npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

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

Xiang Huang[1], Shengluo Ma[1], C. Y. Zhao[1], Hong Wang[2] 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 lightweight[1]. 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 $W\,m^{-1}\,K^{-1}$, preventing the development of organic electronics[4]. Polymers with high TC are urgently demanded in organic energy storage and electronic devices to accommodate revolutionary innovations in organic electronics and optoelectronics[5]. The polymer morphology and topology were found to be closely related to TC[6]. 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 $W\,m^{-1}\,K^{-1}$, over two orders of magnitude greater than that of classical amorphous polymers[7]. Moreover, molecular dynamics (MD) simulations have suggested that individual crystalline PE chains have a very high or even divergent TC[16]. 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 chains[19], and modulation of dihedral energy[20] in MD simulations suggested that ordered chains and large radius of gyration (Rg) 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 chain[21]. 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 space[22]. 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 expensive[23]. 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 informatics[24], 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

[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. ✉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 polymers[36], amorphous polymers[37,38], and copolymers[39]. Most of these studies have employed graph descriptors[37] 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 engineering[40] or by attempting to form general descriptors[41]. 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 task[43]. 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 sites[32]. Graph descriptors have long been applied and validated in the development of drug-like small molecules[44], and the availability of open-source toolkits such as RDKit[45] and Mol2vec[40] 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 evaluation[46]. 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 results[47]. 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 Atom[50], which have been successfully applied in organic chemistry synthesis[51], molecular antibacterial activity prediction[52], 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 experiments[35]. 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 model[53]. 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 understand[54]. 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, respectively[55]. 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 selection[56]. 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 forests[54]. 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 PoLyInfo[59] and PI1M[60] 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—GAFF2[61]. Inspired by drug-like molecular representation and molecular force fields, we obtained 320 physical descriptors by Mordred software[47] 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 analysis[62]. Using the trained ML models, 107 promising polymers with TC greater than 20.00 $W\,m^{-1}\,K^{-1}$ 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 PoLyInfo[59] and PI1M[60] 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 groups[63]. 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 projection[64], where the chemical structure of each monomer was transformed into the Morgan fingerprint[41] 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 consistent[60]. 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 $W m^{-1} K^{-1}$, and only a few polymers have TC greater than 30 $W m^{-1} K^{-1}$ (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_2TC, as the target property for ML models[65].

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

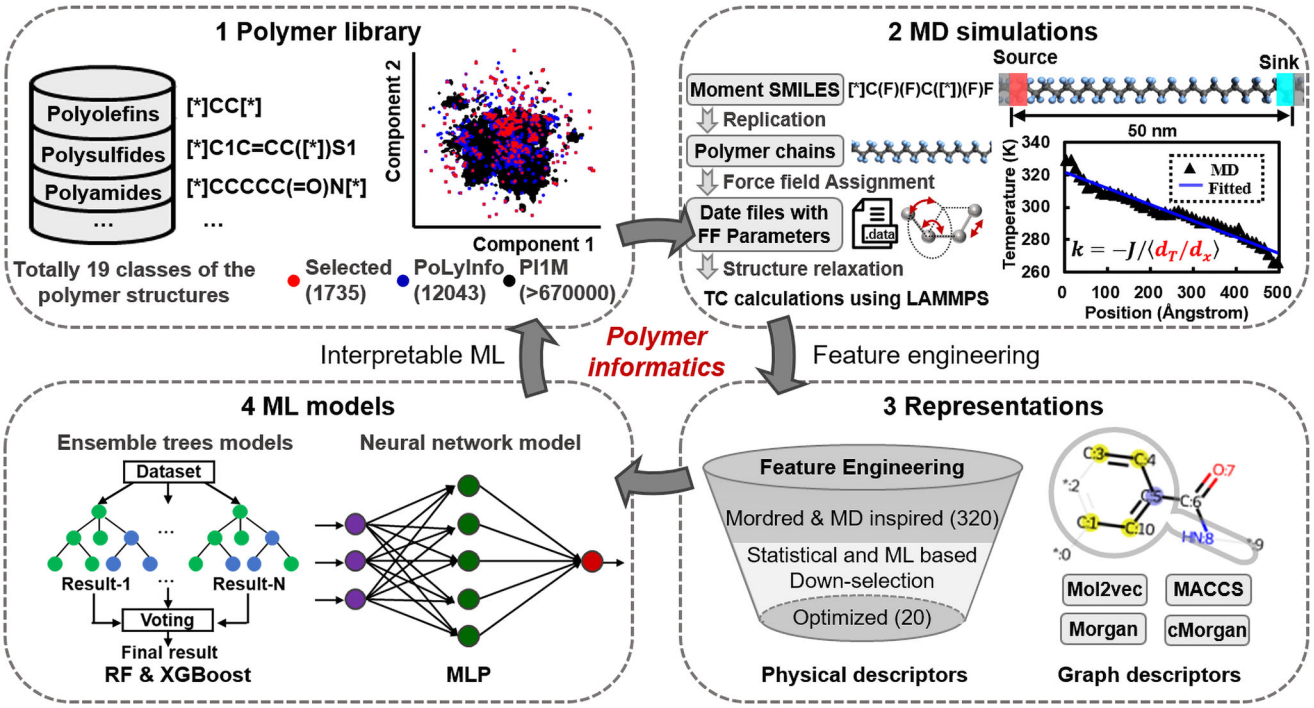

Fig. 1 Schematics of high-throughput screening of polymers with high TC via interpretable machine learning. which is implemented in four components: 1) polymer library construction, 2) MD simulation for the TC of polymers; 3) monomer feature representation and hierarchical down-selection; 4) ML model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

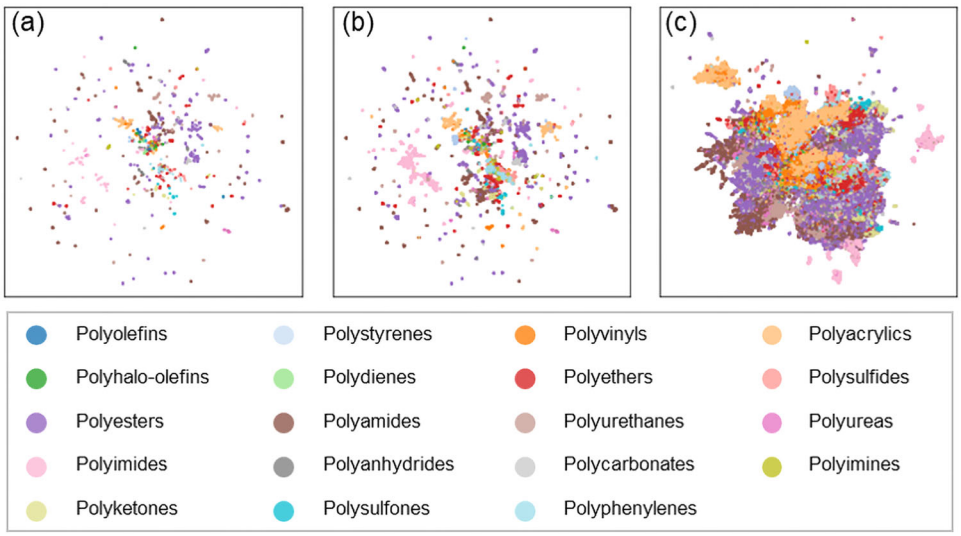

Fig. 2 Visualization of polymer data distribution in a 2D space by UMAP. a, **b** and **c** correspond to the selected, PoLyInfo, and PI1M datasets, respectively.

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_2TC 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 descriptors[66]. 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 Mol2vec[40] 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 elsewhere[60], which was created using the PoLyInfo and PI1M datasets. The MACCS[42] 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 chains[36], it is more intuitive and effective to predict the log_2TC of polymer chains using the associated physical descriptors. But not absolute, which is also related to the parameters such as chain stiffness[67]. 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 Chemprop[68], 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 output[62]. 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

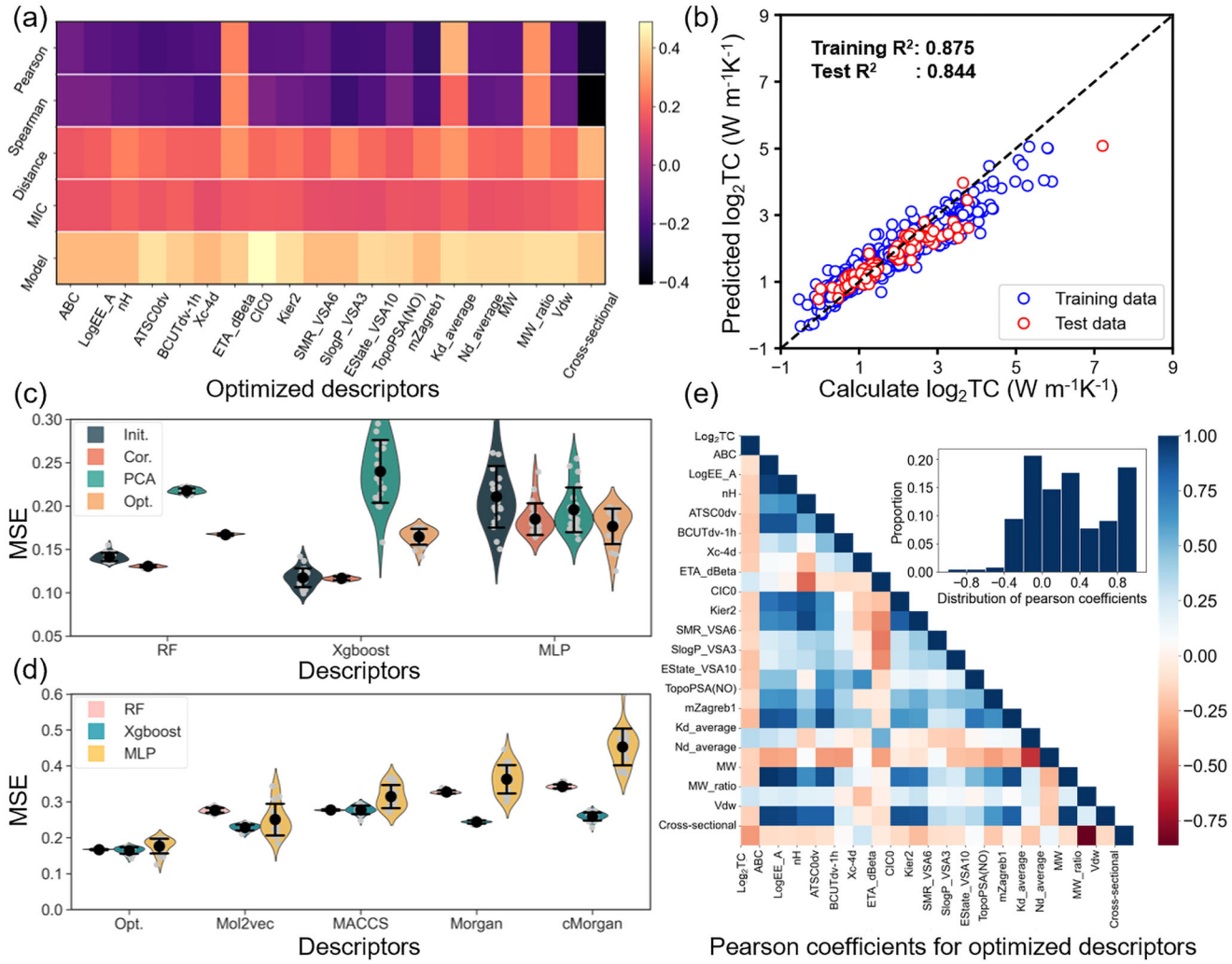

Fig. 3 Polymer descriptors down-selection and ML models training. a Optimized descriptors acquired by down-selection with four coefficients - Pearson, Spearman, Distance, and MIC coefficients - and RF model. **b** Accuracy of RF model based on optimized descriptors, where training R^2 is 0.875 and test 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 TC[69]. 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 simulations[69]. This confirms the significance of our proposed ML framework for predicting the log$_2$TC of polymers.

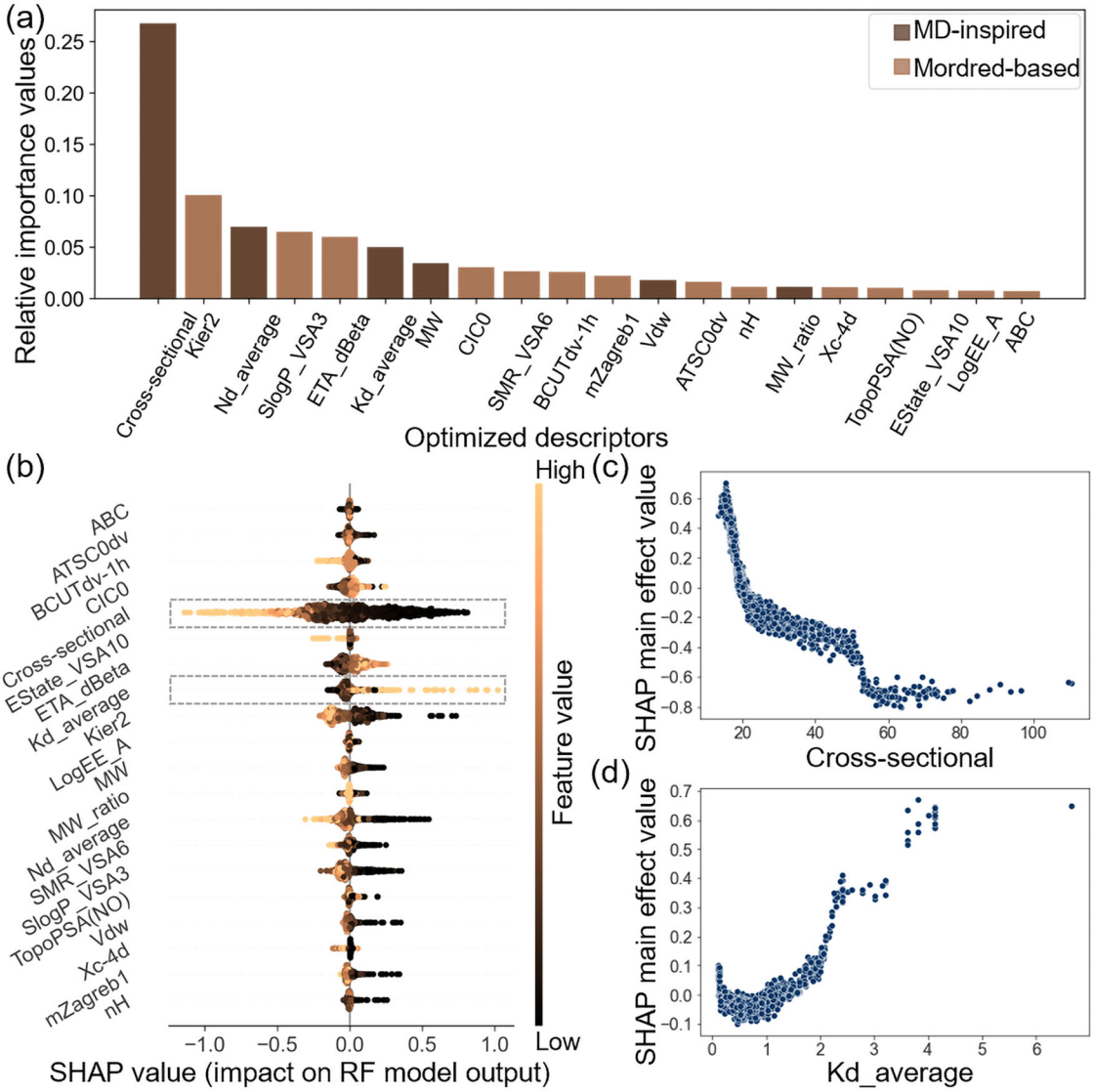

Fig. 4 Analysis of feature importance using SHAP on RF model trained by optimized descriptors. a Average SHAP values for 20 optimized descriptors. **b** Represent the SHAP values of each descriptor related to training data set polymers in a beeswarm diagram. **c**, **d** SHAP values for the Cross-section 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 ML[23]. 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 mechanism[70]. Here, we considered polymer chains with TC greater than 20.00 $W\,m^{-1}\,K^{-1}$ 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

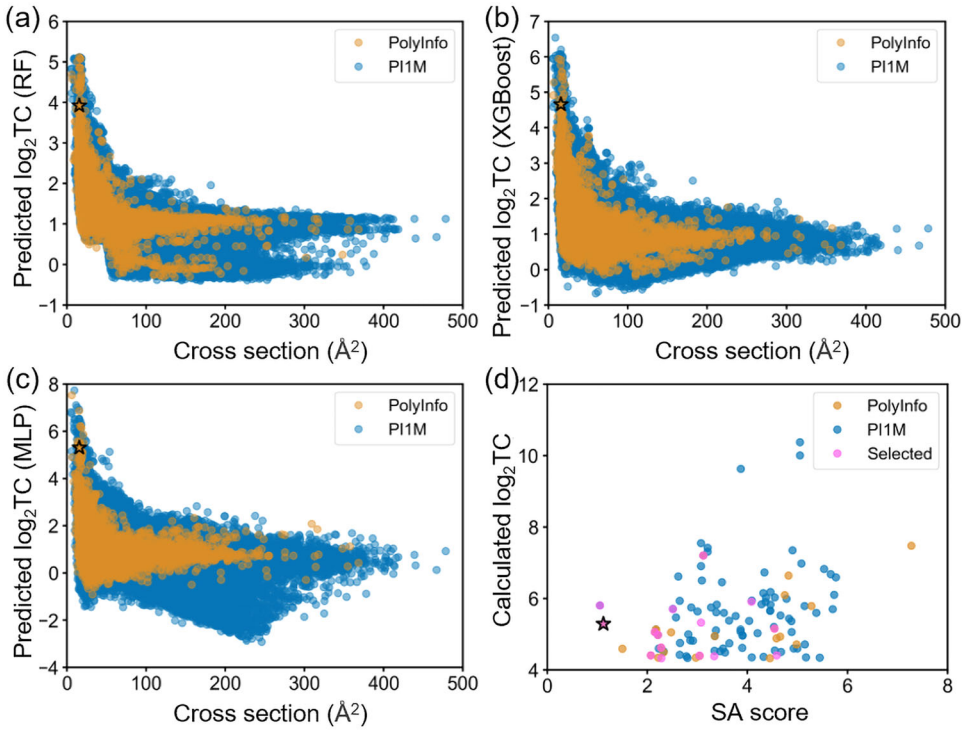

Fig. 5 Prediction of high TC polymers in PoLyInfo and PI1M databases using constructed ML models. a, **b** and **c** based on RF, XGBoost, and MLP models, respectively. **d** Synthetic accessibility (SA) score versus calculated $\log_2$TC of screened high TC polymers (TC > 20.00 W m^{-1} K^{-1}). The 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 $\log_2$TC 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 contributions[71], 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 features[72]. SR does not require massive datasets, as long as a high consistency and accuracy[73]. 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 code[74]. 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 $\geq$0.85 and complexity $\leq$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 $\log_2$TC of polymer structures[75]. 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

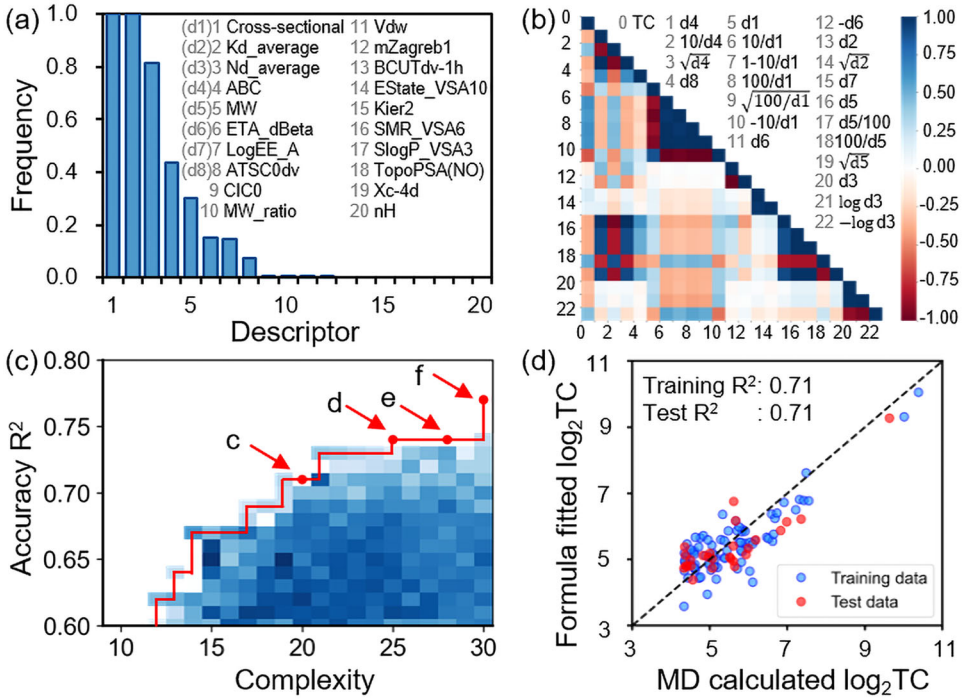

Fig. 6 GPSR for TC prediction of promising polymers. a Frequency of occurrence of optimized descriptors in 158 mathematical formulas (PC values $\geq$ 0.85 and complexity $\leq$ 10). **b** Pearson correlation matrices showing correlations among 22 descriptors and TC, where the descriptors d1–d8 correspond 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 $\log_2$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 $\log_2$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 [*]C=C[*] and [*]N=N[*] 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 backbone[15]. Figure 7b illustrates the phonon dispersion relations, which were obtained by phonon spectral energy density (Phonon-SED) analysis[78], 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 [*]N=N[*], 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 chains[21]. 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 [*]C=C[*]. 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) simulations[79], 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 analysis[63]. 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.

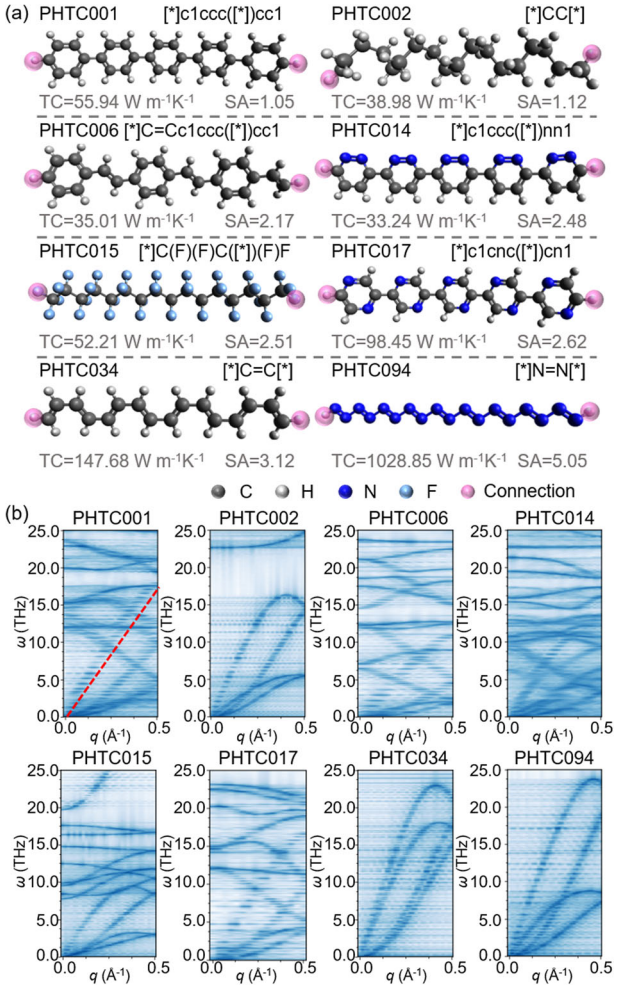

Fig. 7 Structure and phonon dispersion relations for the eight promising polymers. a Polymer chain structures. **b** Phonon dispersion relations. The q is the wavevector, the ω is phonon frequency and the average phonon group velocity of one branch is estimated as the slope of the origin to the maximum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

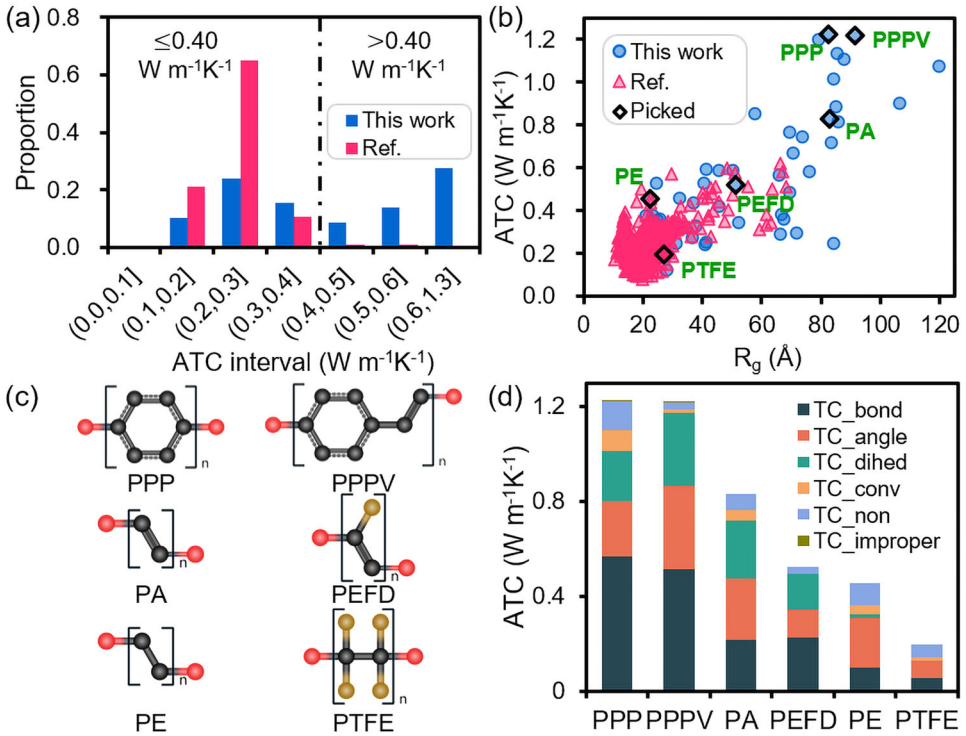

Fig. 8 Thermal conductivity of amorphous polymers (ATC). a ATC of 58 structures randomly selected from data of 107 promising polymers (this work), and the reference (Ref.) data calculated by Hayashi et al. [63], contains 1051 polymers, using the same simulation parameters as set in this work. **b** Radius 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-diyl] (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 polymers[80]. 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 $W m^{-1} K^{-1}$. 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 polymerization[83]. 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 field[84]. 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 files[85].

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 monomer[13]. 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 compounds[86].

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 Simulator[85]. 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 sample[87]. 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 calculations[90].

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 flux[92]. 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 descriptors[47]. 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 package[45]. The Mol2vec fingerprints were embedded via Mol2vec[40]. We referred to the polymer representation model trained using PoLyInfo and PI1M databases for generating Mol2vec fingerprints[60].

The ML models of RF, XGBoost, and MLP were implemented by using Scikit-learn[94]. Hyperparametric optimization for RF, XGBoost, and MLP was operated with the Bayesian Optimization package[95] 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 iterations[52].

To explain the association of optimized descriptors with TC, we used the SHAP toolkit with RF model to evaluate the feature importance[62]. 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 effect[52].

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 gplearn[74]. The 107 polymer structures with TC greater than 20.00 $W\,m^{-1}\,K^{-1}$ 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 approach[77]. 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 approach[78] 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	$\{+, -, \times, \div, \sqrt{x}, \ln x, \lvert x\rvert, -x, 1/x\}$
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

1. Zhang, B. et al. Modulating thermal transport in polymers and interfaces: theories, simulations, and experiments. *ES Energy Environ.* **5**, 37–55 (2019).
2. Ghaffari-Mosanenzadeh, S. et al. A review on high thermally conductive polymeric composites. *Polym. Compos.* **43**, 692–711 (2022).
3. Xu, Q. et al. Recent progress of quantum dots for energy storage applications. *Carb. Neutral.* **1**, 13 (2022).
4. Qian, X., Zhou, J. & Chen, G. Phonon-engineered extreme thermal conductivity materials. *Nat. Mater.* **20**, 1188–1202 (2021).
5. Guo, Y., Zhou, Y. & Xu, Y. Engineering polymers with metal-like thermal conductivity—present status and future perspectives. *Polymer* **233**, 124168 (2021).
6. 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).
7. Xu, Y. et al. Nanostructured polymer films with metal-like thermal conductivity. *Nat. Commun.* **10**, 1771 (2019).
8. Shen, S., Henry, A., Tong, J., Zheng, R. & Chen, G. Polyethylene nanofibres with very high thermal conductivities. *Nat. Nanotechnol.* **5**, 251–255 (2010).
9. Ma, J. et al. Thermal conductivity of electrospun polyethylene nanofibers. *Nanoscale* **7**, 16899–16908 (2015).
10. Lu, C. et al. Thermal conductivity of electrospinning chain-aligned polyethylene oxide (PEO). *Polymer* **115**, 52–59 (2017).
11. Singh, V. et al. High thermal conductivity of chain-oriented amorphous polythiophene. *Nat. Nanotechnol.* **9**, 384–390 (2014).
12. 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).
13. Liu, X., Lin, C. & Rao, Z. Thermal conductivity of straight-chain polytetrafluoroethylene: a molecular dynamics study. *Int. J. Therm. Sci.* **159**, 106646 (2021).
14. 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).
15. 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).
16. Henry, A. & Chen, G. High thermal conductivity of single polyethylene chains using molecular dynamics simulations. *Phys. Rev. Lett.* **101**, 235502 (2008).
17. Liu, J. & Yang, R. Tuning the thermal conductivity of polymers with mechanical strains. *Phys. Rev. B* **81**, 174122 (2010).
18. 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).
19. 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).
20. Zhang, T. & Luo, T. Role of chain morphology and stiffness in thermal conductivity of amorphous. *Polym. J. Phys. Chem. B* **120**, 803–812 (2016).
21. Chen, H. et al. Thermal conductivity of polymer-based composites: fundamentals and applications. *Polym. Sci.* **59**, 41–85 (2016).
22. Kim, S. et al. PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2015).
23. 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).
24. Chen, L. et al. Polymer informatics: current status and critical next steps. *Mater. Sci. Eng. R. Rep.* **144**, 100595 (2021).
25. 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).
26. 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).
27. Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
28. 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).
29. 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).
30. Afzal, M. A. F., Haghighatlari, 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).
31. Wu, C. et al. Flexible temperature-invariant polymer dielectrics with large bandgap. *Adv. Mater.* **32**, 2000499 (2020).
32. 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).
33. Sahu, H. et al. An informatics approach for designing conducting polymers. *ACS Appl. Mater. Interfaces* **13**, 53314–53322 (2021).
34. 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).
35. Bhowmik, R., Sihn, S., Pachter, R. & Vernon, J. P. Prediction of the specific heat of polymers from experimental data and machine learning methods. *Polymer* **220**, 123558 (2021).
36. 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).
37. Ma, R., Zhang, H. & Luo, T. Exploring high thermal conductivity amorphous polymers using reinforcement learning. *ACS Appl. Mater. Interfaces* **14**, 15587–15598 (2022).
38. 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).
39. 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).
40. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35 (2018).
41. 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).
42. 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).
43. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* **12**, 43 (2020).
44. 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).
45. 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. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).
48. Haghighatlari, 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. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.* **126**, 1763–1768 (2018).
56. Chandrashekar, 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., Shiomi, 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 multiobjective 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. Wirnsberger, 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.