JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

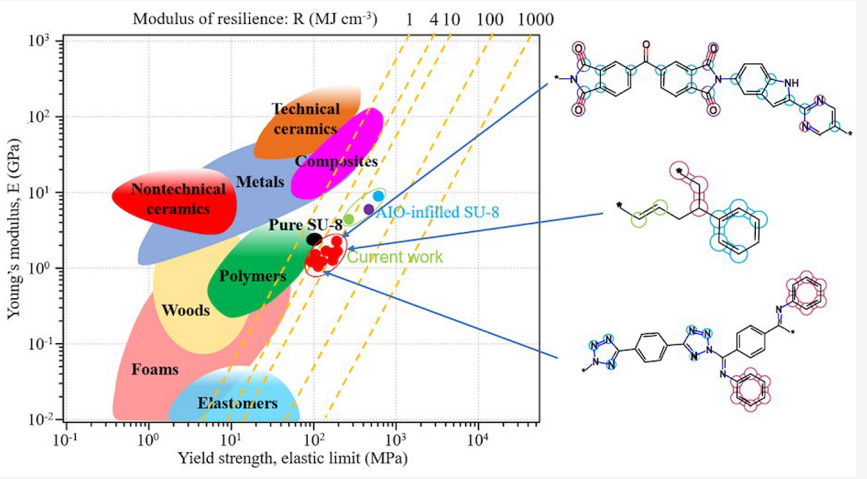
High-Throughput Screening and Prediction of High Modulus of Resilience Polymers Using Explainable Machine Learning

Tianle Yue, Jinlong He, Lei Tao, and Ying Li*

Cite This: J. Chem. Theory Comput. 2023, 19, 4641−4653 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via UNIV OF WISCONSIN-MADISON on October 17, 2024 at 21:28:28 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.



ABSTRACT: The ability to store and release elastic strain energy, as well as mechanical strength, are crucial factors in both natural and man-made mechanical systems. The modulus of resilience (R) indicates a material's capacity to absorb and release elastic strain energy, with the yield strength (σ_y) and Young's modulus (E) as $R = \sigma_y^2/(2E)$ for linear elastic solids. To improve the R in linear elastic solids, a high σ_y and low E combination in materials is sought after. However, achieving this combination is a significant challenge as both properties typically increase together. To address this challenge, we propose a computational method to quickly identify polymers with a high modulus of resilience using machine learning (ML) and validate the predictions through high-fidelity molecular dynamics (MD) simulations. Our approach commences by training single-task ML models, multitask ML models, and Evidential Deep Learning models to forecast the mechanical properties of polymers based on experimentally reported values. Utilizing explainable ML models, we were able to determine the critical substructures that significantly impact the mechanical properties of polymers, such as E and σ_y. This information can be utilized to create and develop new polymers with improved mechanical characteristics. Our single-task and multitask ML models can predict the properties of 12 854 real polymers and 8 million hypothetical polyimides and uncover 10 new real polymers and 10 hypothetical polyimides with exceptional modulus of resilience. The improved modulus of resilience of these novel polymers was validated through MD simulations. Our method efficiently speeds up the discovery of high-performing polymers using ML predictions and MD validation and can be applied to other polymer material discovery challenges, such as polymer membranes, dielectric polymers, and more.

1. INTRODUCTION

The modulus of resilience denotes the maximum amount of energy that can be stored in a material before it undergoes plastic deformation.[1] The storage and release of elastic strain energy in materials, as well as the mechanical strength, play an important role in both natural and engineered mechanical actuation systems.[2] Many emerging technologies also rely on the efficient use of mechanical energy storage and release, such as alternative energy systems.[3−5] Materials with a high modulus of resilience offer exceptional protection against mechanical deformation.[6,7] Additionally, artificial muscles in robots often require a high modulus of resilience as they use large amounts of elastic energy to produce powerful locomotion.[6,8,9] Advanced flexible electronic panels require a high modulus of resilience to balance the demands of durability and flexibility. Materials with a high modulus of

Special Issue: Machine Learning for Molecular Simulation

Received: January 31, 2023
Published: June 20, 2023

© 2023 American Chemical Society

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

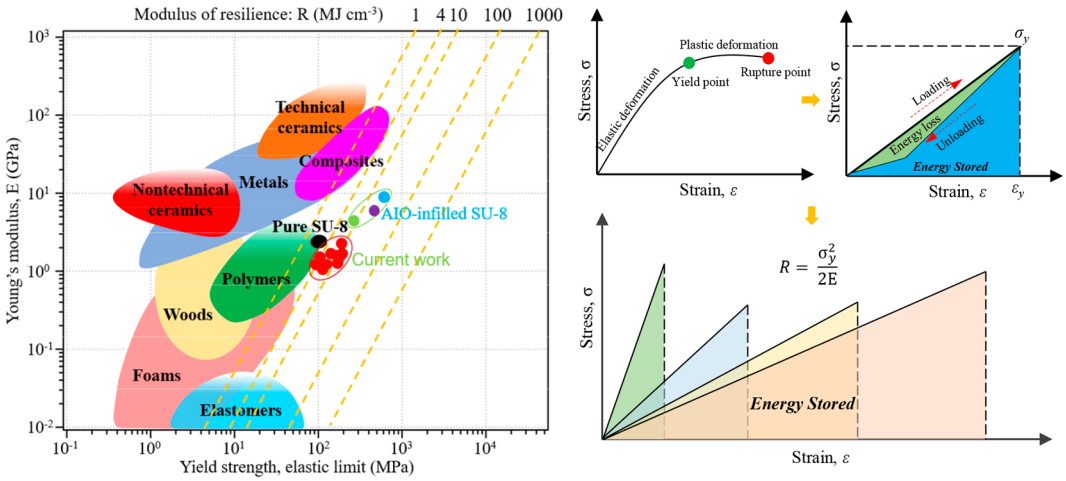


Figure 1. A plot of specific strength versus specific modulus was created to summarize the mechanical properties of polymers discovered through ML and verified through MD, and to compare them with other engineering materials. Additionally, a schematic diagram was included to depict the relationships between stress, strain, and modulus of resilience.

resilience can provide great protection against mechanical deformation while still maintaining the flexibility needed for these electronic devices.[10,11] As shown in Figure 1, the modulus of resilience of isotropic linear elastic solids is mathematically defined as $\sigma_y^2/(2E)$, where σ_y is the yield strength and E is Young's modulus. According to this definition, achieving high σ_y and low E at the same time leads to a high modulus of resilience. However, it is challenging to design and create materials with a high modulus of resilience. Most engineering materials have a strong correlation between σ_y and E, resulting in low $\sigma_y^2/(2E)$ values. Despite this, polymer composites are considered promising candidates for achieving a high modulus of resilience because a polymer matrix can be used as a starting material with a low E, which can be combined with other materials to achieve high yield strength. In this study, computational methods of ML and MD simulations are utilized to accelerate the discovery of high-performance polymers with high modulus of resilience. The results of the discovery can be observed in the modified Ashby Chart (Figure 1) of specific strength vs specific modulus, where the excellent specific modulus of resilience of the discovered polymers will be highlighted. The Ashby chart is a useful tool for evaluating the mechanical properties of materials, by plotting the ratio of specific strength to specific modulus.

Polymers are a versatile class of natural or synthetic materials. Polymers are composed of long-chain molecules, which are widely available in the world. They can be found in biological organisms, such as proteins, cellulose, and nucleic acids, as well as in common man-made materials such as concrete, glass, paper, plastics, and rubbers.[12−15] The polymeric structure consists of several simpler chemical units, which are also known as monomers. Those monomers are covalently bonded together to form long-chain macromolecules. The properties of the polymer, from microstructure to physical and mechanical behavior, are determined by the chemical structure of the monomers and their arrangements.[16−19] This makes polymer materials highly adaptable and useful in a wide range of applications.[16,20−22]

When addressing large-scale data challenges, artificial intelligence (AI) and ML are often considered effective computational methods for finding solutions. ML utilizes existing data, whether experimental or computational and is particularly useful not only in polymer applications but also in materials discovery and characterization.[23−31] However, a majority of the training data currently used for ML algorithms in polymer applications is derived from density functional theory calculations of small, monomeric, or oligomeric substances.[32] Furthermore, the representation of the polymeric structures used in current ML models are often the monomers simplified molecular-input line-entry system (SMILES), which simplifies the input but may not fully capture all information on the polymer.[33−35] SMILES is a widely used method for representing molecules, as it is easy for both humans and machines to read and understand. While SMILES describes the structural information on homopolymers, it does not take into account the influence of polymer chain topology on the target properties, such as cross-linking, branching, and so forth.[36−38] It is difficult to fully understand the macroscopic behaviors of polymer chains, such as self-assembly, polymer crystallization, or knot-type classifications, when they are represented by monomers only. Therefore, using traditional ML methods in combination with current polymeric databases and monomeric representations may not provide a complete understanding of polymer behavior.

To accelerate the discovery of new polymers with high modulus of resilience, we propose an integrated, data-driven approach that leverages ML by utilizing the PoLyInfo database[39] and a large data set of actual polyimides, a hypothetical data set of over 8 million potential polyimides,[40] for screening using custom-built ML models, and MD simulations for validation. Our method starts by gathering 2468 real polymer structures from the PoLyInfo database with their 10 reported physical properties if available, such as density (ρ), Young's modulus (E), tensile yield strength (σ_y), tensile break strength (σ_b), shear modulus (G), shear yield strength (τ_y), shear break strength (τ_b), flexural modulus (E_{flex}), flexural yield strength (σ_{flex}^y), and flexural break strength (σ_{flex}^b). This data set of polymers that have been synthesized and characterized experimentally forms a large data set of real polymers. Using this data set, we train ML models to establish the mapping between structure and properties, gaining insight

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

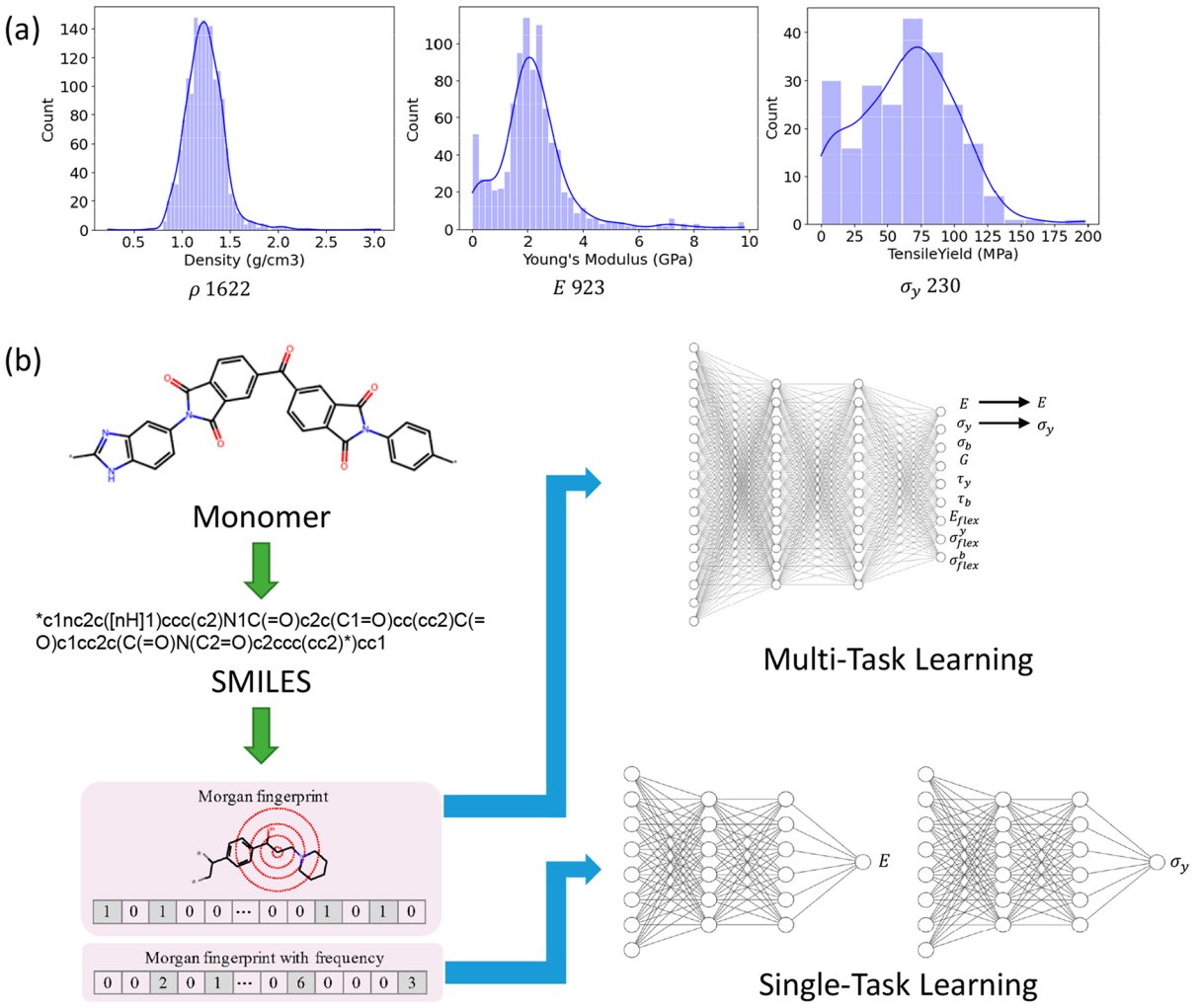


Figure 2. (a) Univariate distribution plots for ρ, E, and σ_y. (b) The structure representation, feature representation, and ML models. The polymer structure representation used in the study was the monomer. Two types of Morgan fingerprints were calculated from the monomer structures: bit vectors and repeat unit notation. The bit vectors were used as input for a multitask feedforward neural network (FFNN) model, while the repeat unit notation vectors were used for training in a single-task model.

into the key structural features that affect different properties. The well-trained ML models are then used as a predictive tool to estimate the properties of new polymers. Using the PoLyInfo and hypothetical data set, over 8 million polymers are screened, and our primary focus is on the properties of σ_y and E, as a high modulus of resilience is highly desirable.

Finally, we identify several multifunctional polymers with improved performance compared to current real polymers and verify their properties through all-atom MD simulations. Our study designs novel multifunctional polymers by narrowing large-scale chemistry space down to promising candidates through ML screening and MD validation. By using physics-based ML predictions to exhaust all possibilities before experimental synthesis, this approach successfully allows for the exploration of the entire design space, making it much more efficient compared to the conventional trial-and-error process. This strategy can also be applied to the molecular design of other polymeric materials.[27]

2. ML MODELS AND APPROACHES

2.1. Training and Screening Data sets. The largest database, PoLyInfo, contains over 18 000 reported polymers, including 12 854 homopolymers with their chemical structures and around 100 types of properties. This homopolymer data set is suitable for training neural networks in our study. Among the entire data set, as shown in Figure 2, 1625 homopolymers have reported values of ρ, 923 homopolymers have reported values of E, and 230 homopolymers have reported values of σ_y. The properties (ρ, E, and σ_y) distributions are shown in Figure 2. Using these reported property values and corresponding monomer structures, ML models can be trained to set up the composition-property mapping of polymers.[27]

Obtaining the ρ, E, and σ_y for all polymers in the PoLyInfo database would be a daunting task, given the amount of labor, cost, and time required. Additionally, the database includes a library of 8 205 087 hypothetical polyimides formed by the polycondensation of known diamines/diisocyanates with dianhydrides from the PubChem database,[41] providing a vast search domain for discovering potential high-performing polymers.

2.2. ML Models. Ensuring an appropriate representation of polymer structure, utilizing a suitable characterize method, and selecting a proper machine learning (ML) algorithm are the three main steps to establishing an ML model for polymer

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

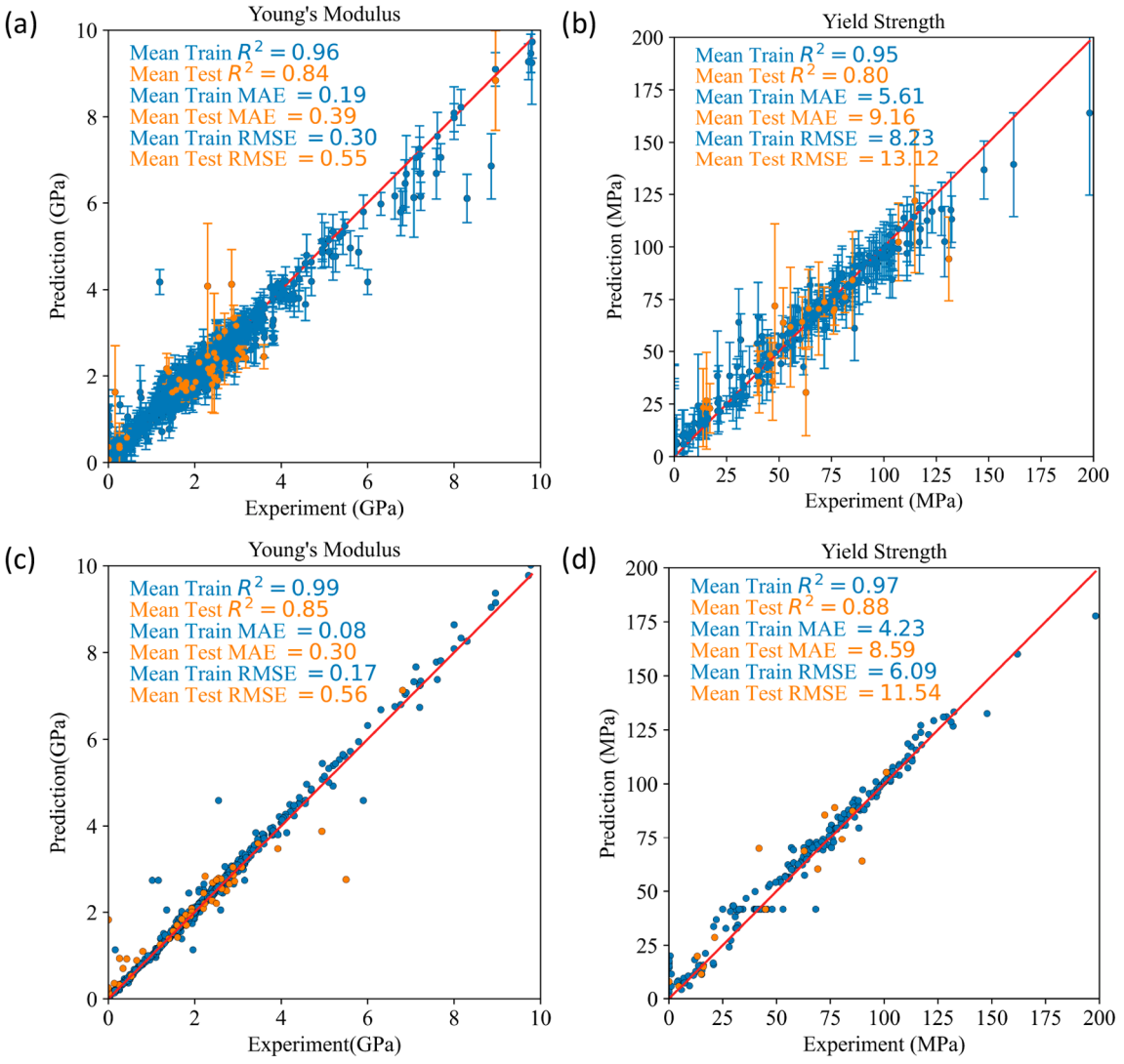


Figure 3. Parity plot of the single-task ensemble FFNN model using improved Morgan fingerprints as input features for (a) E, (b) σ_y, and the parity plot of the multitask ensemble FFNN model using improved Morgan fingerprints as input features for (c) E, (d) σ_y.

informatics.[42] As shown in Figure 2, in this study, to fit the scale of data sets, the SMILES and repeat unit notation are used for structure representation, and single-task and multitask feed-forward neural networks (FFNNs) are implemented as the ML models to process the data.[42−44]

For single-task feedforward neural networks, as shown in Figure 2 (b), Morgan fingerprint (MF)[45] with frequency is utilized for feature representation. The Morgan fingerprint method detects substructures enclosed in a circle of radius R_M and assigns each substructure a numerical identifier. In this study, the SMILES of the repeat unit for each homopolymer were utilized and the fingerprint algorithm was implemented in RDKit with R_M equal to 3. In total, 8831 substructures were detected, but only 129 prominent substructures shared by most polymers were kept for prediction of the values of ρ, E, and σ_y, to reduce the number of dimensions of the input vectors. Finally, for each polymer, vectors were obtained in which each bit represents the presence of a detected substructure. Compared to the default Morgan fingerprints, the improved Morgan fingerprints in this study take into account the frequency of occurrence for each substructure, which carries more physical meaning. This input vector is found to be a proper representation as it indicates both the substructures present and their frequency in the polymer's repeating unit.[42]

Using FFNNs with the improved Morgan fingerprints, two single-task ML models were built for each property individually. An ensemble model that averages the predictions of 12 models was utilized to obtain better prediction performance. Specifically, the single-task ML model for tensile modulus was optimized to have a hidden layer, which has 40 neurons, and the single-task ML model for tensile yield stress was optimized to have four hidden layers with 8, 8, 8, and 16 neurons, respectively. The architecture of these single-task ensemble models can be found in Supporting Information, SI, Figures S1 and S2.

Multitask learning can be applied effectively in the field of polymer informatics due to the fact that there are many properties associated with polymers. Even if the amount of experimental data for a particular property is relatively small, a combination of multiple properties can still form a large data set.[46] In a recent study, Gurnani et al. proposed the use of Multitask Graph Neural Network (MGNN) to predict 23 different properties of polymers, including mechanics-related

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

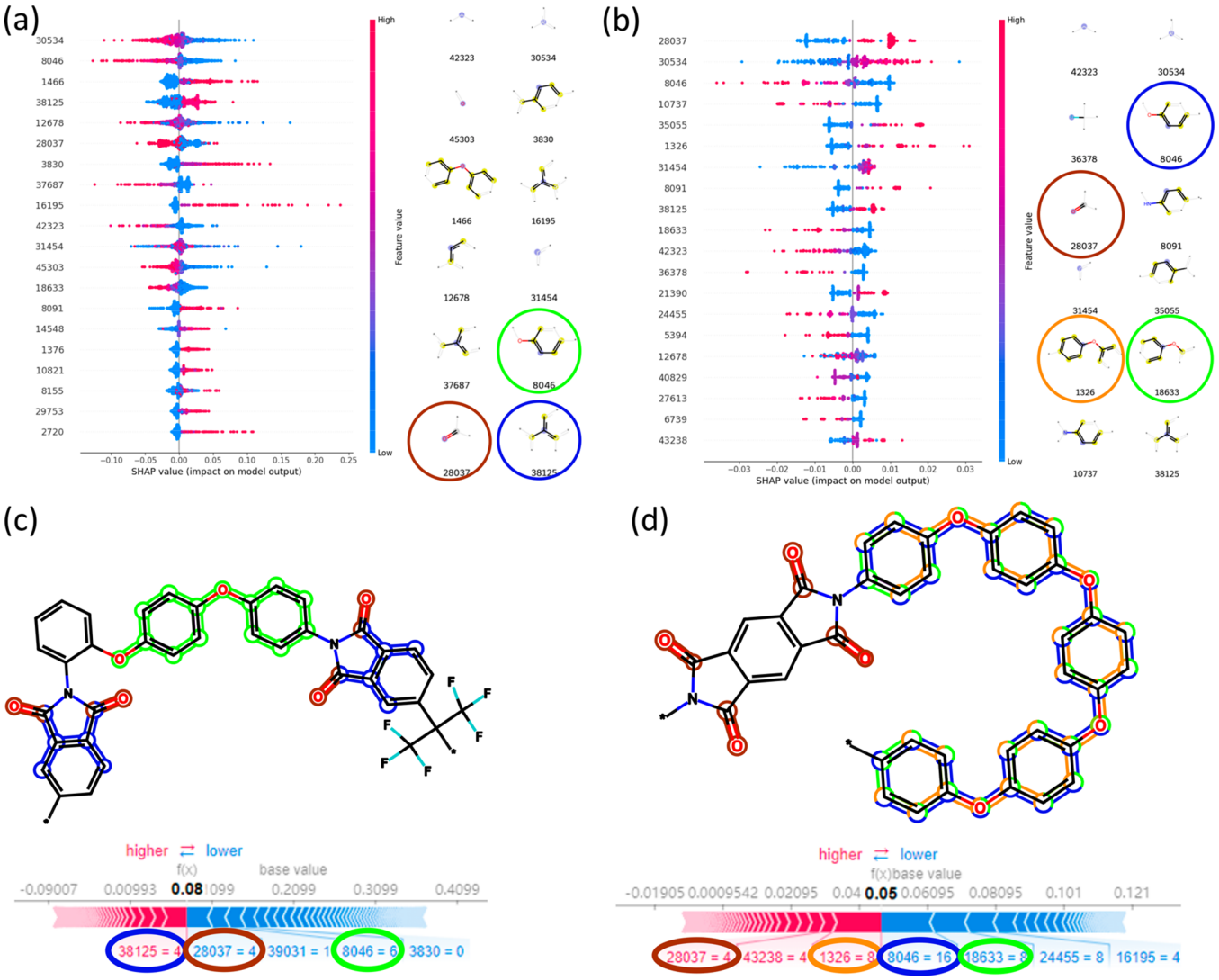


Figure 4. Substructure importance plot for (a) E and (b) σ_y. It lists the most important substructure in descending order and each dot represents the impact from a particular sample in the training set and the most important 12 substructures associated with E and σ_y according to SHAP values. The central atom of the substructures is highlighted in blue. Aromatic atoms are highlighted in yellow. Atoms' connectivity is highlighted in light gray. (c) The individual SHAP value plot for one of the smallest E polymers, and (d) the individual SHAP value plot for the highest σ_y polymers. Red or blue arrows indicate the positive or negative contributions of substructures. 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. The top substructures in this polyimide are highlighted in different colors.

properties such as Young's modulus and tensile strength.[47] MGNN is the first protocol to integrate polymer feature learning from SMILES strings and other relevant features, invariant transformations, data augmentation, and multitask learning. Benchmarking shows that this network has superfast polymer fingerprint recognition and accurate property prediction for the most comprehensive chemical and property arrays studied to date. In our work, we focus on the mechanical properties of polymers, especially E and σ_y. Therefore, we did not include a large number of other polymer properties in the multitask network. Instead, we only used these two properties and seven mechanical properties that are related to E and σ_y in the multitask learning, which is very different from the MGNN study.

An enlarged data set is used for training the ML model, a total of 923, 230, 1009, 19, 6, 6, 67, 6, and 58 real polymers were used with reported experimental values of E, σ_b, G, τ_b, E_{flex}, σ^y_{flex}, and σ^b_{flex}, respectively (see SI Figure S3). As shown in Figure 2(b), the MF algorithm can find and identify all substructures in a monomer. After that, based on the presence of each substructure, all substructures will be marked in a bit vector, which is used for prediction values of E and σ_y. The multitask machine learning model was optimized with three hidden layers, each containing 128, 16, and 64 neurons, respectively (see SI Figure S4 for the architecture of the multitask model). All single-task and multitask ML models in this study were implemented using the Keras package.

The effectiveness of the single-task ML models is demonstrated in Figure 3(a,b). E was determined using a single-task ML model, which was trained using 95% of the 923 real polymers with reported experimental values of tensile modulus. The remaining 5% of data points were used as a test set. R^2 of 0.96 and 0.84 are obtained for the model training and validation, respectively. For σ_y learned through a single-task ML model, there are 230 real polymers with a reported experimental value, and 90% of the data points are selected

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

pseudorandomly for the training set, the other 10% of data points are used for the testing set. Following the same training process as the model for E, the R^2 0.95 and 0.80 for the model training and validation, respectively. The averages of the prediction values generated by 12 models in the ensemble are shown in dots, while the variances from the prediction values of 12 models are illustrated with error bars.

As shown in Figure 3(c,d), the multitask ML model training and validation for E and σ_y obtained R^2 values of 0.99, 0.85, 0.97, and 0.88, respectively. These high R^2 values indicate accurate predictions for these two attributes. It should be noted that due to the small size of the data set for σ_y, which contains only 230 samples, the test set is very small and lacks statistical significance.

As a result, the obtained models were utilized for high-throughput screening of real polymers without reported values of E and σ_y in the PoLyInfo data set, as well as 8 million hypothetical polyimides mentioned in Section 2.1.

Using the established predictive model, we aim to gain insight into the effect of each substructure feature on the final property. To accomplish this, we will calculate SHapley Additive exPlanations (SHAP)[48] values based on single-task ML models to determine the impact of substructures on E and σ_y. The top substructures that have the most influence on E, as determined by the model's output, are shown in Figure 4(a). Figure 4(a) also shows the substructures and their corresponding SHAP values for different polymers. Each row in the figure represents a substructure, and the dots along the same row indicate the SHAP value of that substructure for different polymers. For example, the SHAP values of one of the polymers with the lowest E are highlighted in Figure 4 (c) for further examination. The feature values for the important substructures are presented in Figure 4(a), with one substructure that contributes the most to the high E highlighted in the molecular graph using colored circles. This highlights the impact of the key substructures revealed by the model. The results show that the developed ensemble model not only has good predictive accuracy but also provides clear physical explanations. Similarly, the important substructures for σ_y can be found in Figure 4(b) and one of the polymers with the highest σ_y is highlighted in Figure 4(c) for further examination. The examination of the SHAP values for E and σ_y in Figure 4(a) and 4(b), respectively, shows that some of the key substructures for σ_y match with the top key substructures for E. This suggests that these substructures have a significant impact on both properties and may be important design considerations when developing polymers with high σ_y and E. As seen in Figure 4(a,b), substructures 42332, 30534, 31454, 8046, 28037, and 38125 have been identified as important for both E and σ_y predictions. This makes sense given the correlation between these two properties and highlights the importance of these substructures in the overall mechanical performance of polymers. These substructures may be crucial design considerations when developing polymers with high mechanical properties.

Moreover, SHAP technique was not only used to identify the substructures involved in their mechanical properties but also to provide some physicochemical insights related to their molecular structures and organizations through feature important analysis (SHAP values). Specifically, this technique involves analyzing the relative contribution or importance of different substructures or molecular features in the polymer molecule toward a specific mechanical property. For instance, as observed from Young's modulus in Figure 4c, the substructure "38125" can improve Young's modulus but the substructure "28037" can lower this property. However, for yield strength in Figure 4d, the yield strength can be enhanced by substructure "1326" and weakened by substructure "8046". Clearly, the mechanical properties of polymers exhibit the strongest correlation with specific molecular substructures or motifs in chemical space. These substructures can be identified through their chemical functionality, physical arrangement, and position in the polymer backbone, offering valuable insights into the fundamental relationships between the structure and mechanical properties of polymers.

In addition, as presented in Figure 5, the average feature importance of each substructure for these properties is also

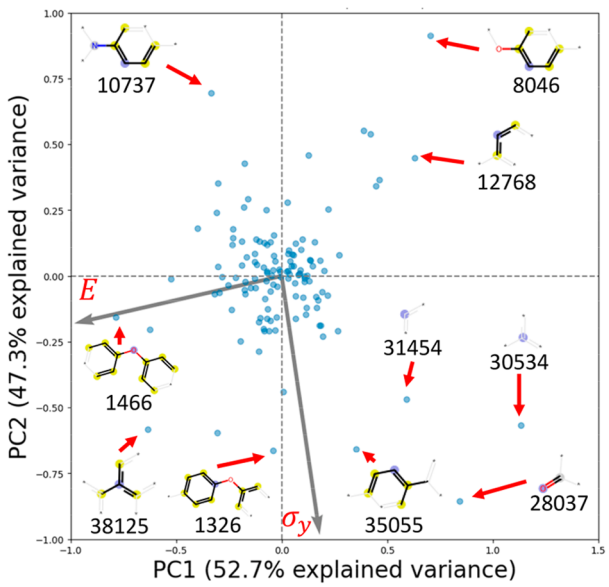


Figure 5. PCA analysis on feature importance for two properties of 129 substructures.

visualized in a principal component analysis (PCA) plot, to illustrate the influence of the 129 substructures on the properties. This provides an indication of the overall contributions of different substructures toward different properties. From the PCA plot of the two properties, it can be observed that the first two principal components (PC1 and PC2) explain most of the variance in the data, offering general guidelines. However, most substructures are located near the origin, indicating the difficulty in differentiating their contributions to different properties. The property σ_y is nearly orthogonal to the other property E, indicating the challenge of adjusting σ_y while adjusting E in a polymer. Key substructures such as "30534" and "28037" that are critical for both properties are located far away from the origin, demonstrating their high impact on the overall performance of a polymer. Compared to the feature importance analysis for a single property, the PCA analysis provides a more comprehensive evaluation of how different substructures are correlated with different properties, allowing for a better understanding of how each substructure affects the polymer's properties.

In addition to using the single-task ML models and the multitask ML model, Evidential Deep Learning (EDL) models[49] have also been used as a benchmark in the study. EDL models provide a way to estimate the uncertainty in the

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

predictions and can be useful for applications where high accuracy is critical. By comparing the performance of the EDL models with that of the other models, it is possible to gain insights into the strengths and limitations of each approach and determine which one is best suited for a particular application. EDL is a powerful machine learning model that can quantify uncertainty for neural network-based molecular structure−property predictions. The EDL model can be trained with a small-scale data set and provides accurate predictions and uncertainty, as shown in SI Figure S5 for the univariate distribution data set as well as the parity plot. However, when the data sets of E and σ_y are used for model training, the EDL models do not show good results, as seen in SI Figure S6 for the parity plots. This may be due to the limitations of the EDL models when dealing with large and complex data sets, or the specific characteristics of the E and σ_y data sets used in the study.

2.3. MD Simulation. All MD simulations in this work are carried out employing the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package.[50] The parametrized polymer consistent force field (PCFF)[51−53] is used to describe the interatomic potentials, which have been extensively applied to calculate materials' properties, such as compressibility, elastic constant, mechanical metrics, thermal properties, and so forth. Nonbonded interactions, including LJ and Coulomb potentials. LJ potential is characterized by the 6/9 types. The six-power mixing criteria[54] are used to calculate the interaction force between different types of atoms. A cutoff distance of 12.0 Å is employed to truncate both van der Waals and electrostatic interactions. The long-range electrostatic interactions are calculated using particle−particle particle−mesh (PPPM) solver, which has a force tolerance of 10^{-4}.

With regard to homopolymers, we first generate a single polymer chain by connecting the head atom with the tail atom of the repeat unit. Each polymer chain has almost 2000 atoms. Energy minimization and geometry optimization are conducted for each chain. On the basis of the optimized single polymer chain, 100 polymer chains are randomly filled into a 3D amorphous box using the self-avoiding random walks method. In a Monte Carlo way, polymer chain configurations are reset during the inserting process so that realistic contacts between atoms are achieved. After that, the initial 3D simulation model for the homopolymer is developed. The detailed procedure for building MD homopolymer models is described in our previous works.[42,55] Then, 40 annealing cycles are performed for the generated homopolymer models to eliminate the system's thermal stress. The annealed homopolymer configurations are used to conduct the following process.

All MD simulations are controlled by the periodic boundary conditions and a 21-step[56] equilibration procedure is implemented to relax each system. During this equilibration process, system relaxation conducts alternatively between high-temperature and high-pressure under isothermal−isobaric ensembles (NPT) and canonical ensembles (NVT) to obtain the full equilibration configuration. The maximum value for temperature and pressure can arrive at 1000 K and 50 000 atm, respectively. After the system is cyclically compressed and stretched, a realistic polymer configuration can be obtained. 0.1 fs is employed for the time step in the whole simulation. The parameters for temperature damping and pressure damping are governed by 100 and 1000 timesteps, respectively. After the relaxing process, the optimized homopolymers are employed to perform the subsequent calculations in mechanical properties.

Prior to conducting mechanical properties, a further relaxation under the NPT ensemble is implemented over 1 ns at 300 K and 0.0 MPa according to the optimized equilibrium configuration above. Then, employing a 0.2 fs time step, a stress-stain loading process is used to tensile the system along the x-direction at 300 K through NEMD simulations. The x-dimension of the simulation box was stretched under the $N\sigma_{ij}\varepsilon_{ij}T$ ensemble during each loading step until an approximate engineering tensional strain realizes 25%. Besides, pressures of the x and y directions are maintained at 0.0 MPa during the tension process, enabling the two dimensions to unrestrained change to attain the Poisson's effect. A constant strain rate of 10^8 s^{-1} is adopted for a stress−strain response during tensile simulation. Polymers' mechanical properties including E, and σ_y are then estimated according to the produced stress−strain curves.

To demonstrate the precision of MD simulations in the prediction of polymers' mechanical properties, we first benchmark the MD simulations by simulating 10 experimental homopolymers and comparing the measured mechanical properties in the experiments with our calculation results. The monitoring stress−strain curves and the extracted E and σ_y for the 10 simulated polymers are shown in SI Figures S7 and S8, respectively. These investigations indicated that physics-based MD simulations can give high-fidelity mechanical properties, allowing us to use MD results to verify the ML predictions.

From MD's perspective, the identified polymers have high modulus of resilience, corresponding to a high yield strength and a low Young's modulus. Observed the stress−strain curves in Figure S10, these polymers presented as a relatively high yield point at a large loading strain location, approximately 0.15. That is to say, these polymers with a high modulus of resilience typically exhibit a relatively low slope in their stress−strain curves as well as high level of external force resistance, indicating a low stiffness or modulus of elasticity. This means that these polymers can undergo significant elastic deformation without permanent deformation or plastic yielding, effectively storing elastic energy during loading. Meanwhile, it has been observed that the remarkable modulus of resilience, characterized by high yield strength and low Young's modulus, is closely correlated with their unique substructures through bridging molecular dynamics (MD) and machine learning (ML) findings (Figure 4).

3. RESULTS AND DISCUSSIONS

3.1. Discovery of High Modulus of Resilience Polymers through Single-Task ML Models. Single-task and multitask ML models can be used to predict the specific properties of new polymers. This is a useful tool for scanning a large design domain that includes two different data sets with approximately 8 million polymers. By applying well-trained ML models to the entire data set, we can estimate the E and σ_y of the polymers, allowing us to identify superior performers.

Our focus is on polymers with high σ_y and low E properties. To verify this, we have selected 10 real polymers with relatively high σ_y and relatively low E for MD verification. These high-performance polymers may not have both properties reported, but we have obtained reasonable estimations of their properties using Single-task ML models, which were used for screening and prediction. On the basis of the tracked stress−strain curves

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

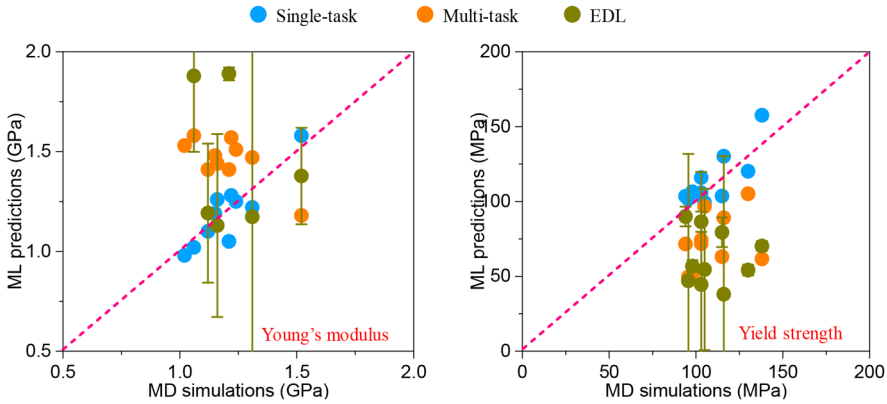


Figure 6. Parity plots of MD simulations vs ML prediction for E and σ_y of 10 novel real polymers selected based on single-task ML models.

Table 1. Comparison of ML Predictions and MD Simulations of 10 Novel Real Polymers

polymer	mechanical properties	MD	ML for single-task	diff[a] (%)	ML for multitask	diff[a] (%)
No. 1 *C=CCC(C*)(c1ccccc1)c1ccccc1	E (GPa)	1.52	1.4	−7.9	1.18	−22.4
	σ_y (MPa)	138	157.61	14.2	61.68	−55.3
No. 2 *C=CCC(C*)c1ccccc1	E (GPa)	1.22	1.19	−2.5	1.57	28.7
	σ_y (MPa)	130	120.2	−7.5	105.12	−19.1
No. 3 *N1C(=O)c2c(C1=O)cc(c(c2)c1ccc(cc1)[Si](C)(C)C)c1cc2c(C(=O)N(C2=O)c2ccc(cc2)Oc2ccc(cc2)*)cc1c1ccc(cc1)[Si](C)(C)C	E (GPa)	1.31	1.07	−18.3	1.47	12.2
	σ_y (MPa)	103	115.97	12.6	71.89	−30.2
No. 4 *c1cc2c(C(=O)N(C2=O)c2cc(cc(c2)C(=O)OC)N2C(=O)c3c(C2=O)cc(c(c3)Cl)C(=O)*)cc1Cl	E (GPa)	1.15	1.09	−5.2	1.48	28.7
	σ_y (MPa)	105	99.20	−5.5	96.67	−7.9
No. 5 *S(=O)(=O)c1ccc(cc1)Nc1ccc(cc1)C1(c2ccccc2-c2ccccc12)c1ccc(cc1)Nc1ccc(cc1)*	E (GPa)	1.16	1.05	−9.5	1.44	24.1
	σ_y (MPa)	94	103.52	10.3	71.58	−23.9
No. 6 *C(=CCCC*)c1ccccc1	E (GPa)	1.12	1.02	−8.9	1.41	25.9
	σ_y	98	106.40	8.6	50.15	−48.8
No. 7 *n1nc(nn1)c1ccc(cc1)c1nn(nn1)C(=Nc1ccccc1)c1ccc(cc1)C(=Nc1ccccc1)*	E (GPa)	1.02	0.90	−11.8	1.53	50.0
	σ_y (MPa)	103	105.79	2.7	74.34	−27.8
No. 8 *c1cc2c(C(=O)N(C2=O)c2cc(ccc2)N2C(=O)c3c(C2=O)cc(c(c3)Cl)C(=O)*)cc1Cl	E (GPa)	1.24	1.09	−12.1	1.51	21.8
	σ_y (MPa)	96	101.60	6.3	49.54	−48.2
No. 9 *c1c(nnc(n1)c1nc(ccc1)c1nc(c(nn1)c1ccccc1)c1ccc(cc1)C(=O)c1ccc(cc1)*)c1ccccc1	E (GPa)	1.06	1.01	−1.6	1.58	49.1
	σ_y (MPa)	116	130.15	12.2	89.09	−23.2
No. 10 *C#CC(=C(*)Cn1cnc2c1cccc2)Cn1cnc2c1cccc2	E (GPa)	1.21	1.14	−5.8	1.41	16.5
	σ_y (MPa)	115	103.64	−9.9	63.13	−45.1

[a]Notably, Diff = (ML − MD)/MD.

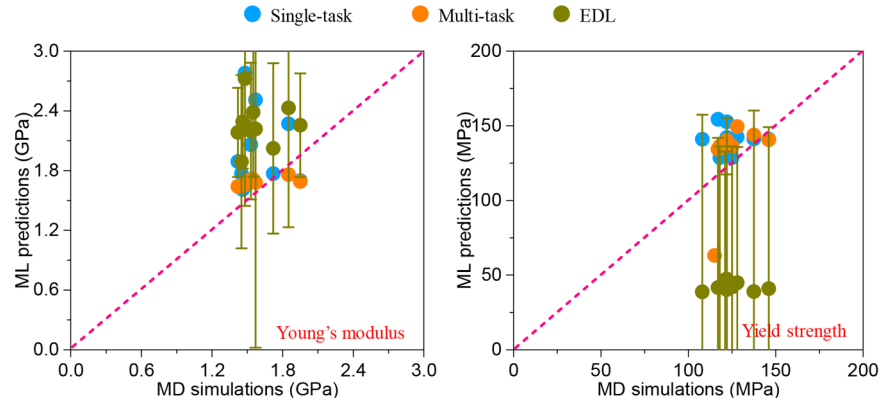


Figure 7. Parity plots of MD simulations vs ML prediction for E and σ_y of 10 novel hypothetical polyimides selected based on multitask ML models.

in Figure S9 and three ML models, Figure 6 compares the MD simulated properties of the ten real polymers with their estimations given by single-task ML models, multitask ML models, and EDL models. The stress−strain curves can be

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

Table 2. Comparison of ML Predictions and MD Simulations of 10 Novel Hypothetical Polyimides

polyimide	mechanical properties	MD	ML for single-task	diff (%)	ML for multitask	diff (%)
No. 11 *c1ccc(c2cc3nc(N4C(=O)c5 ccc(C(=O)c6ccc7c(c6)C(=O)N(*)C7=O)cc5C4=O)ccc3 [nH]2)cc1	E (GPa)	1.85	4.29	132.0	1.76	−4.9
	σ_y (MPa)	128	142.47	11.3	149.42	16.7
No. 12 *c1ccc2nc(c3cccnc3N3C(=O) c4ccc(C(=O)c5ccc6c(c5)C(=O)N(*)C6=O)cc4C3=O) [nH]c2c1	E (GPa)	1.53	4.60	200.8	1.71	11.7
	σ_y (MPa)	137.5	141.21	2.7	143.85	4.6
No. 13 *c1cnc(c2cc3cc(N4C(=O)c5ccc (C(=O)c6ccc7c(c6)C(=O)N(*)C7=O)cc5C4=O)ccc3[nH]2)nc1	E (GPa)	1.95	4.12	111.3	1.69	−13.3
	σ_y (MPa)	146	141.06	−3.4	140.68	−3.6
No. 14 *c1ccc(c2c3ccc(N4C(=O)c5ccc (C(=O)c6ccc7c(c6)C(=O)N(*)C7=O)cc5C4=O)cc3nc3ccc(C)cc23) cc1	E (GPa)	1.57	3.43	118.6	1.68	7.0
	σ_y (MPa)	122	142.16	16.5	140.12	14.9
No. 15 *c1ccc(c2cc3nc(N4C(=O)c5ccc(c6ccc7c(c6)C(=O)N(*)C7=O)cc5C4=O)ccc3[nH]2)cc1	E (GPa)	1.48	4.23	186.0	1.66	12.2
	σ_y (MPa)	121	138.58	14.5	137.42	13.6
No. 16 *c1ccccc1c1nc2ccc(N3C(=O)c4ccc(C(=O)c5ccc6c(c5)C(=O)N(*)C6=O)cc4C3=O)cc2o1	E (GPa)	1.45	3.18	119.2	1.65	13.8
	σ_y (MPa)	125	128.71	3.0	136.4	9.2
No. 17 *c1ccccc1c1nc2cc(N3C(=O)c4ccc(C(=O)c5ccc6c(c5)C(=O)N(*)C6=O)cc4C3=O)ccc2o1	E (GPa)	1.72	3.18	84.8	1.65	−4.1
	σ_y (MPa)	118	128.71	9.1	136.35	15.6
No. 18 *c1ncnc2ccc(c3cccnc3N3C (=O)c4ccc(C(=O)c5ccc6c(c5)C(=O)N(*)C6=O)cc4C3=O)cc12	E (GPa)	1.46	3.21	120.2	1.65	13.0
	σ_y (MPa)	108	140.97	30.5	135.99	25.9
No. 19 *c1ccc(c2nc3cc(N4C(=O)c5ccc (C(=O)c6ccc7c(c6)C(=O)N(*)C7=O)cc5C4=O)ncc3[nH]2)cc1	E (GPa)	1.55	4.94	218.7	1.71	10.3
	σ_y (MPa)	122	152.65	25.1	140.2	14.9
No. 20 *c1ccc2[nH]c(c3ccc(N4C(=O)c5 ccc(C(=O)c6ccc7c(c6)C(=O)N(*)C7=O)cc5C4=O)cc3C)nc2n1	E (GPa)	1.42	4.02	183.2	1.64	15.5
	σ_y (MPa)	117	154.32	31.9	134.34	14.8

found in Figure S9. The agreement between MD simulations and true values can be assessed. It can be seen that the MD simulated E and σ_y are comparable to those estimated by single-task ML models, but there is a lower level of comparability between MD simulation and multitask ML model predictions when compared to single-task ML models. A comparison of the calibrated MD results and the corresponding ML predictions can be found in Table 1. The average RMSE plus/minus one standard deviations for the single-task models are 0.12 ± 0.06 GPa and 11.03 ± 10.00 MPa for E and σ_y, respectively. The average RMSE plus/minus one standard deviations for the multitask model are 0.34 ± 0.23 GPa and 40.78 ± 18.31 MPa for E and σ_y, respectively.

3.2. Discovery of High Modulus of Resilience Polyimides through Multi-Task ML Model. In order to find high-performance polymers in a larger space, a hypothetical data set of more than 8 million possible polyimides is screened, and the multitask ML model is used for the measurement of its performance in this section. This allows us to select 10 potential hypothetical polyimides with relatively high σ_y and relatively low E for MD verification. Through the stress−strain curves shown in Figure S10 and three ML models, Figure 7 compares the MD simulated properties of these ten hypothetical polyimides with their multitask ML estimations, as well as estimations from single-task ML models and EDL models. Similarly, to the previous analysis, the MD simulated E and σ_y are comparable to the multitask ML model predictions, but there are significant gaps between estimations of E and σ_y from the single-task models and the MD simulations. A comparison of the calibrated MD results and the corresponding ML predictions can be found in Table 2. The average RMSE plus/minus one standard deviations for the single-task models are 2.40 ± 0.60 GPa and 21.24 ± 13.21 MPa for E and σ_y, respectively. The average RMSE plus/minus one standard deviations for the multitask model are 0.18 ± 0.16 GPa and 17.34 ± 8.66 MPa for E and σ_y, respectively.

3.3. Discussion. Remarkably, these identified polymers break the barrier of the trade-off between two mechanical properties, Young's modulus, and yield strength. The mechanical properties of polymers are closely linked to their substructures in the chemical space, as governed by physicochemical considerations. Figure 4 illustrates that certain substructures, such as aromatic rings and carbonyl groups, can significantly influence the yield strength and Young's modulus of polymers. These substructures can enhance the material's yield strength while reducing its Young's modulus, ultimately resulting in a high modulus of resilience. This demonstrates how the substructural composition of polymers impacts their mechanical properties, underscoring the intricate relationship between polymer chemistry and mechanical performance. Likewise, from the perspective of mechanics, the existing aromatic rings in the polymers are highly stable, provide strong covalent bonding, and enhance additional intermolecular interactions, which withstand significant levels of external force without permanent deformation, contributing to the high yield strength of polymers. These strong covalent bonds make it difficult for the polymer chains itself to slide or deform under stress, leading to high yield strength. However, the carbonyl groups in the polymers allow for greater molecular mobility, making the polymer chains more flexible during loading, hence lowering Young's modulus. As a result, these identified polymers have successfully overcome the traditional trade-off between the two mechanical properties, breaking the barrier and demonstrating a high modulus of resilience with both high yield strength and low Young's modulus. The key substructures of 20 promising candidates can be found in SI Table S1.

As shown in Section 3.1, the predictions of the single-task ML models match MD simulations better than those of the multitask ML model for both E and σ_y. However, it is important to note that the polymers used for testing in this section were specifically selected by the single-task ML models due to their extreme properties. The goal of our study is to find polymer materials with better performance than existing ones, which means there must be a lack of data in the target searching space to train. For example, the values of E for the selected polymers in Section 3.1 are around 1.0 to 1.5, but

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

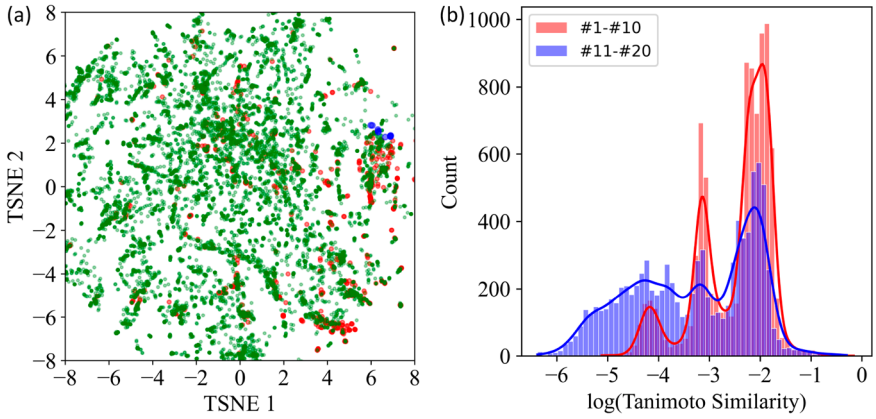


Figure 8. (a) Chemical space visualization of the real polymer data set with reported E (red points), the whole real polymer data set (green points), and the 10 promising hypothetical polyimides (blue points). (b) Histogram plot of the pairwise Tanimoto similarity (Tc) between 20 promising polymers and the real polymer data set with reported E. Two polymers are the same if Tc = 1 and totally different if Tc = 0.

there are only about 100 data points with reported E values between 1.0 and 1.5 out of 923 in the data set, and most of them are not in the target searching space because of a low σ_y. In this situation, it is not fair to compare these two types of ML models with the promising polymers searched by one of them. The results shown in Section 3.2 also support this point. In Section 3.2, the predictions of the multitask ML models match MD simulations better than those of the single-task ML model for both E and σ_y. The polyimides used for testing in Section 3.2 were specifically selected by the multitask ML models due to their extreme properties.

In summary, our study demonstrates that both single-task and multitask ML models can be used to discover high-performance polymers, with predictions of the models having a high level of comparability to MD simulations. However, the lack of data in the target searching space can lead to inaccurate estimates for some polymers. Additionally, the results suggest that the bias of the data set used for training the models can have an impact on the predictions. Overall, the study highlights the potential of ML models for discovering promising polymers, but also emphasizes the need for a larger and more diverse data set to improve the accuracy of predictions.

In Section 3.2, it is evident that there are significant differences between the predictions of single-task ML model and MD simulations for E. The input dimension of single-task models is restricted by the size of the training data set. To reduce the number of parameters in the single-task models and match the small-scale data set, only 129 of the most prevalent substructures were used, as shown in Section 2.2. This leads to a weaker generalization capability compared to the 2048-dimension vectors used in multitask model. Figure 8(a) illustrates the chemical space visualization of a real polymer data set with reported E values (red points), the entire real polymer data set (green points), and 10 promising hypothetical polyimides (blue points). T-distributed Stochastic Neighbor Embedding (TSNE) is used for embedding high-dimensional data into two spaces.[57] TSNE is a popular nonlinear dimensionality reduction and data visualization technique, which can preserve nonlinear similarities between data points. TSNE works by first calculating the similarity between high-dimensional data points based on a Gaussian distribution, then calculating the similarity between data points in the low-dimensional space based on a t-distribution, and minimizing the difference between high-dimensional and low-dimensional similarities. It can be observed that the 10 hypothetical polyimides are structurally similar to each other and are situated in a relatively small area. As shown in Figure 8(b), the pairwise similarity coefficient, the Tanimoto coefficient (Tc), was calculated between 10 promising real polymers selected from the PoLyInfo data set and the training data set with reported E values, as well as 10 promising hypothetical polyimides and the training data set with reported E values. Tc was used to measure the structural similarity between polymers by comparing their fingerprints pairwise. The coefficient of similarity between two polymers is determined by the ratio of common substructures to the total number of substructures present in both. It can be found that compared with ten promising real polymers, the ten promising hypothetical polyimides have lower similarities to the training data set. And as for 10 promising hypothetical polyimides, a Tc of 0.04, indicates that the hypothetical polyimides are dissimilar to the data set used for model training. This lack of similarity is reflected in the poor accuracy results shown in Table 2 (see the SI Figure S11 for the K-means clustering analysis).

4. CONCLUSIONS

Highly resilient materials are in high demand for numerous cutting-edge technologies. To identify better-performing and more promising polymers, we established structure−property relationships through predictive ML models. This allowed us to conduct high-throughput screening of both the PoLyInfo and 8 million hypothetical polyimides data sets. In our study, we collected 10 properties, including ρ, E, σ_y, σ_b, G, τ_y, τ_b, E_{flex}, σ^y_{flex}, and σ^b_{flex}. Among these properties, we focused on two key mechanical properties, E and σ_y. To accomplish this, we built and trained single-task ML models, multitask ML models, and EDL models based on the reported property values. Using the explainable machine learning models, we were able to determine the crucial substructures that significantly affect the mechanical properties of polymers, such as E and σ_y. This information can be utilized to design and manufacture new polymers with enhanced elastic resilience.

By utilizing well-trained machine learning models on the PoLyInfo and 8 million hypothetical polyimides data sets, we identified novel polymers that simultaneously exhibit high σ_y and low E. Through the single-task ML models, we selected ten of the best-performing real polymers. The multitask ML

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653

models were used to discover ten promising hypothetical polyimides. These polymers have a high potential for use in various applications that require high strength and flexibility. These novel polymers show consistent key substructures that contribute to their outstanding performance, surpassing the Ashby frontier of existing polymers as confirmed by MD simulations.

Moreover, the methodology used in this study, which combines explainable machine learning techniques and high-fidelity molecular dynamics simulations, has proven to be a highly efficient approach for handling a vast number of chemical structures. Meanwhile, the neural network demonstrated in this study has shown a promising predictive ability for polymer's mechanical properties, which suggests that it could be applied to address more intricate issues, including the prediction of mechanical properties for branched or cross-linked polymers. Both types of these polymers exhibit different mechanical properties compared to their linear counterparts, and understanding their structure−property relationships is crucial for designing new polymeric materials with specific mechanical functions. The results also demonstrate the influence of data set bias on the predictions of machine learning models. To enhance the accuracy of future predictions, techniques such as active learning can be applied.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00131.

Details on structures of single-task and multitask ML models, pairwise relationships in the data set, parity plots of the EDL models, stress−strain curves using MD simulations, parity plots of experiment vs MD simulations, K-means clustering results, and Kay substructures of 20 promising polymers (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ying Li − *Department of Mechanical Engineering, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0002-1487-3350; Phone: +1 608-265-0577; Email: yli2562@wisc.edu; Fax: +1 608-890-3966

Authors

Tianle Yue − *Department of Mechanical Engineering, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Jinlong He − *Department of Mechanical Engineering, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Lei Tao − *Department of Mechanical Engineering, University of Connecticut, Storrs, Connecticut 06269, United States;* orcid.org/0000-0002-8285-1356

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00131

Author Contributions

T.Y.: Methodology, software, validation, formal analysis, data curation, writing−original draft, visualization. J.H.: Methodology, software, validation, formal analysis, data curation, writing−review and editing, visualization. L.T.: Methodology, software. Y.L.: Conceptualization, methodology, software, writing−review and editing, supervision, project administration, and funding acquisition.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Air Force Office of Scientific Research through the Air Force's Young Investigator Research Program (FA9550-20-1-0183; Program Manager: Dr. Ming-Jen Pan), Air Force Research Laboratory/UES Inc. (FA8650-20-S-5008, PICASSO program), and the National Science Foundation (CMMI-2314424, CMMI-2316200, and CAREER-2323108). Y.L. would also like to thank the support from 3M's Non-Tenured Faculty Award. Any opinions, findings, 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 or National Science Foundation. The authors also acknowledge the National Renewable Energy Laboratory (Eagle Computing System) for providing HPC resources that have contributed to the research results reported within this paper. Support for this research was also provided by the University of Wisconsin−Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

■ REFERENCES

(1) Monroe, B.; Oliviere, D. *Resilience in Palliative Care: Achievement in Adversity*; OUP Oxford: Oxford, 2007.

(2) Dusoe, K. J.; Ye, X.; Kisslinger, K.; Stein, A.; Lee, S.-W.; Nam, C.-Y. Ultrahigh elastic strain energy storage in metal-oxide-infiltrated patterned hybrid polymer nanocomposites. *Nano Lett.* **2017**, *17* (12), 7416−7423.

(3) Zhou, J.; Deng, X.; Yan, Y.; Chen, X.; Liu, Y. Superelasticity and reversible energy absorption of polyurethane cellular structures with sand filler. *Compos. Struct.* **2015**, *131*, 966−974.

(4) Hill, F. A.; Havel, T. F.; Lashmore, D.; Schauer, M.; Livermore, C. Storing energy and powering small systems with mechanical springs made of carbon nanotube yarn. *Energy* **2014**, *76*, 318−325.

(5) Rossi, F.; Castellani, B.; Nicolini, A. Benefits and challenges of mechanical spring systems for energy storage applications. *Energy Procedia* **2015**, *82*, 805−810.

(6) Berkemeier, M. D.; Desai, K. V. Design of a robot leg with elastic energy storage, comparison to biology, and preliminary experimental results. In *Proceedings of IEEE International Conference on Robotics and Automation, 22−28 April 1996*, 1996; Vol. *1*, pp 213−218 vol. 211 DOI: 10.1109/ROBOT.1996.503597.

(7) Costescu, D. M.; Hadăr, A.; Pastramă, Ş. D. Design of an element used to absorb the impact energy for a railway transport vehicle. *Ann. Acad. Romanian Scientists, Ser. Eng. Sci.* **2019**, *11* (1), 17−32.

(8) Scarfogliero, U.; Stefanini, C.; Dario, P. The use of compliant joints and elastic energy storage in bio-inspired legged robots. *Mech Mach Theory* **2009**, *44* (3), 580−590.

(9) Pal, A.; Goswami, D.; Martinez, R. V. Elastic energy storage enables rapid and programmable actuation in soft machines. *Adv. Funct. Mater.* **2020**, *30* (1), 1906603.

(10) Rodriguez, R. D.; Shchadenko, S.; Murastov, G.; Lipovka, A.; Fatkullin, M.; Petrov, I.; Tran, T.-H.; Khalelov, A.; Saqib, M.; Villa, N. E.; Bogoslovskiy, V.; Wang, Y.; Hu, C.-G.; Zinovyev, A.; Sheng, W.; Chen, J.-J.; Amin, I.; Sheremet, E. Flexible Electronics: Ultra-Robust Flexible Electronics by Laser-Driven Polymer-Nanomaterials Integration (Adv. *Funct. Mater. 17/2021). Adv. Funct. Mater.* **2021**, *31* (17), 2170114.

(11) Ji, Z.; Zhang, X.; Yan, C.; Jia, X.; Xia, Y.; Wang, X.; Zhou, F. 3D printing of photocuring elastomers with excellent mechanical strength and resilience. *Macromol. Rapid Commun.* **2019**, *40* (8), 1800873.
(12) Council, N. R. *Polymer Science and Engineering: The Shifting Research Frontiers*; National Academies Press: Washington, D.C., 1994.
(13) Brinson, H. F.; Brinson, L. C. Polymer engineering science and viscoelasticity. *An introduction* **2008**, 99−157.
(14) Sawyer, L.; Grubb, D. T.; Meyers, G. F. *Polymer Microscopy*; Springer Science & Business Media: Berlin/Heidelberg, 2008.
(15) Namazi, H. Polymers in our daily life. *BioImpacts: BI* **2017**, *7* (2), 73.
(16) Lutz, J.-F.; Ouchi, M.; Liu, D. R.; Sawamoto, M. Sequence-controlled polymers. *Science* **2013**, *341* (6146), 1238149.
(17) Li, W.; Liu, J.; Zhao, D. Mesoporous materials for energy conversion and storage devices. *Nat. Rev. Mater.* **2016**, *1* (6), 1−17.
(18) Soroush, M.; Grady, M. C. Polymers, polymerization reactions, and computational quantum chemistry. In *Computational Quantum Chemistry*; Elsevier, 2019; pp 1−16.
(19) Balasubramanian, G.; Munshi, J.; Chen, W.; Chien, T. Towards Improving the Efficiency of Organic Solar Cells by Coarse-Grained Atomistic Modeling of Processing Dependent Morphologies. *Comput. Sci. Eng.* **2021**, *23* (3), 48−55.
(20) Hartmann, L.; Börner, H. Precision Polymers: Monodisperse, Monomer-Sequence-Defined Segments to Target Future Demands of Polymers in Medicine. *Adv. Mater.* **2009**, *21* (32−33), 3425−3431.
(21) Porel, M.; Alabi, C. A. Sequence-defined polymers via orthogonal allyl acrylamide building blocks. *J. Am. Chem. Soc.* **2014**, *136* (38), 13162−13165.
(22) Perry, S. L.; Sing, C. E. 100th anniversary of macromolecular science viewpoint: Opportunities in the physics of sequence-defined polymers. *ACS Macro Lett.* **2020**, *9* (2), 216−225.
(23) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bull.* **2016**, *41* (5), 399−409.
(24) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J. Materiomics* **2017**, *3* (3), 159−177.
(25) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **2017**, *3* (1), 1−13.
(26) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.
(27) Chen, G.; Shen, Z.; Iyer, A.; Ghumman, U. F.; Tang, S.; Bi, J.; Chen, W.; Li, Y. Machine-learning-assisted de novo design of organic molecules and polymers: opportunities and challenges. *Polymers* **2020**, *12* (1), 163.
(28) Lopez-Bezanilla, A.; Littlewood, P. B. Growing field of materials informatics: databases and artificial intelligence. *MRS Commun.* **2020**, *10* (1), 1−10.
(29) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine learning in materials discovery: Confirmed predictions and their underlying approaches. *Annu. Rev. Mater. Res.* 2020, *50* (1).
(30) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports* **2021**, *144*, 100595.
(31) Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **2021**, *6* (8), 655−678.
(32) Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Sci. Adv.* **2020**, *6* (43), No. eabc6216.
(33) Chandrasekaran, A.; Kim, C.; Venkatram, S.; Ramprasad, R. A deep learning solvent-selection paradigm powered by a massive solvent/nonsolvent database for polymers. *Macromolecules* **2020**, *53* (12), 4764−4769.
(34) Zhu, M.-X.; Song, H.-G.; Yu, Q.-C.; Chen, J.-M.; Zhang, H.-Y. Machine-learning-driven discovery of polymers molecular structures with high thermal conductivity. *Int. J. Heat Mass Transfer* **2020**, *162*, 120381.
(35) Nazarova, A. L.; Yang, L.; Liu, K.; Mishra, A.; Kalia, R. K.; Nomura, K.-i.; Nakano, A.; Vashishta, P.; Rajak, P. Dielectric Polymer Property Prediction Using Recurrent Neural Networks with Optimizations. *J. Chem. Inf. Model.* **2021**, *61* (5), 2175−2186.
(36) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59* (12), 5013−5025.
(37) Deacy, A. C.; Gregory, G. L.; Sulley, G. S.; Chen, T. T. D.; Williams, C. K. Sequence Control from Mixtures: Switchable Polymerization Catalysis and Future Materials Applications. *J. Am. Chem. Soc.* **2021**, *143* (27), 10021−10040.
(38) Patel, R. A.; Borca, C. H.; Webb, M. A. Featurization strategies for polymer sequence or composition design by machine learning. *Mol. Syst. Des. Eng.* **2022**, *7* (6), 661−676.
(39) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 7−9 Sept. 2011, *2011*; pp 22−29 DOI: 10.1109/EIDWT.2011.13.
(40) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202−D1213.
(41) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A; Thiessen, P. A; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102−D1109.
(42) Tao, L.; Varshney, V.; Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *J. Chem. Inf. Model.* **2021**, *61* (11), 5395−5413.
(43) Tao, L.; Chen, G.; Li, Y. Machine learning discovery of high-temperature polymers. *Patterns* **2021**, *2* (4), 100225.
(44) Chen, G.; Tao, L.; Li, Y. Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model. *Polymers* **2021**, *13* (11), 1898.
(45) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(46) Kuenneth, C.; Rajan, A. C.; Tran, H.; Chen, L.; Kim, C.; Ramprasad, R. Polymer informatics with multi-task learning. *Patterns* **2021**, *2* (4), 100238.
(47) Gurnani, R.; Kuenneth, C.; Toland, A.; Ramprasad, R. Polymer Informatics at Scale with Multitask Graph Neural Networks. *Chem. Mater.* **2023**, *35* (4), 1560−1567.
(48) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf.* **2017**, *30*.
(49) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **2021**, *7* (8), 1356−1367.
(50) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117* (1), 1−19.
(51) Heinz, H.; Lin, T.-J.; Kishore Mishra, R.; Emami, F. S. Thermodynamically Consistent Force Fields for the Assembly of Inorganic, Organic, and Biological Nanostructures: The INTERFACE Force Field. *Langmuir* **2013**, *29* (6), 1754−1765.
(52) Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T. An ab Initio CFF93 All-Atom Force Field for Polycarbonates. *J. Am. Chem. Soc.* **1994**, *116* (7), 2978−2987.
(53) Sun, H. Ab initio calculations and force field development for computer simulation of polysilanes. *Macromolecules* **1995**, *28* (3), 701−712.
(54) Sun, H. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B* **1998**, *102* (38), 7338−7364.

(55) Tao, L.; He, J.; Arbaugh, T.; McCutcheon, J. R.; Li, Y. Machine learning prediction on the fractional free volume of polymer membranes. *J. Membr. Sci.* **2023**, *665*, 121131.
(56) Abbott, L. J.; Hart, K. E.; Colina, C. M. Polymatic: a generalized simulated polymerization algorithm for amorphous polymers. *Theor. Chem. Acc.* **2013**, *132* (3), 1334.
(57) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *JMLR* **2008**, *9* (11), 2579−2605.

https://doi.org/10.1021/acs.jctc.3c00131
J. Chem. Theory Comput. 2023, 19, 4641−4653