relations for the eight promising polymers. a Polymer chain structures. **b** Phonon dispersion relations. The q is the wavevector, the ω is phonon frequency and the average phonon group velocity of one branch is estimated as the slope of the origin to the maximum frequency point as shown in the red dashed line in the PHTC001 structure.

DISCUSSION

In summary, we have developed an interpretable ML framework for exploring high thermal conductivity polymer chains via high-throughput MD simulations. Inspired by the drug-like small molecule representation and the molecular force field, we reduced the initially calculated/collected 320 physical descriptors to 20 optimized descriptors by hierarchical down-selection. The constructed ML models are capable of effectively reflecting the relationship between optimized descriptors and property, and exhibit high accuracy in TC prediction. All the models of RF, XGBoost and MLP achieved the R^2 of more than 0.80, which is superior to that of represented by conventional graph descriptors. Moreover, the promotion or inhibition of TC by optimized descriptors like cross-sectional area and dihedral stiffness was captured by RF model using SHAP analysis.

Using the trained ML models, we discovered 107 promising polymers with TC greater than 20.00 W m⁻¹ K⁻¹, and 29 of which have SA scores of no more than 3.00. These polymer structures have been validated through high-fidelity MD simulations. Further, we used SR with optimized descriptors to fit the TC of promising polymers, and the derived mathematical formulas enable a preliminary fast screening of high TC polymers without relying

Table 1. Thermal properties for eight promising polymers.

Polymer ID	SMILES	SA	C_v (J cm ⁻³ K ⁻¹)	v_g (m/s)	l (nm)	TC (W m ⁻¹ K ⁻¹)
PHTC001	[*]c1ccc([*])cc1	1.05	3.09	6822.21	2.65	55.94
PHTC002	[*]CC[*]	1.12	3.74	5240.91	1.99	38.98
PHTC006	[*]C=Cc1ccc([*])cc1	2.17	2.97	6295.54	1.87	35.01
PHTC014	[*]c1ccc([*])nn1	2.48	2.70	5927.05	2.08	33.24
PHTC015	[*]C(F)(F)C([*])(F)F	2.51	2.94	2952.11	6.02	52.21
PHTC017	[*]c1cnc([*])cn1	2.62	2.86	7439.50	4.62	98.45
PHTC034	[*]C=C[*]	3.12	2.91	8380.09	6.06	147.68
PHTC094	[*]N=N[*]	5.05	N/A	6378.73	N/A	1028.85

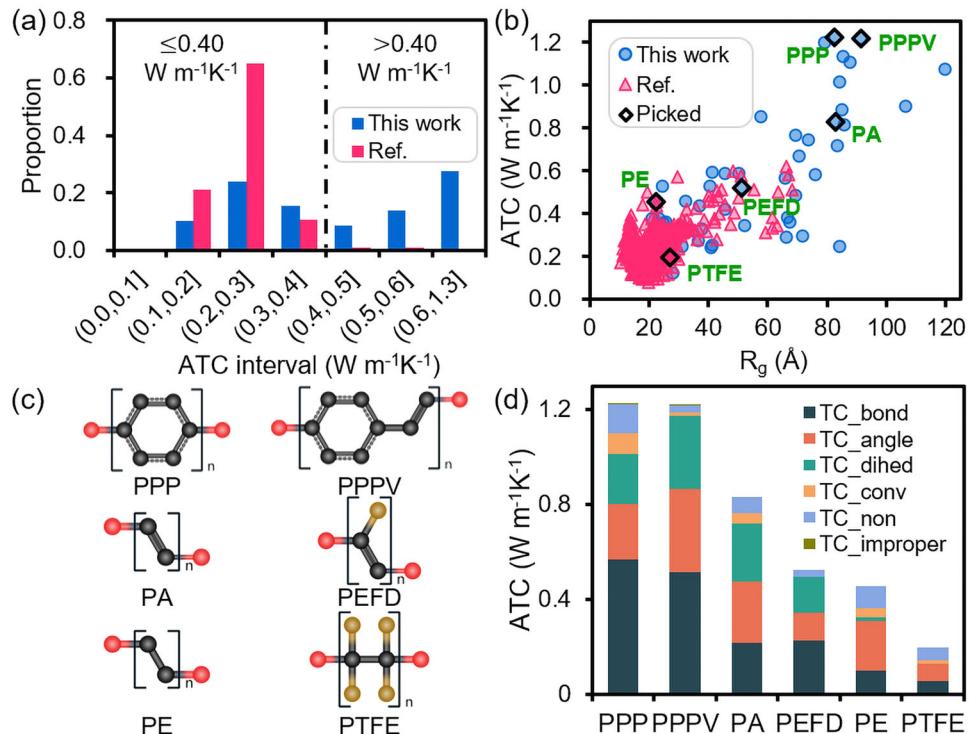


Fig. 8 Thermal conductivity of amorphous polymers (ATC). **a** ATC of 58 structures randomly selected from data of 107 promising polymers (this work), and the reference (Ref.) data calculated by Hayashi et al.⁶³, contains 1051 polymers, using the same simulation parameters as set in this work. **b** Radius of gyration (R_g) versus ATC for polymers of this work and Ref., where the diamond markers indicate the six typical amorphous polymers in **c**, including Poly(p-phenylene) (PPP), Poly(p-phenylenevinylene) (PPPV), Polyacetylene (PA), Poly[(E)-1-fluoroethene-1,2-diy] (PEFD), Polyethylene (PE) and Polytetrafluoroethylene (PTFE), where the black balls indicated the carbon atoms, the golden balls indicated the fluorine atoms, the red balls indicated the connection positions, and the hydrogen atoms were hidden. **d** Contributions of convection and different types of interactions to the ATC of six polymers. The ATC of each amorphous polymer was quantified into six components of the contribution of bond, angle, dihedral (dihed), convection (conv), nonbonded (non), and improper.

on ML models, which is friendly for experimental studies. In closing, we calculated phonon dispersion relations for eight typical polymer structures via phonon spectral energy density analysis to reveal the underlying TC mechanisms. Notably, most of these structures are π -conjugated polymers, whose overlapping p-orbitals enable easy maintenance of strong chain stiffness and large group velocities.

Currently, the pure individual polymer chains are still not accessible by experiments. Remarkably, there have been many efforts to fabricate polymer nanofibers with consistently oriented chains by techniques such as micromechanical stretching, electrostatic spinning, and nanoscale templating, but this has requirements on the inherent properties of the polymers. Although mechanical tensile disentanglement of amorphous polymers may be realized by adjusting conditions such as strain rate and temperature, it is demanding on the mechanical properties of the polymers⁸⁰. Many π -conjugated polymers are not suitable for the stretching process due to their incomparable elastic modulus with PE, whereas electrostatic spinning and nanoscale templating technologies are probably applicable^{81,82}. The conjugated polymer can be dissolved in matching solvents and then prepared to form consistently oriented nanofibers by electrostatic induction or by nanoscale template confinement such as anodic aluminum oxide templates. Moreover, the thermal properties of hierarchical structures of polymers are closely related. We calculated 58 amorphous polymers whose repeating unit was randomly extracted from the set of 107 promising polymers, and half of them have a high ATC of $0.40 \text{ W m}^{-1}\text{K}^{-1}$. Analyzed by R_g data, strong interatomic interactions are also beneficial for obtaining large rigid chain segments in the amorphous system, achieving significant intra-chain thermal

transport and high ATC. The proposed approach may assist in the research of high-performance polymers that are not limited to TC, and aid in understanding the linkage between the properties of different hierarchical structures.

METHODS

Polymer modeling and cross-sectional area calculation

Polymer modeling is a monomer-to-chain process, implemented in the STK tool, with input parameters of monomer SMILES and degree of polymerization⁸³. The length of the polymer chains was set uniformly to 50 nm, and the degree of polymerization was obtained by dividing the chain length by the monomer length and rounding up to an integer. Starting from the polymer SMILES, a molecular chain with polymerization degree 2 was generated by RDKit and optimized using the MMFF force field⁸⁴. Then, the monomer length was determined by measuring the distance between equivalent atoms in two repeating units in the heat transport direction. Following the modeling, a Python pipeline of PYSIMM realized the assignment of GAFF2 force field parameters and the generation of MD simulation input structure data files⁸⁵.

The cross-sectional area is one of the important parameters for thermal conductivity analysis. In molecular dynamics simulations, the calculation of the cross-sectional area is difficult for systems that do not occupy the entire simulation box. The cross-sectional area was estimated by the ratio of the van der Waals volume to the length of the monomer¹³. The Van der Waals volume of the monomer was calculated by the sum of atomic and bond contributions, and has been successfully tested and applied in previous drug compounds⁸⁶.

Calculation of TC by MD simulations

The TC of polymer chains was obtained by NEMD simulations performed in a Large-scale Atomic/Molecular Massively Parallel Simulator⁸⁵. The implementation of NEMD simulations is similar to the steady-state measurement experiments for TC, in which a 1-D steady-state heat transfer is generated by adding the heat source and sink at both ends of the sample⁸⁷. NEMD simulations have been extensively applied in the calculation of TC of low-dimensional systems, as it has been proven to have the ability to identify non-Fourier heat conduction phenomena induced by nanoconfinement^{70,88–90}. As for polymers, Liu et al.⁷⁰ demonstrated that competition between ballistic phonon transport and diffusive phonon transport in single polymer chains leads to a diverging length-dependent thermal conductivity through MEND simulations. Shrestha et al.⁹¹ experimentally examined the temperature dependence of the TC of polymer nanofibers under ultra-high stretch, and indicated that the TC at high-temperature end matched well with the results from NEMD calculations⁹⁰.

In terms of the NEMD method for TC calculation of polymer chains, the heat energy exchange was achieved by an enhanced version of the heat exchange algorithm, which rescales and shifts the velocities of particles inside reservoirs to impose a constant heat flux⁹². The polymer chains were placed in a box of $540 \times 60 \times 60$ ($x \times y \times z$) Å box, where the dimension in the y and z directions was set to 60 Å to avoid interaction with the neighboring polymer chains. Before TC calculation, the polymer chain structures were relaxed to reach a stable conformation. Then, the polymer chain was divided into 50 slabs in the x direction, and the fixed regions at two ends of the chain were set as a heat-insulating walls. In the NEMD simulation, the system was run under NVT (constant number of atoms, volume, and temperature) and NVE (constant number of atoms, volume, and energy) ensembles for 1 ns at 300 K sequentially to release chain stress^{36,93}. After that, the heat was added/extracted to the heat source/sink regions (20 Å of each region) at the end of the polymer chain in a regular rate to create a constant heat flux. The applied heat varies for different polymer chain structures and ranges from 0.01 eV/ps to 0.08 eV/ps. At last, the temperature profile was averaged over the last 2–3 ns and used for TC calculation, solved by $k = -J(d_T/d_x)$, where J is heat flux, d_T/d_x is the temperature gradient. In addition, the ATC of amorphous polymers and the C_V calculations were implemented by an automated pipeline in the Radonpy toolkit (Supplementary Fig. 15)⁶³.

Descriptors calculation and ML models construction

The ideal polymer descriptors are required to minimize and completely represent polymer information, and are one of the key factors in determining the prediction accuracy of ML algorithms. The physical descriptors for this work were sourced from both Mordred software calculations and GAFF2 force field parameters extraction. The Mordred software was initially developed for small molecule characteristics in cheminformatics, which can calculate more than 1800 descriptors⁴⁷. However, since we consider two connecting sites of polymer monomers, only 286 valid descriptors were obtained. Therefore, as a complement, we additionally extracted parameters from each polymer force field file as the descriptor. For graph descriptors, MACCS, Morgan, and cMorgan fingerprints were calculated in the RDKit package⁴⁵. The Mol2vec fingerprints were embedded via Mol2vec⁴⁰. We referred to the polymer representation model trained using PoLyInfo and PI1M databases for generating Mol2vec fingerprints⁶⁰.

The ML models of RF, XGBoost, and MLP were implemented by using Scikit-learn⁹⁴. Hyperparametric optimization for RF, XGBoost, and MLP was operated with the Bayesian Optimization package⁹⁵ which is a global optimization tool to achieve good prediction accuracy R^2 . The Gaussian regression process and acquisition

function with ten random pairs of parameters were selected for initial training, and the ideal parameters for each ML model were determined after 100 optimization iterations⁵².

To explain the association of optimized descriptors with TC, we used the SHAP toolkit with RF model to evaluate the feature importance⁶². The SHAP analysis is based on a game-theoretic approach that associates the optimal credit allocation with the local explanations of the model, which considers the model performance by neglecting each feature and provides the direction of each descriptor effect⁵².

Mathematical formulas for TC fitted by symbolic regression (SR)

The mathematical formulae were acquired and selected using an efficient stepwise strategy with GPSR as implemented in gplearn⁷⁴. The 107 polymer structures with TC greater than $20.00 \text{ W m}^{-1} \text{ K}^{-1}$ were randomly divided into 3:1 as training and test sets, respectively. At first, Pearson coefficients were used as evaluation metrics of training fitness to filter optimized descriptors and generate sub-descriptors, and a new dataset containing 22 descriptors was generated. Further, the grid search strategy with the hyperparameters and metric R^2 as listed in Table 2 was applied to determine the mathematical formulas. We ultimately discussed four formulas at the Pareto front that were identified by Latin hypercube sampling approach⁷⁷. More information about SR can be found in the Supplementary Note 9.

Analysis of phonon dispersion relations by phonon spectral energy density (Phonon-SED)

To understand the TC mechanism of polymers, MD simulations coupled with Phonon-SED approach⁷⁸ were employed to calculate the dispersion relations of polymers. The polymer chain with a length of 100 Å was constructed as an input of SMILES and placed into a box with the cross section of 60×60 Å. After energy minimization, the system was run under the NVT (constant number of atoms, volume, and temperature) ensemble for 0.25 ns at 2 K sequentially to release chain stress. Subsequently, the system was run under the NVE (constant number of atoms, volume, and energy) ensemble for 2 million steps with the timestep of 0.25 fs. During this period, the velocity and position of each atom in the polymer backbone were recorded with intervals of 20 steps. The Phonon-SED converted the time domain information of atomic velocities and positions into wave vectors versus angular frequencies via two-dimensional Fourier transform,

Table 2. Setup of hyperparameters in gplearn toolkit for GPSR.

Parameter	Value
Generations	300
Population size in every generation	5000
Probability of crossover (pc)	[0.30,0.90], step = 0.05
Probability of subtree mutation (ps)	[(1-pc)/3,(1-pc)/2] (step = 0.01)
Probability of hoist mutation (ph)	[(1-pc)/3,(1-pc)/2] (step = 0.01)
Probability of point mutation (pp)	1-pc-ps-ph
Function set	{+, -, ×, ÷, \sqrt{x} , ln x , x , − x , 1/ x }
Parsimony coefficient	0.001, 0.003, 0.005
Metric	R^2
Stopping criteria	0.900
Random_state	0, 1, 2, 3, 4
Init_depth	[2, 6], [4, 8], [6, 10], [2, 10]

expressed as

$$\Phi(q, \omega) = \frac{1}{4\pi\tau_0 N_T} \sum_a^{\{x,y,z\}} \sum_b^B m_b \left| \int_0^{\tau_0} \sum_n^{N_T} \dot{u}_a(n, b; t) \times e^{iq \cdot r(n, 0; t) - i\omega t} dt \right|^2 \quad (1)$$

Where q is the wavevector, ω is the frequency, τ_0 is the simulation time, m_b is the mass of atom b , a is the cartesian direction, N_T is the number of the unit cell in the polymer chain, $\dot{u}_a(n, b; t)$ is the velocity of atom b in the unit cell n at time t in the a direction, and $r(n, 0; t)$ is the equilibrium position of unit cell n .

DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available in the GitHub repository: <https://github.com/SJTU-MI/IMLforPTC> or from the corresponding authors on reasonable request.

CODE AVAILABILITY

The codes for physical feature down-selection and running the comparisons between the polymer representation and machine learning models are available in the GitHub repository: <https://github.com/SJTU-MI/IMLforPTC>. Detailed descriptions can be found in the Methods Section and Supplementary Information.

Received: 13 June 2023; Accepted: 4 October 2023;

Published online: 14 October 2023

REFERENCES

- Zhang, B. et al. Modulating thermal transport in polymers and interfaces: theories, simulations, and experiments. *ES Energy Environ.* **5**, 37–55 (2019).
- Ghaffari-Mosanenzedeh, S. et al. A review on high thermally conductive polymeric composites. *Polym. Compos.* **43**, 692–711 (2022).
- Xu, Q. et al. Recent progress of quantum dots for energy storage applications. *Carb. Neutral.* **1**, 13 (2022).
- Qian, X., Zhou, J. & Chen, G. Phonon-engineered extreme thermal conductivity materials. *Nat. Mater.* **20**, 1188–1202 (2021).
- Guo, Y., Zhou, Y. & Xu, Y. Engineering polymers with metal-like thermal conductivity—present status and future perspectives. *Polymer* **233**, 124168 (2021).
- Ma, H. & Tian, Z. Effects of polymer topology and morphology on thermal transport: a molecular dynamics study of bottlebrush polymers. *Appl. Phys. Lett.* **110**, 091903 (2017).
- Xu, Y. et al. Nanostructured polymer films with metal-like thermal conductivity. *Nat. Commun.* **10**, 1771 (2019).
- Shen, S., Henry, A., Tong, J., Zheng, R. & Chen, G. Polyethylene nanofibres with very high thermal conductivities. *Nat. Nanotechnol.* **5**, 251–255 (2010).
- Ma, J. et al. Thermal conductivity of electrospun polyethylene nanofibers. *Nanoscale* **7**, 16899–16908 (2015).
- Lu, C. et al. Thermal conductivity of electrospinning chain-aligned polyethylene oxide (PEO). *Polymer* **115**, 52–59 (2017).
- Singh, V. et al. High thermal conductivity of chain-oriented amorphous poly-thiophene. *Nat. Nanotechnol.* **9**, 384–390 (2014).
- Cao, B.-Y. et al. High thermal conductivity of polyethylene nanowire arrays fabricated by an improved nanoporous template wetting technique. *Polymer* **52**, 1711–1715 (2011).
- Liu, X., Lin, C. & Rao, Z. Thermal conductivity of straight-chain polytetra-fluoroethylene: a molecular dynamics study. *Int. J. Therm. Sci.* **159**, 106646 (2021).
- Crnjar, A., Melis, C. & Colombo, L. Assessing the anomalous superdiffusive heat transport in a single one-dimensional PEDOT chain. *Phys. Rev. Mater.* **2**, 015603 (2018).
- Zhang, T., Wu, X. & Luo, T. Polymer nanofibers with outstanding thermal conductivity and thermal stability: fundamental linkage between molecular characteristics and macroscopic thermal properties. *J. Phys. Chem. C* **118**, 21148–21159 (2014).
- Henry, A. & Chen, G. High thermal conductivity of single polyethylene chains using molecular dynamics simulations. *Phys. Rev. Lett.* **101**, 235502 (2008).
- Liu, J. & Yang, R. Tuning the thermal conductivity of polymers with mechanical strains. *Phys. Rev. B* **81**, 174122 (2010).
- Lin, S., Cai, Z., Wang, Y., Zhao, L. & Zhai, C. Tailored morphology and highly enhanced phonon transport in polymer fibers: a multiscale computational framework. *npj Comput. Mater.* **5**, 126 (2019).
- Li, S., Yu, X., Bao, H. & Yang, N. High thermal conductivity of bulk epoxy resin by bottom-up parallel-linking and strain: a molecular dynamics study. *J. Phys. Chem. C* **122**, 13140–13147 (2018).
- Zhang, T. & Luo, T. Role of chain morphology and stiffness in thermal conductivity of amorphous. *Polym. J. Phys. Chem. B* **120**, 803–812 (2016).
- Chen, H. et al. Thermal conductivity of polymer-based composites: fundamentals and applications. *Polym. Sci.* **59**, 41–85 (2016).
- Kim, S. et al. PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2015).
- Yang, J., Tao, L., He, J., McCutcheon, J. R. & Li, Y. Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Sci. Adv.* **8**, eabn9545 (2022).
- Chen, L. et al. Polymer informatics: current status and critical next steps. *Mater. Sci. Eng. R. Rep.* **144**, 100595 (2021).
- Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
- Wu, S., Yamada, H., Hayashi, Y., Zamengo, M. & Yoshida, R. Potentials and challenges of polymer informatics: exploiting machine learning for polymer design. Preprint at <https://doi.org/10.48550/arXiv.2010.07683> (2020).
- Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
- Huang, Y. et al. Structure–property correlation study for organic photovoltaic polymer materials using data science approach. *J. Phys. Chem. C* **124**, 12871–12882 (2020).
- Nagasawa, S., Al-Naamani, E. & Saeki, A. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *J. Phys. Chem. Lett.* **9**, 2639–2646 (2018).
- Afzal, M. A. F., Haghiglilar, M., Ganesh, S. P., Cheng, C. & Hachmann, J. Accelerated discovery of high-refractive-index polyimides via first-principles molecular modeling, virtual high-throughput screening, and data mining. *J. Phys. Chem. C* **123**, 14610–14618 (2019).
- Wu, C. et al. Flexible temperature-invariant polymer dielectrics with large bandgap. *Adv. Mater.* **32**, 2000499 (2020).
- Wu, K. et al. Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: toward optimized dielectric polymeric materials. *J. Polym. Sci., Part B: Polym. Phys.* **54**, 2082–2091 (2016).
- Sahu, H. et al. An informatics approach for designing conducting polymers. *ACS Appl. Mater. Interfaces* **13**, 53314–53322 (2021).
- Tao, L., Varshney, V. & Li, Y. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *J. Chem. Inf. Model.* **61**, 5395–5413 (2021).
- Bhowmik, R., Sih, S., Pachter, R. & Vernon, J. P. Prediction of the specific heat of polymers from experimental data and machine learning methods. *Polymer* **220**, 123558 (2021).
- Zhu, M.-X., Song, H.-G., Yu, Q.-C., Chen, J.-M. & Zhang, H.-Y. Machine-learning-driven discovery of polymers molecular structures with high thermal conductivity. *Int. J. Heat. Mass Transf.* **162**, 120381 (2020).
- Ma, R., Zhang, H. & Luo, T. Exploring high thermal conductivity amorphous polymers using reinforcement learning. *ACS Appl. Mater. Interfaces* **14**, 15587–15598 (2022).
- Wu, S. et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **5**, 66 (2019).
- Zhou, T., Wu, Z., Chilukoti, H. K. & Müller-Plathe, F. Sequence-engineering polyethylene–polypropylene copolymers with high thermal conductivity using a molecular-dynamics-based genetic algorithm. *J. Chem. Theory Comput.* **17**, 3772–3782 (2021).
- Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35 (2018).
- Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
- Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).
- Capocchi, A., Probst, D. & Raymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* **12**, 43 (2020).
- D’Souza, S., Prema, K. V. & Balaji, S. Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discov. Today* **25**, 748–756 (2020).
- Landrum, G. RDKit: open-source cheminformatics software. <https://www.rdkit.org/> (2020).

46. Karelson, M., Lobanov, V. S. & Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **96**, 1027–1044 (1996).
47. Moriaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).
48. Haghightatlari, M. et al. ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *WIREs Comput. Mol. Sci.* **10**, e1458 (2020).
49. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon software: an easy approach to molecular descriptor calculations. *Match* **56**, 237–248 (2006).
50. Roy, K. & Das, R. N. QSTR with extended topochemical atom (ETA) indices. 16. Development of predictive classification and regression models for toxicity of ionic liquids towards Daphnia magna. *J. Hazard. Mater.* **254–255**, 166–178 (2013).
51. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
52. Tiihonen, A. et al. Predicting antimicrobial activity of conjugated oligoelectrolyte molecules via machine learning. *J. Am. Chem. Soc.* **143**, 18917–18931 (2021).
53. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
54. Marchetti, F., Moroni, E., Pandini, A. & Colombo, G. Machine learning prediction of allosteric drug activity from molecular dynamics. *J. Phys. Chem. Lett.* **12**, 3724–3732 (2021).
55. Schobert, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.* **126**, 1763–1768 (2018).
56. Chandrashekhar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28 (2014).
57. Kuenneth, C. et al. Polymer informatics with multi-task learning. *Patterns* **2**, 100238 (2021).
58. Kamal, D. et al. Novel high voltage polymer insulators using computational and data-driven techniques. *J. Chem. Phys.* **154**, 174906 (2021).
59. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. in Proc. International Conference on Emerging Intelligent Data and Web Technologies 22–29 (IEEE, 2011).
60. Ma, R. & Luo, T. PI1M: a benchmark database for polymer informatics. *J. Chem. Inf. Model.* **60**, 4684–4690 (2020).
61. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
62. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. Preprint at <https://doi.org/10.48550/arXiv.1705.07874> (2017).
63. Hayashi, Y., Shioi, J., Morikawa, J. & Yoshida, R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Comput. Mater.* **8**, 222 (2022).
64. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* **3**, 861 (2018).
65. Chen, L., Tran, H., Batra, R., Kim, C. & Ramprasad, R. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Comput. Mater. Sci.* **170**, 109155 (2019).
66. Kavzoglu, T. & Mather, P. M. The role of feature selection in artificial neural network applications. *Int. J. Remote Sens.* **23**, 2919–2937 (2002).
67. Ma, H. & Tian, Z. Chain rotation significantly reduces thermal conductivity of single-chain polymers. *J. Mater. Res.* **34**, 126–133 (2019).
68. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
69. Robbins, A. B. & Minnich, A. J. Crystalline polymers with exceptionally low thermal conductivity studied using molecular dynamics. *Appl. Phys. Lett.* **107**, 201908 (2015).
70. Liu, J. & Yang, R. Length-dependent thermal conductivity of single extended polymer chains. *Phys. Rev. B* **86**, 104307 (2012).
71. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
72. Weng, B. et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat. Commun.* **11**, 3513 (2020).
73. Wang, Y., Wagner, N. & Rondinelli, J. M. Symbolic regression in materials science. *MRS Commun.* **9**, 793–805 (2019).
74. Stephens, T. Genetic programming in Python, with a scikit-learn inspired API: gplearn. <https://gplearn.readthedocs.io/en/stable/>.
75. Zhou, Y., Rao, Y., Zhang, L., Ju, S. & Wang, H. Machine-learning prediction of Vegard's law factor and volume size factor for binary substitutional metallic solid solutions. *Acta Mater.* **237**, 118166 (2022).
76. Paulson, N. H., Libera, J. A. & Stan, M. Flame spray pyrolysis optimization via statistics and machine learning. *Mater. Des.* **196**, 108972 (2020).
77. Agarwal, G., Doan, H. A., Robertson, L. A., Zhang, L. & Assary, R. S. Discovery of energy storage molecular materials using quantum chemistry-guided multi-objective bayesian optimization. *Chem. Mater.* **33**, 8133–8144 (2021).
78. Thomas, J. A., Turney, J. E., Iutzi, R. M., Amon, C. H. & McGaughey, A. J. H. Predicting phonon dispersion relations and lifetimes from the spectral energy density. *Phys. Rev. B* **81**, 081411 (2010).
79. Müller-Plathe, F. A simple nonequilibrium molecular dynamics method for calculating the thermal conductivity. *J. Chem. Phys.* **106**, 6082–6085 (1997).
80. Pawlak, A. The entanglements of macromolecules and their influence on the properties of polymers. *Macromol. Chem. Phys.* **220**, 1900043 (2019).
81. Di Benedetto, F. et al. Patterning of light-emitting conjugated polymer nanofibres. *Nat. Nanotechnol.* **3**, 614–619 (2008).
82. Hagaman, D. et al. Block copolymer supramolecular assembly using a precursor to a novel conjugated polymer. *Polym. Chem.* **4**, 1482–1490 (2013).
83. Turcani, L., Berardo, E. & Jelfs, K. E. stk: a python toolkit for supramolecular assembly. *J. Comput. Chem.* **39**, 1931–1942 (2018).
84. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminform.* **6**, 37 (2014).
85. Fortunato, M. E. & Colina, C. M. pysimm: a python package for simulation of molecular systems. *SoftwareX* **6**, 7–12 (2017).
86. Zhao, Y. H., Abraham, M. H. & Zissimos, A. M. Fast calculation of van der waals volume as a sum of atomic and bond contributions and its application to drug compounds. *J. Org. Chem.* **68**, 7368–7373 (2003).
87. Hu, Y. et al. Unification of nonequilibrium molecular dynamics and the mode-resolved phonon Boltzmann equation for thermal transport simulations. *Phys. Rev. B* **101**, 155308 (2020).
88. Das, S. & Muthukumar, M. Thermal conduction and phonon transport in folded polyethylene chains. *Macromolecules* **56**, 393–403 (2023).
89. Yang, N., Zhang, G. & Li, B. Violation of Fourier's law and anomalous heat diffusion in silicon nanowires. *Nano Today* **5**, 85–90 (2010).
90. Zhang, T. & Luo, T. Morphology-influenced thermal conductivity of polyethylene single chains and crystalline fibers. *J. Appl. Phys.* **112**, 094304 (2012).
91. Shrestha, R. et al. Crystalline polymer nanofibers with ultra-high strength and thermal conductivity. *Nat. Commun.* **9**, 1664 (2018).
92. Wirsberger, P., Frenkel, D. & Dellago, C. An enhanced version of the heat exchange algorithm with excellent energy conservation properties. *J. Chem. Phys.* **143**, 124104 (2015).
93. Ju, S., Palpant, B. & Chalopin, Y. Adverse effects of polymer coating on heat transport at the solid–liquid. *Interface J. Phys. Chem. C* **121**, 13474–13480 (2017).
94. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
95. Nogueira, F. Bayesian optimization: open source constrained global optimization tool for Python. <https://github.com/fmfn/BayesianOptimization> (2014).

ACKNOWLEDGEMENTS

This work was supported by the Shanghai Pujiang Program (No. 20PJ1407500), the National Natural Science Foundation of China (No. 52006134), the Shanghai Key Fundamental Research Grant (No. 21JC1403300), and the SJTU Global Strategic Partnership Fund (2022 SJTU-Warwick). The computations in this paper were run on the π 2.0 cluster supported by the Center for High-Performance Computing at Shanghai Jiao Tong University.

AUTHOR CONTRIBUTIONS

S.J. conceived the idea and supervised the project. X.H. designed the research, trained and evaluated the machine learning models, as well as performed the molecular dynamics simulations. S.M. executed the feature engineering and symbolic regression. C.Y.Z. and H.W. supervised the project. all authors modified and approved the manuscript. X.H. and S.M. are to be considered co-first authors.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-023-01154-w>.

Correspondence and requests for materials should be addressed to Shenghong Ju.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Supplemental Materials for

Exploring high thermal conductivity polymers via interpretable machine learning with physical descriptors

Xiang Huang^{1, †}, Shengluo Ma^{1, †}, C. Y. Zhao¹, Hong Wang², and Shenghong Ju^{1, 2, *}

¹ China-UK Low Carbon College, Shanghai Jiao Tong University, Shanghai, China

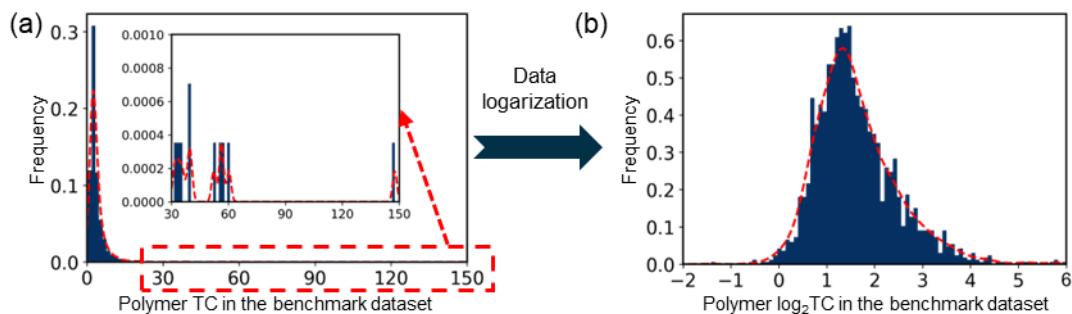
² Materials Genome Initiative Center, School of Material Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[†] Co-first author

* Corresponding email: shenghong.ju@sjtu.edu.cn

Supplementary Note 1. Thermal conductivity distribution of the benchmark data

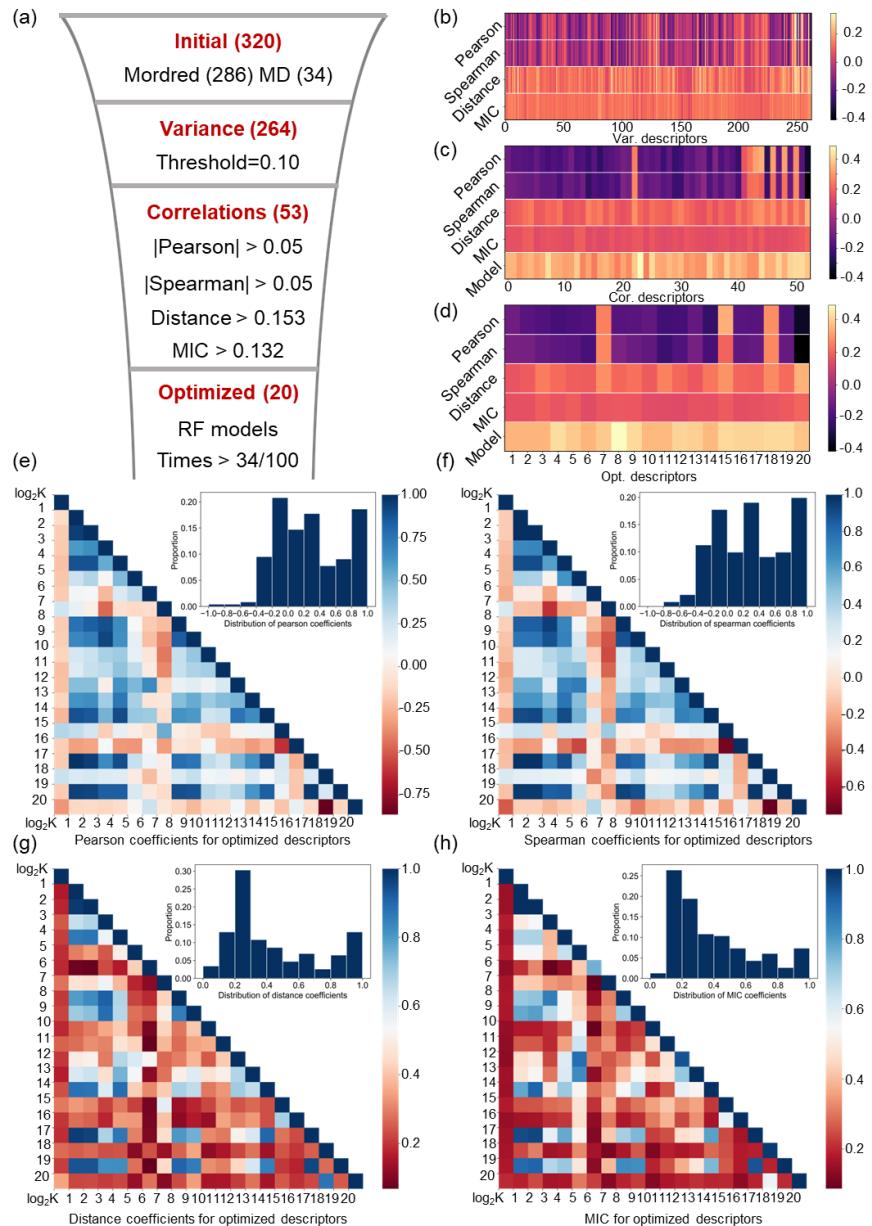
Supplementary Figure 1 illustrates the distributions of raw thermal conductivity (TC) and log-transformed thermal conductivity ($\log_2\text{TC}$) in the benchmark dataset. Given the wide TC distribution of the original data, the log-transformed TC of $\log_2\text{TC}$ is closer to the normal distribution, which contributes to the generalization ability of the machine learning model across the thermal conductivity range .



Supplementary Figure 1: Distribution of (a) TC and (b) $\log_2\text{TC}$ in the benchmark dataset

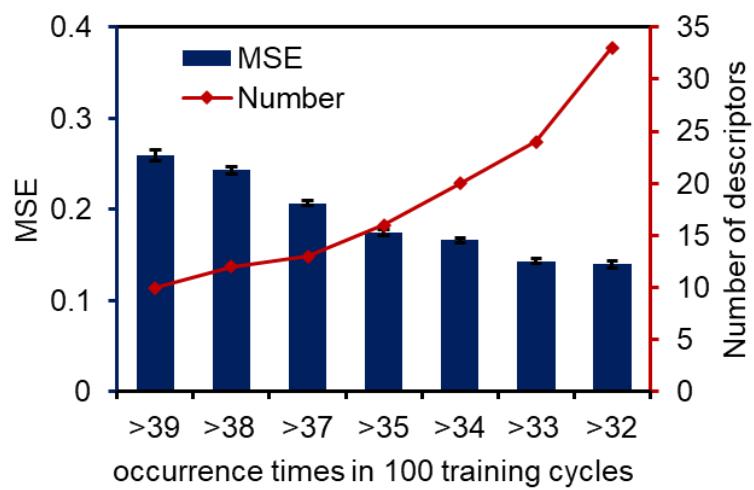
Supplementary Note 2. Downselection of polymer descriptors

Polymer descriptors were downselected in three stages: removing features with low variance (Var.), primary filtering referred to different correlation coefficients (Cor.), and final selection assisted with ML model (Opt.), as shown in Supplementary Figure 2.



Supplementary Figure 2: Downselection and analysis of descriptors. (a) Process of descriptors downselection. (b), (c) and (d) Relationship between descriptors at different stages (Var., Cor. and Opt.) and thermal conductivity based on various metrics or frequency of occurrence of descriptors through 100 Random Forest (RF) model selections. (e), (f), (g) and (h) Pearson, Spearman, Distance and MIC correlation matrices showing correlations among optimized descriptors and TC. The inset shows the distribution statistics of the corresponding correlation coefficient values respectively. The description of the 20 optimized descriptors is available in Supplementary Note S3.

We trained the RF model 100 cycles during the feature selection stage, and randomly disrupted the input descriptors before each running. The optimized descriptors were determined by a statistical threshold, whereby the occurrence of descriptors with a frequency greater than this threshold is retained. Supplementary Figure 3 shows the performance of the RF model with different physical descriptor ensembles determined by various thresholds from 0.39 to 0.32. Overall, as the threshold decreases, the number of reserved descriptors increases and its performance becomes better. Balancing the mean-square error (MSE) of the ML model and the number of descriptors, 20 optimized descriptors were retained with a threshold of 0.34.



Supplementary Figure 3: Performance of ML models trained with different ensemble thresholds-determined physical descriptors

Supplementary Note 3. List of MD-inspired and Mordred-based fingerprints

The MD-inspired descriptors include force field related (FF-related) descriptors and thermal conductivity related (TC- related) descriptors, listed in Supplementary Table 1.

Supplementary Table 1. Descriptions of all MD-inspired descriptors

Name	Description
Epsilon_max	Maximum value of the depth of the energy potential (Lennard–Jones parameter)
Epsilon_min	Minimum value of the depth of the energy potential (Lennard–Jones parameter)
Epsilon_average	Average value of the depth of the energy potential (Lennard–Jones parameter)
Sigma_max	Maximum value of the equilibrium distance (Lennard–Jones parameter)
Sigma_min	Minimum value of the equilibrium distance (Lennard–Jones parameter)
Sigma_average	Average value of the equilibrium distance (Lennard–Jones parameter)
Kb_max	Maximum value of force constants of the bond
Kb_min	Minimum value of force constants of the bond
Kb_average	Average value of force constants of the bond
R0_max	Maximum value of equilibration structural parameters of the bond
R0_min	Minimum value of equilibration structural parameters of the bond
R0_average	Average value of equilibration structural parameters of the bond
Ka_max	Maximum value of force constants of the bond angle
Ka_min	Minimum value of force constants of the bond angle
Ka_average	Average value of force constants of the bond angle
Theta0_max	Maximum value of equilibration structural parameters of the bond angle
Theta0_min	Minimum value of equilibration structural parameters of the bond angle
Theta0_average	Average value of equilibration structural parameters of the bond angle
Kd_max	Maximum value of force constants of the dihedral angle
Kd_min	Minimum value of force constants of the dihedral angle
Kd_average	Average value of force constants of the dihedral angle
Nd_max	Maximum value of multiplicity for the torsional angle parameters
Nd_min	Minimum value of multiplicity for the torsional angle parameters
Nd_average	Average value of multiplicity for the torsional angle parameters
Delta_max	Maximum value of phase angle for the torsional angle parameters
Delta_min	Minimum value of phase angle for the torsional angle parameters
Delta_average	Average value of phase angle for the torsional angle parameters
Ki_max	Maximum value of force constants of the improper angle
Ki_min	Minimum value of force constants of the improper angle
Ki_average	Average value of force constants of the improper angle
MW	Molecular weight of monomer
MW_ratio	The ratio of the molecular weight of the main chain to that of the monomer
Vdw	Van der Waals volume of the monomer
Cross-sectional	Equivalent cross-sectional area of polymer chain

The Mordred-based descriptors were calculated by Mordred software , the 286 Mordred-based descriptors in this work are listed in Supplementary Table 2, and the detailed meaning could be found at <https://mordred-descriptor.github.io/documentation/master/descriptors.html> (accessed Aug 10, 2022).

Supplementary Table 2. Descriptions of 286 Mordred-based descriptors

ABC	nAcid	nBase	SpMax_A	SpMAD_A
LogEE_A	VE1_A	VR1_A	VR3_A	nAromAtom
nAtom	nBridgehead	nHetero	nH	nN
nO	nS	nF	nCl	nBr
ATS0dv	ATS0Z	AATS0dv	AATS0d	AATS2d
AATS0Z	AATS1Z	AATS2Z	AATS3Z	ATSC0dv
ATSC1dv	ATSC2dv	ATSC3dv	ATSC4dv	ATSC5dv
ATSC6dv	ATSC7dv	ATSC8dv	ATSC1d	ATSC2d
ATSC3d	ATSC4d	ATSC5d	ATSC6d	ATSC7d
ATSC8d	ATSC0Z	ATSC1Z	ATSC2Z	ATSC3Z
ATSC4Z	ATSC5Z	ATSC6Z	ATSC7Z	ATSC8Z
AATSC0dv	AATSC1dv	AATSC2dv	AATSC3dv	AATSC0d
AATSC1d	AATSC2d	AATSC3d	AATSC0Z	AATSC1Z
AATSC2Z	AATSC3Z	MATS1dv	MATS2dv	MATS3dv
MATS1d	MATS2d	MATS3d	MATS1Z	MATS2Z
MATS3Z	GATS1dv	GATS2dv	GATS3dv	GATS1d
GATS2d	GATS3d	GATS1Z	GATS2Z	GATS3Z
BCUTdv-1h	BCUTdv-11	BCUTd-1h	BCUTd-11	BCUTZ-1h
BCUTZ-11	BalabanJ	BertzCT	nBondsD	nBondsT
C1SP1	C2SP1	C1SP2	C3SP2	C1SP3
C2SP3	C3SP3	C4SP3	HybRatio	Xch-4d
Xch-5d	Xch-6d	Xch-4dv	Xc-3d	Xc-4d
Xc-5d	Xc-6d	Xpc-4d	AXp-0d	AXp-1d
AXp-2d	AXp-3d	NsCH3	NdCH2	NssCH2
NdsCH	NsssCH	NddC	NdssC	NaaaC
NssssC	NsNH2	NssNH	NaaNH	NdsN
NaaN	NsssN	NaasN	NsOH	NdO
NssO	NaaO	NdS	NssS	NaaS
NddssS	SsssCH	SdssC	SaasC	SaaaC
SsssN	SaasN	AETA_beta	AETA_beta_s	ETA_beta_ns_d
AETA_beta_ns_d	AETA_eta_RL	AETA_eta_BR	ETA_epsilon_3	ETA_dBeta
fragCpx	fMF	nHBAcc	IC0	IC1
IC2	IC3	SIC0	SIC1	SIC2
SIC3	SIC4	BIC5	CIC0	MIC0
MIC1	MIC2	Kier2	Kier3	FilterItLogS
PEOE_VSA1	PEOE_VSA2	PEOE_VSA3	PEOE_VSA4	PEOE_VSA6
PEOE_VSA7	PEOE_VSA8	PEOE_VSA9	PEOE_VSA10	PEOE_VSA11

PEOE_VSA12	PEOE_VSA13	SMR_VSA1	SMR_VSA3	SMR_VSA4
SMR_VSA6	SMR_VSA9	SlogP_VSA1	SlogP_VSA2	SlogP_VSA3
SlogP_VSA4	SlogP_VSA5	SlogP_VSA7	SlogP_VSA8	SlogP_VSA10
SlogP_VSA11	EState_VSA1	EState_VSA2	EState_VSA3	EState_VSA4
EState_VSA5	EState_VSA6	EState_VSA7	EState_VSA8	EState_VSA9
EState_VSA10	VSA_EState1	VSA_EState2	VSA_EState3	VSA_EState4
VSA_EState5	VSA_EState7	VSA_EState8	VSA_EState9	AMID_h
AMID_N	AMID_O	AMID_X	nRing	n5Ring
n6Ring	n7Ring	nHRing	n4HRing	n5HRing
n6HRing	n5aRing	naHRing	n6aHRing	nARing
n5ARing	n6ARing	nAHRing	n5AHRing	n6AHRing
nFRing	n7FRing	n8FRing	n9FRing	n10FRing
n11FRing	n12FRing	nG12FRing	nFHRing	n7FHRing
n8FHRing	n10FHRing	nG12FHRing	nFaRing	n8FaRing
n9FaRing	n10FaRing	nG12FaRing	nFaHRing	nFARing
n10FARing	n12FARing	nG12FARing	nFAHRing	nG12FAHRing
RotRatio	SLogP	TopoPSA(NO)	TopoPSA	GGI4
GGI5	GGI6	GGI7	JGI1	JGI2
JGI3	JGI4	JGI5	JGI6	JGI7
JGI8	JGI9	JGI10	JGT10	TopoShapeIndex
mZagreb1				

Supplementary Note 4. Descriptions of optimized descriptors for polymer thermal conductivity (TC) prediction

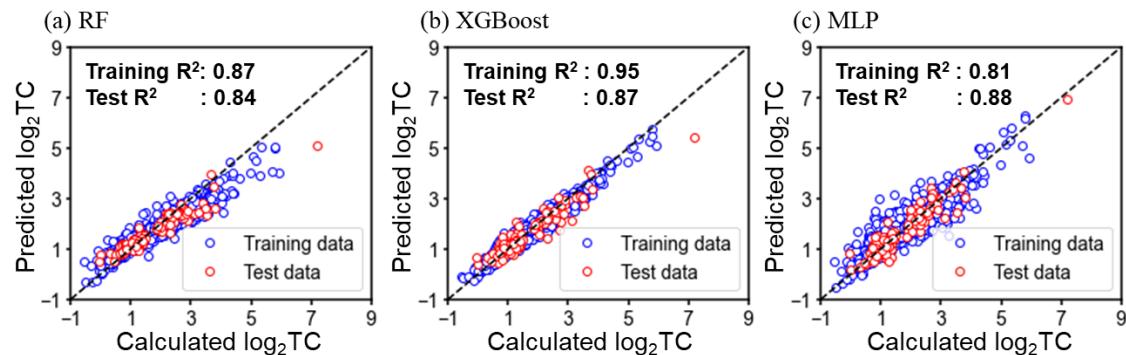
Polymer optimized descriptors were obtained by downselection, listed in Supplementary Table 3.

Supplementary Table 3. Descriptions of all Optimized descriptors

No.	Name	Description	Block
1	ABC	Atom-bond connectivity index	Atom-bond connectivity index descriptor
2	LogEE_A	LogEE of adjacency matrix	Adjacency matrix descriptor
3	Nh	Number of H atoms	Atom count descriptor
4	ATSC0dv	Centered moreau-broto autocorrelation of lag 0 weighted by valence electrons	Centered Autocorrelation of Topological Structure descriptor
5	BCUTdv-1h	First highest eigenvalue of Burden matrix weighted by valence electrons	BCUT descriptor
6	Xc-4d	4-ordered Chi cluster weighted by sigma electrons	Chi descriptor
7	ETA_dBeta	ETA delta beta	ETA delta beta descriptor
8	CIC0	0-ordered complementary information content	Complementary information content descriptor
9	Kier2	kappa shape index 2	Kappa shape index 2 descriptor
10	SMR_VSA6	MOE MR VSA Descriptor 6 (2.75 <= x < 3.05)	MOE type descriptors using Wildman-Crippen MR and surface area contribution
11	SlogP_VSA3	MOE logP VSA Descriptor 3 (-0.20 <= x < 0.00)	MOE type descriptors using Wildman-Crippen LogP and surface area contribution
12	EState_VSA10	EState VSA Descriptor 10 (9.17 <= x < 15.00)	MOE type descriptors using EState indices and surface area contribution
13	TopoPSA(NO)	Topological polar surface area (use only nitrogen and oxygen)	Topological polar surface area descriptor
14	mZagreb1	modified Zagreb index (version 1)	Zagreb index descriptor
15	Kd_average	Average value of force constants of the dihedral angle	FF-related descriptor
16	Nd_average	Average value of multiplicity for the torsional angle parameters	FF-related descriptor
17	MW	Molecular weight of monomer	TC- related descriptor
18	MW_ratio	The ratio of the molecular weight of the main chain to that of the monomer	TC- related descriptor
19	Vdw	Van der Waals volume of the monomer	TC- related descriptor
20	Cross-sectional	Equivalent cross-sectional area of polymer chain	TC- related descriptor

Supplementary Note 5. Construction of different ML models with optimized descriptors

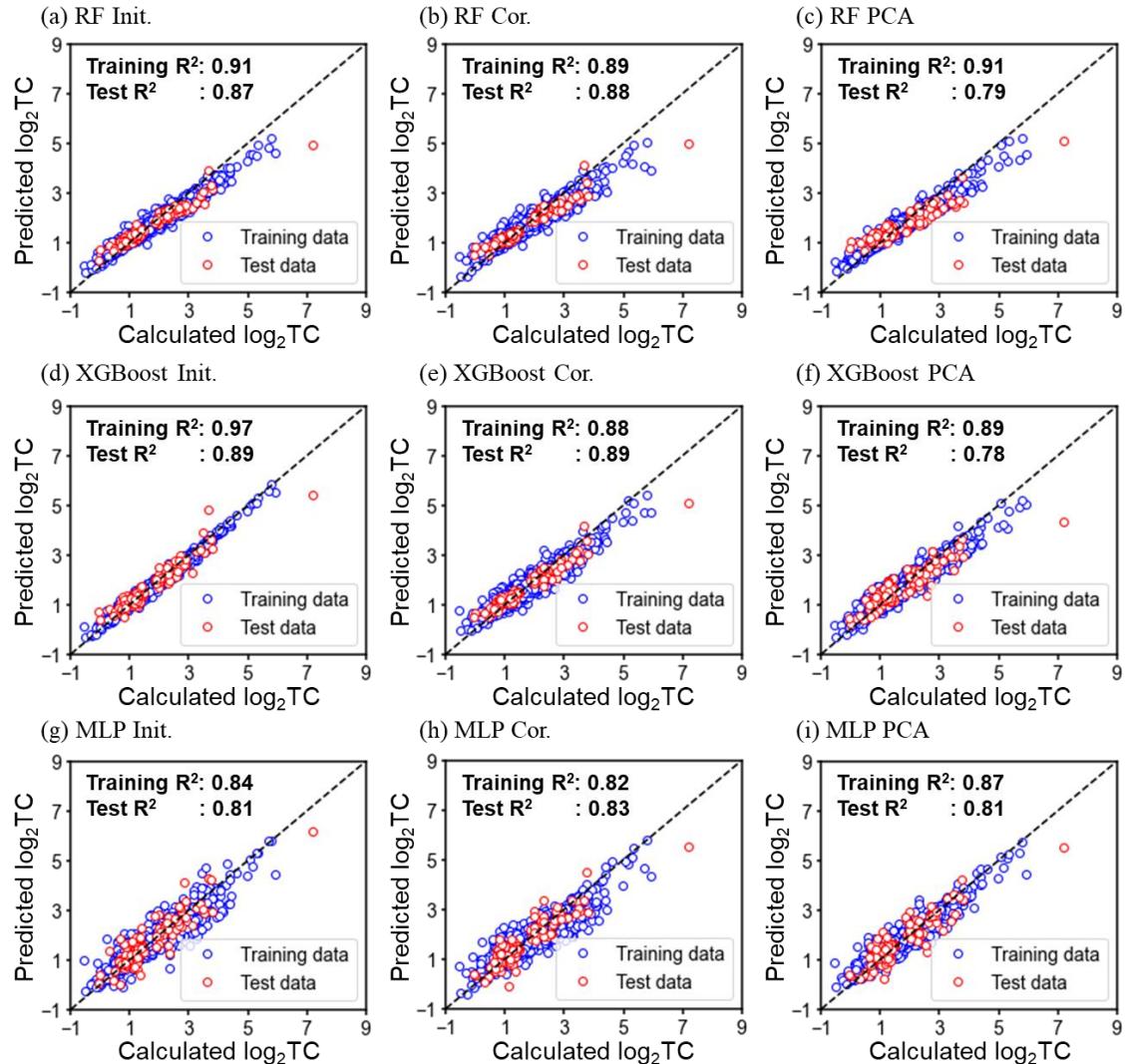
Supplementary Figure 4 shows the RF, XGBoost and MLP models trained with the optimized descriptors. The test set consisted of 100 randomly selected polymer structures from the benchmark dataset.



Supplementary Figure 4: Construction of different ML models with optimized descriptors

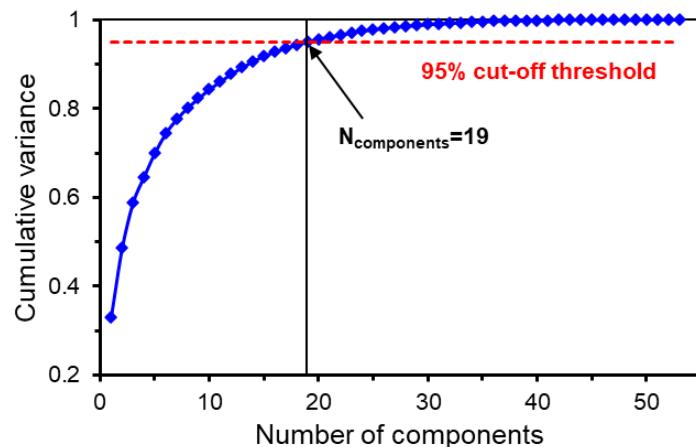
Supplementary Note 6. Accuracy of ML models trained by different sets of descriptors

Supplementary Figure 5 shows the accuracy of ML models at different descriptors down-selection stages. The extra principal component analysis (PCA)¹ technique with 95% variance was used to compare.



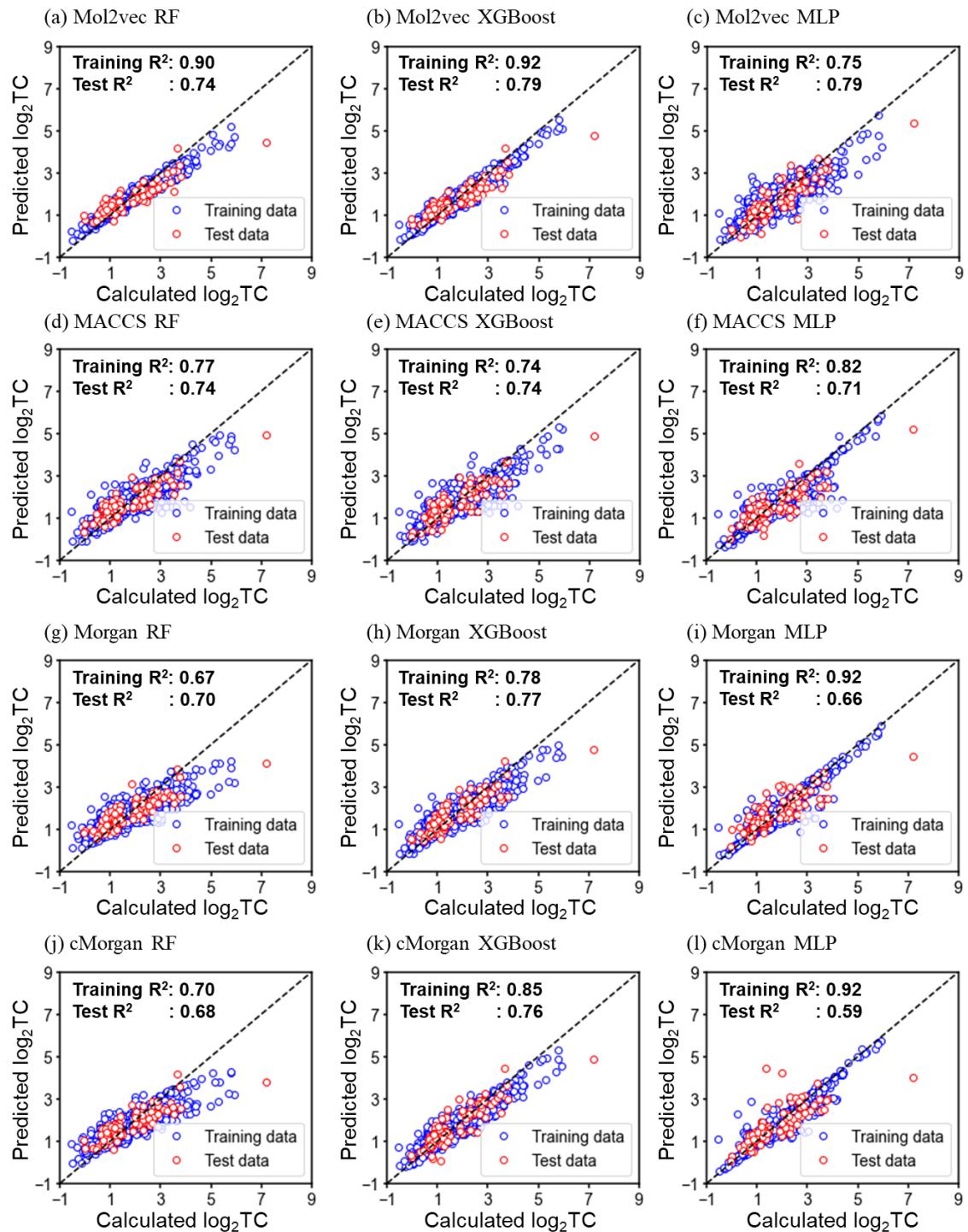
Supplementary Figure 5: Accuracy of ML models at different descriptors downselection stages

Supplementary Figure 6 represents the relationship between the number of principal components and cumulative variance. Here, the PCA with 19 principal components of 95% variance was performed.



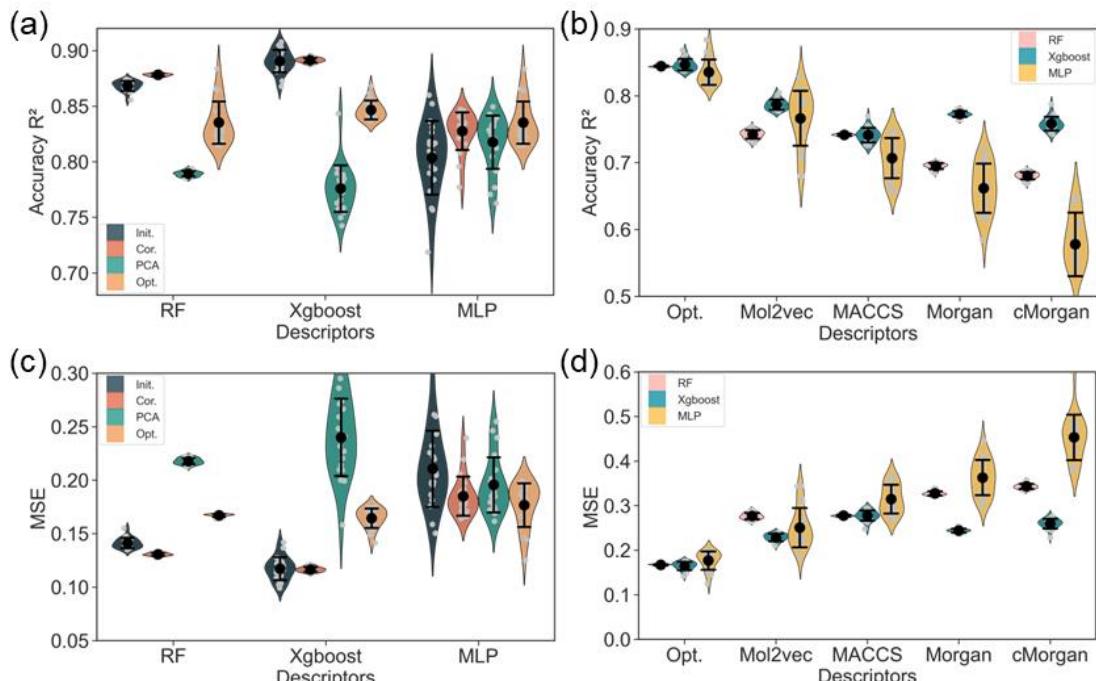
Supplementary Figure 6: Relationship between the number of principal components and cumulative variance

Supplementary Figure 7 shows the training and testing accuracy of ML models using Mol2vec , MACCS , Morgan and Morgan count (cMorgan) graph descriptors.



Supplementary Figure 7: Accuracy of ML models with various graph descriptors

Supplementary Figure 8 displays the accuracy of the trained ML models with physical descriptors at different down-selection stages or with graph descriptors. Combining the results of R^2 and mean-square error (MSE), the performance of machine learning has a slight reduction accompanied by a reduction in the dimensionality of the physical descriptors, but the ML models trained by optimized descriptors are always better than 95% PCA-derived descriptors (Supplementary Figure 8a and 8c). Further, optimized descriptors have significant advantages over the other graph descriptors (Supplementary Figure 8b and 8d).

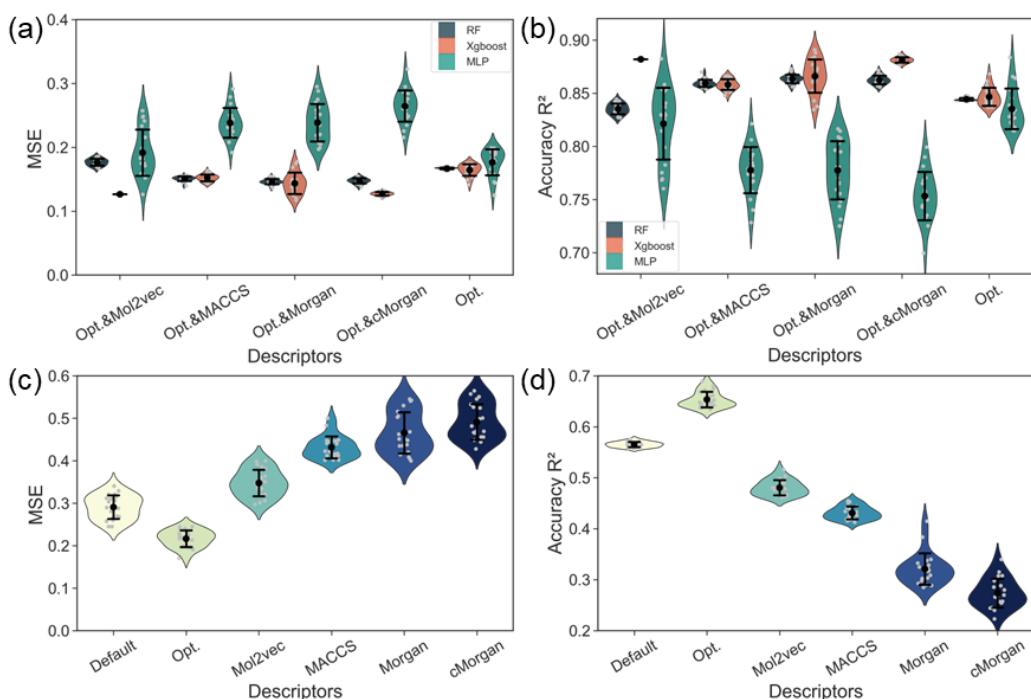


Supplementary Figure 8: Accuracy of the trained ML models with different types of descriptors. (a) and (c) Accuracy R^2 and mean-square error (MSE) of ML models at different down-selection stages. (b) and (d) Accuracy R^2 and MSE of ML models with different polymer representation approaches.

We additionally evaluated ML models with a hybrid descriptor set composed of the optimized descriptors and one of the graph descriptors in Supplementary Figure 9a and 9b. The performance of RF and XGBoost models trained with hybrid descriptors is only a small improvement or even comparable to that of trained with optimized descriptors, which reflects the fact that the optimized descriptors cover relatively complete information about the polymer structures. Meanwhile, due to the redundant information introduced by the hybrid descriptors, the trained MLP models are overfitted and the overall performance is worse than that of the model trained with the optimized descriptors.

Furthermore, we applied the optimized descriptors or graph descriptors to train the directed message passing neural network (DMPNN) models in Chemprop . The benchmark dataset containing

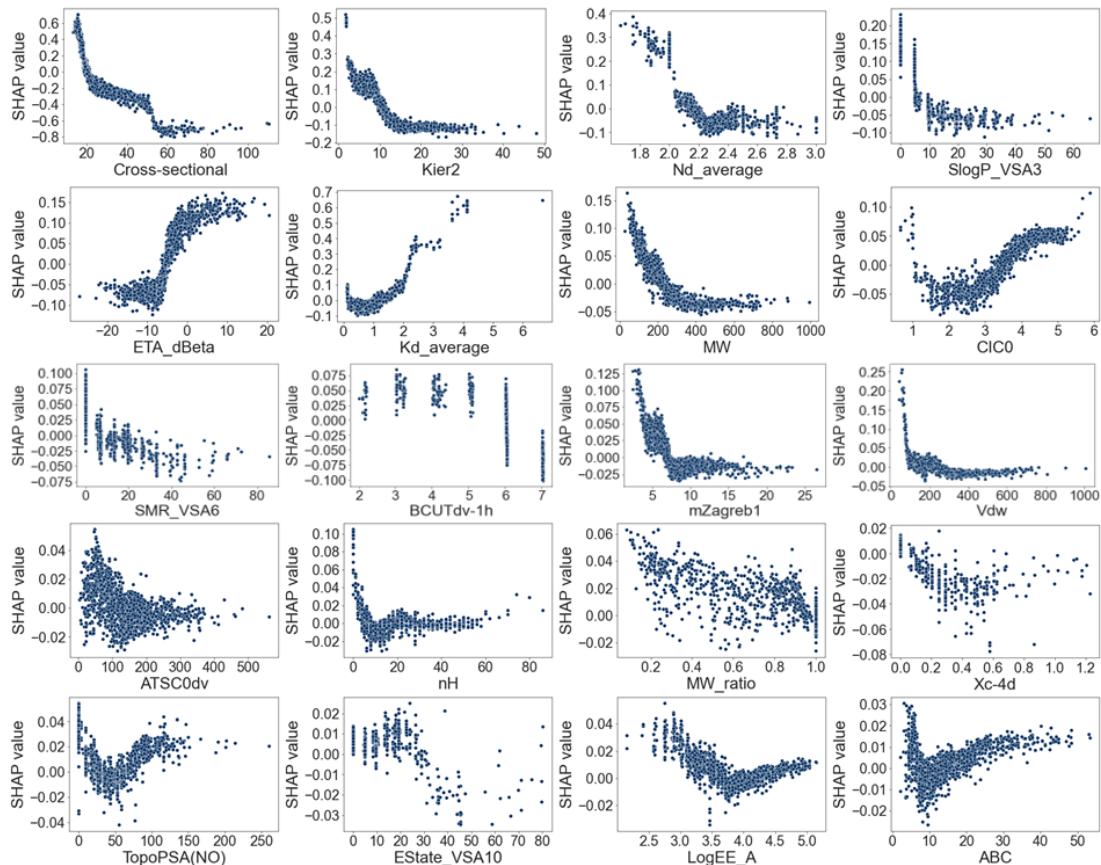
1735 polymer structures was randomly divided by training data, test data and validation data of 8:1:1, and the hyperparameters for the DMPNN models were optimized using the internal implementation of Chemprop, which relies on Bayesian optimization. The Chemprop toolkit has a built-in 2D molecular graph representation for training ML models, which we call the default descriptors. As shown in Supplementary Figure 9c and 9d. despite the limited amount of data in the available benchmark dataset makes it difficult to output a high-performance DMPNN model, the performance of the optimized descriptors is the best compared to other descriptors. This illustrates the potential of optimized descriptors for applications in diverse and complex ML models.



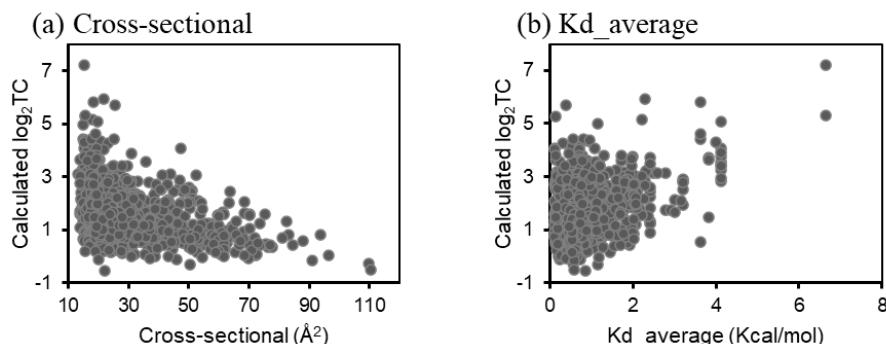
Supplementary Figure 9: Evaluation of ML models with various types of descriptors. (a) MSE and R^2 of ML models which trained by a hybrid descriptor set composed of the optimized descriptors and one of the graph descriptors. (b) MSE and R^2 of directed message passing neural network (DMPNN) models trained with different types of descriptors.

Supplementary Note 7. Analysis of the feature importance

Supplementary Figure 10 exhibits the SHAP values for each optimized descriptor of the training data set polymers as a function of the corresponding descriptor value. The subplots are sorted by the average SHAP value. The SHAP values of the top-ranked features corresponding to the training dataset polymers show a monotonic relationship with the feature values overall, while the lower-ranked features have difficulty in capturing this rule. Wherein, the relationships between two important MD-inspired descriptors and TC are shown in Supplementary Figure 11.



Supplementary Figure 10: Distribution of SHAP values for each optimized descriptor



Supplementary Figure 11: Relationships between two important MD-inspired descriptors and TC of 1735 benchmark polymers.

Supplementary Note 8 Discovery of polymers with high thermal conductivity (PHTC) by proposed ML workflow

We performed virtual screening of polymer data in PolyInfo and PI1M based on the trained RF, XGBoost and MLP machine learning models, respectively. In this work, a total of 107 polymer structures with thermal conductivity greater than $20 \text{ W m}^{-1}\text{K}^{-1}$ were identified and validated by molecular dynamics simulations, as listed in Supplementary Table 4, and the relevant monomer structure can be viewed in Supplementary Figure 12 by polymer ID (PID). The TC obtained from RF, XGBoost and MLP model predictions and MD simulations are $\log_2\text{TC}$ in $\text{W m}^{-1}\text{K}^{-1}$. The effective cross-sectional area of the polymer is given in \AA^2 . In addition, we calculated the thermal conductivity of 58 of these structures in the amorphous state using nonequilibrium molecular dynamics. The TC of these amorphous polymers (ATC, $\text{W m}^{-1}\text{K}^{-1}$) easily results in a large value, owing to the large chain stiffness.

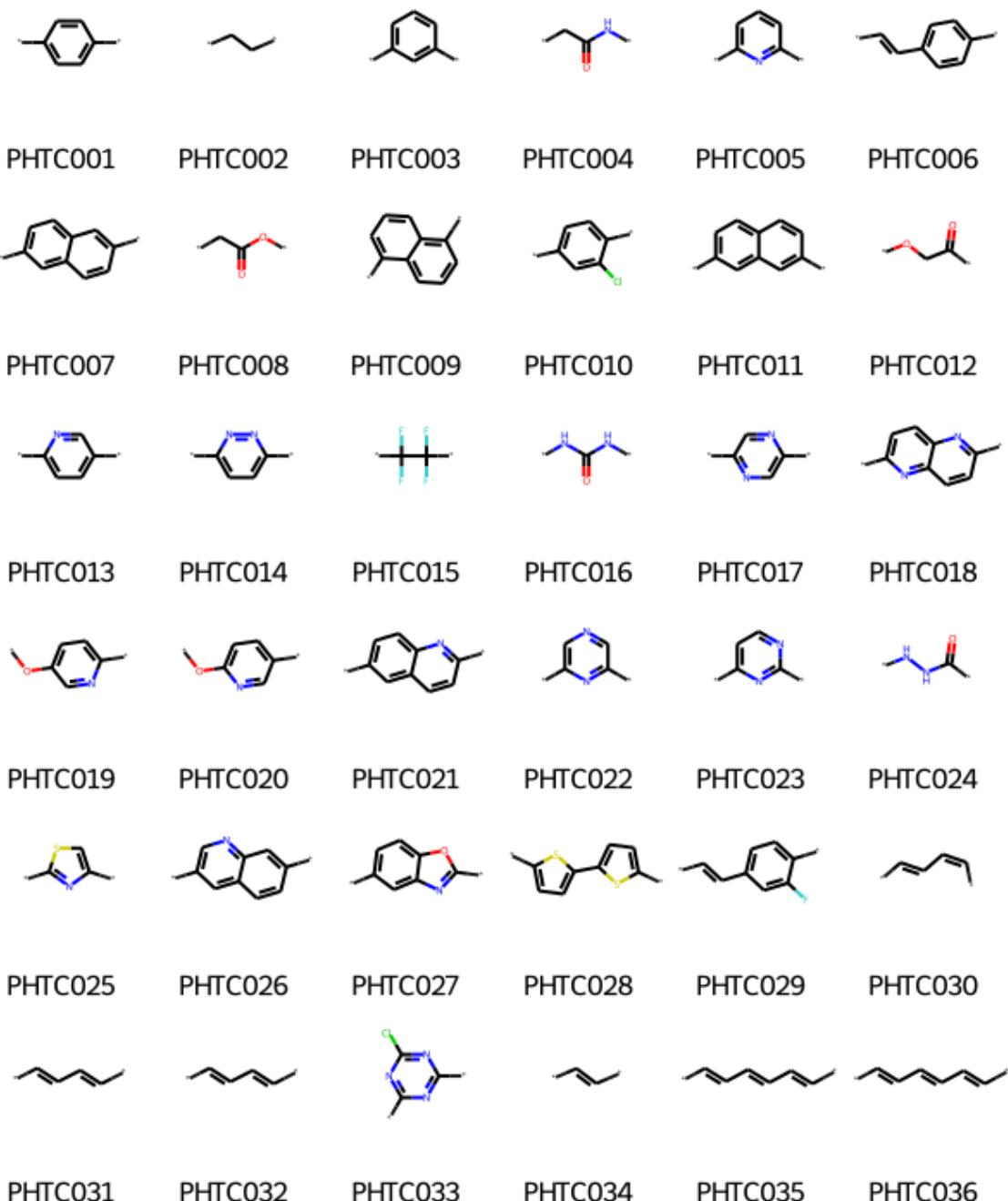
Supplementary Table 4. MD-validated polymers with thermal conductivity greater than $20 \text{ W m}^{-1}\text{K}^{-1}$

PID	SMILES	SA	Cross	RF	XGBoost	MLP	MD	ATC
PHTC001	[*]c1ccc([*])cc1	1.05	18.32	5.00	5.77	6.28	5.81	1.22
PHTC002	[*]CC[*]	1.12	15.69	3.91	4.66	5.30	5.28	0.46
PHTC003	[*]c1cccc([*])c1	1.50	20.74	4.02	5.21	5.86	4.59	0.28
PHTC004	[*]CC(=O)N[*]	2.07	15.13	4.00	4.34	4.05	4.41	0.38
PHTC005	[*]c1cccc([*])n1	2.15	19.72	4.65	4.98	5.65	5.06	0.38
PHTC006	[*]C=Cc1ccc([*])cc1	2.17	16.76	4.29	4.36	5.48	5.13	1.22
PHTC007	[*]c1ccc2cc([*])ccc2c1	2.17	18.28	3.55	4.37	5.01	5.12	1.20
PHTC008	[*]CC(=O)O[*]	2.21	15.09	4.01	4.45	4.36	4.98	0.23
PHTC009	[*]c1cccc2c([*])cccc12	2.22	23.86	3.17	3.81	4.36	4.35	
PHTC010	[*]c1ccc([*])c(Cl)c1	2.23	21.89	3.73	4.06	5.34	4.61	0.74
PHTC011	[*]c1ccc2ccc([*])cc2c1	2.28	19.21	3.81	4.49	5.05	4.62	0.35
PHTC012	[*]OCC([*])=O	2.28	20.86	3.31	4.17	3.70	4.33	0.27
PHTC013	[*]c1ccc([*])nc1	2.33	16.98	4.85	5.08	5.45	4.50	
PHTC014	[*]c1ccc([*])nn1	2.48	15.68	4.87	5.91	4.97	5.05	
PHTC015	[*]C(F)(F)C([*])(F)F	2.51	25.57	4.05	5.27	5.07	5.71	0.20
PHTC016	[*]NC(=O)N[*]	2.57	14.34	4.53	4.91	3.76	5.47	0.53
PHTC017	[*]c1cnc([*])cn1	2.62	15.75	5.11	6.08	5.85	6.62	
PHTC018	[*]c1ccc2nc([*])ccc2n1	2.64	16.79	3.96	4.00	5.62	5.93	
PHTC019	[*]Oc1ccc([*])nc1	2.65	14.79	4.16	4.46	4.17	4.43	
PHTC020	[*]Oc1ccc([*])cn1	2.65	14.74	4.16	4.52	4.17	4.76	0.23
PHTC021	[*]c1ccc2nc([*])ccc2c1	2.78	17.62	3.96	4.19	5.59	4.38	
PHTC022	[*]c1cncc([*])n1	2.80	18.06	5.08	5.84	5.53	4.34	
PHTC023	[*]c1ccn([*])n1	2.82	18.06	5.08	5.82	5.52	5.00	0.25

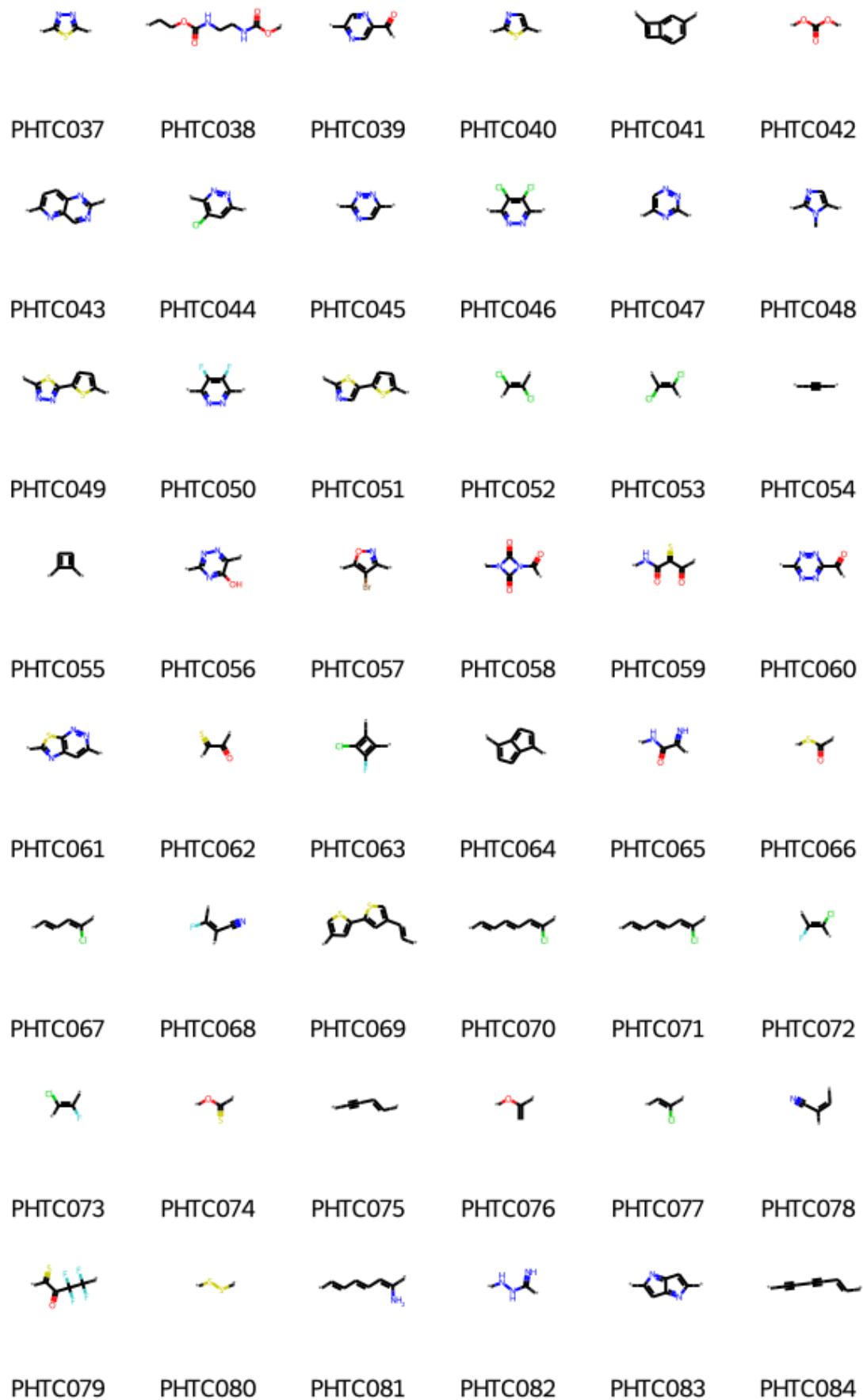
PID	SMILES	SA	Cross	RF	XGBoost	MLP	MD	ATC
PHTC024	[*]NNC([*])=O	2.86	14.28	4.19	4.56	3.32	4.83	0.46
PHTC025	[*]c1csc([*])n1	2.87	15.82	5.11	6.20	5.76	5.75	0.27
PHTC026	[*]c1cnc2cc([*])ccc2c1	2.91	17.29	3.99	4.24	5.63	4.82	0.72
PHTC027	[*]c1ccc2oc([*])nc2c1	2.97	15.17	3.91	3.55	3.59	4.34	0.44
PHTC028	[*]c1ccc(-c2ccc([*])s2)s1	3.00	15.98	4.06	4.26	4.22	5.64	0.42
PHTC029	[*]C=Cc1ccc([*])c(F)c1	3.04	17.60	3.89	4.34	4.80	4.41	
PHTC030	[*]C=C/C=C\[*]	3.07	15.68	5.06	5.35	5.79	5.33	
PHTC031	[*]C=C/C=C/[*]	3.07	14.06	5.05	5.10	5.97	7.55	1.13
PHTC032	[*]/C=C/C=C/[*]	3.07	14.26	5.05	5.10	5.95	6.91	0.89
PHTC033	[*]c1nc([*])nc(Cl)n1	3.09	13.85	4.53	5.64	5.63	6.51	0.15
PHTC034	[*]C=C[*]	3.12	15.35	5.08	5.44	6.92	7.21	0.83
PHTC035	[*]C=C/C=C/C=C/[*]	3.20	13.73	4.65	4.34	4.86	7.42	1.02
PHTC036	[*]/C=C/C=C/C=C/[*]	3.20	13.60	4.65	4.34	4.87	7.31	1.11
PHTC037	[*]c1nn([*])s1	3.27	8.76	5.09	6.04	6.14	5.75	0.59
PHTC038	[*]CCOC(=O)NCCNC(=O)O[*]	3.33	15.55	3.20	4.10	3.65	4.38	
PHTC039	[*]C(=O)c1cnc([*])cn1	3.34	15.80	4.11	4.34	4.35	4.54	
PHTC040	[*]c1cnc([*])s1	3.35	15.34	4.86	5.77	5.40	4.95	
PHTC041	[*]C1=Cc2ccc([*])cc21	3.37	18.78	4.37	4.64	5.50	5.11	
PHTC042	[*]OC(=O)O[*]	3.38	11.93	3.97	4.05	3.96	5.61	0.21
PHTC043	[*]c1ccc2nc([*])ncc2n1	3.39	15.82	3.98	3.95	5.54	5.83	0.82
PHTC044	[*]c1cc(Cl)c([*])nn1	3.45	19.29	4.19	5.08	4.97	4.52	
PHTC045	[*]c1cnc([*])nn1	3.49	14.33	5.06	5.97	5.57	6.46	0.90
PHTC046	[*]c1nn([*])c(Cl)c1Cl	3.51	15.39	4.11	5.12	5.41	4.60	
PHTC047	[*]c1cnnc([*])n1	3.53	16.38	5.04	5.92	5.33	5.60	0.17
PHTC048	[*]c1cnc([*])n1C	3.55	19.00	4.25	4.46	5.63	4.49	
PHTC049	[*]c1ccc(-c2nn([*])s2)s1	3.63	14.53	4.15	4.39	4.83	5.39	0.33
PHTC050	[*]c1nn([*])c(F)c1F	3.64	12.41	4.40	5.32	4.51	4.75	
PHTC051	[*]c1ccc(-c2cnc([*])s2)s1	3.76	15.25	4.17	4.44	4.85	5.51	0.59
PHTC052	[*]C(Cl)=C([*])Cl	3.86	20.41	4.17	4.38	4.81	5.13	0.29
PHTC053	[*]/C(Cl)=C([*])Cl	3.86	20.56	4.13	4.42	4.80	5.08	0.25
PHTC054	[*]C#C[*]	3.87	5.85	4.23	4.58	5.15	9.63	
PHTC055	[*]C1=CC=C1[*]	3.90	22.30	4.36	5.72	5.65	4.62	
PHTC056	[*]c1nn([*])c(O)n1	3.90	16.45	4.50	5.18	4.20	4.96	0.76
PHTC057	[*]c1noc([*])c1Br	3.90	11.67	4.61	5.55	4.77	4.97	
PHTC058	[*]C(=O)N1C(=O)N([*])C1=O	4.04	16.85	3.66	4.36	5.49	5.37	0.36
PHTC059	[*]NC(=O)C(=S)C([*])=O	4.08	21.85	4.01	5.07	4.60	5.91	0.18
PHTC060	[*]C(=O)c1nn([*])nn1	4.09	11.89	4.36	5.35	4.15	5.95	0.24
PHTC061	[*]c1cc2nc([*])sc2nn1	4.09	14.51	4.38	4.81	5.11	4.35	
PHTC062	[*]C(=O)C([*])=S	4.20	16.81	3.78	4.92	4.41	5.03	
PHTC063	[*]C1=C(F)C(Cl)=C1[*]	4.28	17.13	4.69	4.59	5.00	5.84	

PID	SMILES	SA	Cross	RF	XGBoost	MLP	MD	ATC
PHTC064	[*]C1=CC=C2C([*])=CC=C12	4.32	22.02	3.94	3.73	4.63	6.16	0.67
PHTC065	[*]NC(=O)C([*])=N	4.32	18.67	3.86	5.17	3.12	4.37	
PHTC066	[*]SC([*])=O	4.34	8.22	4.70	5.29	4.52	6.73	0.25
PHTC067	[*]/C=C/C=C([*])Cl	4.36	17.41	4.65	3.72	4.63	5.90	
PHTC068	[*]/C(F)=C([*])C#N	4.36	20.54	3.77	5.11	3.47	6.02	0.38
PHTC069	[*]C=Cc1csc(-c2cc([*])cs2)c1	4.45	15.88	3.69	3.61	3.08	4.33	
PHTC070	[*]/C=C/C=C/C=C([*])Cl	4.45	15.96	4.28	4.78	3.99	5.66	
PHTC071	[*]/C=C/C=C/C=C([*])Cl	4.45	15.96	4.28	4.78	3.99	5.66	
PHTC072	[*]/C(F)=C([*])Cl	4.46	18.39	4.44	5.10	4.90	6.18	0.30
PHTC073	[*]C(F)=C([*])Cl	4.46	17.50	4.69	5.25	4.94	5.83	0.36
PHTC074	[*]OC([*])=S	4.53	14.49	4.63	5.07	3.76	5.20	
PHTC075	[*]C#CC=C[*]	4.54	18.03	4.51	5.02	5.18	5.16	1.07
PHTC076	[*]OC([*])=C	4.54	19.19	4.07	4.36	3.57	4.55	
PHTC077	[*]/C=C([*])Cl	4.55	21.62	4.28	4.80	5.14	6.03	0.35
PHTC078	[*]C=C([*])C#N	4.58	24.93	3.53	4.00	2.70	4.88	
PHTC079	[*]C(=S)C(=O)C(F)(F)C([*])(F)F	4.59	25.30	3.19	4.28	3.92	4.41	
PHTC080	[*]SS[*]	4.66	13.04	3.77	3.91	0.95	4.93	
PHTC081	[*]/C=C/C=C/C=C([*])N	4.66	15.44	4.20	4.41	4.69	5.53	0.57
PHTC082	[*]NNC([*])=N	4.67	15.08	4.45	4.61	4.41	4.51	
PHTC083	[*]C1=CC2=NC([*])=CC2=N1	4.75	17.77	4.55	4.07	4.51	6.10	0.59
PHTC084	[*]C#CC#CC=C[*]	4.78	12.75	4.11	3.79	4.61	5.99	
PHTC085	[*]C=C([*])F	4.82	17.93	4.71	5.35	4.81	6.64	0.52
PHTC086	[*]/C=C/N[*]	4.83	15.68	5.02	5.28	6.52	5.44	0.53
PHTC087	[*]C#C/C(C#N)=C([*])C#N	4.88	15.14	3.46	5.00	4.71	5.37	0.49
PHTC088	[*]/N=C/C(=S)N[*]	4.88	18.77	4.37	4.29	5.00	5.06	0.24
PHTC089	[*]/N=N/C([*])=O	4.90	8.82	4.55	5.76	6.04	7.35	
PHTC090	[*]/C=C/O[*]	4.92	15.71	3.94	3.83	3.43	4.60	
PHTC091	[*]=CC=NN=[*]	4.98	5.95	4.69	4.92	7.53	4.71	
PHTC092	[*]NN[*]	5.00	11.88	4.79	5.37	6.33	4.60	
PHTC093	[*]/N=N/[*]	5.05	5.21	4.83	5.95	6.81	10.38	
PHTC094	[*]N=N[*]	5.05	5.93	4.83	5.95	6.76	10.01	
PHTC095	[*]/N=C([*])Cl	5.08	9.62	5.06	5.71	6.50	6.98	0.12
PHTC096	[*]C#C/C(C#N)=C([*])Cl	5.14	14.06	3.63	5.06	4.58	5.79	0.58
PHTC097	[*]/C=N/C=C([*])C	5.14	17.30	4.02	3.64	4.36	4.72	
PHTC098	[*]/C=C/C=CS[*]	5.17	14.14	4.14	3.28	4.13	4.36	
PHTC099	[*]C1=CC2=CC([*])=NC2=N1	5.28	17.50	4.62	4.10	4.55	5.79	
PHTC100	[*]/C=C/N=C([*])Cl	5.38	19.65	4.16	4.37	4.61	5.60	0.25
PHTC101	[*]/C=C/S[*]	5.44	14.40	4.57	4.20	5.74	4.35	
PHTC102	[*]/C=N[*]	5.53	14.02	5.10	6.07	6.07	6.82	0.31
PHTC103	[*]/N=N/N=C([*])C#N	5.66	11.24	3.69	5.07	4.33	6.68	0.19

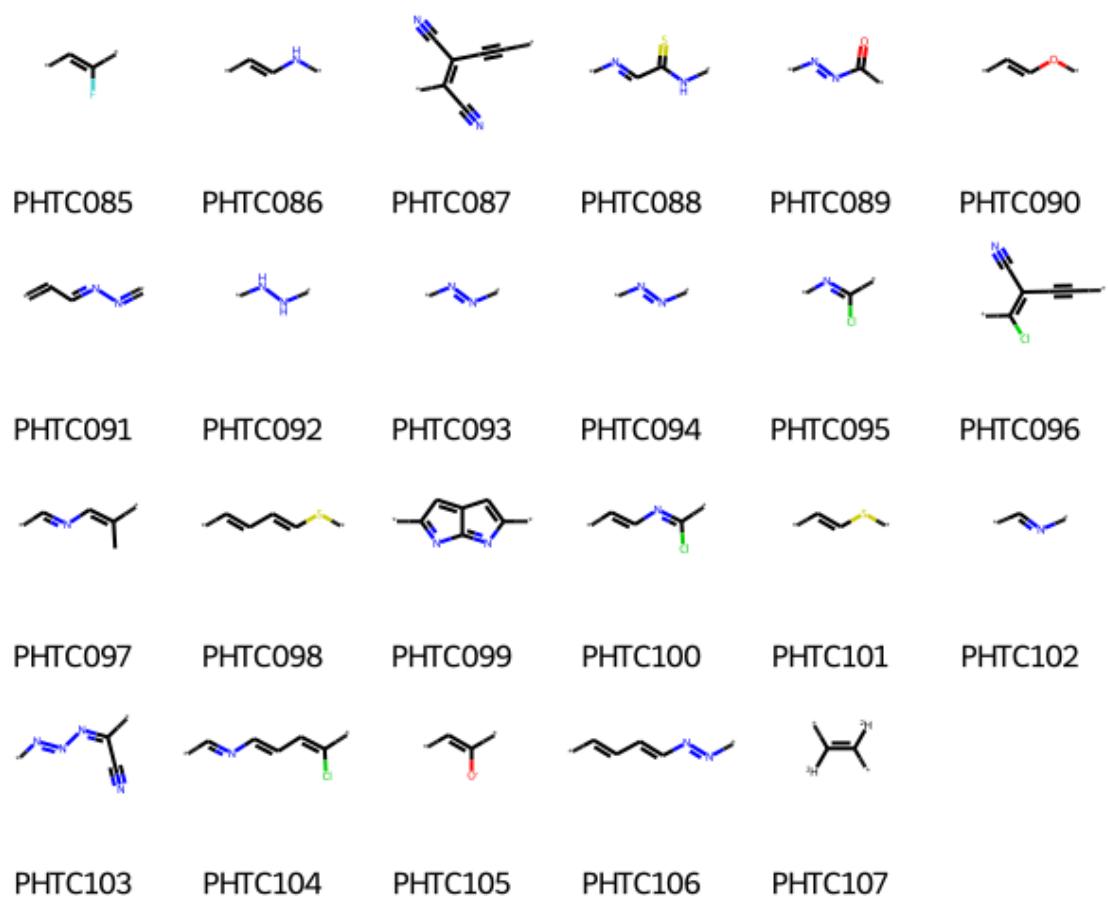
PID	SMILES	SA	Cross	RF	XGBoost	MLP	MD	ATC
PHTC104	[*]/C=N/C=C/C=C([/*])Cl	5.70	15.50	4.33	4.46	4.07	5.70	
PHTC105	[*]/C=C([/*])[O-]	5.73	18.54	4.07	4.66	2.87	6.10	
PHTC106	[*]/C=C/C=C/N=N/[*]	5.76	12.71	4.74	4.11	7.01	6.59	0.85
PHTC107	[*]C([2H])=C([*])[2H]	7.27	11.30	4.32	4.33	1.07	7.48	



Supplementary Figure 12: Continued



Supplementary Figure 12: Continued



Supplementary Figure 12: Polymer monomer structures with TC greater than $20 \text{ W m}^{-1}\text{K}^{-1}$ verified by MD simulations. Polymer properties can be found in Supplementary Table 4 using the PID.

Supplementary Note 9. Symbolic regression for characterizing the TC of promising polymers

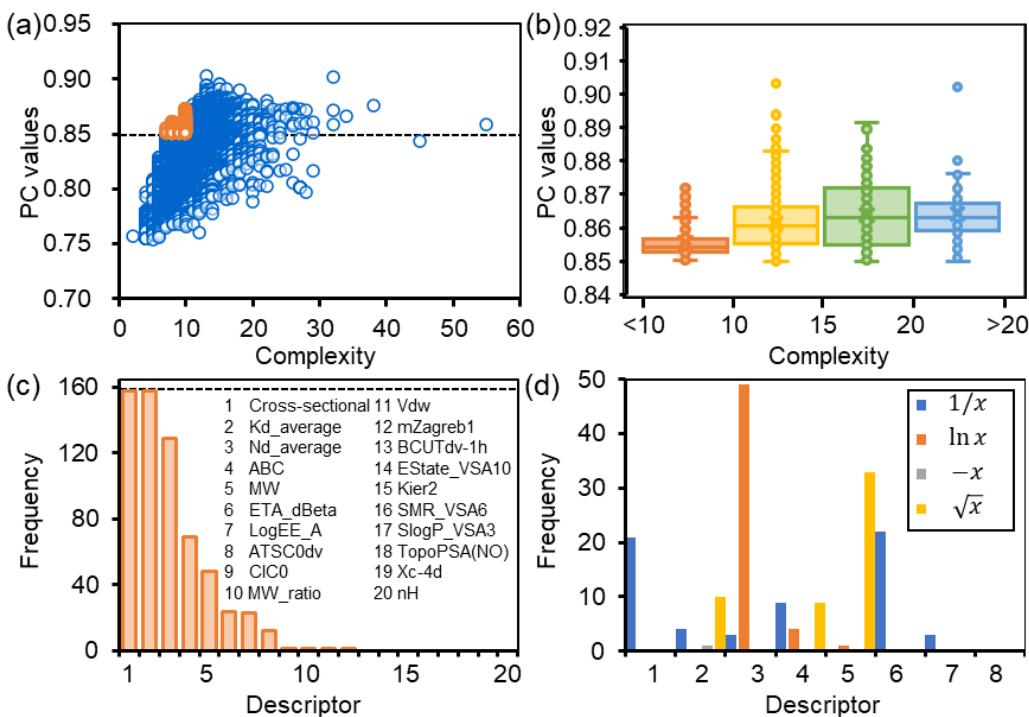
The 107 promising polymer structures ($\text{TC} > 20 \text{ W m}^{-1}\text{K}^{-1}$) with optimized descriptors were utilized for symbolic regression (SR), where 20% structures were randomly selected for testing set. The mathematical formulae were acquired and selected using an efficient stepwise strategy with SR based on a genetic programming (GPSR) as implemented in the gplearn code . At first, Pearson coefficients were used as evaluation metrics of training fitness to filter optimized descriptors and generate sub-descriptors, and the grid search strategy with the hyperparameters listed in Supplementary Table 5 was applied in GPSR. The hyperparametric combinations corresponding to the 22380 formulas are characterized by complexity and training fitness of Pearson coefficients (PC) as shown in Supplementary Figure 13a. The complexity is equivalent to the formula length in gplearn. Among these formulas, 4365 of them with PC values not less than 0.85 are statistically by complexity in Supplementary Figure 13b. Generally, the formulas with large complexity have a higher training fitness. But the formulas are also usually quite long and do not facilitate the calculation and analysis. Only formulas with large fitness and low complexity are appropriate, so we selected 158 formulas with the PC value not less than 0.85 and complexity within 10 for subsequent analysis. By counting the frequency of occurrence of the 20 optimized descriptors in 158 formulas, the first 8 descriptors were finally retained, as presented in Supplementary Figure 13c. It is worth emphasizing that the MD-inspired descriptors of cross-sectional area (cross-sectional) and dihedral force constants (Kd_average) appeared in each of the formulas. And, we additionally estimated the frequency of occurrence of sub-descriptors created by taking the inverse ($-x$), logarithm ($\ln x$), and so on, as shown in Supplementary Figure 13d. As a result, a new ensemble containing 22 descriptors which have a strong association with TC is listed in Supplementary Table 6 for retraining in gplearn.

Further, we reset the grid search hyperparameters listed in Supplementary Table 7, and utilized the accuracy R^2 as the evaluation metric. The 19580 formulas with fitting accuracy greater than 0.6 versus complexity are visualized in Supplementary Figure 14a. The effective accuracy is the average of the training and testing accuracy. Despite the fact that the effective accuracy of some formulas exceeds 0.8, and their complexity is greater than 50. The Pareto front of accuracy R^2 vs. complexity (no more than 30) of 9073 mathematical formulas shown via density plot in Supplementary Figure 14b, and the points of c, d, e and f were identified by Latin hypercube sampling approach². The formulas for the four points are listed in Supplementary Table 8, and their fitting results compared with MD calculated $\log_2\text{TC}$ are

displayed in Supplementary Figure c-f. The definitions of the variables x_0 to x_{21} can be found in Supplementary Table 6.

Supplementary Table 5. Setup of hyperparameters in gplearn for filter and generate new sub-descriptors

Parameter	Value	Combination
Generations	300	1
Population size in every generation	1000,2000	2
Probability of crossover (pc)	[0.30,0.90], step=0.05	
Subtree mutation (ps)	[(1-pc)/3,(1-pc)/2] (step= 0.01)	
Hoist mutation (ph)	[(1-pc)/3,(1-pc)/2] (step = 0.01)	746
Point mutation (pp)	1-pc-ps-ph	
Function set	{+,-,×,÷, \sqrt{x} , $\ln x$, $ x $, $-x$, $1/x$ }	1
Parsimony coefficient	auto	1
Metric	Pearson coefficient	1
Stopping criteria	0.900	1
Random_state	0, 1, 2, 3, 4	5
Init_depth	[2, 6], [4, 8], [6, 10]	3



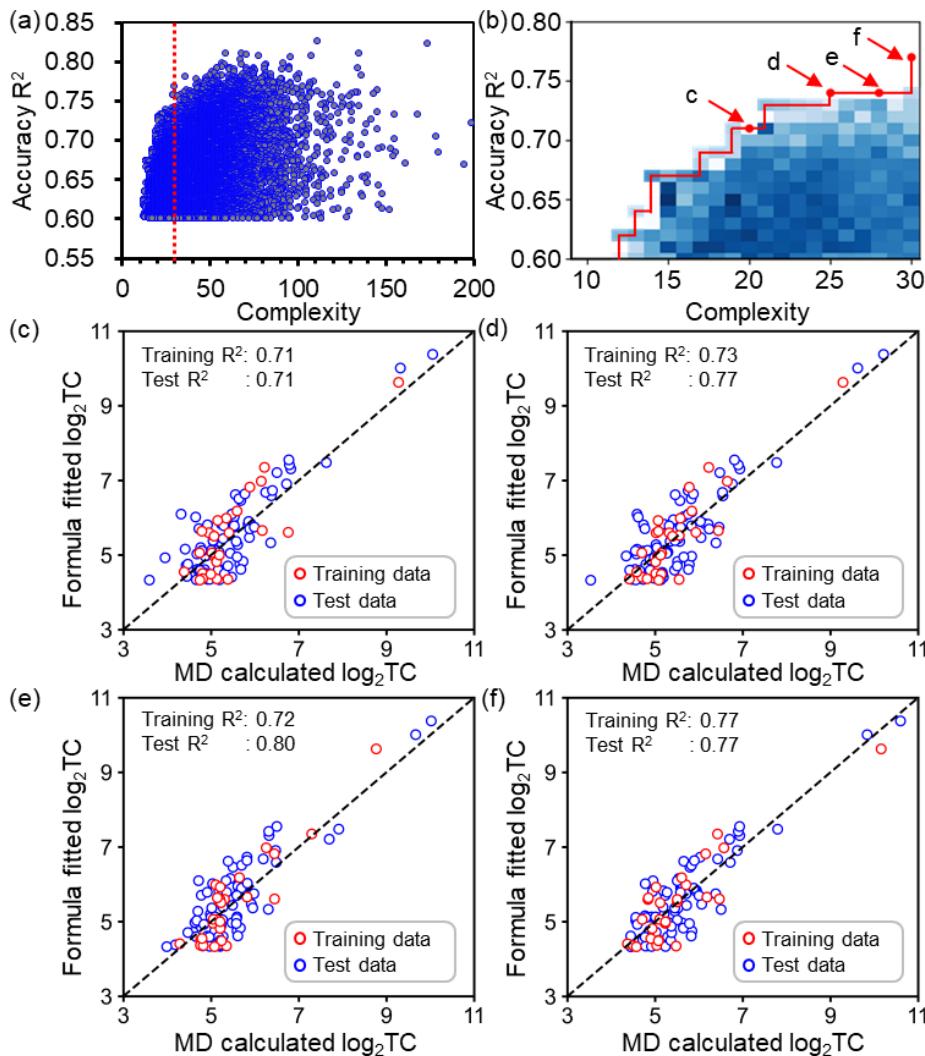
Supplementary Figure 13: Filtering optimized descriptors and creating sub-descriptors in gplearn through Pearson coefficient (PC) values. (a) Mathematical formula complexity versus PC values. (b) Statistics of formulas with Pearson coefficient not less than 0.85 by complexity. (c) and (d) Frequency of occurrence of optimized descriptors and sub-descriptors in 158 mathematical formulas (PC values ≥ 0.85 and complexity ≤ 10).

Supplementary Table 6. New descriptors ensemble for GPSR

Variable	Format	PC values	Variable	Format	PC values
x_0	ABC	-0.37	x_{11}	-ETA_dBeta	0.11
x_1	10/ABC	0.44	x_{12}	Kd_average	0.20
x_2	\sqrt{ABC}	-0.39	x_{13}	$\sqrt{Kd_average}$	0.12
x_3	ATSC0dv	-0.32	x_{14}	LogEE_A	-0.40
x_4	Cross-sectional	-0.52	x_{15}	MW	-0.39
x_5	10/Cross-sectional	0.63	x_{16}	MW/100	-0.39
x_6	1-10/Cross-sectional	0.59	x_{17}	100/MW	0.52
x_7	100/Cross-sectional	0.63	x_{18}	\sqrt{MW}	-0.43
x_8	$\sqrt{100/Cross - sectional}$	0.62	x_{19}	Nd_average	-0.31
x_9	-10/Cross-sectional	-0.63	x_{20}	log Nd_average	-0.35
x_{10}	ETA_dBeta	-0.11	x_{21}	- log Nd_average	0.44

Supplementary Table 7. Reset of hyperparameters in gplearn for GPSR

Parameter	Value	Combination
Generations	300	1
Population size in every generation	5000	1
Probability of crossover (pc)	[0.30,0.90], step=0.05	
Subtree mutation (ps)	[(1-pc)/3,(1-pc)/2] (step= 0.01)	
Hoist mutation (ph)	[(1-pc)/3,(1-pc)/2] (step = 0.01)	746
Point mutation (pp)	1-pc-ps-ph	
Function set	{+, -, ×, ÷, \sqrt{x} , $\ln x$, $ x $, $-x$, $1/x$ }	1
Parsimony coefficient	0.001, 0.003, 0.005	3
Metric	R ²	1
Stopping criterial	0.900	1
Random_state	0, 1, 2, 3, 4	5
Init_depth	[2, 6], [4, 8], [6, 10], [2, 10]	4



Supplementary Figure 14: GPSR for TC prediction of promising polymers. (a) 19580 formulas with fitting accuracy greater than 0.6 versus complexity. (b) Pareto front of accuracy R² vs. complexity of 9073 mathematical formulas shown via density plot. The four points of c, d, e and f represent the four formulas at the Pareto front, whose fitting results vs. MD calculated TC are plotted in (c), (d), (e) and (f) respectively.

Supplementary Table 8. The four mathematical formulas at the Pareto front in Supplementary Figure 14b

Point	Formulas	R ²	Complexity
c	$\log_2 TC = \sqrt{(x_{12} - x_8 - x_{16} + 0.49) \times x_9 \times (x_6 + x_{20})} + \sqrt{\ln(x_{12})} + x_{21}$	0.71	20
d	$\log_2 TC = \sqrt{x_{12} - \ln \left[(x_8 - x_{11}) \times \frac{x_8}{\ln(\sqrt{x_{19}})} \right] \times (x_6 + x_{20}) \times x_7} + x_{21} + \sqrt{\ln x_{12}}$	0.74	25
e	$\log_2 TC = \ln \left[\frac{1}{x_{10} \times (x_1 - x_2)} - x_{12} \right] \times (x_5 - 0.824) + \frac{x_{17}}{\sqrt{\frac{x_3}{(x_{12} - x_{11}) \times x_5}}} + x_{21} + x_{13}$	0.74	28
f	$\log_2 TC = \sqrt{x_8 + \left[\frac{x_{10}}{\ln(\ln x_{17} - x_{12}) + x_8} - x_{12} \right] \times (x_6 + x_{20}) \times \frac{x_8}{x_{20}}} + x_{21} + \sqrt{x_{19} - x_{16}}$	0.77	30

Supplementary Note 10. Investigation of thermal transport in different polymer hierarchical structures including polymer chains and amorphous polymers

An automated pipeline implemented in Radonpy toolkit for the modeling and simulation of amorphous polymers, as shown in Supplementary Figure 15. For each polymer simulation, the monomer was entered into the pipeline in the form of SMILES, and a polymer chain containing about 1000 atoms was generated by polymerization of the repeating unit using a self-avoiding random walk algorithm. Next, the GAFF2 force field with the information of bonds, angles, dihedrals, impropers and inter-chain interactions was assigned for the polymer chain. A simulation cell with 10 non-overlapping chains was formed by copying, arranging and rotating the individual chain. Then, the temperature of the simulation cell was increased from 300K to 700K, and maintained at 700K for 1ns in the NVT ensemble with a Nosé–Hoover thermostat for packing the system. After the packing simulation, the system was equalized following a 21-step equilibrium protocol³ at alternating temperatures of 300K and 600K for approximately 1.5 ns of each stage while the system was compressed to 50,000 atm and then decompression to 1 atm by combining the NVT and NPT simulations.

The equilibrium amorphous model contains approximately 10,000 atoms, and the NPT simulation was performed for 5ns with a timestep of 1fs to record the parameters for calculating properties of equilibrium systems such as radius of gyration, specific heat capacity and volumetric heat capacity. The radius of gyration R_g was calculated by:

$$R_g = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - r_{mean})^2} \quad (1)$$

where N is the number of repeating units, r_{rep} is the position of repeating unit i , r_{mean} is the mean position of the repeating unit in a polymer chain.

The specific heat capacity C_p was calculated from the fluctuations in the enthalpy H :

$$C_p = \frac{\langle \delta H^2 \rangle}{m k_B T^2} \quad (2)$$

where $\langle \cdot \rangle$ represent time averaging, m is the mass of atoms in the system, k_B is the Boltzmann constant, and T is the temperature.

The volumetric thermal expansion coefficient α_V was calculated by the covariance of volume V of enthalpy H ⁴:

$$\alpha_V = \frac{\langle \delta V \delta H \rangle}{k_B T^2 \langle V \rangle} \quad (3)$$

The isothermal compressibility β_T and isothermal bulk modulus K_T were calculated based on the fluctuations in the volume ⁴:

$$\beta_T = \frac{\langle \delta V^2 \rangle}{k_B T \langle V \rangle} \quad (4)$$

$$K_T = \frac{1}{\beta_T} \quad (5)$$

Hence, the volumetric heat capacity may be expressed as ⁴:

$$C_V = C_P - \frac{\alpha_V^2 T \langle V \rangle}{\beta_T m} \quad (6)$$

Since the volumetric heat capacity is mainly dependent on the characteristics of the repeating units of the polymer, the C_V of polymer chains and amorphous polymers formed from the identical repeating unit is almost the same.

Applying the strain enables the de-entanglement of disordered chains in amorphous polymers to obtain oriented and ordered polymer chain segments. Moreover, the stretching process is available in the MD simulations. The pipeline provided the interface for applying various strains to the polymer system at constant temperature and strain rate. The stretched polymer model was required to relax in the NVT ensemble for 1 ns with a timestep of 0.2 fs to release the internal stress.

To the ATC of amorphous polymers, the reverse non-equilibrium molecular dynamics (NEMD) simulation was performed with the equilibrium model with 3x replication (~30,000 atoms) in thermal transport heat transform direction (x -direction). The system was divided into N ($N=20$ in this work) slabs along the x -direction, and the temperature gradients were generated by exchanging the velocity between the coldest atom in slab $N/2$ and the hottest atom in slab 0, as shown in the right part of Supplementary Figure 15. To prevent temperature shifts after cell replication, the system was initially run at 300K for 2ps. After that, the reverse NEMD simulation was run at 300K for 1 ns with 5000 cycles of velocity swaps. According to the exchanged energy obtained using the Müller-Plathe algorithm ΔE and the temperature gradient $\partial T / \partial x$ output by reverse NEMD simulation, the ATC λ can be expressed by:

$$\lambda = \frac{J_Q}{(\partial T / \partial x)} = \frac{\Delta E}{2A\Delta t(\partial T / \partial x)} \quad (7)$$

where J_Q is heat flux, A is the cross-sectional area of the amorphous system in the direction of heat flux, Δt is the simulation time.

Additional thermal conductivity decomposition analysis was employed to quantify the contribution of six components including of bond, angle, dihedral (dihed), convection (conv), nonbonded (non) and

improper, to the thermal conductivity. The energy flux along the direction of unit vectors J_u can be expressed as the contribution of convection (first term) and interatomic interactions (second term)⁵:

$$J_u = \frac{1}{V} \left\{ \sum_{i \in V} e_i v_{i,u} + \sum_{i \in V} (S_i v_i)_u \right\} \quad (8)$$

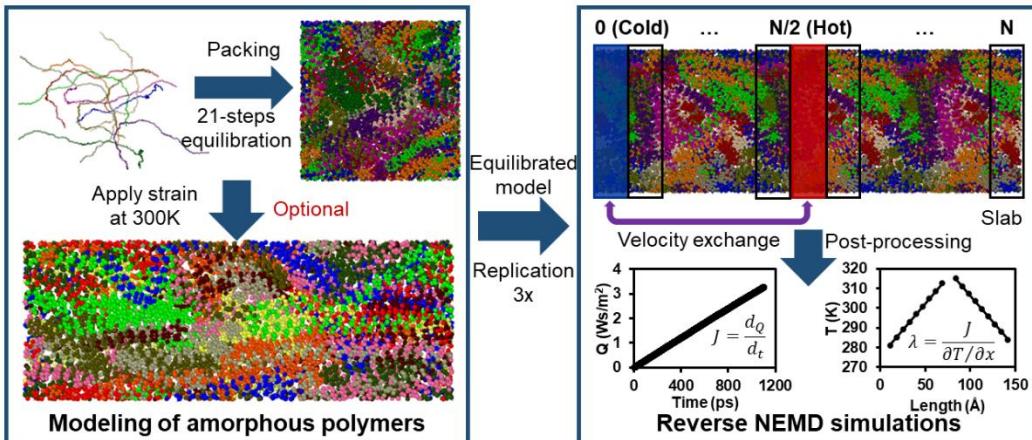
where e_i is potential and kinetic energy of each atom, $v_{i,u}$ is the velocity of the atom, S_i is the stress tensor and i is the index of atoms. The second term can be further divided into each component of the interactions. The stress tensor for component (a, b) is:

$$\begin{aligned} S_{ab} = & \sum_{n=1}^{N_p} r_{io,a} F_{i,b} + \sum_{n=1}^{N_b} r_{io,a} F_{i,b} + \sum_{n=1}^{N_a} r_{io,a} F_{i,b} + \sum_{n=1}^{N_d} r_{i,a} F_{i,b} + \sum_{n=1}^{N_i} r_{io,a} F_{i,b} \\ & + Kspace(n_{i,a}, F_{i,b}) \end{aligned} \quad (9)$$

The stress tensor was divided into six parts of the contributions pairwise, bond, angle, dihedral, improper and K-space, sequentially. where r_{io} is the relative position of atom i to the geometric center of its interacting atoms, F_i is the force acting on atom i . Therefore, the thermal for each part λ_{par} can be written as:

$$\lambda_{par} = \frac{J_{par}}{J} \lambda \quad (10)$$

where J is the total heat flux calculated by Equation (8), and J_{par} is partial heat flux of each part in Equations (8) and (9), λ is the total thermal conductivity of polymers.



Supplementary Figure 15: Automated pipeline for the calculation of thermal conductivity of amorphous polymers (ATC) by reverse non-equilibrium molecular dynamics (NEMD) simulations.

To acquire a unified understanding of the effect of dihedral angle on the thermal conductivity of polymers with different hierarchical structures, we chose polyethylene (PE) as the object, due to its excellent mechanical properties and single backbone dihedral type of C-C-C-C (C denotes carbon atom).

Supplementary Figure 16a demonstrates the ATC of amorphous polymers under various strains ranging from 0 to 100%. The strain rate was unified as 1×10^9 s⁻¹ and the temperature at 300K. It is noteworthy to mention that this work mainly focused on the relationships between polymer microstructures under different strains and ATC, rather than the length effect of the models. Therefore, we first estimated the ATC of the stretched amorphous polymer with finite length using equilibrium molecular dynamics (EMD) simulations. In EMD simulation, the ATC calculation was carried out under an NVT ensemble for 1 ns using the Green–Kubo formula:

$$\lambda = \frac{V}{k_B T^2} \int_0^\infty \langle J_x(0)J_x(t) \rangle dt \quad (11)$$

where J_x is the heat flux along the x -direction, which is consistent with the direction of stretching.

The orientational order parameter of the backbone was used for characterized the polymer chain alignment⁶:

$$P_{2b} = 1.5\langle(e_i \cdot e_x)^2\rangle - 0.5 \quad (12)$$

where e_x is the unit vector in the direction of applied strain, and the local chain orientation at atom i was computed from the chord connecting second neighbor atoms⁷: $e_i = (r_{i+1} - r_{i-1})/|r_{i+1} - r_{i-1}|$.

In Supplementary Figure 16a, with the increased strain applied to the amorphous PE, the orientational order parameters are subsequently raised, resulting in a larger ATC. The two blue stars indicate the ATC of the amorphous polymers without strain and with 100% strain, which calculated by reverse NEMD simulations. This work was mainly concerned with the effect of dihedral energy on the thermal conductivity of the identical polymer structure, and the difference between the ATC calculated by reverse NEMD and EMD was acceptable, therefore the NEMD simulations were adopted in the subsequent research. The R_g calculated from amorphous polymers under different strains is displayed in Supplementary Figure 16b, with a close positive correlation with the persistent length ξ , namely⁸:

$$\xi = \frac{R_g^2 \times 6}{2(2N - 1) \times l} + \frac{l}{2} \quad (13)$$

where l is the bond length, $l=1.526$ Å for PE. The large ξ suggests that the amorphous system has large rigid chain segments, which is conducive to intra-chain heat transport and generates high ATC. Evidently, the stretching effect is favorable for achieving large rigid chain segments.

Through GAFF2 force field, the dihedral energy $E_{dihedral}$ is computed using:

$$E_{dihedral} = \sum_{dihedrals} K_d [1 + co s(n_d \varphi - \delta_d)] \quad (14)$$

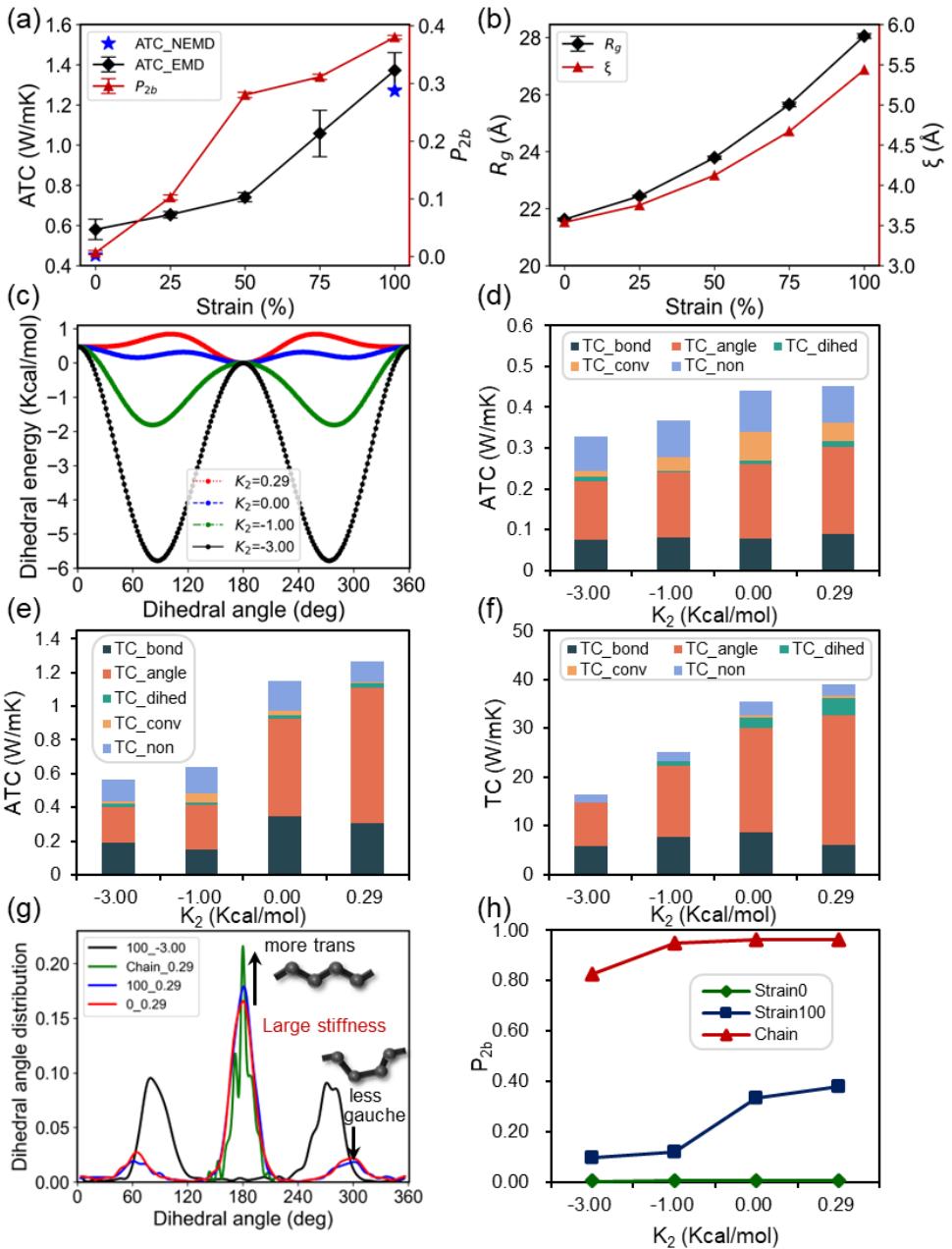
where K_d is the force constant of dihedral, δ is the phase angle for the torsional angle parameters, n is the multiplicity. For the C-C-C-C dihedral, there are three sets of parameters, as listed in

Supplementary Table 9. The dihedral energy curve can be adjusted by modifying the force constant so as to explore the effect of disorder on thermal conductivity. We increased the disorder of polymer systems by switching the force constant of K_2 from 0.29 to 0.00, -1.00 and -3.00 kcal/mol, respectively. The modified dihedral energy curves are demonstrated in Supplementary Figure 16c.

Supplementary Figure 16d-f illustrates the thermal conductivity of strain-free amorphous PE (0-strained PE), 100% strained amorphous PE (100%-strained PE), and PE chains, with different modified dihedral energy, respectively. Here, ATC was specifically defined as the thermal conductivity of the amorphous polymer to distinguish it from that of polymer chains. Each ATC or TC was quantified into five components of the contribution of bond, angle, dihedral, convection and nonbonded, since the contribution of improper was negligible. The intra-chain heat transport associated with bonds, angles and dihedrals is dominant for the contribution to thermal conductivity. Combined with Supplementary Figure 16h, we concluded that low dihedral energy leads to a decrease in the orientational order parameter and an increase in disorder for the PE with different hierarchical structures, which not only directly weakens the dihedral contribution, but also accounts for the lower angular contribution. For different hierarchical structures, the 0-strained PE model is relatively insensitive to the reduction in dihedral energy, because its orientational order parameter is close to 0. While the original chain orientation of 100% strained PE and individual PE chains are kept at a certain level separately, thereby the change of dihedral energy makes the reduction of ATC/TC more significant. When the K_2 shifts from 0.29 to -3.00 kcal/mol, the ATC of 100% strained PE decreases from 1.27 to 0.56 W m⁻¹K⁻¹, and the TC of PE chain decreases from 38.98 to 0.56 W m⁻¹K⁻¹. The impact of the dihedral term on thermal conductivity is more significant with increasing polymer orientational order parameters. In addition, we counted the dihedral distribution of different PE systems in Supplementary Figure 16g. The dihedral angle of 180 deg corresponds to trans conformation, while dihedral angles of 60 deg and 300 deg correspond to gauche transformation. Increasing the dihedral energy or applying the stain makes more gauche conformations transform to trans conformations, thus resulting in large chain stiffness and high thermal conductivity.

Table S9. List of parameters of the C-C-C-C dihedral

Dihedral	K_d (kcal/mol)	n_d	δ_d (deg)
C-C-C-C	$K_3=0.13$	$n_3 = 3$	$\delta_3=0$
	$K_2=0.29$	$n_2 = 2$	$\delta_2=180$
	$K_1=0.11$	$n_1 = 1$	$\delta_1=0$



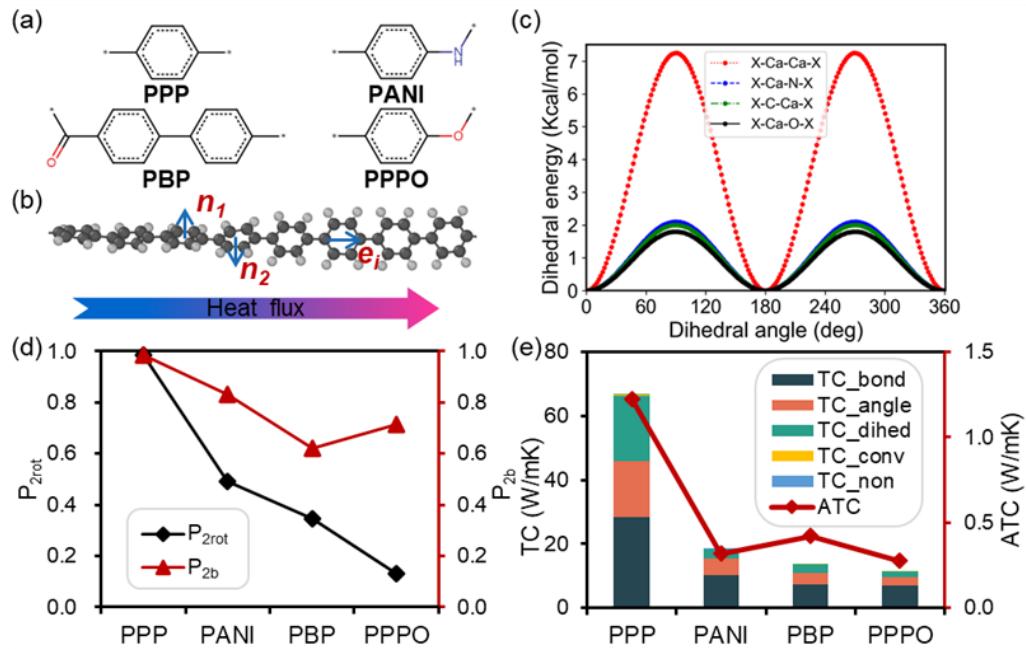
Supplementary Figure 16: Investigation of thermal conductivity of PE with different hierarchical structures by changing dihedral energy. (a) ATC and orientational order parameters (P_{2b}) of amorphous PE with varying strains. (b) Radius of gyration (R_g) and persistent length (ξ) of amorphous PE with varying strains. (c) dihedral energy curve. (d), (e) and (f) thermal conductivity of amorphous PE, 100% strained amorphous PE and PE chains with various dihedral energies, in which each thermal conductivity was quantified into five components of the contribution of bond, angle, dihedral (dihed), convection (conv) and nonbonded (non). (g) dihedral angle distribution. The legend depicts the pairs of polymer structure and dihedral energy force constant K_2 , e.g., 100_-3.00 means 100% strained amorphous polymer with modified K_2 of -3.00 kcal/mol. (h) orientational order parameters of different PE hierarchical structures.

The low dihedral energy increases the disorder of the polymer systems, which is mainly attributed to the severe chain rotation. For the study of the behavior of polymer chain rotation, a well-defined reference plane in the structure is required. This is facilitated by the coplanarity of the carbon atoms in the benzene ring of aromatic polymers, and the polymers of Poly(p-phenylene) (PPP), Polyaniline (PANI), Poly(benzophenone) (PBP) and Poly(p-phenylene oxide) (PPPO) were picked in this study, as shown in Supplementary Figure 17a. The orientational parameter P_{2rot} was defined as:

$$P_{2rot} = 1.5 \left\langle \left(\mathbf{n}_j \cdot \mathbf{n}_k \right)^2 \right\rangle - 0.5 \quad (15)$$

where \mathbf{n}_j and \mathbf{n}_k are the unit normal vectors of two planes of adjacent benzene rings along the heat flux transport direction in Supplementary Figure 17b.

The energy curves of critical dihedrals of four polymers are exhibited in Supplementary Figure 17c, where Ca indicates the carbon atom in the aromatic ring, C indicates the rest carbon atom, N indicates the nitrogen atom, O indicates the oxygen atom and X indicates any atom. Overall, the X-Ca-Ca-X has the largest dihedral energy, then X-Ca-N-X, X-C-Ca-X and X-Ca-O-X by decreasing order of dihedral energy. The large dihedral energy corresponds to a strong chain stiffness to suppress the chain rotation, allowing for a large rotational parameter. The ranking of the four polymers by orientational parameters (Supplementary Figure 17d) and chain TC (Supplementary Figure 17e) from largest to smallest is PPP, PANI, PBP and PPPO. however, the ranking by orientational order parameters is different, where PBP has the smallest value. As PBP has two dihedral angles X-Ca-Ca-X and X-C-Ca-X, and great dihedral energy differences may lead to poor stiffness matching of the chain. This does not determine the length of locally rigid chain segments, where the corresponding amorphous system has a larger ATC compared to that of PANI and PPPO in Supplementary Figure 17e.



Supplementary Figure 17: Four aromatic polymers for probing the effect of chain rotation phenomenon on thermal conductivity. (a) repeating units of aromatic polymers of Poly(p-phenylene) (PPP), Polyaniline (PANI), Poly(benzophenone) (PBP) and Poly(p-phenylene oxide) (PPPO). (b) schematic drawing of the plane normal vector n_1 and n_2 , and the backbone vector in e_i . (c) dihedral energy curve, where Ca indicates the carbon atom in the aromatic ring, C indicates the rest carbon atom, N indicates the nitrogen atom, O indicates the oxygen atom and X indicates any atom. (d) Rotational order parameters (P_{2rot}) and orientational order parameters (P_{2b}). (e) TC of polymers and ATC of amorphous polymers corresponding to four repeating units.

Supplementary References

1. Abdi, H. & Williams, L. J. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. **2**, 433-459 (2010).
2. Paulson, N. H., Libera, J. A. & Stan, M. Flame spray pyrolysis optimization via statistics and machine learning. Mater. Des. **196**, 108972 (2020).
3. Larsen, G. S., Lin, P., Hart, K. E. & Colina, C. M. Molecular Simulations of PIM-1-like Polymers of Intrinsic Microporosity. Macromol. **44**, 6944-6951 (2011).
4. Allen, M. P. & Tildesley, D. J. Computer Simulation of Liquids. Vol. 1 (Oxford University Press, 2017).
5. Torii, D., Nakano, T. & Ohara, T. Contribution of inter- and intramolecular energy transfers to heat conduction in liquids. J. Chem. Phys. **128**, 044504 (2008).
6. Muthaiah, R. & Garg, J. Temperature effects in the thermal conductivity of aligned amorphous polyethylene—A molecular dynamics study. J. Appl. Phys. **124**, 105102 (2018).
7. Lavine, M. S., Waheed, N. & Rutledge, G. C. Molecular dynamics simulation of orientation and crystallization of polyethylene during uniaxial extension. Polymer **44**, 1771-1779 (2003).
8. Wei, X. & Luo, T. Chain length effect on thermal transport in amorphous polymers and a structure–thermal conductivity relation. Phys. Chem. Chem. Phys. **21**, 15523-15530 (2019).