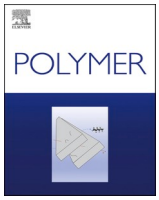
Polymer 320 (2025) 128081

Contents lists available at ScienceDirect

Polymer

journal homepage: www.elsevier.com/locate/polymer



COMFO: Integrated deep learning model facilitates discovery of multifunctional polyimide materials

Bo Zhang [a], Xueqing Li [b], Ming Zeng [a,**], Jingguo Cao [b,*]

[a] College of Marine and Environmental Sciences, Tianjin University of Science & Technology, 300457, Tianjin, China
[b] College of Chemical Engineering and Materials Science, Tianjin University of Science & Technology, 300457, Tianjin, China

ARTICLE INFO

Keywords:
Polyimide
Deep learning
Dielectric materials
Structure-property relationship

ABSTRACT

Rapid development in 5G and electronic appliance fields places higher demands on the dielectric, thermal and mechanical properties of materials. Dielectric materials with excellent comprehensive properties have become an urgent need of the times. Emerging machine learning techniques could greatly accelerate the discovery of high-performance dielectric materials. However, it remains unknown whether traditional 2D fingerprints or descriptors can extract molecular structure information more completely. In this study, data on four types of properties of polyimide (PI), including dielectric constant, glass transition temperature, tensile modulus and coefficient of thermal expansion, were collected to construct a deep learning model-COMFO to explore polyimide dielectric materials with excellent comprehensive performance. Our COMFO model could extract the key feature information in the molecule from three perspective learning tasks as well as process and learn them. Specifically, the three learning tasks include extracting the feature information in the SMILES sequence using a large language model, the bidirectional encoder Transformer; extracting the information about the atoms and bonds of polymer molecules from molecular graph using the Attentive FP network; and extracting the information about the substructures of polymer molecules through molecular fingerprints. The multi-perspective feature extraction task gave our model a more excellent performance ($R^2 > 0.90$). The performance of the model was confirmed by various ways, including experimental validation, MD simulation validation, and comparison with 12 other models. Design guidelines for low dielectric constant PIs were discovered by monomer structure analysis. High-throughput virtual screening of 158,022 unknown PIs was performed and three PIs with excellent comprehensive properties (especially dielectric properties) were identified. MD and DFT approaches verified and analyzed the properties of these three potential high-performance PIs. In the future, this research could also contribute to the forward development of materials in other fields.

1. Introduction

In recent years, with the rapid development of electronic appliance fields and the advent of the 5G era, materials with excellent dielectric properties, good heat resistance, and high mechanical strength have attracted extensive attention [1]. In order to adapt to the development of the industry, these excellent materials need to possess some key properties, such as high glass transition temperature (Tg) and tensile strength (TM). In addition to this, the materials were also required to have low linear coefficient of thermal expansion (CTE) and dielectric constant (ε) when applied to electronic packaging and 5G signal transmission [2]. The commonly used dielectric materials in traditional research were glass, ceramics and their composites [3,4]. However, compared with polymer dielectrics, these traditional materials have the disadvantages of high dielectric loss, poor flexibility and difficult processing. Currently, commonly used polymer dielectric materials in the 5G era included polypropylene [5], polyphenylene sulfide [6], polycarbonate [7], and polyethersulfone [8]. Electronic appliances might experience heat generation during use and the processing temperature of these materials might also exceed 250 °C during primary fabrication. Therefore, the selected polymer dielectric materials should have good heat resistance, with glass transition temperatures of at least 300 °C or higher [9]. However, none of these commonly used polymer materials could overcome the problem of high-temperature tolerance faced by the

* Corresponding author.
** Corresponding author.
E-mail addresses: ming.zeng@tust.edu.cn (M. Zeng), cjg@tust.edu.cn (J. Cao).

https://doi.org/10.1016/j.polymer.2025.128081
Received 4 October 2024; Received in revised form 12 January 2025; Accepted 22 January 2025
Available online 24 January 2025

material during processing and use. Therefore, high heat-resistant and low dielectric polymer materials have become an urgent need for the development of the times.

As a potential high-performance engineering material, polyimide(PI) has good mechanical properties, excellent heat resistance (Tg > 300 °C), and outstanding solvent resistance due to its rigid imide groups and strong intermolecular interaction forces [10]. Owing to its superior overall performance, PI was widely used in many fields such as flexible display [11], aerospace, gas separation [12] and fuel cells [13]. However, the dielectric constant of PI was relatively high, which greatly restricted their development in the field of 5G and microelectronics. Polyimides were usually synthesized by polycondensation of dianhydrides with diamines or diisocyanates. The multivariability of the monomer structure made for flexibility in molecular design. Through the design of the molecular structure of monomers and the combination between monomers, many polyimides have been designed for different demand situations [14,15]. For example, introducing trapezoidal and twisted spirocyclic units into the monomer could effectively reduce the stacking of PI molecular chains and thus improve its air permeability [16]. However, there was often a "constrained" relationship between the properties of polyimide materials [17]. Moreover, the synthesis of conventional polyimide materials has been a trial-and-error process that has always relied on empirical explorations, so it could be a lengthy endeavor to obtain a PI dielectric material with excellent comprehensive performance.

Fortunately, the development of computational simulation and machine learning techniques has brought a new light to this work. The use of computational simulation, including molecular dynamics simulation (MD) and density functional theory (DFT) could quantitatively establish the complex relationship between polymer structure and properties [18]. In addition, the MD and DFT methods could interpret the polymer structure from the microscopic scale such as molecular free volume, intermolecular interactions etc. and thus guide the design of molecular structures [19]. Currently, a number of studies have applied MD and DFT to predict and analyze the key properties of polyimides [20,21]. However, while the computational simulation approach could alleviate the challenges of conventional material design to some extent, its time-consuming computational cost also limited its exploration within the broader polymer chemistry space [22]. It was worth noting that the computational accuracy of MD and DFT remained reliable, and the use of a combination of experiments and MD/DFT to validate the ML findings was a highly promising approach to materials design. In recent years, machine learning (ML) and deep learning (DL) techniques have received increasing attention in the field of materials design. It was an efficient method driven by data from traditional experimental tests or computational simulations. Researchers could rely on this technique to establish quantitative structure-property relationships of polymer materials [23,24]. The computational speed of machine learning was much faster than experimental and computational simulations. Moreover, since the training of machine learning was based on a large and rich material structure-property database, the obtained conclusions were more widely generalized than computational simulations.ML and DL have been successfully applied to the prediction of polymer properties such as glass transition temperature (Tg) [25], cutoff wavelength, and gas separation [26]. For example, Wang et al. [16] successfully predicted the gas separation properties of PI using active learning techniques combined with MD. Due to the limited database of gas separation, they built up the data with the help of MD simulation on one hand, and on the other hand, in order to save the cost of calculations they chose the FFV and AVS, which are highly correlated with the performance and selectivity of gas separation, as a substitute to perform the calculations. Finally, after 8 rounds of active learning training, they screened out several high-performance PI membranes, and used MD to verify the conclusions drawn. All the previous studies have proved the huge potential and unstoppable development trend of machine learning technology in the field of material science.

Feature extraction is a key aspect of polymer machine learning, and the effectiveness of the extraction is one of the decisive factors in the performance and generalization ability of the model. At present, there are many kinds of feature representations, such as molecular fingerprints, molecular graphs, descriptors, etc., which have different focuses when extracting polymer molecular information. However, it was still unknown whether these feature extraction methods with different focuses could extract the hidden information in the polymer molecular structure more completely.

In this study, we collected data on four types of properties of polyimide, including dielectric constant (ε), glass transition temperature (Tg), tensile strength (TM) and linear coefficient of thermal expansion (CTE), and constructed a deep learning model-COMFO to explore polyimide dielectric materials with excellent comprehensive performance. Our COMFO model could extract the key feature information in the molecule from three perspective learning tasks as well as process and learn them. The feature extraction tasks from these three perspectives included extracting feature information in SMILES sequences using a large language model, the bidirectional encoder Transformer; extracting information about atoms and bonds of polymer molecule from molecular graph using an Attentive FP network; and extracting information about the substructure of polymer molecule through molecular fingerprints. The multi-angle feature extraction task gave our COMFO model an even better performance. The performance of the model was confirmed by various ways, including experimental validation, MD simulation validation, and comparison with 12 other models. Finally, a high-throughput virtual screening of 158022 PIs was performed and three classes of PIs with excellent comprehensive properties (especially dielectric properties) were identified. MD simulations and DFT calculations further validated and analyzed the properties of the three PIs.

2. Methods

2.1. Data preparation

In this study, we focused on four types of properties of polyimide, including dielectric constant (ε), glass transition temperature (Tg), linear coefficient of thermal expansion (CTE), and tensile modulus (TM). Structure-property data for 875 PIs were collected from 109 relevant papers (see Table S1 for source literature), constituting dataset A.(Note: The PI dielectric constant data collected in this study were mainly from research on flexible displays, batteries and high performance materials, where the thickness of the material was between 5 and 100 μm. The influence of thickness on the distribution of electric field in the material and the polarization effect is quite weak under this thickness, and the dielectric properties of the materials in this range depended more on their molecular structure than on their thickness.)The related data distribution was illustrated in Fig. 1c and d. The specific steps for data collection and cleaning were as follows:

(1) Structures of polyimides were collected from the literature;
(2) The structures of PI with number-average molecular weight (Mn) > 6000 g/mol and weight-average molecular weight (Mw) > 10000 g/mol were screened;
(3) ChemDraw 19.0 software was used to draw the repeat units of the PIs (to indicate the presence of bonds at the linkages of the repeat units, we denote them with *);
(4) Convert the repeat units of PI into SMILES characters uniformly.
(5) Considering the influence of different testing instruments and testing methods on the properties of the final PI, we took the mean value of the properties of the repetitive data and deleted the repetitive structures (See Note S6 for more detailed data collection).

In order to screen potential high-performance PIs from a larger polymer chemical space, we collected a series of real existing diamine,

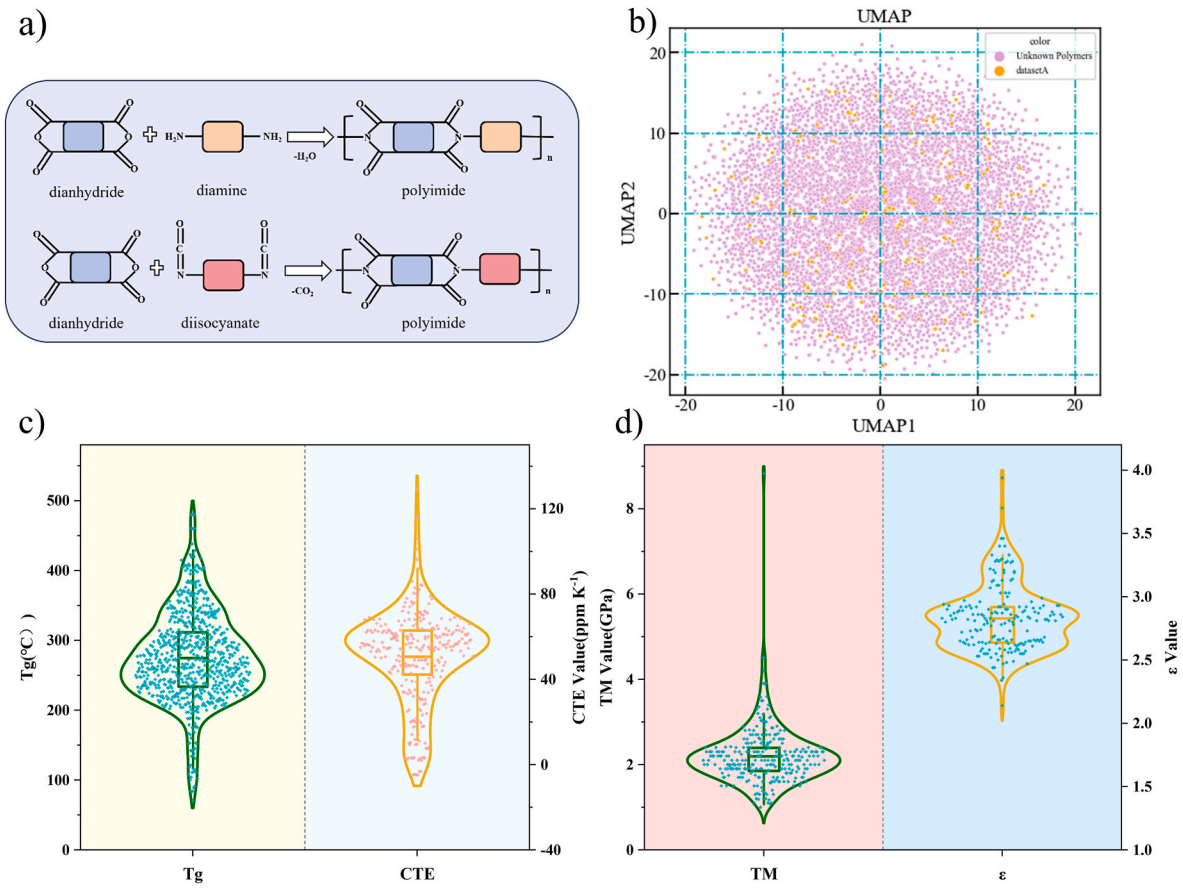


Fig. 1. (a) Two synthesis routes of PI; (b) UMAP plots of unknown polymers versus dataset A; (c)–(d) Plots of data distributions for four predicted properties.

dianhydride, and diisocyanate monomers from the PubChem database and constructed 158022 PIs (screening space) based on the two types of synthesized modes of PIs (Fig. 1a). Fig. b shows the distribution of the 158022 unknown PIs and the PIs of dataset A in chemical space. According to the UMAP plot, the unknown PIs uniformly occupied a larger chemical space compared to dataset A, indicating the rationality and effectiveness of the constructed screening space.

2.2. Feature extraction

This study aimed to construct an integrated deep learning model-COMFO, which was based on taking the simplified chemical language SMILES as input. The complete framework of the COMFO model consisted of 1) 3 major feature extraction, 2) feature integration, and 3) model training and prediction. In the first step of feature extraction, we utilized the SMILES language as input to construct three learning tasks for molecular information extraction. The SMILES-based feature extraction task was designed to extract the feature information contained in SMILES sequences using a large language model, the molecular graph-based feature extraction task was designed to introduce the Attentive FP network to extract the information of atoms and bonds embedded in the molecular graph, and the molecular fingerprint-based feature extraction task was designed to extract the information of substructures in polymer molecules.

2.2.1. SMILES-based feature extraction task

The SMILES-based feature extraction task was based on the recently released Self-Supervised Learning (SSL) platform [27], which was built on millions to billions of datasets. The bidirectional encoder Transformer (BET) implemented the self-supervised learning. The BET model contained an attentional mechanism, which allowed it to capture the hidden information in the SMILES sequence more accurately. Previous autoencoder models contained two layers, the encoder and the decoder. The encoder aimed to read the information and convert it into a latent space representation, while the decoder was used to convert the representation into a target probability distribution. The purpose of this study was only to extract the embedding information from SMILES using the DL model, thus the BET model consisted of only Transformer-based bidirectional encoders, which greatly reduced the training time of the deep learning model.

The BET model contained 8 Transformer encoders, each of which in turn consisted of a fully-connected feed-forward layer with an embedding size of 1024 and a self-attention layer with a header of 8. The learning strategy of the BET model was to inferentially reproduce real masked SMILES pairs by self-supervised learning of unmasked SMILES pairs, thus achieving the goal of understanding the information embedded in the SMILES language. To construct real SMILES pairs and masked SMILES pairs, 61 symbols with special materialized meanings were inserted into the SMILES sequence input to the BET model. The maximum sequence length was required to be 256, and the head and tail of the SMILES sequence were supplementally set to '<s>' and '<\s>' symbols, respectively. Therefore, if the input SMILES length was over 256, it was deleted, and the shortfall was replaced with '<pad>' symbols. The embedding dimension of each symbol was 512. 15 % of the symbols in the SMILES were extracted for masking process during training, which 80 % were masked, 10 % were without change, and the remaining 10 % were done with random replacement. The model was trained using Adam as optimizer, 0.1 as weight decay and with 0.0001 as maximum learning rate. The fully trained BET model could provide new DL-based fingerprints for downstream prediction tasks. In this study, we use a BET model pre-trained based on the ChEMBL database to extract the feature information of SMILES sequences with an extracted feature

embedding dimension of 512 (Fig. 2b).

2.2.2. Molecular graph-based feature extraction task

To extract the hidden information in the molecular structure based on molecular graphs, we introduced the Attentive FP network proposed by Xiong et al. [28]. Compared with the traditional GNN model, the introduction of the graph attention mechanism makes it easier to learn the atomic and molecular property information contained in the molecular graph in depth. The SMILES characters input to the model were first converted into molecular graphs using the RDKit package, which mainly composed of atomic feature matrices and bond feature matrices (see Tables S2–S3 for the definition rules). Since the task of this part was only to extract the embedding information based on the molecular graph using the Attentive FP, the model only consisted of the input, Attentive and readout layers, and the readout information was not considered for the time being to be fed into the regression model for training. The input layer reads the input molecular graph information. In the Attentive layer, we introduced two Attentive FP convolutional layers to extract the atomic features of the input molecules, and the initial number of neuron nodes was set to the default value of 300. The readout layer, also known as the pooling layer, with the number of layers of 1, was used to read out the molecular embedding information extracted by the graph-attention mechanism. In this study, we use the output of the pooling layer as the molecular graph-based embedding feature information (Fig. 2c).

2.2.3. Molecular fingerprint-based feature extraction task

MorGan fingerprints were used as an algorithm for fingerprint feature extraction. Compared to physicochemical descriptors, MorGan fingerprints were dynamic in the sense that they could change their recognized substructures and connectivity information between repeat units in response to changes in the input data. Inspired by the work of Yang et al. [29] and in order to minimize the information loss during fingerprint recognition, we focused on the frequency information of the occurrence of each substructure during the iteration of the MorGan algorithm. The improved MorGan fingerprint(IMF)not only recognized the presence or absence of substructures but also recorded the frequency of occurrence of the already existing substructures. The initial iteration radius of the MorGan fingerprint used in this study was set to 2. When the IMF was calculated, we then used two major methods, the variance test method and RF-based feature importance scores, to train and adjust the output IMF feature dimensions. The molecular fingerprints themselves could be fed into the regression model for training, but we set up two more control groups to determine whether model should be further employed to extract the IMF fingerprint information. Control group 1: IMF fingerprints were extracted using the LSTM model; Control group 2: IMF fingerprints were extracted using the model FFNN model. The readout feature dimensions of the FFNN and LSTM models were still consistent with the original IMF fingerprints (Fig. 2a).

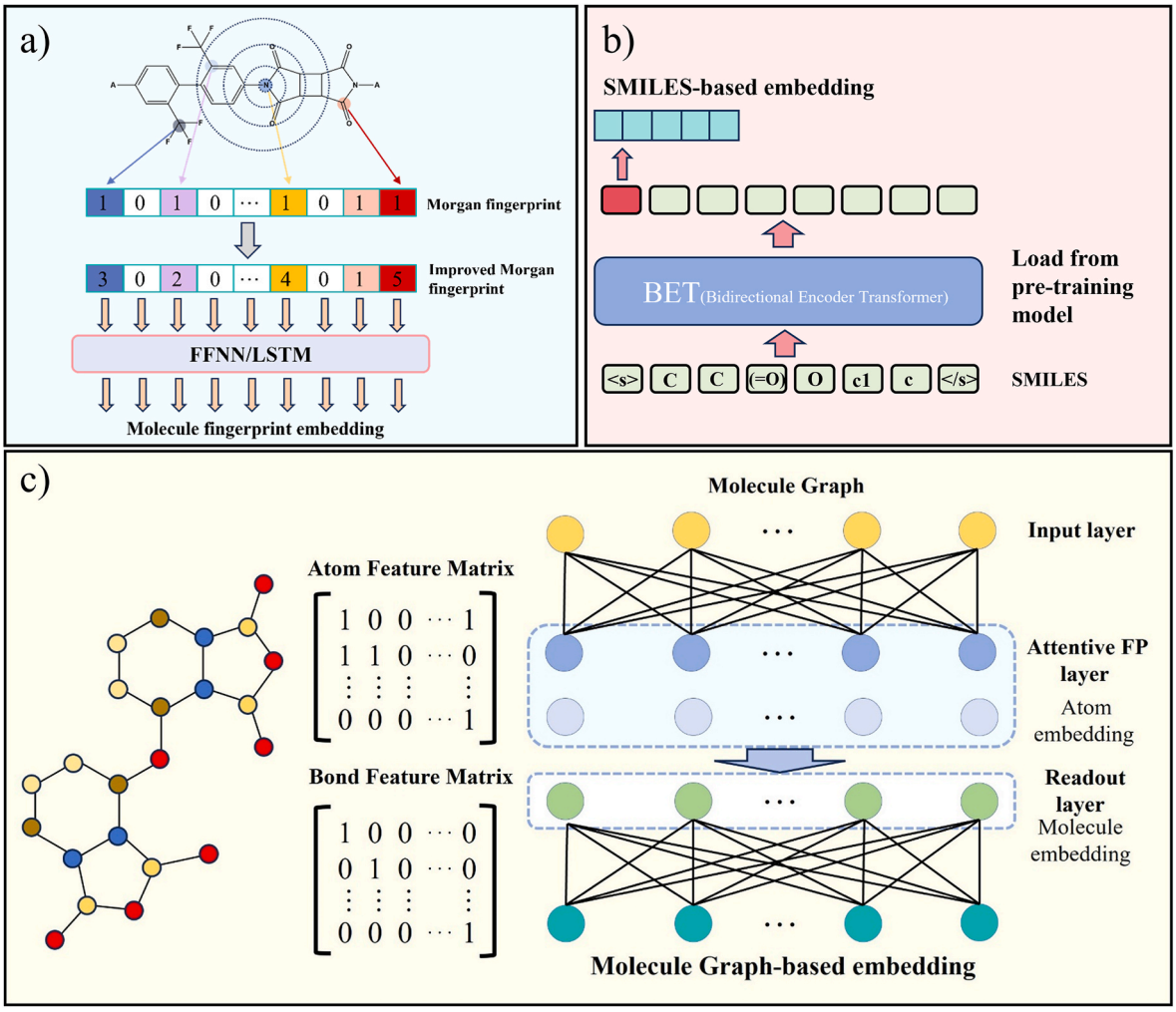


Fig. 2. Three major feature learning tasks. a) molecular fingerprint-based feature learning task; b) SMILES-based feature learning task; c) molecular graph-based feature learning task.

2.3. COMFO model development

2.3.1. COMFO model construction

After the embedding features were extracted from each of the three major feature learning tasks mentioned above, the role of the integrated module was to merge these embedding information. In this module, we adopted vector splicing to integrate these embedding information. The integrated features would be fed into the pooling layer for further average pooling. Finally, the features read out from the pooling layer would be input into the fully connected layer for training to construct the regression model (COMFO) (Fig. 3). A total of four fully connected layers were constructed, each of which was connected to a normalization layer and an activation function (ReLU), respectively. According to the different extraction methods of molecular fingerprints, three types of COMFO models were constructed, namely, COMFO_ORI, COMFO_FFNN and COMFO_LSTM. Among them, COMFO_ORI was the model built by using the original fingerprints, COMFO_FFNN was the model built by using the molecular fingerprints extracted from the FFNN model, and COMFO_LSTM was the model built by using the molecular fingerprints extracted from the LSTM model.

A ratio of 7:3 was used to divide the training set and test set of the model during training. The global random seed is set to 42. The batch_size was set to 30, the shuffle was set to Ture, and the dropout was set to 0.3. The Adam optimizer was used, the initial learning rate was set to 0.0001, and the weight decay was set to 0.000001. The initial number of iterations was set to 501. The COMFO model was built based on the Pytorch library. Meanwhile, to check the model performance, we also built three comparison models based on scikit-learn library, including DNN, RF and XGBoost respectively. The coefficient of determination squared (R^2), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to comprehensively evaluate the performance and generalization ability of the models.

2.3.2. Training and hyperparameter optimization of COMFO models

In the feature extraction module of molecular fingerprints, we mainly adjusted the parameters such as the iteration radius of IMF fingerprints and the dimension of output features. In the feature extraction module of molecular graph, to prevent overfitting, we fixed the number of layers of Attentive FP convolutional layers to 2, and mainly adjusted the number of nodes of its neurons as well as the feature embedding dimension of the readout layer. To simplify the parameter adjustment process, the output feature embedding dimensions of these two types of learning tasks were set to the same value and adjusted synchronously. Due to the complexity of deep learning model hyperparameter tuning, the hyperparameters of all models were not set arbitrarily but were randomly selected based on a large amount of parameter selection experience from previous studies (The hyperparameter adjustment process was shown in Fig. S2). The learning rate, weight decay and the number of iterations for model training were gradually scaled up from the initial values based on the predicted properties.

2.4. Simulation details

To validate the prediction of the model, four properties namely Tg, CTE, ε and TM were calculated by molecular dynamics simulation in this study. The molecular dynamics simulation was carried out by using the Forcite module in Material Studio software. The model building process was as follows: 1) firstly, a PI molecular chain with 20 repeat units was built and geometrically optimized; 2) a cubic simulation Cell was built using the Amorphous Cell module. The number of PI molecular chains in the Cell was 10 and the initial density was 0.1 g/cm^3. The constructed Cell needs to be continued for dynamic equilibrium (see NoteS1 for detailed steps).

The free volume fraction (FFV) and cohesive energy density (CED) of the three types of PIs were calculated with the MD method. The electrostatic potential (ESP) and the HOMO-LUMO orbitals of the repeat unit were calculated with the Dmol3 module.

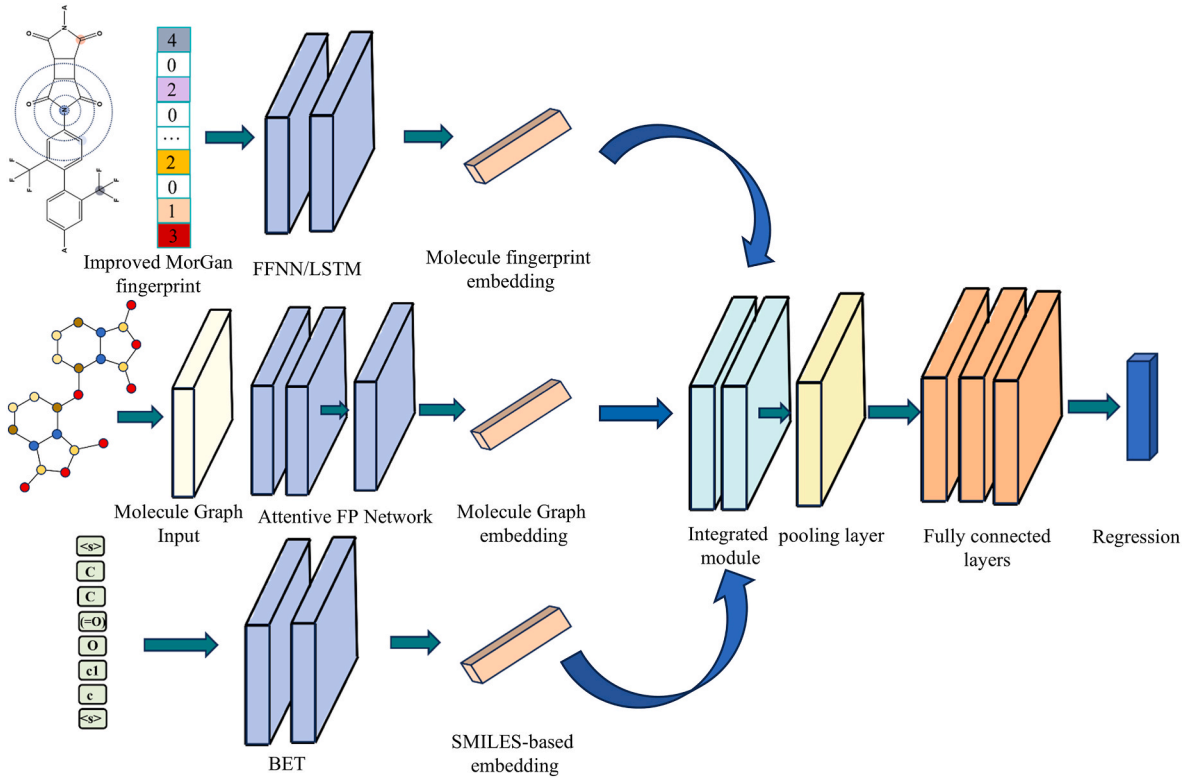


Fig. 3. A framework for the COMFO model.

3. Results and discussion

3.1. Performance of COMFO models

In this study, we have focused on four PI properties, i.e., Tg, TM, CTE, and ε. By comparing the three types of COMFO models as well as adjusting the parameters of the initial learning rate, weight decay and the number of iterations for model training, we trained structure-property relationship models for each property (see Table S4 for the hyperparameter statistics of each optimal model).

Fig. 4a–d and Table 1 show the performance of the four properties prediction models on the training set and test set, respectively. The results show that most of the property prediction models have R^2 higher than 0.90. Among them, the optimal model for Tg prediction was COMFO_FFNN, whose R^2 on the training/testing set was 0.982/0.915. The highest MAE, RMSE, and MAPE of this model were 13.542, 17.723, and 5.015, respectively, on the training or testing set, which indicated that the Tg prediction model that we built had excellent performance. Otherwise, the optimal models for TM, CTE and ε prediction had R^2 of 0.936/0.876, 0.975/0.918 and 0.971/0.903 on the training/testing set, respectively. Combined with the performance evaluation indexes of RMSE, MAE and MAPE of each model, the results showed that the COMFO model built for four types of property prediction in this study has excellent performance and good generalization ability. To verify this conclusion, we built three comparison models based on scikit-learn library, such as DNN, RF and XGBoost respectively. The glass transition temperature (Tg) is a comprehensive indicator of polymer performance, which could reflect the polymer's rigidity, mechanical properties and heat resistance to a certain extent. The dielectric properties of polymers are also closely related to their rigidity. Therefore, to simplify the number of training times for the model, in this study we only built the comparison model with Tg predictions as a representation. In terms of feature representation, we compared four commonly used molecular fingerprints/descriptors for each model, i.e., MACCS fingerprints, ECFP fingerprints, IMF fingerprints, and RDKit descriptors. The hyperparameters of DNN, RF and XGBoost models were determined using Bayesian optimization. Fig. 4e shows the performance of each comparative model and the three types of COMFO models. The results showed that the three types of COMFO models performed significantly better than the remaining 12 comparison models, the R^2 of most COMFO models were higher than 0.90. This was mainly due to two reasons. a) Feature extraction was one of the most important factors in determining the performance and generalization ability of the polymer machine learning model. Our COMFO model established three feature learning tasks, including learning the information contained inside SMILES sequences using the large language model-BET, learning the information of atoms and bonds contained inside molecular graphs using the Attentive FP network, and extracting the information of substructures contained inside polymer molecules through molecular fingerprints. The multi-perspective learning tasks have enabled our model to extract the information contained in the molecular structure more broadly. b) The key feature information extracted from these learning tasks was integrated by vector splicing. And this was not a simple feature integration, to further enable the model to learn this integrated information and prevent overfitting, we did average pooling of this integrated information.

Table 1
Performance of the test set of four property prediction models.

Property	Model	R^2(train/test)	MAE (train/test)	RMSE (train/test)	MAPE (train/test)
Tg	COMFO_FFNN	0.982/0.915	6.445/13.542	8.805/17.723	2.426/5.015
TM	COMFO_ORI	0.936/0.876	0.154/0.139	0.184/0.190	8.133/6.369
CTE	COMFO_FFNN	0.975/0.918	2.303/3.707	3.367/5.198	7.027/8.579
ε	COMFO_LSTM	0.971/0.903	0.026/0.044	0.044/0.078	0.899/1.488

The output layer of the COMFO model before exiting was used as its integrated fingerprints. To further examine the ability of COMFO model to extract molecular features, we compared the integrated fingerprints extracted by the COMFO model with the three commonly used feature representations (MACCS fingerprints, ECFP fingerprints, and RDKit descriptors) through feeding into DNN and XGBoost model. Fig. 4f shows the results of the performance comparison between the integrated fingerprint and the three commonly used feature representations. The results showed that the RMSE of the DNN and XGBoost models built

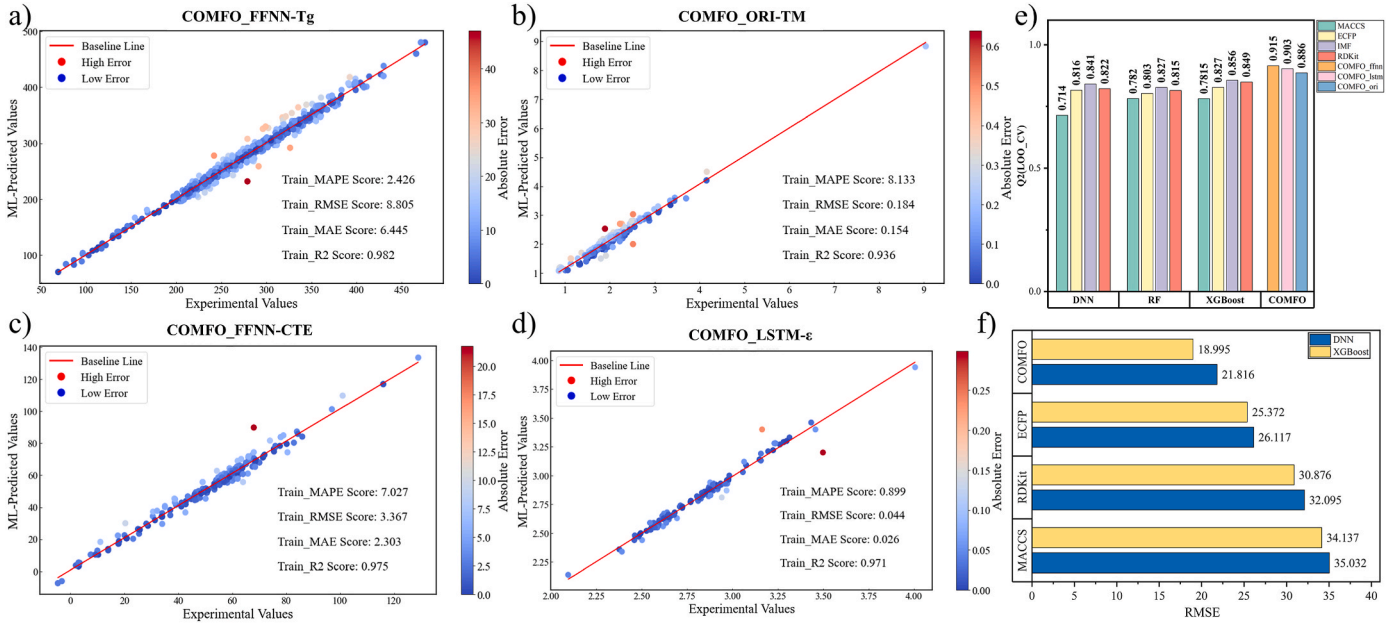


Fig. 4. a)-d) Training set performance plots for Tg, TM, CTE and ε prediction models; The predicted ε value was at the temperature of 298K. e) Performance of COMFO model vs. other models; f) Performance of COMFO fingerprints vs. other fingerprints.

Table 2
Comparison of ML-predicted and MD simulated values of the four properties of three real PIs.

Property	PI-1			PI-2			PI-3		
	ML	MD	Diff	ML	MD	Diff	ML	MD	Diff
Tg/°C	318.77	310	2.83 %	330.83	312	6.04 %	331.9	317	4.7 %
TM/GPa	1.82	1.75	4.0 %	1.89	1.80	5.0 %	2.22	2.46	9.76 %
CTE/ppmK^{-1}	53.17	59.2	10.19 %	50.54	54.39	7.08 %	58.15	63.83	8.90 %
ε	3.20	3.0	6.67 %	3.24	3.43	5.54 %	3.18	3.01	5.65 %

Notes: Diff = |(MD-ML)/MD| × 100 %.

Table 3
Comparison of ML-predicted and MD simulated values of the four properties of three high-performance PIs.

Property	HPPI-1			HPPI-2			HPPI-3		
	ML	MD	Diff	ML	MD	Diff	ML	MD	Diff
Tg/°C	418.2	401	4.54 %	411.22	383	7.31 %	412.3	377	9.36 %
TM/GPa	3.37	3.21	4.98 %	3.0	3.19	5.96 %	2.97	3.14	5.41 %
CTE/ppmK^{-1}	29.5	32.12	8.16 %	34.3	37.84	9.36 %	35.11	33.73	4.09 %
ε	2.26	2.32	2.59 %	2.33	2.41	3.32 %	2.41	2.52	4.37 %

based on COMFO fingerprints was the lowest, ECFP was the second highest, and MACCS performed the worst. Table 4 shows the results of comparing the performance of single embedded fingerprints combined with the full connectivity layer model and that of the COMFO model based on the complete integrated fingerprint. All these comparisons fully demonstrated the rationality and excellent performance of our established multi-angle feature extraction strategy.

3.2. Validation of COMFO models

We selected three real existing PIs and calculated their Tg, TM, CTE and ε by MD simulation to further examine the performance of the COMFO model. Notably, our MD simulation approach has been validated with five PIs reported in the literature. The values of the four properties of these five PIs have been tested in full (the molecular structures of the five PIs, and the results of the comparison between the tested values and the MD simulation values were shown in Table S5 and Fig. S3), and the results indicated that our MD simulations fitted well with the literature-reported values. Fig. 5 exhibited the property values of the three real PIs calculated by MD simulations. The properties of these three PIs have not been fully reported, but we aimed to employ the validated MD to further test the performance of the COMFO model. ML has predicted the four property values of the three types of PIs in full. Table 2 showed values of ML predictions and MD simulations for the four properties of the three PIs, and the results showed that the errors between ML predictions and MD simulations for the four properties were still within reasonable limits.

Considering that there was still a difference between the MD simulation and the ML prediction, we found that the error of the Tg prediction was relatively minimal and the Tg test was relatively convenient because the sample did not need to be made into a specified shape. Therefore, we synthesized these three PIs experimentally and measured their Tg values by DMA in a nitrogen atmosphere at a heating rate of 5 °C/min (see NoteS2 and Fig. S1 for the synthesis details of the PIs and the test results), and the results showed that our MD simulations and ML predictions were in very good agreement with the experiments. Therefore, our COMFO model and MD simulation approach were reliable. In addition to this, we also tabulated and compared this research work with the previous research results (Table S9), and the results showed that the performance of our model is still superior.

Table 4
Comparison of the performance of single embedded fingerprints combined with the full connectivity layer model to that of the COMFO model based on integrated fingerprints.

Model	RMSE on Test	R^2 on Test
SMILES-based embedding-FCNN	26.758	0.826
Molecule Graph embedding-FCNN	28.332	0.805
Molecule Fingerprint embedding-FCNN	25.137	0.834
COMFO_FFNN	17.723	0.915
COMFO_LSTM	18.016	0.903
COMFO_ORI	18.886	0.886

3.3. Discovery and analysis of high-performance PIs

When the models with excellent performance have been established, we would prefer to search for potential PIs with excellent comprehensive performance by means of high-throughput virtual screening. The Tg, TM, CTE and ε of 158022 PIs in the screening space were predicted using the final trained COMFO model. Due to the competition between multiple properties and the complexity of multi-dimensional spatial screening, here we have established a three-dimensional design space with Tg, TM and ε only, as shown in Fig. 6a. Also to analyze the ease of synthesis of potential PIs, we introduced SAscore values [30] (see NoteS3 for the concepts and calculations). The SAscore values of 1–10 sequentially indicated a hierarchy of compounds synthesized from easy to difficult. Each point in the 3D space represented one PI, and the color was mapped to the SAscore value.

In Fig. 6b, we visualized monomers of the top 20 PIs with low dielectric constants. This figure showed the most frequent monomer structures of dianhydrides, diamines, and diisocyanates in PIs with low dielectric constants. Among the top monomers, fluorine-containing monomers (A4, A5, B1, B3) and aliphatic ring-containing dianhydrides (A1, A2, A3, A5) were common low-dielectric-constant units, which was consistent with the literature's strategy for designing PIs with low dielectric constants [31,32]. In addition to this, monomer B2 was found to be rich in methyl groups, the presence of which contributes to the rigidity of the molecular chain, which could increase the free volume. The large free volume could weaken the polarization interactions per unit volume, which in turn could lower the dielectric constant. The non-coplanar structures (A1, A2, A6, C1, C4) have large bulk and spatial site resistances, and their presence effectively reduced the packing density of the PI molecular chains, which in turn increased their free volume. Therefore, the introduction of non-coplanar structures was beneficial in decreasing the dielectric constant [33,34]. In addition, monomers A3 and A7 were also found to reduce the dielectric constant of PI, mainly due to their inclusion of torsional and asymmetric structures, which could effectively reduce the stacking density of the PI molecular chains, and thus reduce the dielectric constant. The C3 monomer contained the C–Cl group, and the presence of the chlorine

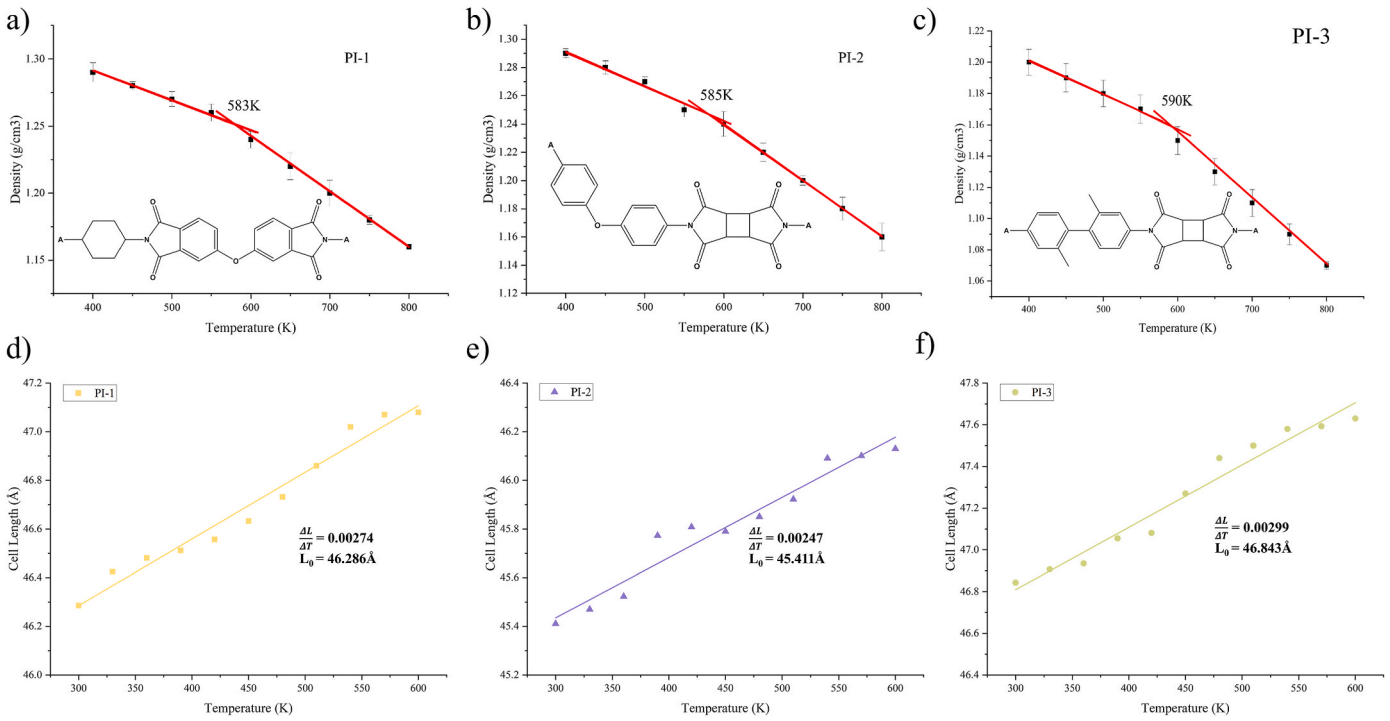


Fig. 5. a)-c) Tg of three real PIs through MD simulation; d)-f) CTE of three real PIs through MD simulation.

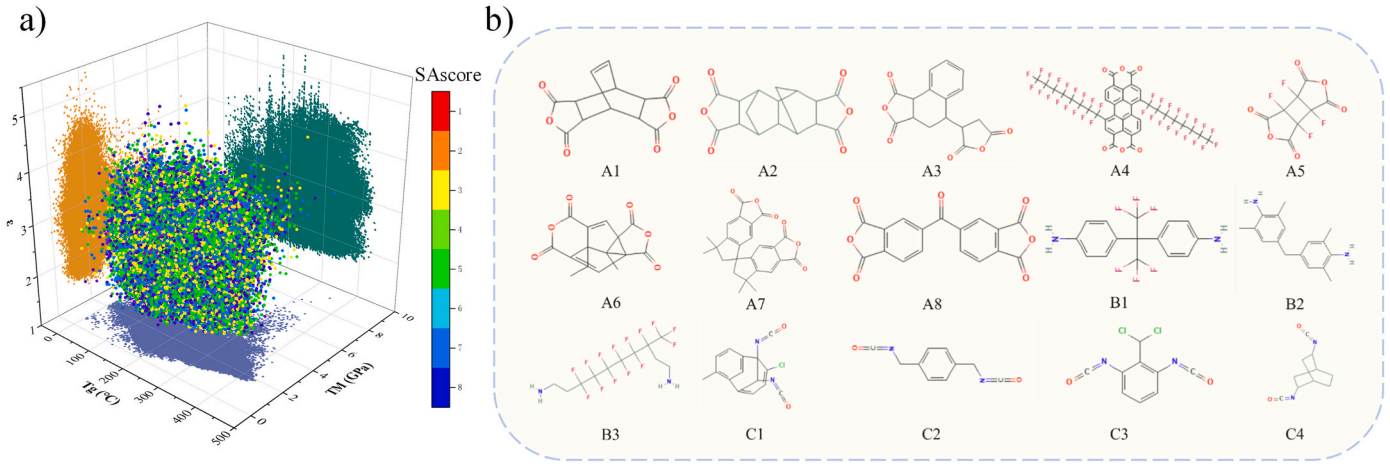


Fig. 6. a) High-throughput virtual screening plot; b) Monomer visualization of PIs with low dielectric constants.

atom increased the polarity of the molecule and could slightly enhance the dielectric constant of the material. However, the symmetry of the chlorine atom substitution position resulted in a smaller dipole moment of the whole molecule, thus weakening its contribution to the increase of the dielectric constant. Moreover, the substitution of chlorine atoms and isocyanate groups in the para- and neighboring positions of the benzene ring produced a spatial site-blocking effect, limiting the flexibility of the molecular chain segments. This highly rigidized structure further reduced the likelihood of polarization of the molecular chain, thus lowering the dielectric constant. At the same time, the benzene ring backbone also imparted large rigidity and conjugation, which further reduced its dielectric constant. And it is to be noted that the property of PI is a synergistic interaction between its multiple substructures, the primordial dianhydride linked to C3 is A7, which has a large free volume and high rigidity and thus could lower the dielectric constant of the overall PI. In conclusion, combining the prediction of COMFO model and the analysis of high-performance monomers, we have identified potential high-performance PIs and their design guidelines.

To further explore more valuable and meaningful guidance for low dielectric constants, we further extracted 15 high-performance dianhydride, diamine, or diisocyanate monomers each from the screening space predicted by the COMFO model (it is important to point out that all of these monomers were obtained after screening in order of low dielectric constant PI. Their detailed visualization and analysis were placed in Note S4).

When applied to 5G and electronic appliance fields, the PIs were required to have Tg $>$ 300 °C, TM $>$ 2.5 GPa, and $\varepsilon < 3$. Therefore, here, we focused on these three criteria, and at the same time, we paid attention to the synthesis difficulty of the PIs and ensured that their CTEs were lower than 45 ppmK^{-1}, so that we finally screened out three PIs with excellent comprehensive performances (See Note S9 for detailed screening). The structure of these three types of PIs and the MD simulation results were shown in Fig. 7a and b. Table 3 and Fig. 7e show the values and deviations of the ML predictions and MD simulations of these

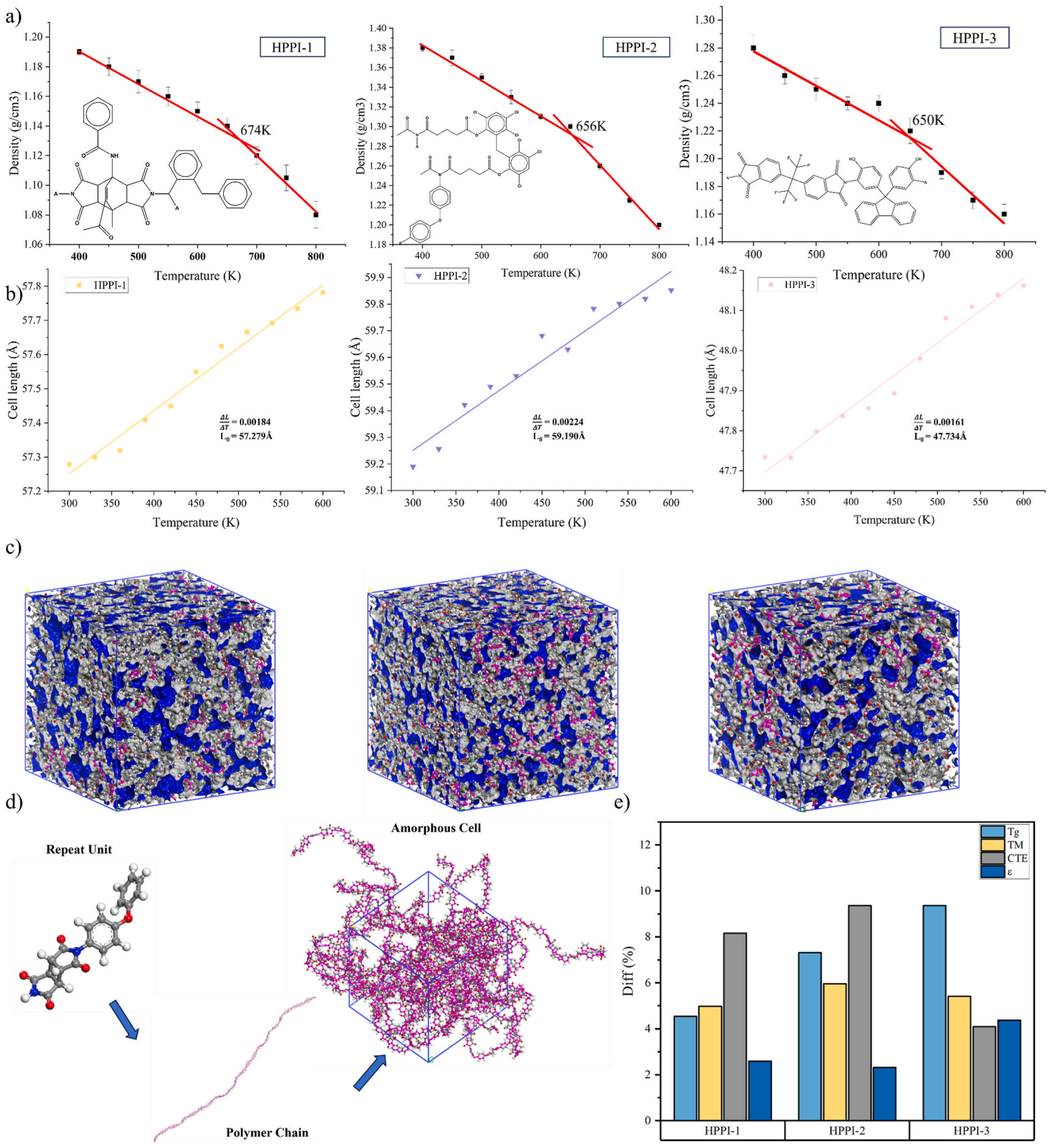


Fig. 7. a) Tg for the three high-performance PIs through MD simulation; b) CTE for the three high-performance PIs through MD simulation; c) FFV for the three types of high-performance PIs; d) Schematic diagram of the MD modeling process (PI-2); and e) Errors of MD simulated values versus the predicted values of ML.

three PIs. The results showed that the values of MD simulations were very close to the values predicted by ML, and their errors were all within acceptable limits. According to Table 3, all three types of potential PIs possess excellent thermo-mechanical properties, which makes them more competent to fulfill the performance needs of the materials during initial processing and utilization. The dielectric constants of HPPI-1, HPPI-2, and HPPI-3 were 2.26, 2.33, and 2.41, respectively, which proved the excellent comprehensive performance of the discovered PIs. The low dielectric constants could reduce the energy loss and capacitive effects during 5G signal transmission and electronic packaging [35]. To further verify the reliability of the model, we designed experiments to synthesize HPPI-3, and tabulated and compared the experimental test results with those of ML prediction and MD simulation (HPPI-3 synthesis method, test conditions, experimental test plots, and the comparison table were placed in the Note S5). And we also synthesized two other PIs for simultaneous validation (Note S8). The results demonstrated that the

error of our ML prediction does not exceed 5 %, thus, overall, the COMFO model is reliable and valuable. Fig. 8 shows the repeat unit structures of the three potential PIs after DFT geometry optimization. It could be found that the molecular chains of the 3 types of PIs contained non-coplanar structures, which was consistent with our previous conclusion of low dielectric constant found during the monomer structure analysis.

To further analyze the dielectric mechanism of the three types of PIs, the free volume fraction (FFV) and cohesive energy density (CED) of the three types of PIs were calculated with the MD method, and the electrostatic potential (ESP) and HOMO-LUMO orbitals of their repeat units were calculated with the DFT method. Fig. 8a exhibited the electrostatic potential maps of the three types of potential PIs, where the blue and green colors were the negative potential regions. It could be observed that the non-coplanar structures in all three types of PIs were in the negative potential region, which affected the electron distribution around them, leading to a decrease in the potential of some regions around them. In conjunction with Fig. 8b, due to the dianhydride fragments containing abundant non-coplanar structures, both HOMO and LUMO of HPPI-1 and HPPI-2 were concentrated in the dianhydride and exhibited larger energy gaps (Eg), especially for HPPI-1. The large Eg could help to reduce the charge-transfer-complex (CTC) effect on the molecular chain, which in turn lowered its dipole moment [36]. This could lead to a decrease in the dielectric constant. However, purely electronic structure analysis was not decisive and should be discussed in conjunction with other perspectives. Fig. 7c and Table S8 showed the FFV and CED of the three potential PIs and PI-1. The results showed that the three potential PIs had significantly larger FFV and lower CED compared to PI-1. The cohesion energy density (CED) could be used to reflect the magnitude of the intermolecular forces. HPPI-1 contained non-coplanar side groups, which resulted in weaker intermolecular forces and in turn led to a larger FFV, whereas HPPI-2 contained an abundance of ether bonds, which increased the flexibility of the molecular chain, leading to a tendency for the molecular chain to torsion, and in turn to a large FFV. It was generally believed that a large FFV reduced the polarization per unit volume. Furthermore, a large FFV would reduce the packing density of the molecular chains, making it easier for cavities or air to form between the molecular chains. Both the reduction of polarization and the formation of cavities or air could lead to a lower dielectric constant [37]. In the future, data on dielectric loss properties, including dielectric loss factor and dielectric loss tangent (tanδ), should be collected using more research tools for building

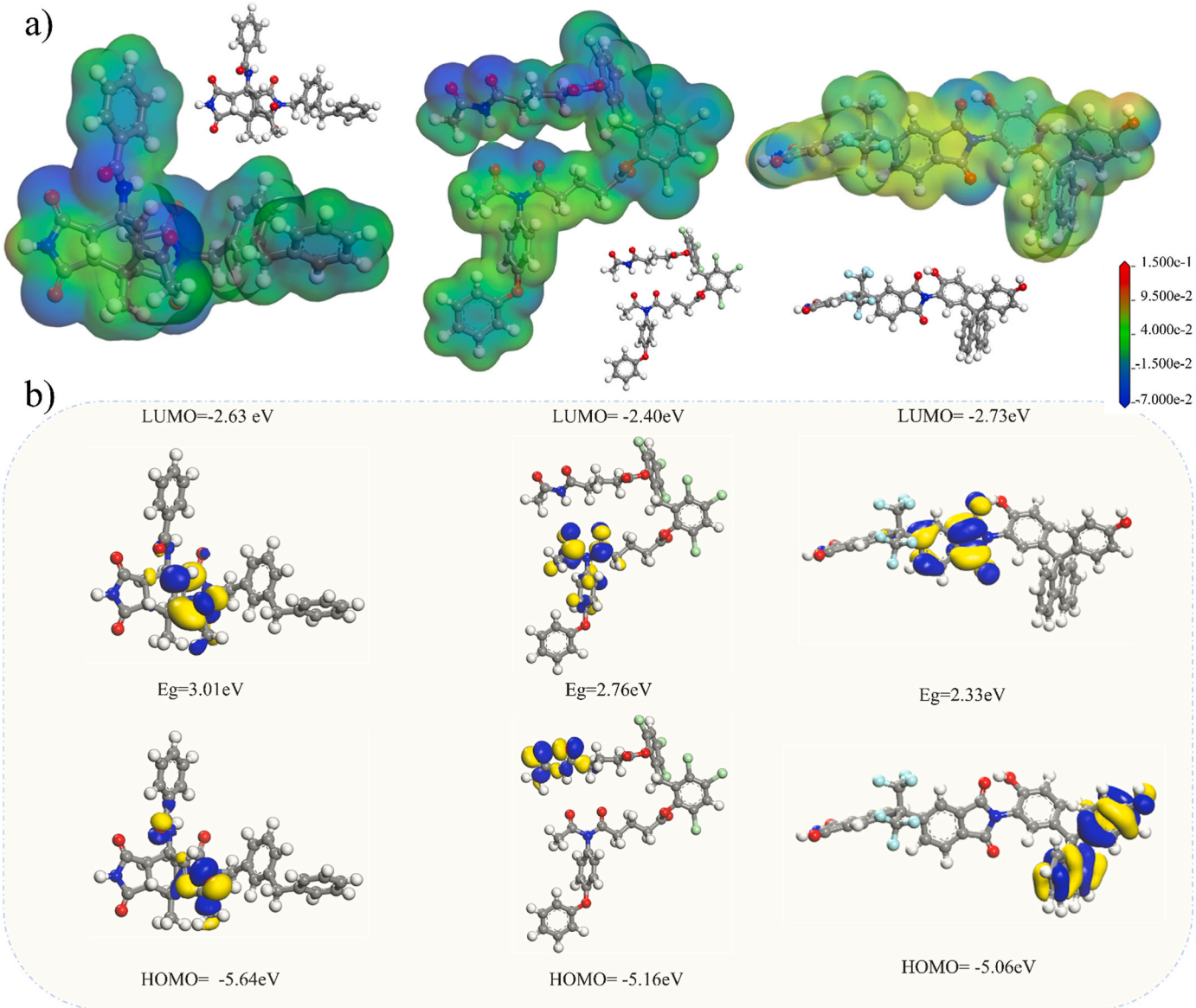


Fig. 8. a) Electrostatic potential maps of three high-performance PIs; b) HOMO/LUMO orbits of three high-performance PIs.

machine learning models. Meanwhile, it will be very meaningful to combine the research with the dielectric constant property and cooperate with the interpretable analysis of ML to discover more guidance about the dielectric properties of materials.

This deep learning model reduced the lengthy time loss in monomer selection and structure design for high performance materials, and screened three PIs with excellent comprehensive performance for 5G and electronic appliance fields. The polymerization reaction routes of these three types of PIs have been determined and these three PIs were synthesizable according to the Synthesis Accessibility score (SAscore). It was undeniable that the synthesis of polymeric materials was strongly influenced by the choice of their solvents and this difficulty should not be underestimated. This was not the focus of this study. We hope that in the future the three potential PIs could be actually synthesized by researchers in need. Therefore, in addition to the validation of MD simulations prior to the experimental study, we also screened suitable solvents and non-solvents for these three potential comprehensive high-performance materials (see NoteS6-S7 for the monomer structure, SMILES, ease of synthesis (SAscore) and solvents of the three potential PIs). The prediction and analysis of the performance of aggregated structures is a very critical issue, and in the future, we should develop more reliable molecular descriptors or other feature representations to accurately represent this critical information, or, through collecting enough information about the aggregated structures and their influence on the performance, we could ultimately build a more comprehensive and reliable MI model to guide the design of the materials further.

4. Conclusion

In order to solve the material design challenges, four types of properties of polyimide, including glass transition temperature, tensile modulus, linear coefficient of thermal expansion and dielectric constant, were collected in this study, and a deep learning model-COMFO was built for discovering multifunctional polyimide dielectric materials. Three learning tasks for feature extraction were built, including extracting feature information in SMILES sequences using a large language model-BET, extracting the information about the atoms and bonds of molecules from molecular graph using the Attentive FP network, and extracting information about substructures in polymer molecules through molecular fingerprints. The COMFO model could integrate, process and learn the key feature information extracted from the three types of learning tasks. The multi-perspective feature extraction task gave our COMFO model superior performance. The performance of the model was confirmed using multiple approaches, including experimental validation, MD simulation validation, and comparison with 12 other models. A screening space containing 158022 PIs was constructed based on real synthesis monomers and synthesis methods of PI. From the monomer structure analysis of high-performance PIs, key substructures affecting the dielectric constant, such as fluorine and aliphatic groups, were identified. Ultimately, three potential polyimide dielectric materials with excellent comprehensive performance were identified based on high-throughput screening. MD simulation further confirmed that these potential PIs have excellent thermal-mechanical properties and lower dielectric constants. Finally, DFT calculations and MD simulations were used to analyze the dielectric mechanism of these three types of PIs. The combination of efficient monomer structure analysis, deep learning prediction, and MD simulation validation will open up new avenues for the design and synthesis of high-performance materials, and these studies could also contribute to the forward development of materials in aerospace, microelectronics, and autonomous driving. In the future, we hope that researchers could consider adopting AI to learn the eight synthesis laws of compounds, and then to reverse the prediction of the synthesis of specific structures, including synthesized monomers, reagents and processing parameters, which would be our way forward. Alternatively, in the short term, we hope that researchers could explore more reasonable and accurate ways to collect and represent polymer processing parameters to further improve QSPR work.

CRediT authorship contribution statement

Bo Zhang: Writing – original draft, Validation, Methodology, Investigation, Data curation. **Xueqing Li:** Visualization, Validation, Project administration, Investigation. **Ming Zeng:** Writing – review & editing, Visualization, Supervision, Project administration, Conceptualization. **Jingguo Cao:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.polymer.2025.128081.

Data availability

The DOIs of the data source literature used for machine learning model training, testing, and validation in this paper are included in the Supporting Information. Hyperparameters for machine learning model training are also detailed in the Supporting Information. The SIMLES-BET based feature extraction task has been built as a pre-training model, which could greatly reduce the running time of the model when used.The computation of IMF fingerprints was fast and could be completed in about 10min. Based on our current hardware specifications (i9-12900H 2.50 GHz, 16 GB DDR5) and NVIDIA RTX 3060 graphics card, the training time for the whole set of deep learning models was about 50min, and the processing of the 158022 PIs and the prediction of all the properties took about 5.5h. The related code has been put on Github (https://github.com/ZBPHD10248/Deep-Learning-COMFO.git). The raw data of the model training and testing set and 15022 PIs in the screening space have also been put on Github.

References

[1] X. Chen, J. Sun, L. Fang, Y. Tao, X. Chen, J. Zhou, Q. Fang, Cross-linkable fluorinated polynorbornene with high thermostability and low dielectric constant at high frequency, ACS Appl. Polym. Mater. 2 (2) (2020) 768–774.
[2] J. Hou, L. Fang, G. Huang, M. Dai, F. Liu, C. Wang, M. Li, H. Zhang, J. Sun, Q. Fang, Low-dielectric polymers derived from biomass, ACS Appl. Polym. Mater. 3 (6) (2021) 2835–2848.
[3] D. Zhang, W. Liu, R. Guo, K. Zhou, H. Luo, High discharge energy density at low electric field using an aligned titanium dioxide/lead zirconate titanate nanowire array, Adv. Sci. 5 (2) (2018).
[4] Y. Zhou, H. Luo, S. Chen, X. Han, D. Zhang, Optimising the dielectric property of carbon nanotubes/P(VDF-CTFE) nanocomposites by tailoring the shell thickness of liquid crystalline polymer modified layerInspec keywordsOther keywords, Iet Nanodielectrics 2 (4) (2019) 142–150.
[5] X.P. Yuan, Y. Matsuyama, T.C.M. Chung, Synthesis of functionalized isotactic polypropylene dielectrics for electric energy storage applications, Macromolecules 43 (9) (2010) 4011–4015.
[6] L.S. Zhang, D.Y. Chen, B.H. Jin, B.W. Zhang, P.J. Gong, B. Zhang, C.B. Park, G.X. Li, Ultrahigh electromagnetic wave transmitting polyphenylene sulfide microcellular foams based on molecular structure design for 5G communication, Ind. Eng. Chem. Res. 62 (14) (2023) 5850–5863.
[7] M. Chawla, N. Shekhawat, S. Aggarwal, A. Sharma, K.G.M. Nair, Cole-cole analysis and electrical conduction mechanism of N^+ implanted polycarbonate, J. Appl. Phys. 115 (18) (2014).
[8] Y.F. Wei, Q.Y. Jiang, J.M. Hao, J.X. Mu, Comparative synthesis and properties of POSS-based fluorinated poly(ether sulfone) random terpolymers, RSC Adv. 7 (7) (2017) 3914–3920.
[9] X. Yan, F. Dai, Z. Ke, K. Yan, C. Chen, G. Qian, H. Li, Synthesis of colorless polyimides with high Tg from asymmetric twisted benzimidazole diamines, Eur. Polym. J. 164 (2022).

[10] P. Xiao, X.J. He, F. Zheng, Q.H. Lu, Super-heat resistant, transparent and low dielectric polyimides based on spirocyclic bisbenzoxazole diamines with $T_g > 450$ °C, Polym. Chem. 13 (24) (2022) 3660–3669.
[11] M. Lian, F. Zheng, X.M. Lu, Q.H. Lu, Tuning the heat resistance properties of polyimides by intermolecular interaction strengthening for flexible substrate application, Polymer 173 (2019) 205–214.
[12] Y.S. Shi, Z.G. Wang, Y.P. Shi, S.W. Zhu, Y.T. Zhang, J. Jin, Synergistic design of enhanced $^{\pi-\pi}$ interaction and decarboxylation cross-linking of polyimide membranes for natural gas separation, Macromolecules 55 (7) (2022) 2970–2982.
[13] J. Yang, L.Z. Tong, A.S. Alsubaie, K.H. Mahmoud, Y.Y. Guo, L. Liu, L. Guo, Z. H. Sun, C. Wang, Hybrid proton exchange membrane used in fuel cell with amino-functionalized metal-organic framework in sulfonated polyimide to construct efficient ion transport channel, Adv. Compos. Hybrid Mater. 5 (2) (2022) 834–842.
[14] Z. Xu, Z.L. Croft, D. Guo, K. Cao, G.L. Liu, Recent development of polyimides: synthesis, processing, and application in gas separation, J. Polym. Sci. 59 (11) (2021) 943–962.
[15] H. Yeganeh, B. Tamami, I. Ghazi, A novel direct method for preparation of aromatic polyimides via microwave-assisted polycondensation of aromatic dianhydrides and diisocyanates, Eur. Polym. J. 40 (9) (2004) 2059–2064.
[16] M. Wang, J. Jiang, Accelerating discovery of polyimides with intrinsic microporosity for membrane-based gas separation: synergizing physics-informed performance metrics and active learning, Adv. Funct. Mater. 23 (2024).
[17] W.-F. Peng, H.-Y. Lei, X.-X. Zhang, L.-H. Qiu, M.-J. Huang, Fluorine substitution effect on the material properties in transparent aromatic polyimides, Chin. J. Polym. Sci. 40 (7) (2022) 781–788.
[18] M. Zhao, C.L. Zhang, Y.X. Weng, Improved artificial neural networks (ANNs) for predicting the gas separation performance of polyimides, J. Membr. Sci. 681 (2023).
[19] L. Tao, J. He, N.E. Munyaneza, V. Varshney, W. Chen, G. Liu, Y. Li, Discovery of multi-functional polyimides through high-throughput screening using explainable machine learning, Chem. Eng. J. 465 (2023).
[20] S.V. Lyulin, A.A. Gurtovenko, S.V. Larin, V.M. Nazarychev, A.V. Lyulin, Microsecond atomic-scale molecular dynamics simulations of polyimides, Macromolecules 46 (15) (2013) 6357–6363.
[21] M. Mohammadi, H. Fazli, M. Karevan, J. Davoodi, The glass transition temperature of PMMA: a molecular dynamics study and comparison of various determination methods, Eur. Polym. J. 91 (2017) 121–133.
[22] L. Tao, G. Chen, Y. Li, Machine learning discovery of high-temperature polymers, Patterns 2 (4) (2021).
[23] M.X. Yu, Y.J. Shi, X. Liu, Q.Z. Jia, Q. Wang, Z.H. Luo, F.Y. Yan, Y.N. Zhou, Quantitative structure-property relationship (QSPR) framework assists in rapid mining of highly Thermostable polyimides, Chem. Eng. J. 465 (2023).
[24] M. Ramprasad, C. Kim, A. Jha, Assessing and improving machine learning model predictions of polymer glass transition temperatures, Abstr. Pap. Am. Chem. Soc. 257 (2019).
[25] A. Jha, A. Chandrasekaran, C. Kim, R. Ramprasad, Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures, Model. Simulat. Mater. Sci. Eng. 27 (2) (2019).
[26] Q. Yuan, M. Longo, A.W. Thornton, N.B. McKeown, B. Comesaña-Gándara, J. C. Jansen, K.E. Jelfs, Imputation of missing gas permeability data for polymer membranes using machine learning, J. Membr. Sci. 627 (2021).
[27] D. Chen, J.X. Zheng, G.W. Wei, F. Pan, Extracting predictive representations from hundreds of millions of molecules, J. Phys. Chem. Lett. 12 (44) (2021) 10793–10801.
[28] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, M. Zheng, Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism, J. Med. Chem. 63 (16) (2020) 8749–8760.
[29] J. Yang, L. Tao, J.L. He, J.R. McCutcheon, Y. Li, Machine learning enables interpretable discovery of innovative polymers for gas separation membranes, Sci. Adv. 8 (29) (2022).
[30] P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, J. Cheminf. 1 (2009).
[31] T. Matsumoto, S. Kawabata, R. Takahashi, Alicyclic Polyimides based on bicyclo [2.2.1]heptane- 2,3,5,6-tetracarboxylic 2,3-5,6-dianhydrides, High Perform. Polym. 18 (5) (2006) 719–726.
[32] C.D. Simone, E. Vaccaro, D.A. Scola, The synthesis and characterization of highly fluorinated aromatic polyimides, J. Fluor. Chem. 224 (2019) 100–112.
[33] C. Qian, Z.-G. Fan, W.-W. Zheng, R.-X. Bei, T.-W. Zhu, S.-W. Liu, Z.-G. Chi, M. P. Aldred, X.-D. Chen, Y. Zhang, J.-R. Xu, A facile strategy for non-fluorinated intrinsic low-k and low-loss dielectric polymers: valid exploitation of secondary relaxation behaviors, Chin. J. Polym. Sci. 38 (3) (2020) 213–219.
[34] Y. Li, Y. Chu, R. Fang, S. Ding, Y. Wang, Y. Shen, A. Zheng, Synthesis and memory characteristics of polyimides containing noncoplanar aryl pendant groups, Polymer 53 (1) (2012) 229–240.
[35] C.-C. Kuo, Y.-C. Lin, Y.-C. Chen, P.-H. Wu, S. Ando, M. Ueda, W.-C. Chen, Correlating the molecular structure of polyimides with the dielectric constant and dissipation factor at a high frequency of 10 GHz, ACS Appl. Polym. Mater. 3 (1) (2021) 362–371.
[36] Z.Y. Peng, A. Ye, L. Zhang, X.L. Li, C. Lian, C.Z. Li, Micro-crosslinked polyimide nanocomposites with low dielectric constant and low dielectric loss for microwave antenna with molecular dynamics, Compos. Commun. 46 (2024).
[37] C. Qian, R.X. Bei, T.W. Zhu, W.W. Zheng, S.W. Liu, Z.G. Chi, M.P. Aldred, X. D. Chen, Y. Zhang, J.R. Xu, Facile strategy for intrinsic low-*k* dielectric polymers: molecular design based on secondary relaxation behavior, Macromolecules 52 (12) (2019) 4601–4609.