Chemical Engineering Journal 465 (2023) 142949

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Discovery of multi-functional polyimides through high-throughput screening using explainable machine learning

Lei Tao [a,1], Jinlong He [b,1], Nuwayo Eric Munyaneza [c,1], Vikas Varshney [d], Wei Chen [e], Guoliang Liu [c,f], Ying Li [b,*]

[a] Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, United States
[b] Department of Mechanical Engineering, University of Wisconsin-Madison, Madison, WI 53706, United States
[c] Department of Chemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, United States
[d] Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433, United States
[e] Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, United States
[f] Department of Chemical Engineering, Department of Materials Science and Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, United States

ARTICLE INFO

Keywords:
Multi-functional polyimides
Explainable machine learning
Molecular dynamics
Ultrahigh thermomechanical properties

ABSTRACT

Polyimides have been widely used in modern industries because of their excellent mechanical and thermal properties, e.g., high-temperature fuel cells, displays, and aerospace composites. However, it usually takes decades of experimental efforts to develop a successful product. Aiming to expedite the discovery of high-performance polyimides, we utilize computational methods of machine learning (ML) and molecular dynamics (MD) simulations. Our study provides compelling evidence for the effectiveness of a data-driven approach in discovering novel polyimides. We first build a comprehensive library of more than 8 million hypothetical polyimides based on the polycondensation of existing dianhydride and diamine/diisocyanate molecules. Then we establish multiple ML models for the thermal and mechanical properties of polyimides based on their experimentally reported values, including glass transition temperature, Young's modulus, and tensile yield strength. The obtained ML models demonstrate excellent predictive performance in identifying the key chemical substructures influencing the thermal and mechanical properties of polyimides. The use of explainable machine learning describes the effect of chemical substructures on individual properties, from which human experts can understand the cause of the ML model decision. Applying the well-trained ML models, we obtain property predictions of the 8 million hypothetical polyimides. Then, we screen the whole hypothetical dataset and identify three (3) best-performing novel polyimides that have better-combined properties than existing ones through Pareto frontier analysis. For an easy query of the discovered high-performing polyimides, we also create an online platform https://polyimide-explorer.herokuapp.com/ that embeds the developed ML model with interactive visualization. Furthermore, we validate the ML predictions through all-atom MD simulations and examine their synthesizability. The MD simulations are in good agreement with the ML predictions and the three novel polyimides are predicted to be easy to synthesize via Schuffenhauer's synthetic accessibility score. Following the proposed ML guidance, we successfully synthesized a novel polyimide and the experimentally obtained high glass transition/thermal decomposition temperature demonstrated its excellent thermal stability. Our study demonstrates an efficient way to expedite the discovery of novel polymers using ML prediction and MD validation. The high-throughput screening of a large computational dataset can serve as a general approach for new material discovery in other polymeric material exploration problems, such as organic photovoltaics, polymer membranes, and dielectrics.

* Corresponding author.
E-mail address: yli2562@wisc.edu (Y. Li).
[1] Equal contribution.

https://doi.org/10.1016/j.cej.2023.142949
Received 11 February 2023; Received in revised form 31 March 2023; Accepted 10 April 2023
Available online 13 April 2023

1. Introduction

Polyimides are high-performance engineering plastics that have excellent strength and stiffness, exceptional heat resistance, and chemical stability. Their attractive mechanical and thermal properties are widely utilized in applications for aerospace, automobile, and electronics industries [1–6]. For example, Kapton^{Tm} – a polyimide film product developed by DuPont Company in the 1950 s has been used till now as an excellent electrical insulating material. It can withstand temperatures of up to 400 °C and maintain excellent mechanical properties across a broad temperature range (-269–400 °C). Such outstanding thermomechanical properties are of great value in meeting multi-functional requirements for many other applications, e.g., high-temperature fuel cells, displays, and membrane separations. Several polymerization methods are developed to synthesize polyimides, such as cycloaddition [7], diesters of tetracarboxylic acids [8], nucleophilic substitution [9], etc. Commercially, the DuPont synthesis method utilizes the condensation reaction between a dianhydride and a diamine upon the elimination of water molecules [10]. An alternative method yielding identical products as the DuPont route utilizes the reaction between a dianhydride and a diisocyanate with the elimination of carbon dioxide [11–13] molecule. Different dianhydride, diamines, or diisocyanates are selected to tailor the properties of the final polyimide polymers. For instance, pyromellitic dianhydride (PMDA) + 4,4′-oxydianiline (ODA) produces Kapton^{Tm} with high thermal resistance and mechanical strength [14], diacid anhydride + m-phenylene diamine yields polyetherimide that has excellent toughness and rigidity [15], while cyclobutane tetracarboxylic dianhydride (CBDA) + ODA leads to a polyimide that is colorless with good electrical insulating properties [16]. Such experimental studies of dianhydride and diamine/diisocyanate reactions have resulted in several multi-functional polyimides that meet different requirements. Nevertheless, it is impossible for experimentalists to exhaust all possible two-component reactions between dianhydride and diamine/diisocyanate (that have not been experimentally synthesized) for materials design and discovery of novel polyimides. Designing experiments can be a challenging task that requires a high level of expertise, and it is often difficult to develop an optimal experimental design without extensive knowledge in the field [17]. Therefore, researchers have tried to use computational approaches to efficiently predict the properties of new polyimides.

Most current computational efforts have been devoted to the estimation of glass transition temperature T_g as polyimides are among the most heat-resistant polymers. Researchers have analyzed the T_g of polyimides in various ways like theoretical analysis, molecular dynamics (MD) simulations, and using machine learning (ML) techniques. For example, using theoretical analysis, Ronova et al. [18] developed a method to calculate the conformational parameters of 26 polyimides and studied the influence of chemical structure on T_g. They found the connecting bridge between two imide rings has varying impacts on the rigidity of the polyimide chain, and consequently the T_g. When the imide rings are connected by flexible bridges such as ether, carbonyl or silicon, the rigidity of the diamine component has a significant impact on the rigidity of the resulting polyimide and its T_g. This method is helpful to some extent in selecting appropriate diamines and dianhydrides for desired T_g, but it is only a qualitative derivation based on a few polyimides and cannot be applied to estimate the T_g for all other possible polyimides. Due to the lack of quantitative theory to correlate the chemical structure and T_g, researchers have turn to MD simulations to give reasonable judgment of T_g of polyimides. The MD simulated T_g is demonstrated to agree well with experimental values for different polyimides like Ultem^{Tm} and Extem^{TM} [19], (β – CN)APB/ODPA polyimide [20], R-BAPS polyimide [21], Isomeric polyimides [22,23], etc. In addition, MD allows a careful evaluation of the dependence of T_g on the structural factors like chain rigidity, intermolecular interactions, fractional free volume, etc. However, it is still unfeasible to simulate all possible polyimides with the time-consuming all-atom MD simulations. To determine the T_g of polyimides in a more effective way, ML techniques have been employed recently to build a predictive model of T_g. Wen et al. [24] collected T_g of 225 polyimides, obtained the simplified molecular-input line-entry system (SMILES) of their monomers, and generated 1342 molecular descriptors as feature inputs for the ML model. Their LASSO (least absolute shrinkage and selection operator) model together with bagging approach has an average error 18 K in T_g prediction, demonstrating a good prediction power of the ML model. However, their use of a small dataset brings concerns of the generalization ability of the obtained ML model, and a further improvement of the model using a larger dataset is greatly desired [25].

The current challenge for designing new polyimides lies in the fact that while polyimides are multi-functional materials that can be polymerized with two components —dianhydride and diamine/diisocyanate— there is no efficient way to evaluate the multi-functional properties of all possible polyimides synthesized from different two-component reactions. From the above discussion, all the previous studies encounter various issues when assessing the single property (T_g) of polyimides. In addition, there is very little work devoting effort to exploring other properties simultaneously. Wu et al. [26] estimated T_g and tensile modulus (E) for 6 polyimides through MD simulations, and Wang et al. [27] synthesized copolyimide for both high strength and low dielectric constant. When multiple properties are involved, the evaluation of polyimides takes even more effort, and it is understandable that the current multi-property studies only examine the simple case of two properties at the same time. Facing these difficulties and challenges, we realize that the discovery of novel polyimides with excellent multi-functional properties needs to be addressed through a new strategy.

To expedite the discovery of new polyimide with multiple tailored properties, we propose an integrated data-driven method that takes advantage of ML using a large dataset of real polyimides, a hypothetical dataset of more than 8 million possible polyimides for screening using the custom-built ML models, and MD simulations for validation. Our proposed ML-assisted approach consists of several key steps that are carefully designed to maximize the efficiency and effectiveness of the material discovery process. These steps include: (1) Historical data collection, which involves gathering and organizing a large dataset of polyimide structures and their associated properties; (2) ML model training, which uses this dataset to develop a predictive model that can identify new polyimide candidates with desirable properties; (3) Hypothetical structure creation, which involves generating new polyimide structures using the predictive model; (4) Promising candidate screening, which involves evaluating these new structures based on their predicted properties and selecting the most promising candidates for further analysis; (5) MD simulation validation, which uses high fidelity MD simulations to validate the stability and performance of the selected candidates; and (6) Experimental synthesis and measurements, which involves synthesizing the most promising candidates in the laboratory and measuring their properties to validate predictions of the ML model. Together, these steps form a comprehensive and effective workflow for discovering new polyimide materials with desirable properties.

We first collect 2233 real polyimides structures from PoLyInfo database [28], with their 7 reported physical properties if there is any available, such as density (ρ), glass transition temperature (T_g), melting temperature (T_m), decomposition temperature (T_d), Young's modulus (E), tensile yield strength (σ_y), and tensile break strength (σ_b). PolyInfo dataset contains more than 18,000 homopolymers and more than 494,837 properties. Its information is collected from public literature. These polyimides that have been synthesized and characterized experimentally constitute a large dataset of real polyimides. Based on this large dataset, we train ML models to establish the structure–property relationship for each of the seven properties and obtain physical insight into the key structural features of different properties. Next, the well-trained ML models will work as the predictive tool to estimate

multiple properties of new polyimides. We mainly focus on the σ_y, E, and T_g properties as high-temperature mechanical properties are most desired for industries like aerospace, automobile, and electronics. In accordance with the two polymerization routes of "dianhydride + diamine" or "dianhydride + diisocyanate" for polyimide synthesis, we collect all the existing dianhydride, diamine, and diisocyanate molecules from PubChem [29] database, and carry out the polycondensation computationally to establish a large dataset of 8 million hypothetical polyimides (excluding 2233 existing polyimides). Finally, we identify several multi-functional polyimides that outperform the current real polyimides, accompanied by validating their properties through all-atom MD simulations. Our study designs novel multi-functional polyimides by significantly expanding the chemical space of polyimides and then narrowing them down to promising candidates through ML screening and MD validation. Exhausting all possibilities using physics-based ML prediction before experimental synthesis successfully allows the exploration of the whole design space. Our proposed methodologies, with their carefully designed and integrated processes, represent a significant step forward in the field of polymer informatics. By providing a standardized workflow for implementing ML models in the discovery of novel polyimides, we aim to accelerate the pace of materials discovery and enhance the efficiency of research efforts. Moreover, our study demonstrates the critical importance of comprehensive data-driven analysis in the successful discovery of novel materials. By leveraging historical data and applying advanced ML techniques, we were able to identify promising candidates for further development and achieve significant improvements in the efficiency and accuracy of the discovery process. We believe our findings will have broad implications for the field of materials science and inspire further research in the development of advanced data-driven techniques for materials discovery. Such a design strategy is much more efficient compared to the conventional trial-and-error process and can be applied to the molecular design of other polymeric materials [30].

2. Results and discussion

2.1. Expanding chemical space of polyimides

Polyimides are mainly formed by the polycondensation either of a dianhydride and a diamine with the removal of water molecules (Fig. 1**a**), or of a dianhydride and a diisocyanate with the release of carbon dioxide molecules (Fig. 1**b**). The final product contains two functional imide groups that are the signature of polyimides (highlighted in the red shade in Fig. 1**a-d**). Fig. 1**c** illustrates some examples of the 2233 real polyimides collected from the PoLyInfo [28] – a database of experimentally reported polymers. When locating these 2233 real polyimides in a 2D chemical space plot, we find most of them are close to each other based on their structural similarity (green points in Fig. 1**e**). It indicates the diversity of the real polyimides is still limited, far from covering all possible chemical structures of polyimides.

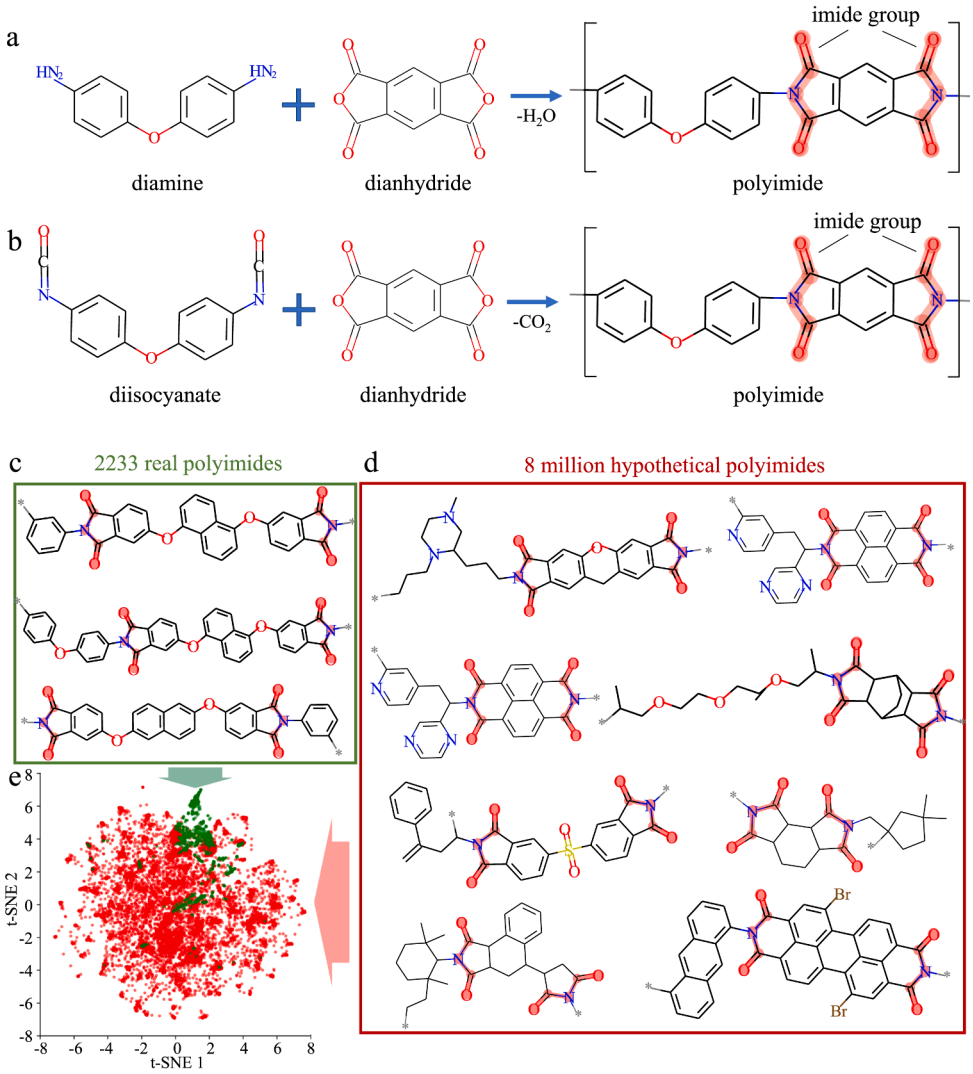


Fig. 1. Comparison of real polyimides and hypothetical polyimides. (a) Synthesis of polyimide from a diamine and a dianhydride. (b) Synthesis of polyimide from a diisocyanate and a dianhydride. (c) Example structures from 2233 real polyimides in PoLyInfo. (d) Example structures from 8 million hypothetical polyimides based on reactions shown in subplots a and b. (e) Chemical space visualization of the real polyimide dataset (green points) and the hypothetical polyimide dataset (red points). The 2D visualization is based on fingerprints using t-SNE algorithm. The comparison indicates that hypothetical polyimides (red points) cover a much broader chemical space while the real polyimide (green points) primarily concentrate in a local region. It suggests that the hypothetical polyimides are much more diverse in chemical structures, as shown in subplot d. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The synthesis route illustrated in Fig. 1**a-b** has led to the synthesis of successful polyimides like Kapton^{Tm}, UPILEX^{Tm}, Avimid^{Tm}, etc., using different combinations of two components: dianhydride and diamine/diisocyanate. Although there are plenty of options for dianhydride, diamine, and diisocyanate, it is unfeasible to synthesize all possible combinations experimentally. For example, in the PubChem [29] database - the world's most extensive collection of existing organic molecules containing more than 299 million substances and various properties such as toxicity, stereochemistry, topological polar surface area, etc. The numbers of dianhydride, diamine, and diisocyanate are more than 200, 30000, and 10000, respectively. Thus, the combination of dianhydride and diamine will lead to more than 6 million possible polyimides, while dianhydride and diisocyanate will lead to more than 2 million possible polyimides. Although it is impossible to realize these 8 million polyimides experimentally, from a computational point of view, the chemical structures of these 8 million polyimides can be obtained when the imide functional group is formed correctly. Fig. 1**d** illustrates several chemical structures of the obtained hypothetical polyimides. This hypothetical dataset incorporates more diverse structures that occupy a broad area in the chemical space (red points in Fig. 1**e**). It is beneficial for us to expand the chemical space from a local region of real polyimides to a broader area of hypothetical polyimides, possibly leading to a greater chance to discover novel polyimides with desired multi-functional properties. An exhausting search of the 8 million hypothetical polyimides is unrealistic to be achieved experimentally but can only successfully be done using fast-to-compute ML models.

To evaluate these hypothetical polyimides computationally, the first and the most crucial step is to establish a reliable structure–property relationship from the experimentally studied polyimides. Along with the reported chemical structures, most of the 2233 real polyimides have more than one property measured like thermal, electronic, mechanical, or dielectric properties. Table 1 summarizes the 7 properties we collected, including density (ρ), glass transition temperature (T_g), melting temperature (T_m), thermal decomposition temperature (T_d), Young's modulus (E), tensile yield strength (σ_y), and tensile break strength (σ_b). Experimental values come with uncertainties in terms of different samples, different instruments, different measurement methods, etc. Taking the average or median value as the representative value of a property is a typical approach, and using the median of a distribution of property values is found the best way to address uncertainty in data sources. Data coming from unstandardized tests remains a challenge in the field. To make the best use of the scarce experimental data, we examine their test methods and include all reasonable data points, and take the median value if needed.

The establishment of the structure–property relationship for each property is a well-defined task that can be addressed with ML techniques [30]. The input is the chemical structure of real polyimides (with the improved Morgan fingerprints as feature representations in Supporting Information S1), while the output is the property. It would be ideal if a single ML model deals with 7 properties simultaneously, namely a multi-task ML model that predicts 7 properties given a polyimide structure. Unfortunately, the problem of task compatibility often leads to inferior overall performance when different tasks compete in multi-task learning [31]. Fig. 2 plots the 7 properties' pairwise relationships for the real polyimide dataset. The non-diagonal plots display the pairwise relationships between any two different properties. We hardly notice correlations between mechanical properties and thermal properties such as σ_y vs. T_g, although some correlations are obvious between two thermal properties such as T_m vs. T_g. Overall, a multi-task learning of 7 properties doesn't show an ideal model performance (see Supporting Information S2 for the performance of the multi-task learning), associated with architecture rigidity to predict multiple properties which aren't well-correlated.

Table 1
7 physical properties collected for 2233 real polyimides from the PoLyInfo.

	ρ	T_g	T_m	T_d	E	σ_y	σ_b
Number of available data points	337	1870	249	1663	447	110	480
Property range	0.81 ~ 1.54	−9 ~ 490	−22 ~ 517	60 ~ 640	2.12e-3 ~ 18.6	5.03e-5 ~ 0.25	8.83e-5 ~ 0.41
Property unit	g/cm^3	°C	°C	°C	GPa	GPa	GPa

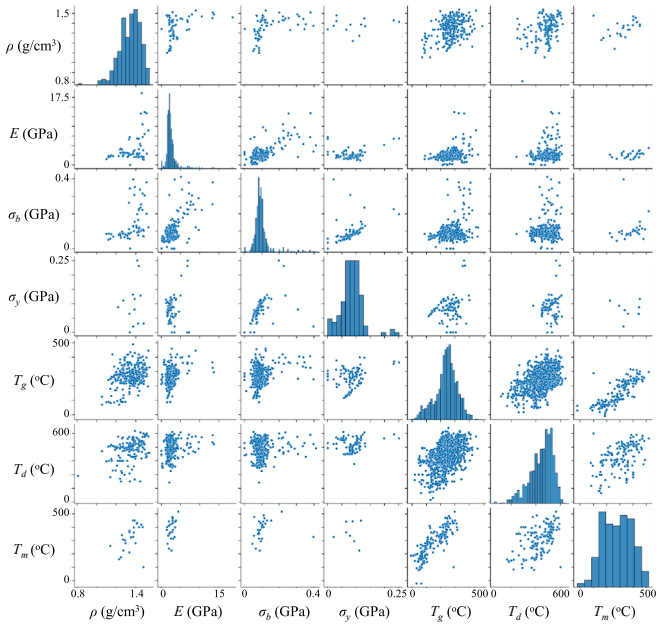


Fig. 2. Pairwise relationships in the real polyimide dataset. The diagonal plots are univariate distribution plots for each property. The non-diagonal plots display the pairwise relationships between any two different properties. It is hard to notice correlations between mechanical properties and thermal properties such as σ_y vs. T_g, although some correlations are obvious between two thermal properties, such as T_m vs. T_g.

2.2. Explainable ML models and physical insights based on real polyimides

ML algorithms applicable for polymer's structure–property relationships include feed-forward neural networks (FFNN), recurrent neural networks (RNN), graph convolutional neural networks (CNN), Gaussian process regression (GPR), random forests (RF), etc. We have tested different ML algorithms on the single task T_g of polymers, and found the FFNN using improved Morgan fingerprints was among the best models [25,32,33]. Morgan fingerprint method detects substructures enclosed in a circle of radius R, and assigns each detected substructure a numerical identifier. We use the SMILES of the repeat unit for each polyimide and implement the fingerprint algorithm in RDKit with R equals 3. A large number of substructures is detected, but we only keep the 121 prominent substructures shared by most polyimides. Finally, for each polyimide, we obtain a vector of size 121 in which each bit represents the number of a detected substructure. Compared to the default Morgan fingerprints, our improved Morgan fingerprints consider the frequency of occurrence for each substructure, carrying more physical meaning. This input vector is found to be a proper representation as it indicates both what substructures and how many of them exist in the polymer's repeating unit [32]. In the default Morgan feature representation 1/0 is used to indicate the existence of a certain substructure and different substructures can be hashed into the same vector bit. This so-call collision problem can lead to more molecules being represented in the same way. There are 9% molecules being duplicately

represented through the default Morgan fingerprint. But after using our improved Morgan fingerprint, there are only 5% molecules being duplicately represented. As 5% is only a small portion of the whole dataset, it has little effect on the final model training. With improved Morgan fingerprints, the collision problem can be addressed. However, labeling the number of occurrences for substructures still doesn't encode the microscale level feature of polymers such as average chain length or molecular weight. Considering the limitation of the experimental dataset on the MW information, the effect of molecular weight is not explicitly represented in the fingerprints.

After examining the good performance of FFNN with the improved Morgan fingerprints on each of the focused properties, we build 7 single-task ML models for each property individually. The architecture of each ML model, such as the number of layers and number of neurons in each layer, is optimized through hyperparameter tuning. Based on the training set, we utilize 5-fold cross-validation to fine-tune the hyperparameters and select the optimal model. Subsequently, we train the optimized model on the entire training set and assess its performance on the test set. It is noteworthy that we conducted dozens of trials for hyperparameter tuning, and in the final ensemble model, we trained three models simultaneously. To obtain a better prediction performance, we build an ensemble model that averages a few models to get the final prediction. For example, the single-task ML model for T_g is optimized to have 4 hidden layers, and each layer has 34, 16, 8 neurons, respectively. We train this model architecture three times on the same dataset then average the three model predictions to a single-task ensemble model for T_g. The same strategy is applied to other property tasks (see Supporting Information S2 for the architecture of each single-task ensemble model).

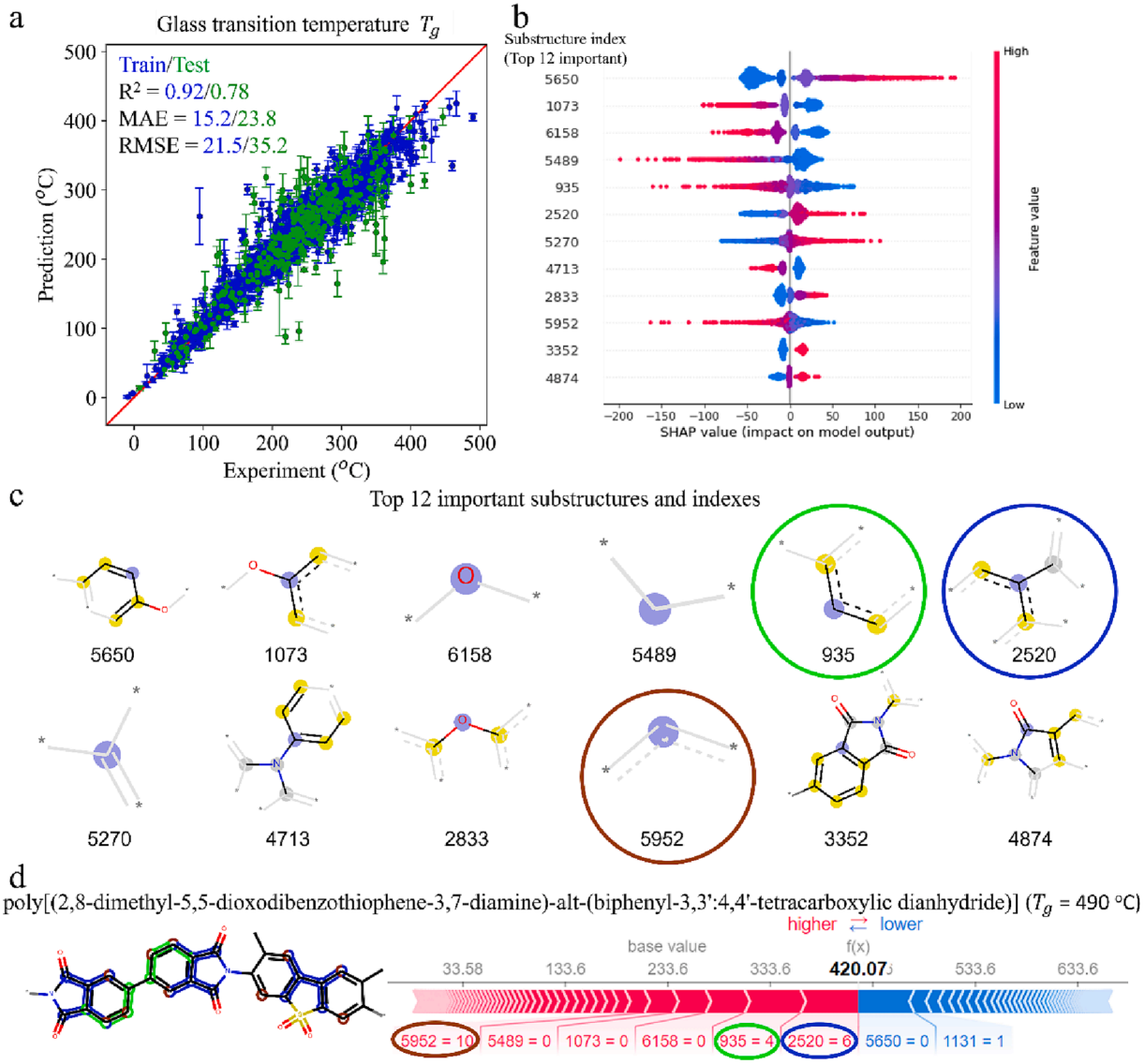


Fig. 3. Performance and feature importance of the single-task ensemble model for T_g. (a) The parity plot of the single-task ensemble FFNN model using improved Morgan fingerprints as input features. The ensemble average predictions are illustrated in dots, while the variance from the three models in the ensemble is illustrated with error bars. (b) Substructure importance plot. It lists the most important substructure in descending order and each dot represents the impact from a particular sample in the training set. (c) The most important 12 substructures associated with T_g according to SHAP values. The central atom of the substructures is highlighted in blue. Aromatic atoms are highlighted in yellow. Atoms' connectivity is highlighted in light gray. (d) The individual SHAP value plot for the highest T_g polyimide poly[(2,8-dimethyl-5,5-dioxodibenzothiophene-3,7-diamine)-alt-(biphenyl-3,3′:4,4′-tetracarboxylic dianhydride)]. Its T_g prediction f(x) is based on the mean T_g of all real polyimides (base value) and contributions of its substructures. Red or blue arrows indicate positive or negative contribution of each substructure. The feature value of a substructure can be "0" meaning the absence of the substructure in the molecule, but its feature importance is still a valid value indicated by the length of the arrow. Top substructures in this polyimide are highlighted in different colors on the left. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

All our models are implemented using the Keras package [34]. Among the 7 properties, σ_y-E combination characterizes mechanical strength and modulus of resilience of a material, and T_g is a key representative of thermal property. This study aims at a multi-functional objective where we intend to discover novel polyimides that have high performance for σ_y, E, and T_g simultaneously.

Firstly for T_g, the model performance and feature importance are illustrated in Fig. 3. Among the 1870 real polyimides whose T_g is experimentally reported, with a random splitting, 90% of the data points are used as the training set, and the other 10% data points are held out as the test set. R^2 as well as the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) are employed to quantify the performance of the ensemble model. The error bar indicates the variance of the predictions from the three models contained in the ensemble, suggesting the difference in the prediction performance from the single model to the ensemble model. It shows that R^2 of 0.92 and 0.78 are obtained for the model training and validation, respectively (Fig. 3**a**). Such prediction performance is comparable to other T_g models trained on different polymer datasets [8]. When a predictive model is established, we prefer to get insights into how each substructure feature affects the final property. Therefore, we calculate SHapley Additive exPlanations (SHAP) values to evaluate the impact of substructures on the T_g (see Supporting Information S3 for the details of SHAP). The top substructures influencing the model's output are listed in Fig. 3**b-c**. Each row in Fig. 3**b** represents a substructure, and dots along the same row indicates the SHAP value of that substructure from different polyimides.

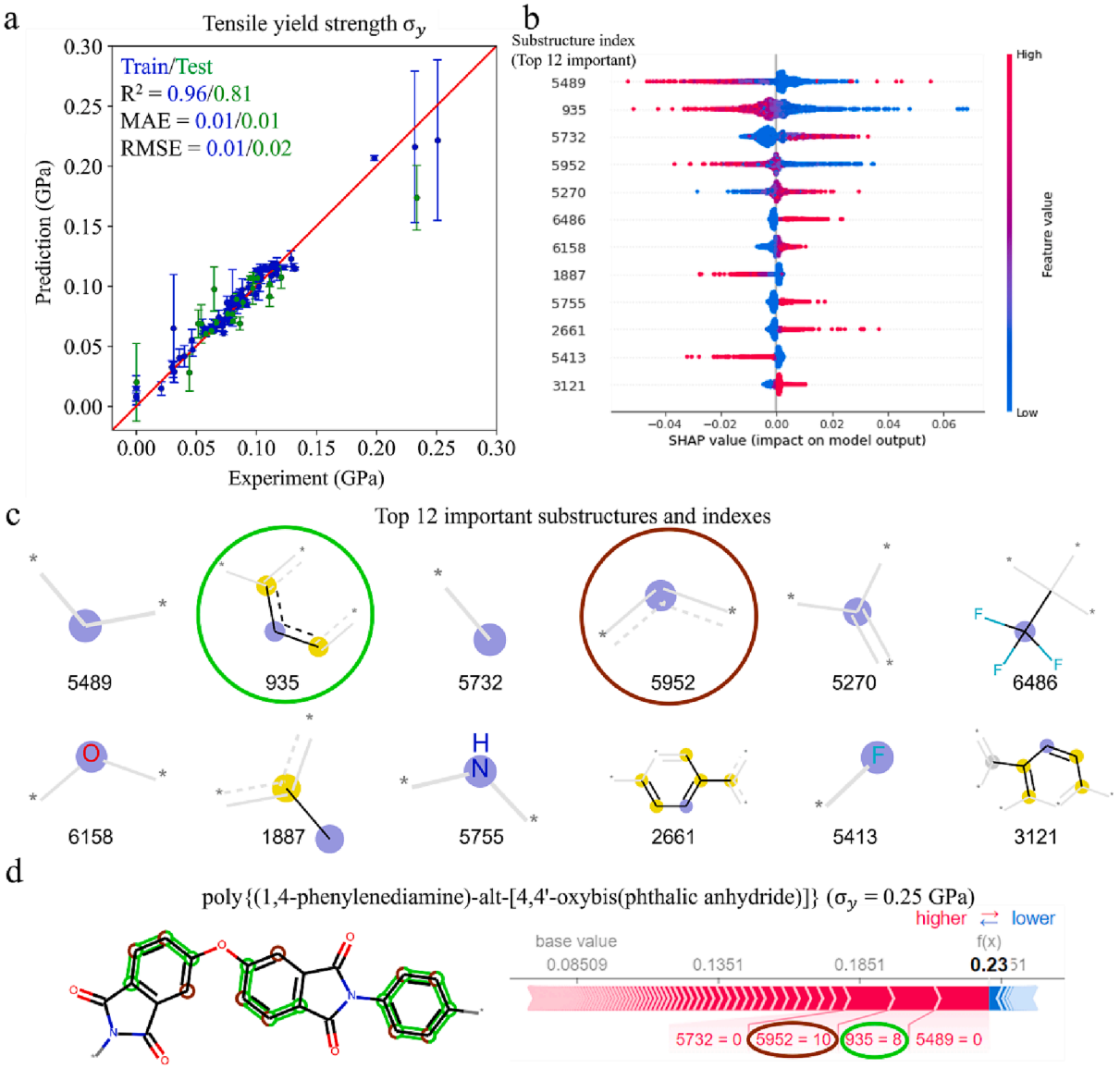


Fig. 4. Performance and feature importance of the single-task ensemble model for σ_y. (a) The parity plot of the single-task ensemble FFNN model using improved Morgan fingerprints as input features. The ensemble average predictions are illustrated in dots, while the variance from the three models in the ensemble is illustrated with error bars. (b) Substructure importance plot. It lists the most important substructures in descending order and each dot represents the impact of a particular sample in the training set. (c) The most important substructures associated with σ_y according to SHAP values. The central atom of the substructures is highlighted in blue. Aromatic atoms are highlighted in yellow. Atoms' connectivity is highlighted in light gray. (d) The individual SHAP value plot for the highest σ_y polyimide poly(1,4-phenylenediamine)-alt-[4,4′-oxybis(phthalic anhydride)]}. Its σ_y prediction f(x) is based on the mean σ_y of all real polyimides (base value) and contributions of its substructures. Red or blue arrows indicate positive or negative contribution of each substructure. The feature value of a substructure can be "0" meaning the absence of the substructure in the molecule, but its feature importance is still a valid value indicated by the length of the arrow. Top substructures in this polyimide are highlighted in different colors on the left. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

From the perspective of functional groups, high glass transition temperatures usually correlate to heteroaromatic units, rigid aromatic units, or inflexible linkages [10,32]. These features are reflected by the key substructures like numerical identifies "4874", "3352", and "2833", etc. The most important substructure in the dataset is "5650", which comprises aromatic rings and oxygen rings. The revealed high impact of aromatic rings and oxygen linkages on T_g is consistent with previous studies [32,35]. Examination of SHAP values of the highest T_g polyimide, poly[(2,8-dimethyl-5,5-dioxodibenzothiophene-3,7-diamine)-alt-(biphenyl-3,3':4,4'-tetracarboxylic dianhydride)], is highlighted in Fig. 3**d**. The feature values for important substructures are presented. Three substructures contributing most to the high T_g are highlighted in the molecular graph with colored circles, showing the impact of the revealed key substructures. Our developed ensemble model is not only with good predictive accuracy, but also with clear physical explanations. Therefore, we will apply the obtained model to make high-throughput screening of the 8 million hypothetical polyimides.

For σ_y, there are 110 real polyimides whose σ_y values are experimentally reported. Following the same training process as T_g we obtain R^2 of 0.94 and 0.85 for the model training and validation (Fig. 4a). It is shown that without ensembling, a single model would give a large error bar at some points, but after with ensembling, the ensemble average prediction match much better with the true value. It well demonstrates the advantage of the ensemble model in reducing the prediction

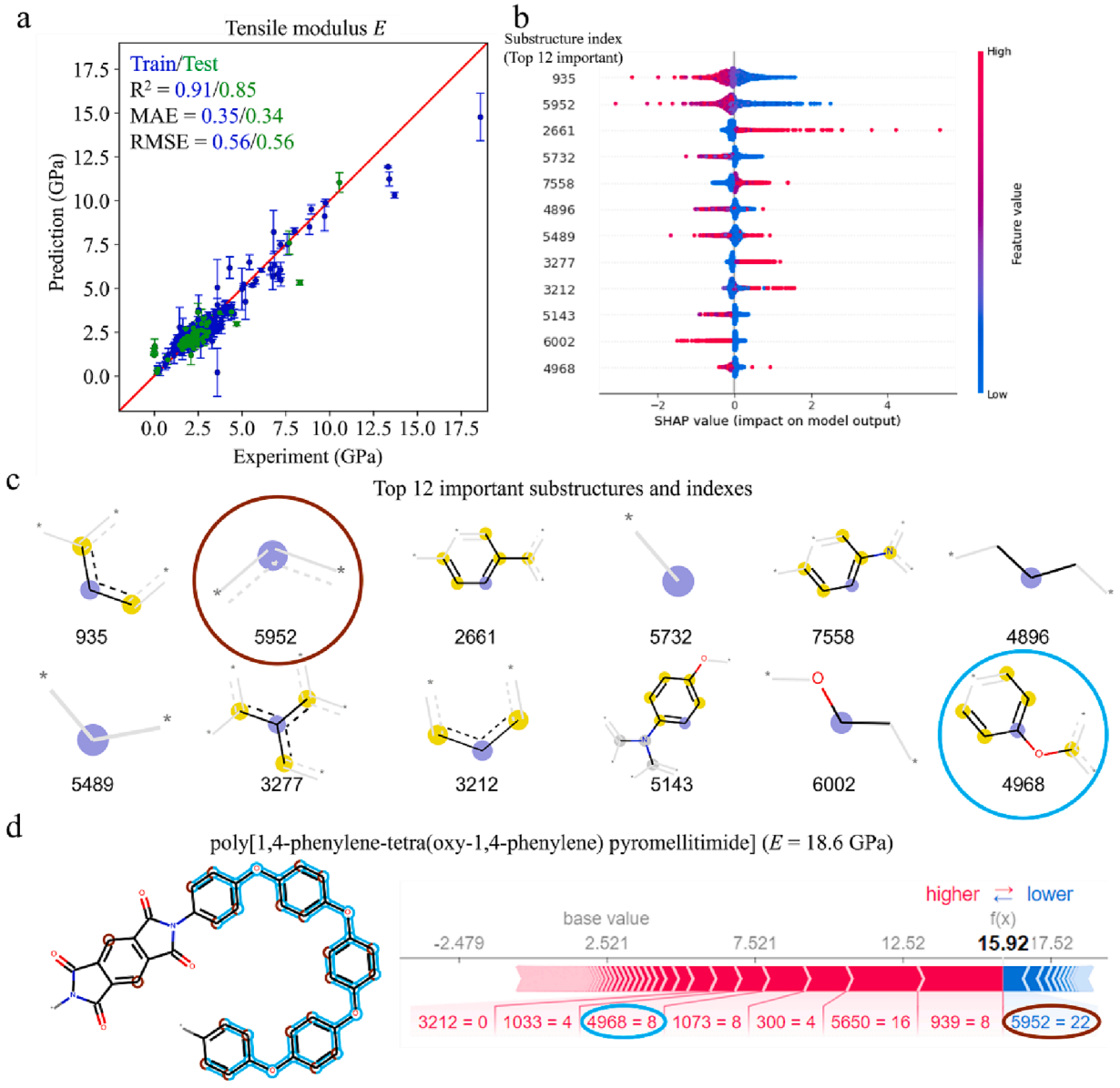


Fig. 5. Performance and feature importance of the single-task ensemble model for *E*. (a) The parity plot of the single-task ensemble FFNN model using improved Morgan fingerprints as input features. The ensemble average predictions are illustrated in dots, while the variance from the three models in the ensemble is illustrated with error bars. (b) Substructure importance plot. It lists the most important substructures in descending order and each dot represents the impact of a particular sample in the training set. (c) The most important substructures associated with *E* according to SHAP values. The central atom of the substructures is highlighted in blue. Aromatic atoms are highlighted in yellow. Atoms' connectivity is highlighted in light gray. (d) The individual SHAP value plot for the highest *E* polyimide poly[1,4-phenylene-tetra(oxy-1,4-phenylene) pyromellitimide]. Its *E* prediction f(x) is based on the mean *E* of all real polyimides (base value) and contributions of its substructures. Red or blue arrows indicate the positive or negative contribution of each substructure. The feature value of a substructure can be "0" meaning the absence of the substructure in the molecule, but its feature importance is still a valid value indicated by the length of the arrow. Top substructures in this polyimide are highlighted in different colors on the left. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variance. In addition, this single-task ensemble model outperforms the multi-task model, whose R^2 reaches only 0.5 ~ 0.86 (See Supporting Information S2 and Ref. [35]). It is evident that multi-task learning compromises the σ_y prediction performance because of the minimal correlation between these properties. From the perspective of key substructures based on SHAP values, most of the important ones for σ_y in Fig. 4**b-c** don't match the top key substructures for T_g (Fig. 3**b-c**). Some substructures like "935" and "5952" related to aromatic rings are the common critical features for both properties, but their synergy with other substructures still requires careful examination case by case. For example, both features make positive contributions to the σ_y of the poly (1,4-phenylenediamine)-alt-[4,4'-oxybis(phthalic anhydride)]} as shown in Fig. 4**d.** This polyimide processes the highest σ_y of 0.25 GPa among all real polyimides. Intuitively, we realize that certain aromatic rings play important roles in improving T_g and σ_y properties, and the obtained ensemble models are able to reveal the inherent correlations and provide quantitative estimations directly.

Not limited to the multi-functionality of two properties T_g and σ_y, we also incorporate a third property, Young's modulus E, to consider the stiffness of polyimides. We train and validate the FFNN single-task ensemble model using 447 real polyimides with available E values. The obtained R^2 of 0.92 and 0.81 for the model training and validation (Fig. 5**a**), respectively, are superior to the value of 0.44 ~ 0.74 for the multi-task model (See supporting information S2 and Ref. [35]). Similar to σ_y, the property E has no clear correlations with other properties, which makes the single-task learning a better choice. The most important substructure of the ensemble model for E is presented in Fig. 5**b-c.** When the contributions of substructures are shown for the highest E polyimide poly[1,4-phenylene-tetra(oxy-1,4-phenylene) pyromellitimide], the substructure "5952" that has a positive impact in the previous highest T_g and highest σ_y conditions is now making a negative contribution. We can infer that the influence of a substructure is not an invariant. For example, among the substructures of a polyimide, a particular substructure can have a positive impact on a property; while among the substructures of another polyimide, the same substructure may have a negative impact on this property. Therefore, a substructure's influence is based on the synergistic interaction of all substructures for the polyimide. Fig. 5**d** highlights two key substructures governing E, but it is worth noticing that their contributions are part of the combined effect of all substructures.

Because of the synergistic effect of these substructures, each substructure's feature importance for a property is different according to different polymers. To have a general evaluation of a substructure's contribution to a given property, the corresponding values obtained from analyzing different polymers can be averaged as an index. For a certain substructure, three indexes can be obtained to roughly characterize its contribution to the three properties. Fig. 6 shows the average feature importance of each substructure for three properties in a principal component analysis (PCA) plot. It indicates the overall contributions of different substructures toward different properties. From the PCA plot of the three properties, the first two principal components PC1 and PC2 explained most of the variance in the data, and some general guidelines can be obtained. Firstly, it is clear that most substructures locate near the origin, demonstrating the difficulty in differentiating their contributions to different properties. Secondly, the property σ_y as a variable vector is nearly orthogonal to the other two properties T_g and E, indicating the challenge to adjust σ_y of a polyimide while adjusting its T_g and E. Lastly, some key substructures aforementioned like "935" that is critical for all three properties are far away from the origin, demonstrating their high impact on the overall performance of a polyimide. Compared to the feature importance analysis for one property, the PCA analysis provides a more general evaluation on how different substructures are correlated with respect to different properties, which help us gain more insights on how each substructure affects the polyimide's properties.

2.3. *Discovery of multifunctional polyimides through Pareto frontier analysis*

Single-task ML models allow us to evaluate a specific property given a new polyimide. When we apply the well-trained ML models on the 8 million hypothetical polyimides, we obtain estimations of their σ_y, E, and T_g so that we can discover better performers regarding the multifunctionalities. As there are three properties to compete against each other, the design space becomes three-dimensional, as illustrated in Fig. 7. In the real polyimide dataset, some polyimides only have one or two properties reported experimentally. However, all three properties are needed in this 3D space. Thus, for those properties that are not

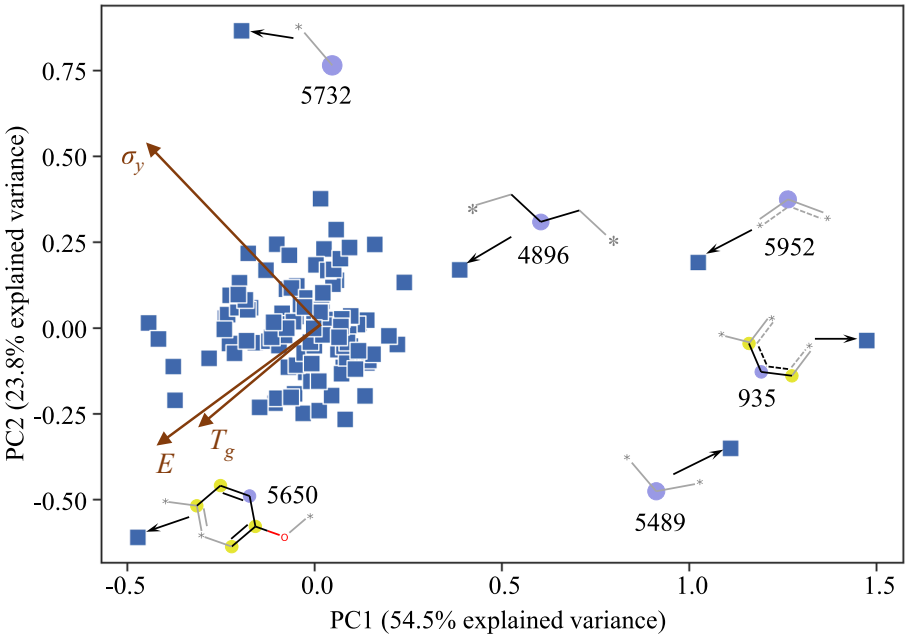
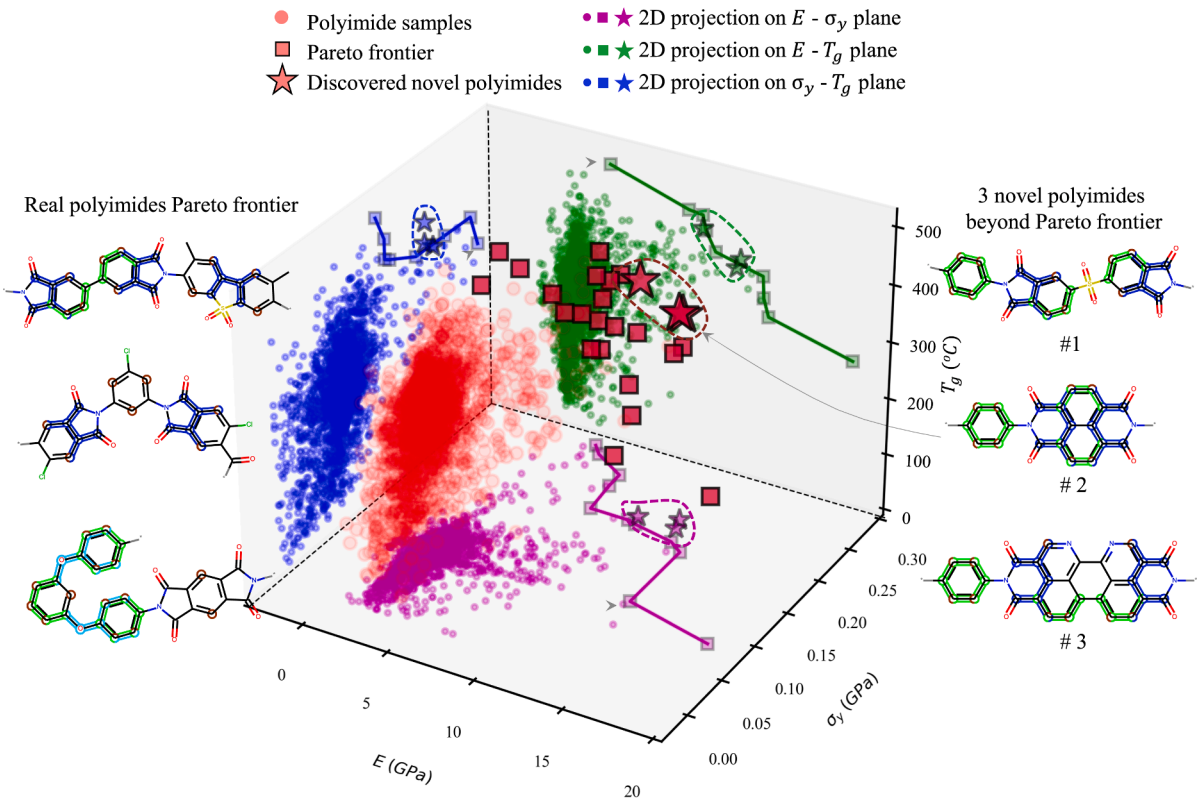


Fig. 6. PCA analysis on feature importance for three properties of 121 substructures.



Note: The highlighted substructures of the 3 real polyimides and 3 novel polyimides can be referred to Figure 3-5.

Fig. 7. Comparisons of the three properties of real polyimides. Red dots and squares correspond to the performance of real polyimides in 3D coordinates. Red squares are polyimides whose σ_y, *E,* and T_g define a boundary to which no other real polyimides can reach (Pareto frontier). The red stars, however, are hypothetical polyimides we discover whose properties are beyond the Pareto frontier. The projections of these red markers on three planes are shown in different colors. The boundaries on each 2D plane are better illustrated with the lines passing Pareto frontiers. The chemical structures of three real polyimides near the Pareto frontier are shown on the left, and the chemical structures of the three hypothetical polyimides that most beyond the Pareto frontier are shown on the right. The key substructures identified in Figs. 3-5 are highlighted in different colors for both three real polyimides and three hypothetical polyimides. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reported, we complement them with the ML model predictions. Using the completed data of σ_y, *E,* T_g, all real polyimides are positioned accordingly in Fig. 7. Although the resultant design space is based on a mixture of experimental values and ML predictions, we consider it well constructed and reliable, given the good predictive performance of the ML models (Figs. 3-5).

When examining the ML predicted σ_y, *E,* and T_g of all hypothetical polyimides, some of them can be referred to as the Pareto frontier [36], for which none of the properties can be improved without degrading other properties. This set of Pareto frontier of hypothetical polyimides defines an envelope boundary for the three properties (σ_y, *E,* and T_g). Among the Pareto frontier of hypothetical polyimides, the three best-performing hypothetical polyimides (high σ_y, *E,* and T_g at the same time) are selected (see Supporting Information S4 for the 3D property space of hypothetical polyimides). Similarly, based on the three properties of all real polyimides, a new set of Pareto frontier can be identified, which defines the property boundary of real polyimides (shown in Fig. 7). Fig. 7 also illustrates the projections of all real polyimides on three planes. The 2D Pareto frontier line is more straightforward on each plane and the *E* -T_g, *E* - σ_y, and σ_y -T_g combinations are three special cases from the 3D design space (see Supporting Information S4 for the individual 2D projection figures). When the three best-performing novel polyimides discovered from the 8 million hypothetical polyimides (indicated by red stars) are superposed in Fig. 7, it is found that the discovered hypothetical polyimides are beyond the Pareto frontier boundary of real polyimides, suggesting superior performances. Their chemical structures are shown on the right side of Fig. 7. Compared to the chemical structures of three real polyimides near Pareto frontiers on the left side of Fig. 7, we find common structural features such as aromatic rings in the backbone of the main chain and the sulfonyl functional of two double bonds between the sulfur and oxygen. This similarity suggests a successful pattern captured by our ML models (Figs. 3-5). In addition, features like pyridine rings are hardly found in the real polyimide structures and are observed in the discovered novel polyimides. Pyridine rings increase the aromaticity of the polymer structures and help to maintain mechanical properties at high temperatures [37–43]. The important substructures influencing the thermomechanical properties of polyimides are embodied in these discovered novel polyimides. They locate beyond the Pareto frontier boundary in the real polyimides design space, denoting improved multifunctionalities and, therefore, better options for experimental synthesis. If based on more thermal/mechanical properties, the radar charts for 7 properties (shown in Supporting Information S5) also demonstrate a more balanced performance of the three discovered hypothetical polyimides. In terms of their locations in the t-SNE chemical space, the three novel polyimides are not far away from the real polyimides. In this respect, they are similar to real polyimides but possess extraordinary properties that have not been discovered (see Supporting Information S6 for the t-SNE chemical space location of the three novel polyimides).

2.4. MD validation of thermal and mechanical properties

To validate the thermomechanical properties of the discovered novel polyimides, we carry out all-atom MD simulations to analyze their three properties, σ_y, *E,* and T_g. We build all-atoms models to simulate the novel

polyimides made of two components (dianhydride + diamine/diisocyanate). The polymer consistent force field (PCFF) [44–47] is used to define interatomic interactions. It is a second-generation force field [45,48–51], parameterized for organic compounds containing H, C, N, O, S, P, halogen atoms, and ions. PCFF has a broad coverage of organic polymers in calculations of cohesive energies, mechanical properties, compressibilities, heat capacities, and elastic constants. We employ a multi-step strategy [52] to simulate the cross-linking reactions of polyimide. Reactive atoms are first assigned to monomers (dianhydride) and crosslinkers (diamine/diisocyanate), and then covalent bonds (elastic springs) are formed between reactive atoms within a cutoff distance. After relaxing the cross-linked network for a while, extra hydrogen atoms are removed, and partial charges are adjusted to follow the charge-neutral principle. Based on the relaxed network, the second round of cross-linking continues with an increased cutoff distance. When the curing degree is satisfied, the cross-linking simulation stops further bond breaking and formation [53–55]. We pack 500 of each component within the 3D-periodic amorphous cell. The curing target is to reach more than 90 percent of reactive atoms on monomers. The final cross-linked polymer models contain ~ 20,000 atoms and have a box side length of 65 Å (see Supporting Information S7 for the details of cross-linking steps). Periodic boundary conditions are set along with all three directions. LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) package is used for MD simulations.

Before simulations of thermal or mechanical properties, polyimides are equilibrated first through a 21-step MD equilibration protocol (see

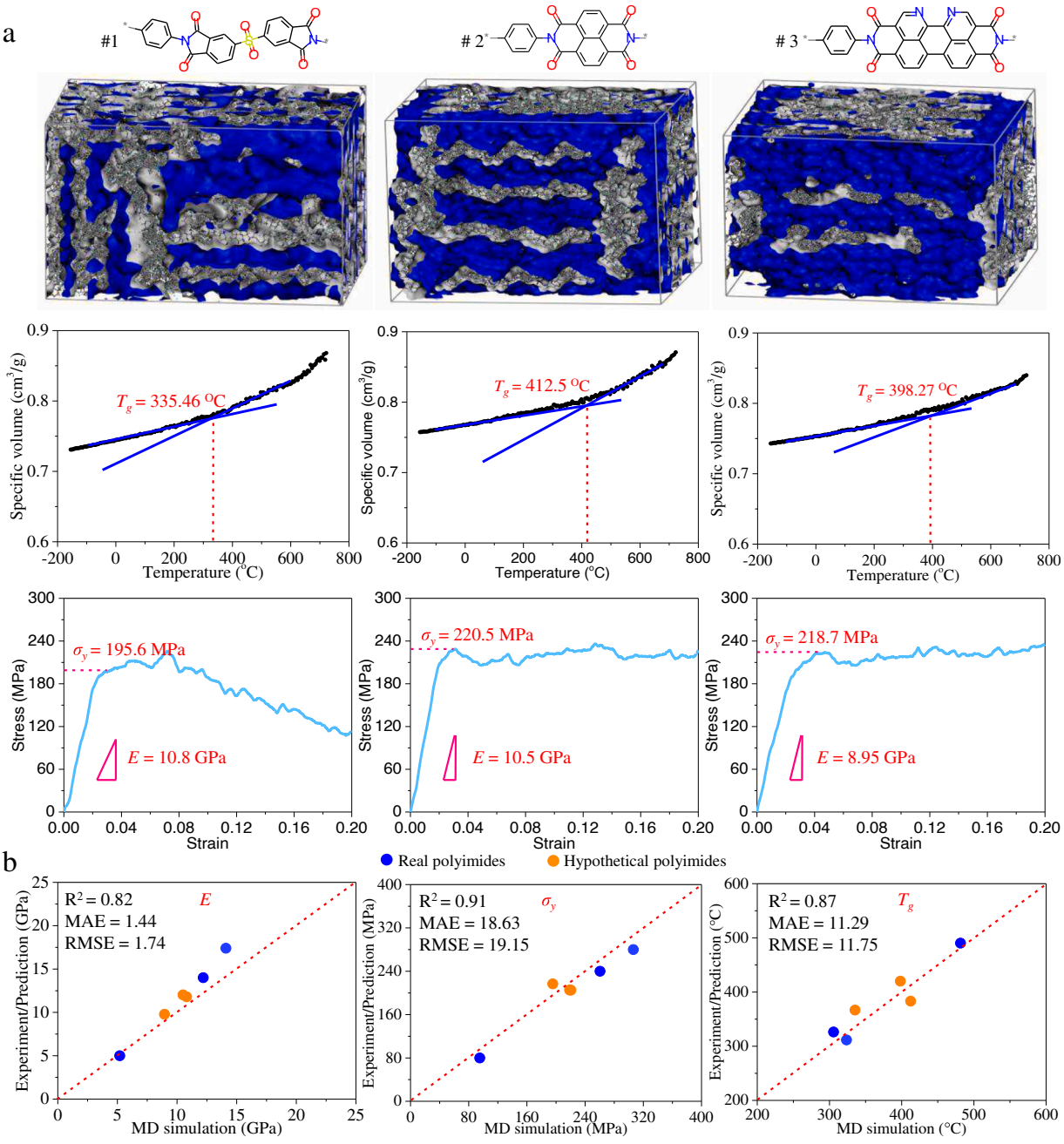


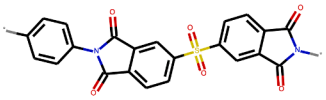
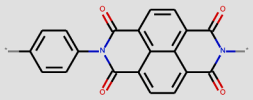
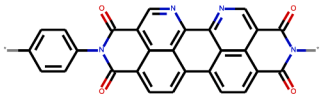
Fig. 8. MD simulations of three novel polyimides beyond the current Pareto frontier. (a) The four rows show the molecular structure of the repeat unit, MD simulation box of semi-crystalline structures, the specific volume vs. temperature curve for T_g extraction, and stress vs. strain curves of the three semi-crystalline structures for E and σ_y extraction. (b) The parity plots of MD simulation vs. experiment/ML prediction for E, σ_y, and T_g. All MD simulated T_g, E and σ_y values are comparable to experimental measurements or ML predictions.

Supporting Information S7 for the details of the 21-step equilibration). To obtain the T_g of the system, we carry out a cooling process simulation by gradually decreasing the temperature from 1000 K to 100 K. The simulated specific volume vs. temperature curves is shown in Fig. 8. The segments at both sides of the curve have a constant slope, which represents two different phases (rubbery and glassy). The least-square fitted lines are also plotted in Fig. 8, and their interception represents the T_g [54,56,57]. It is worth noting that the time scale of MD simulation is around nanoseconds, so the modeled cooling rate is much faster than that of the experiments [57–60]. Although the simulated cooling rate is not exactly consistent with experiments, the T_g estimation from MD simulation is still proven to reasonably agree with the experimental value [58–62]. The tensile simulation for stress–strain response is achieved by changing the simulation box with a constant strain rate. Like the condition of high cooling rate in T_g simulation, the high strain rate in MD simulation differs from the true values in experiments. Nevertheless, a low strain-rate sensitivity has been demonstrated with a relatively small strain rate, for example, around 1×10^8 s^{-1}.

Since we are interested in multi-functional polyimides with high T_g, σ_y, and E properties, we first select 3 real polyimides whose T_g, σ_y, and E are relatively high for MD verification. To obtain reliable simulation results, we first benchmarked our MD method by simulating ten experimental polyimides with amorphous structures (see Supporting Information S7 for their chemical structures, molecular structures, SMILES, specific volume vs. temperature curves, the stress vs. strain curves, and the extracted thermal and mechanical properties). The resulting thermal and mechanical properties, including T_g, E, and σ_y, are in excellent agreement with their corresponding measurements as reported in the experiments. For these selected 3 high-performance real polyimides, they may not have all three properties reported, but we have obtained reasonable ML estimations as approximate true values from Section 2.2. Given the planar nature of monomers due to abundant benzene rings, the polyimide cases in consideration are expected to be semi-crystalline or have an ordered phase. That is to say, polymers have different molecular states, such as amorphous, semi-crystalline, and full-crystalline. Polymers with a high degree of crystallinity have higher mechanical properties than their corresponding amorphous phase. As described in PolyInfo dataset, these selected polyimides are semi-crystalline structures. Likewise, molecules-based ML models are also implicitly related to polyimides' aggregation state through their thermal and mechanical properties. To verify these points, we first simulate the mechanical properties of the selected high-performance polyimides with the amorphous state. Additionally, polymer structure predictor (PSP) [63], as a tool proposed to predict the polymer crystal structure models, is also employed to build extra full-crystalline and semi-crystalline structures for these polyimides. In addition, the degree of crystallization for semi-crystalline structures is 0.55, which is determined using the Polymer Genome platform [63]. Results indicated that the mechanical properties of full-crystalline structures are highest, followed by semi-crystalline structures, and amorphous structures (see Supporting Information S9 for the details of MD simulations and model setup for full-crystalline and semi-crystalline structures). The results obtained in Fig. 8**b** and Table 2 show the mechanical properties of semi-crystalline structures. Obviously, when the selected polyimides are semi-crystalline structures, Young's modulus and yield strength of three real polyimides (experimental measurements) and three novel polyimides (ML predictions) are well consistent with MD simulations. In addition, MD simulated T_g is comparable to experimental measurements. In addition to the focused three properties, density as a fundamental intrinsic property of the material is also listed for comparison. In short, these investigations show that the discrepancies between the MD-estimated properties and the ML-predicted properties are within an acceptable range, considering the uncertainties in MD simulations and ML predictions.

In addition to MD validation before the future experimental study, we further examine the synthesizability of these polyimides. We calculate Schuffenhauer's synthetic accessibility (SA) score of the 3 novel polyimides' two reacting components. The SA score of 1 ~ 10 indicates the accessibility from easy to difficult. The highest SA score is 3.81 for the discovered 3 polyimides, which suggests easy synthesizability (see Supporting Information S8 for the SA score analysis).

Table 2
Comparison of ML predictions and MD simulations of three novel polyimides.

Polyimide	Property	MD	ML	Diff
#1  *c1ccc(N2C(=O)c3ccc(S(=O)(=O)c4ccc5c(c4)C(=O)N(*)C5=O)cc3C2=O)cc1	T_g (°C)	329.72	366.7	-11.22%
	σ_y (MPa)	195.6	216.68	-10.78%
	E (GPa)	10.8	11.8	-9.26%
	$\boldsymbol{\rho}$ (g/cm^3)	1.30	1.44	-10.77%
#2  *c1ccc(N2C(=O)c3ccc4c5c(ccc(c35)C2=O)C(=O)N(*)C4=O)cc1	T_g (°C)	409.5	383.16	6.43%
	σ_y (MPa)	220.5	205.12	6.98%
	E (GPa)	10.5	12.02	-14.48%
	$\boldsymbol{\rho}$ (g/cm^3)	1.32	1.42	-7.58%
#3  *c1ccc(N2C(=O)c3ccc4c5ccc6c7c(cnc(c8ncc(c3c48)C2=O)c75)C(=O)N(*)C6=O)cc1	T_g (°C)	390.24	420.03	-7.63%
	σ_y (MPa)	218.7	205.39	6.09%
	E (GPa)	8.95	9.78	-9.27%
	$\boldsymbol{\rho}$ (g/cm^3)	1.30	1.46	-12.31%

Note: Diff = (MD - ML) / MD * 100%.

2.5. Experimental validation of the discovered polyimide #2

Considering all the differences between MD simulation results and ML predictions listed in Table 2, we found that the novel polyimide #2 had relatively smaller differences. Therefore, we selected the #2 candidate for further experimental synthesis and measurement. Guided by the aforementioned reaction template and reacting components, we purchased all required chemicals from Sigma Aldrich unless otherwise stated. The reacting components for the discovered polyimide #2 are 1,4,5,8-naphthalenetetracarboxylic dianhydride (1,4,5,8 NTDA) and *p*-Phenylenediamine (*p*-PDA). In addition, *p*-Cresol or *p*-chlorophenol was used to improve the solubility of the reacting components (referred to as *p*-Cresol reaction or *p*-chlorophenol reaction), as shown in Fig. 9**a**. *p*-Cresol (99%) was supplied by Thermo Fisher Scientific. The solvents and 1,4,5,8 NTDA were used as received without further purification, while *p*-PDA was sublimated at 150 °C before use. On one hand, even though synthetic accessibilities scores indicate that the syntheses of these promising polymers are feasible, it is still difficult to find some excellent

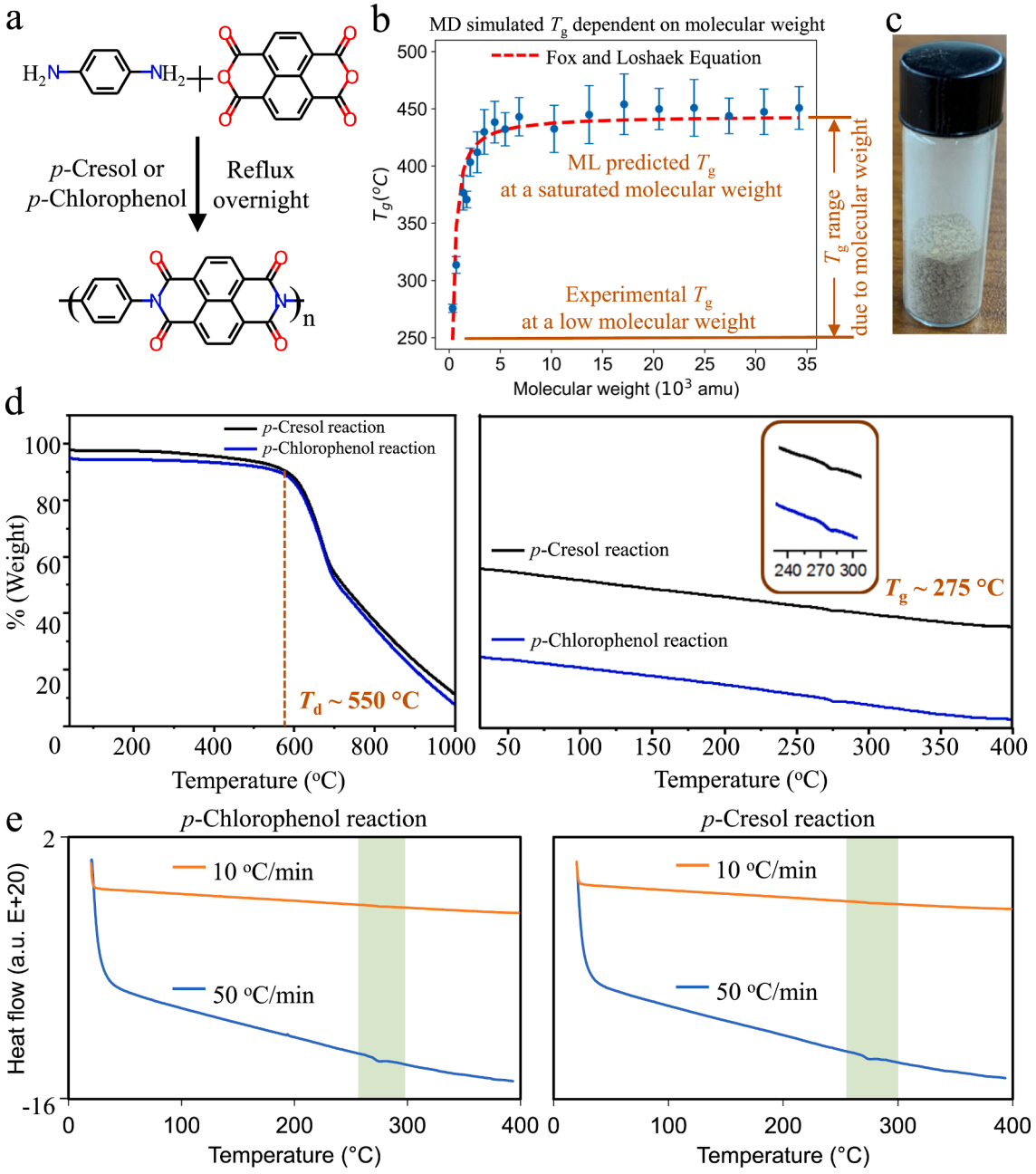


Fig. 9. Experimental synthesis and measurement of the discovered polyimide #2. (a) The reaction between the two reacting components 1,4,5,8-naphthalenetetracarboxylic dianhydride (1,4,5,8 NTDA) and *p*-Phenylenediamine (*p*-PDA). *p*-Cresol or *p*-chlorophenol was used to improve the solubility of the reacting components. (b) The effect of molecular weight on the T_g from MD simulations. Models with different molecular weights are simulated. The T_g estimations and the corresponding standard deviations are plotted in dots and error bars. The Fox and Loshaek Equation fits well with the obtained data, describing the T_g dependency with molecular weight. (c) The reacted sample of the discovered polyimide #2. The product precipitates only 20 min into the reaction, leading to a low molecular weight of the final sample. (d) Thermogravimetric analysis (TGA) results of the samples by either *p*-Cresol reaction or *p*-chlorophenol reaction. T_d = 550 °C is obtained and in good agreement of the ML prediction of 536 °C. T_g = 275 °C is obtained and lower than the ML prediction due to the low molecular weight of the final product. (e) Differential scanning calorimetry (DSC) results of the samples by either *p*-Cresol reaction or *p*-chlorophenol reaction. When the heating rate is changed from 10 °C to 50° C/min, T_g response at 275 °C becomes more obvious.

solvents to dissolve the reactants; on the other hand, mechanical tests require more samples and further processing of the sample into a dog-bone specimen. Therefore, in this experimental part, we focus on measuring the T_g.

In a typical reaction, 1.00 g (3.73 mmol) of 1,4,5,8 NTDA, 0.403 g (3.73 mmol) of *p*-PDA, and 50.0 g of *p*-chlorophenol were charged into an oven-dried two-neck round bottom flask equipped with a magnetic stirrer. The flask was fitted with a Dean–Stark trap and a reflux tube, and then transferred to an oil bath heated to 240 °C. The color of the solution changed from brown to crimson and eventually dark purple as the reaction progressed. To ensure inert conditions throughout the polymerization, the reaction was refluxed under a slow stream of nitrogen (10 mL min^{-1}). After 15 h, the reaction mixture was allowed to cool to room temperature and centrifuged at 8000 rpm to isolate the polymer precipitates. Subsequently, the residue was stirred in *p*-chlorophenol (20 g × 3) at 50 °C and centrifuged again to remove any unreacted monomers. Finally, the polymer was dried in a vacuum oven at 240 °C for 24 h. A similar procedure was followed for the polymerization reaction in *p*-cresol. Yield: 34.0 % for the *p*-chlorophenol reaction and 31.8 % for the *p*-cresol reaction. Fig. 9**c** shows the obtained sample of the polyimide #2. The product precipitates ~ 20 min into the reaction, leading to a low molecular weight of the final sample.

To measure the thermal properties of the final product, differential scanning calorimetry (DSC) and thermogravimetric analysis (TGA) were performed. DSC was performed on a Discovery DSC2500 (TA Instruments) between 20 and 400 °C at a heating rate of 10 °C min^{-1} or 50 °C min^{-1}. TGA was performed on a Discovery TGA5500 thermogravimetric analyzer (TA Instruments) between 30 and 1000 °C at a heating rate of 10 °C min^{-1}. Fig. 9**d** shows the TGA results, giving T_d = 550 °C which is in good agreement with the ML prediction of 536 °C. T_g = 275 °C is obtained which is lower than the ML prediction of 383 °C due to the low molecular weight of the final product. The Fox and Loshaek Equation develops an empirical formula for the prediction of the molecular weight dependence of the T_g [64]. It describes that at a lower molecular weight or degree of crosslinking, a lower T_g is resulted. With further MD simulations for models with different molecular weights, the T_g performance of the sample is well-fitted according to the Fox and Loshaek Equation, as shown in Fig. 9**b**. When the ML model predicts the limiting value of the glass transition temperature at a very high molecular weight, the actual molecular weight or degree of crosslinking of the obtained sample affects its T_g in a significant way. One limitation of our ML model is that the experimental related parameters (e.g., molecular weight) are difficult to incorporate in the ML model training because it was reported for only a few of the ~ 1800 datasets. Fig. 9**e** is for the further experimental measurement of T_g from DSC results. It demonstrates a similar T_g response at 275 °C. It is also observed that when the heating rate is changed from 10 °C to 50° C/min, the T_g response at 275 °C becomes more obvious, which is a typical situation for measuring the glass transition of rigid polymers like polymers with intrinsic microporosity (PIMs) [65].

3. Conclusion

Multi-functional polyimide is a key technology enabler for diverse applications, such as high-temperature fuel cells, polymer composites, and membranes. However, successful products of polyimides are limited to a few of them, like Kapton^{Tm}. To discover more promising polyimides with better performance, we build a large hypothetical polyimide dataset for high-throughput screening. 8 million possible polyimides are obtained computationally based on the polycondensation of existing dianhydride and diamine/diisocyanate molecules. This hypothetical dataset significantly expands the chemical space of existing polyimides, offering a great opportunity for materials discovery and design. However, it is infeasible to synthesize all of them for experimental analysis. To this end, we establish structure–property relationships through predictive ML models to do the high throughput screening of these 8 million hypothetical polyimides. Among the collected 7 properties (ρ, T_g, T_m, T_d, E, σ_y, and σ_b), we focus on 3 properties, T_g, E, and σ_y, and find single-task ML model outperforms the multi-task ML model. With the help of explainable ML models, we identify the key substructures influencing the thermal and mechanical properties of polyimides, such as aromatic rings and oxygen linkages. Applying the well-trained ML models to the 8 million hypothetical polyimides, we identify 3 best-performing novel polyimides with simultaneous high values of T_g, E, and σ_y. Consistent key substructures are also present in the discovered novel polyimides contributing to the high performances. Their thermomechanical properties are found beyond the Pareto frontier of existing polyimides, further confirmed by MD simulations. Although their chemical structures have not been reported experimentally, their polymerization route is well established, and corresponding reacting components are found easy to synthesize with a low synthetic accessibility score. Using the ML-guided reaction template and reacting components, we synthesize the discovered polyimide #2 successfully and measure its thermal properties. Due to the low molecular weight of the sample obtained, its measured T_g follows the Fox and Loshaek Equation and the experimentally obtained value of thermal decomposition temperature 550 °C demonstrates excellent thermal stability. Our study has successfully identified several novel polyimides with exceptional thermomechanical properties, offering promising directions for the synthesis of innovative materials with a wide range of potential applications. Through the use of advanced ML techniques and a carefully designed workflow, we were able to rapidly screen a large number of potential candidates and identify those with the most desirable properties. These novel polyimides demonstrate excellent mechanical strength, thermal stability, and other important characteristics, making them highly attractive candidates for use in a variety of fields, including aerospace, electronics, and automotive industries. We believe our findings will make a significant contribution to the development of new materials and inspire further research in this exciting and rapidly evolving field.

This study discovers novel polyimides with promising thermomechanical properties and guides the further experimental synthesis of innovative polyimides. More importantly, our method of utilizing explainable ML techniques and high-fidelity MD simulations demonstrates an efficient way to deal with a daunting number of chemical structures. It is important to note that our proposed method is designed specifically to provide guidance for the selection of promising candidates and corresponding raw materials, rather than to evaluate the entire experimental synthesis process. While factors such as solvent selection, reaction time, temperature control, toxicity, and other conditions are undoubtedly critical aspects of the experimental synthesis process, they are outside the scope of this study. Nonetheless, we believe that our ML-assisted workflow represents a significant step forward in the field of polymer informatics and offers exciting possibilities for future research. By leveraging the power of advanced ML techniques and carefully designed workflows, we can rapidly identify promising candidates for further development and enhance the efficiency and effectiveness of materials discovery efforts. Looking ahead, we believe that similar workflows could be established to evaluate the entire experimental synthesis process and further enhance the reproducibility and reliability of materials discovery research. Additionally, our method could be further applied to the high throughput screening for other polymeric material problems, such as organic photovoltaics, polymer membranes, and dielectrics.

4. Data and code availability

Data and code are available at https://github.com/figotj/Polyimide_explorer. Based on our hardware specifications (12th Gen Intel(R) Core(TM) i7-12700 K 3.60 GHz, 32 GB DDR5 RAM) and NVIDIA RTX A4000 Graphics Card, the training of the machine learning model takes

less than 1 h to finish. To process and screen the 8 million hypothetical polyimides, it takes ~ 10 h to complete. And we develop an online interactive platform https://polyimide-explorer.herokuapp.com/ for better visualization of more than 77,000 high-performing hypothetical polyimides. Detailed information is illustrated in the platform including polyimide's molecular structures, properties, polymerization route, and the corresponding reacting components that are commercially available. The developed machine learning model is also embedded in the platform for easy application.

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.

Data availability

Data will be made available on request.

Acknowledgments

We gratefully acknowledge financial support from the Air Force Office of Scientific Research (AFOSR) through the Air Force's Young Investigator Research Program (FA9550-20-1-0183; Program Manager: Dr. Ming-Jen Pan), AFOSR (FA9550-18-1-0381), Air Force Research Laboratory/UES Inc. (FA8650-20-S-5008, PICASSO program), and the National Science Foundation (CMMI-2316200, CAREER-2323108 and DMR-1752611). Y.L. would like to thank the support from 3M's Non-Tenured Faculty Award. This research also benefited in part from the computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Defense and National Science Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2023.142949.

References

[1] S. Diaham, Polyimide in electronics: Applications and processability overview, Polyimide Electron. Electr. Eng. Appl. (2021) 2020–2021, https://doi.org/10.5772/intechopen.92629.
[2] Y.S. Negi, S.R. Damkale, S. Ansari, Photosensitive polyimides, J. Macromol. Sci. Part C Polym. Rev. 41 (1–2) (2001) 119–138, https://doi.org/10.1081/MC-100002057.
[3] M. Hasegawa, N. Sensui, Y. Shindo, R. yokota,, Structure and Properties of Novel Asymmetric Biphenyl Type Polyimides, J. Photopolym. Sci. Technol. 9 (2) (1996) 367–378, https://doi.org/10.2494/photopolymer.9.367.
[4] I. Gouzman, E. Grossman, R. Verker, N. Atar, A. Bolker, N. Eliaz, Advances in Polyimide-Based Materials for Space Applications, Adv. Mater. 31 (18) (2019) 1807738, https://doi.org/10.1002/adma.201807738.
[5] S. Ghaffari-Mosanenzadeh, O. Aghababaei Tafreshi, S. Karamikamkar, Z. Saadatnia, E. Rad, M. Meysami, H.E. Naguib, Recent advances in tailoring and improving the properties of polyimide aerogels and their application, Adv. Colloid Interface Sci. 304 (2022), 102646, https://doi.org/10.1016/j.cis.2022.102646.
[6] E.P. Favvas, F.K. Katsaros, S.K. Papageorgiou, A.A. Sapalidis, A.C. Mitropoulos, A review of the latest development of polyimide based membranes for CO2 separations, React. Funct. Polym. 120 (2017) 104–130, https://doi.org/10.1016/j.reactfunctpolym.2017.09.002.
[7] G.W. Goodall, W. Hayes, Advances in cycloaddition polymerizations, Chem. Soc. Rev. 35 (3) (2006) 280–312, https://doi.org/10.1039/B507209N.
[8] W.M. Alvino, L.E. Edelman, Polyimides from diisocyanates, dianhydrides, and tetracarboxylic acids, J. Appl. Polym. Sci. 19 (11) (1975) 2961–2980, https://doi.org/10.1002/app.1975.070191103.
[9] J. Liu, G. Chen, N. Mushtaq, X. Fang, Synthesis of organosoluble and light-colored cardo polyimides via aromatic nucleophilic substitution polymerization, Polym. Adv. Technol. 26 (12) (2015) 1519–1527, https://doi.org/10.1002/pat.3574.
[10] D.-J. Liaw, K.-L. Wang, Y.-C. Huang, K.-R. Lee, J.-Y. Lai, C.-S. Ha, Advanced polyimide materials: Syntheses, physical properties and applications, Prog. Polym. Sci. 37 (7) (2012) 907–974.
[11] P. Paraskevopoulou, D. Chriti, G. Raptopoulos, G.C. Anyfantis, Synthetic polymer aerogels in particulate form, Materials 12 (9) (2019) 1543, https://doi.org/10.3390/ma12091543.
[12] J.Q. Pan, W.W. Lau, Z. Zhang, X. Hu, Synthesis and properties of new copolymers containing hindered amine, J. Appl. Polym. Sci. 61 (8) (1996) 1405–1412, https://doi.org/10.1002/(SICI)1097-4628(19960822)61:8<1405::AID-APP22>3.0.CO;2-W.
[13] 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.
[14] T. Takekoshi, Polyimides, Kirk-Othmer Encycl. Chem. Technol. (2000), https://doi.org/10.1002/0471238961.1615122520011105.a01.
[15] A. Sezer Hicyilmaz, A. Celik Bedeloglu, Applications of polyimide coatings: A review, SN, Appl. Sci. 3 (2021) 1–22, https://doi.org/10.1007/s42452-021-04362-5.
[16] C. Yi, W. Li, S. Shi, K. He, P. Ma, M. Chen, C. Yang, High-temperature-resistant and colorless polyimide: Preparations, properties, and applications, Sol. Energy 195 (2020) 340–354.
[17] O.A. Tafreshi, Z. Saadatnia, S. Ghaffari-Mosanenzadeh, S. Okhovatian, C.B. Park, H.E. Naguib, Machine learning-based model for predicting the material properties of nanostructured aerogels, SPE Polym. 4 (1) (2023) 24–37, https://doi.org/10.1002/pls2.10082.
[18] I.A. Ronova, M. Bruma, Influence of chemical structure on glass transition temperature of polyimides, Struct. Chem. 21 (2010) 1013–1020, https://doi.org/10.1007/s11224-010-9639-1.
[19] S.G. Falkovich, S.V. Lyulin, V.M. Nazarychev, S.V. Larin, A.A. Gurtovenko, N. V. Lukasheva, A.V. Lyulin, Influence of the electrostatic interactions on thermophysical properties of polyimides: molecular-dynamics simulations, J Polym. Sci. B 52 (9) (2014) 640–646, https://doi.org/10.1002/polb.23460.
[20] A. Chakrabarty, T. Cagin, Coarse grain modeling of polyimide copolymers, Polymer 51 (12) (2010) 2786–2794.
[21] S. Lyulin, A. Gurtovenko, S. Larin, V. Nazarychev, A. Lyulin, Microsecond atomic-scale molecular dynamics simulations of polyimides, Macromolecules 46 (15) (2013) 6357–6363, https://doi.org/10.1021/ma4011632.
[22] X. Ma, F. Zheng, C.G. van Sittert, Q. Lu, Role of intrinsic factors of polyimides in glass transition temperature: An atomistic investigation, J. Phys. Chem. B 123 (40) (2019) 8569–8579, https://doi.org/10.1021/acs.jpcb.9b06585.
[23] M. Li, X. Liu, J. Qin, Y. Gu, Molecular dynamics simulation on glass transition temperature of isomeric polyimide, Express Polym. Lett. 3 (10) (2009) 665–675, https://doi.org/10.3144/expresspolymlett.2009.83.
[24] C. Wen, B. Liu, J. Wolfgang, T.E. Long, R. Odle, S. Cheng, Determination of glass transition temperature of polyimides from atomistic molecular dynamics simulations and machine-learning algorithms, J. Polym. Sci. 58 (11) (2020) 1521–1534, https://doi.org/10.1002/pol.20200050.
[25] L. Tao, G. Chen, Y. Li, Machine learning discovery of high-temperature polymers, Patterns 2 (4) (2021), 100225, https://doi.org/10.1016/j.patter.2021.100225.
[26] H. Lei, S. Qi, D. Wu, Hierarchical multiscale analysis of polyimide films by molecular dynamics simulation: Investigation of thermo-mechanical properties, Polymer 179 (2019), 121645.
[27] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, X. Chen, A comparison for dimensionality reduction methods of single-cell RNA-seq data, Front. Genet. 12 (2021), 646936, https://doi.org/10.3389/fgene.2021.646936.
[28] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, M. Yamazaki, PoLyInfo: Polymer database for polymeric materials design, in: 2011 International Conference on Emerging Intelligent Data and Web Technologies, IEEE, 2011, pp. 22–29.
[29] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B.o. Yu, J. Zhang, S.H. Bryant, PubChem substance and compound databases, Nucleic Acids Res. 44 (D1) (2016) D1202–D1213.
[30] G. Chen, Z. Shen, A. Iyer, U.F. Ghumman, S. Tang, J. Bi, W. Chen, Y. Li, Machine-learning-assisted de novo design of organic molecules and polymers: Opportunities and challenges, Polymers 12 (1) (2020) 163, https://doi.org/10.3390/polym12010163.
[31] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, S. Savarese, Which tasks should be learned together in multi-task learning? Int. Conf. Mach. Learning PMLR (2020) 9120–9132.
[32] G. Chen, L. Tao, Y. Li, Predicting polymers' glass transition temperature by a chemical language processing model, Polymers 13 (11) (2021) 1898, https://doi.org/10.3390/polym13111898.
[33] L. Tao, V. Varshney, Y. Li, Benchmarking machine learning models for polymer informatics: an example of glass transition temperature, J. Chem. Inf. Model. 61 (11) (2021) 5395–5413, https://doi.org/10.1021/acs.jcim.1c01031.
[34] F. Chollet, keras (2015). https://doi.org/https://github.com/fchollet/keras.
[35] C. Kuenneth, A.C. Rajan, H. Tran, L. Chen, C. Kim, R. Ramprasad, Polymer informatics with multi-task learning, Patterns 2 (4) (2021), 100238.
[36] K.M. Jablonka, G.M. Jothiappan, S. Wang, B. Smit, B. Yoo, Bias free multiobjective active learning for materials design and discovery, Nat. Commun. 12 (1) (2021) 2312, https://doi.org/10.1038/s41467-021-22437-0.
[37] D.-J. Liaw, F.-C. Chang, M.-K. Leung, M.-Y. Chou, K. Muellen, High thermal stability and rigid rod of novel organosoluble polyimides and polyamides based on bulky and noncoplanar naphthalene– biphenyldiamine, Macromolecules 38 (9) (2005) 4024–4029, https://doi.org/10.1021/ma048559x.
[38] D.-J. Liaw, K.-L. Wang, F.-C. Chang, Novel organosoluble poly (pyridine– imide) with pendent pyrene group: Synthesis, thermal, optical, electrochemical,

electrochromic, and protonation characterization, Macromolecules 40 (10) (2007) 3568–3574, https://doi.org/10.1021/ma062546x.
[39] D.J. Liaw, K.L. Wang, F.C. Chang, K.R. Lee, J.Y. Lai, Novel poly (pyridine imide) with pendent naphthalene groups: synthesis and thermal, optical, electrochemical, electrochromic, and protonation characterization, J. Polym. Sci. A Polym. Chem. 45 (12) (2007) 2367–2374, https://doi.org/10.1002/pola.21997.
[40] K.-L. Wang, W.-T. Liou, D.-J. Liaw, W.-T. Chen, A novel fluorescent poly (pyridine-imide) acid chemosensor, Dyes Pigm. 78 (2) (2008) 93–100.
[41] K.-L. Wang, W.-T. Liou, D.-J. Liaw, S.-T. Huang, High glass transition and thermal stability of new pyridine-containing polyimides: effect of protonation on fluorescence, Polymer 49 (6) (2008) 1538–1546.
[42] S. Zhang, Y. Li, D. Yin, X. Wang, X. Zhao, Y. Shao, S. Yang, Study on synthesis and characterization of novel polyimides derived from 2, 6-bis (3-aminobenzoyl) pyridine, Eur. Polym. J. 41 (5) (2005) 1097–1107.
[43] X. Wang, Y. Li, S. Zhang, T. Ma, Y. Shao, X. Zhao, Synthesis and characterization of novel polyimides derived from pyridine-bridged aromatic dianhydride and various diamines, Eur. Polym. J. 42 (6) (2006) 1229–1239.
[44] H. Sun, S.J. Mumby, J.R. Maple, A.T. Hagler, An ab initio CFF93 all-atom force field for polycarbonates, J. Am. Chem. Soc. 116 (7) (1994) 2978–2987, https://doi.org/10.1021/ja00086a030.
[45] H. Sun, P. Ren, J. Fried, The COMPASS force field: parameterization and validation for phosphazenes, Comput. Theor. Polym. Sci. 8 (1–2) (1998) 229–246, https://doi.org/10.1016/S1089-3156(98)00042-7.
[46] H. Sun, Ab initio calculations and force field development for computer simulation of polysilanes, Macromolecules 28 (3) (1995) 701–712, https://doi.org/10.1021/ma00107a006.
[47] H. Heinz, T.-J. Lin, R. Kishore Mishra, F.S. Emami, Thermodynamically Consistent Force Fields for the Assembly of Inorganic, Organic, and Biological Nanostructures: The INTERFACE Force Field, Langmuir 29 (6) (2013) 1754–1765.
[48] S.W. Bunte, H. Sun, Molecular Modeling of Energetic Materials: The Parameterization and Validation of Nitrate Esters in the COMPASS Force Field, J. Phys. Chem. B 104 (11) (2000) 2477–2489, https://doi.org/10.1021/jp991786u.
[49] H. Sun, COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds, J. Phys. Chem. B 102 (38) (1998) 7338–7364, https://doi.org/10.1021/jp980939v.
[50] M.J. McQuaid, H. Sun, D. Rigby, Development and validation of COMPASS force field parameters for molecules with aliphatic azide chains, J. Comput. Chem. 25 (1) (2004) 61–71, https://doi.org/10.1002/jcc.10316.
[51] N.D. Kondratyuk, V.V. Pisarev, Calculation of viscosities of branched alkanes from 0.1 to 1000 MPa by molecular dynamics methods using COMPASS force field, Fluid Phase Equilibria 498 (2019) 151–159, https://doi.org/10.1016/j.fluid.2019.06.023.
[52] V. Varshney, S.S. Patnaik, A.K. Roy, B.L. Farmer, A Molecular Dynamics Study of Epoxy-Based Networks: Cross-Linking Procedure and Prediction of Molecular and Material Properties, Macromolecules 41 (18) (2008) 6837–6842, https://doi.org/10.1021/ma801153e.
[53] C. Jang, T.W. Sirk, J.W. Andzelm, C.F. Abrams, Comparison of Crosslinking Algorithms in Molecular Dynamics Simulation of Thermosetting Polymers, Macromol. Theory Simul. 24 (3) (2015) 260–270, https://doi.org/10.1002/mats.201400094.
[54] C.C. L. Abbott, Polymatic: A Simulated Polymerization Algorithm, (2013). https://doi.org/metabase.uaem.mx//handle/123456789/2185.
[55] L.J. Abbott, J.E. Hughes, C.M. Colina, Virtual Synthesis of Thermally Cross-Linked Copolymers from a Novel Implementation of Polymatic, J. Phys. Chem. B 118 (7) (2014) 1916–1924, https://doi.org/10.1021/jp409664d.
[56] D. Rigby, R.-J. Roe, Molecular dynamics simulation of polymer liquid and glass. I. Glass transition, J. Chem. Phys. 87 (12) (1987) 7285–7292.
[57] K.-Q. Yu, Z.-S. Li, J. Sun, Polymer Structures and Glass Transition: A Molecular Dynamics Simulation Study, Macromol. Theory Simul. 10 (6) (2001) 624–633, https://doi.org/10.1002/1521-3919(20010701)10:6<624::AID-MATS624>3.0.CO;2-K.
[58] J. Buchholz, W. Paul, F. Varnik, K. Binder, Cooling rate dependence of the glass transition temperature of polymer melts: Molecular dynamics study, J. Chem. Phys. 117 (15) (2002) 7364–7372, https://doi.org/10.1063/1.1508366.
[59] C. Li, G.A. Medvedev, E.-W. Lee, J. Kim, J.M. Caruthers, A. Strachan, Molecular dynamics simulations and experimental studies of the thermomechanical response of an epoxy thermoset polymer, Polymer 53 (19) (2012) 4222–4230, https://doi.org/10.1016/j.polymer.2012.07.026.
[60] 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.
[61] C. Li, A. Strachan, Molecular dynamics predictions of thermal and mechanical properties of thermoset polymer EPON862/DETDA, Polymer 52 (13) (2011) 2920–2928, https://doi.org/10.1016/j.polymer.2011.04.041.
[62] J. Han, R.H. Gee, R.H. Boyd, Glass Transition Temperatures of Polymers from Molecular Dynamics Simulations, Macromolecules 27 (26) (1994) 7781–7784, https://doi.org/10.1021/ma00104a036.
[63] H. Sahu, K.-H. Shen, J.H. Montoya, H. Tran, R. Ramprasad, Polymer Structure Predictor (PSP): A Python Toolkit for Predicting Atomic-Level Structural Models for a Range of Polymer Geometries, J. Chem. Theory Comput. 18 (4) (2022) 2737–2748, https://doi.org/10.1021/acs.jctc.2c00022.
[64] T.G. Fox, S. Loshaek, Influence of molecular weight and degree of crosslinking on the specific volume and glass temperature of polymers, J. Polym. Sci. 15 (80) (1955) 371–390, https://doi.org/10.1002/pol.1955.120158006.
[65] H. Yin, Y.Z. Chua, B. Yang, C. Schick, W.J. Harrison, P.M. Budd, M. Böhning, A. Schönhals, First Clear-Cut Experimental Evidence of a Glass Transition in a Polymer with Intrinsic Microporosity: PIM-1, J. Phys. Chem. Lett. 9 (8) (2018) 2003–2008, https://doi.org/10.1021/acs.jpclett.8b00422.