SUPPLEMENTARY INFORMATION

PolyUniverse: Generation of a Large-scale Polymer Library Using Rule-Based Polymerization Reactions for Polymer Informatics

Tianle Yue,[1] Jianxin He, [1] Ying Li[1*]

[1]Department of Mechanical Engineering, University of Wisconsin-Madison, Madison, WI 53706, United States

*Corresponding author: yli2562@wisc.edu

Table S1. Number of small molecule compounds within the three datasets, including amino acids, cyclic olefins, epoxides, hydroxy carboxylic acids, lactams, lactones, poly carboxylic acids and acid halides, polyamines, polycarboxylic acid anhydrides, polyisocyanates, polyols and thiols, and vinylidenes.

Monomers	Count (GDB-13)	Count (GDB-17)	Count (PubChem)
Amino acid	4662	0	9253
Cyclic olefin	180414699	10357888	17133014
Epoxide	188757121	61644	4701127
Hydroxy carboxylic acid	194381403	6041868	4239342
Lactam	11465098	1129110	679436
Lactone	1898490	392362	1937567
Poly carboxylic acid and acid halide	9001267	1024466	1073964
Polyamine	6940	3095	542024
Polycarboxilic acid anhydride	14268280	13653	256271
Polyisocyanate	9606974	911417	1115284
Polyol and thiol	3961429	524249	2773404
Vinylidene	0	0	17634

Table S2. Number of small molecule compounds with certain functional groups in the PubChem dataset.

Functional group	Count
phenol, aliphatic prim- and sec-alcohol	1073964
prim-amine (aliphatic and aromatic)	401107
prim- and sec-amine (aliphatic and aromatic)	17133014
phenol, aliphatic prim- and sec-alcohol	2331368
aliphatic prim- and sec-carboxylic acid, aromatic carboxylic acid	542024
epoxide (poly)	56730
hindered phenol and thiophenol	50107
amino acid	2773404
hydroxy carboxilic acid	1937567
lactam	1115284
lactone except gamma-butyrolactone	679436
cyclic olefin	4239342
epoxide（mono and poly）	256271
vinyl (terminal olefin) include acrylate	4701127

Table S3. Number of small molecule compounds with certain functional groups in the GDB-13 dataset.

Functional group	Count
phenol, aliphatic prim- and sec-alcohol	9001267
prim-amine (aliphatic and aromatic)	20194525
prim- and sec-amine (aliphatic and aromatic)	180414699
phenol, aliphatic prim- and sec-alcohol	10994073
aliphatic prim- and sec-carboxylic acid, aromatic carboxylic acid	6940
epoxide (poly)	662909
hindered phenol and thiophenol	5950
amino acid	3961429
hydroxy carboxilic acid	1898490
lactam	9606974
lactone except gamma-butyrolactone	11465098
cyclic olefin	194381403
epoxide（mono and poly）	14268280
vinyl (terminal olefin) include acrylate	188757121

Table S4. Number of small molecule compounds with certain functional groups in the GDB-17 dataset.

Functional group	Count
phenol, aliphatic prim- and sec-alcohol	1024466
prim-amine (aliphatic and aromatic)	2018364
prim- and sec-amine (aliphatic and aromatic)	10357888
phenol, aliphatic prim- and sec-alcohol	1368370
aliphatic prim- and sec-carboxylic acid, aromatic carboxylic acid	3095
epoxide (poly)	13653
hindered phenol and thiophenol	827
amino acid	524249
hydroxy carboxilic acid	392362
lactam	911417
lactone except gamma-butyrolactone	1129100
cyclic olefin	6041868
epoxide （mono and poly）	320173
vinyl (terminal olefin) include acrylate	61644

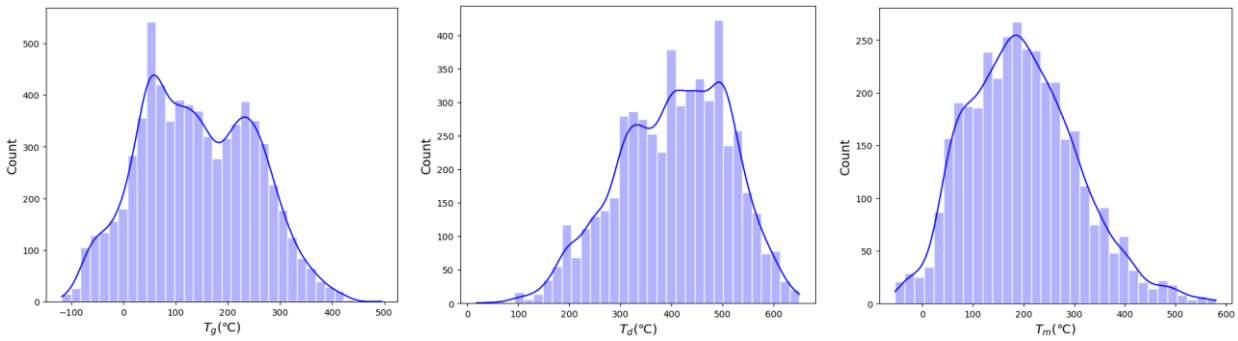


Figure S1. Univariate distribution plots for T_g, T_m, and T_d.

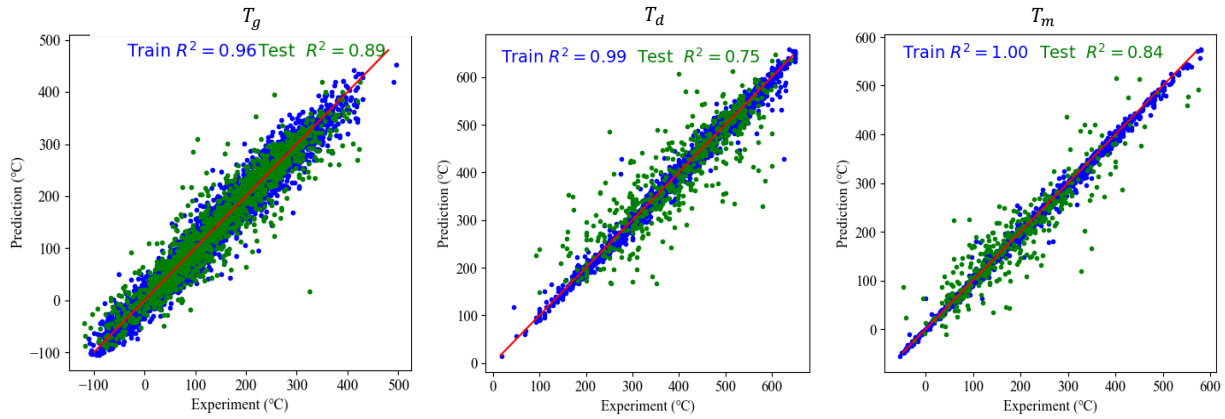


Figure S2. Parity plot of FNN models for T_g, T_m, and T_d.

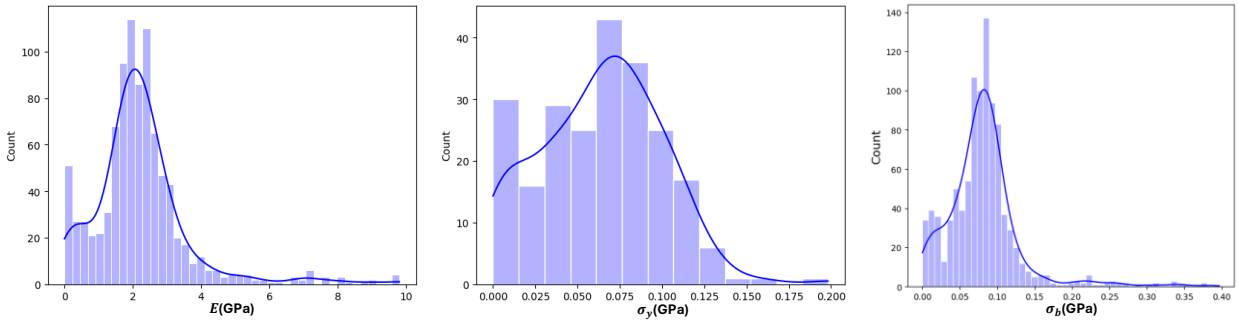


Figure S3. Univariate distribution plots for E, σ_y, and σ_b.

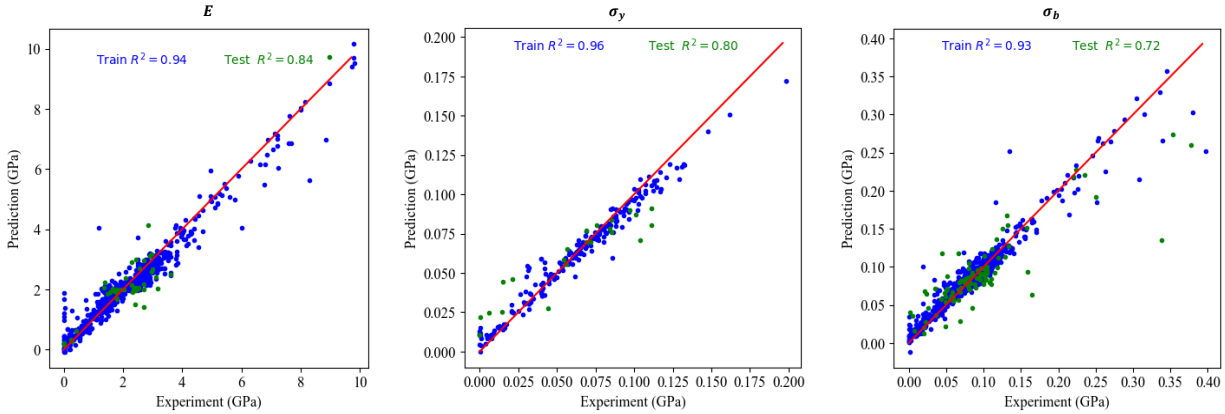


Figure S4. Parity plot of FNN models for E, σ_y, and σ_b.

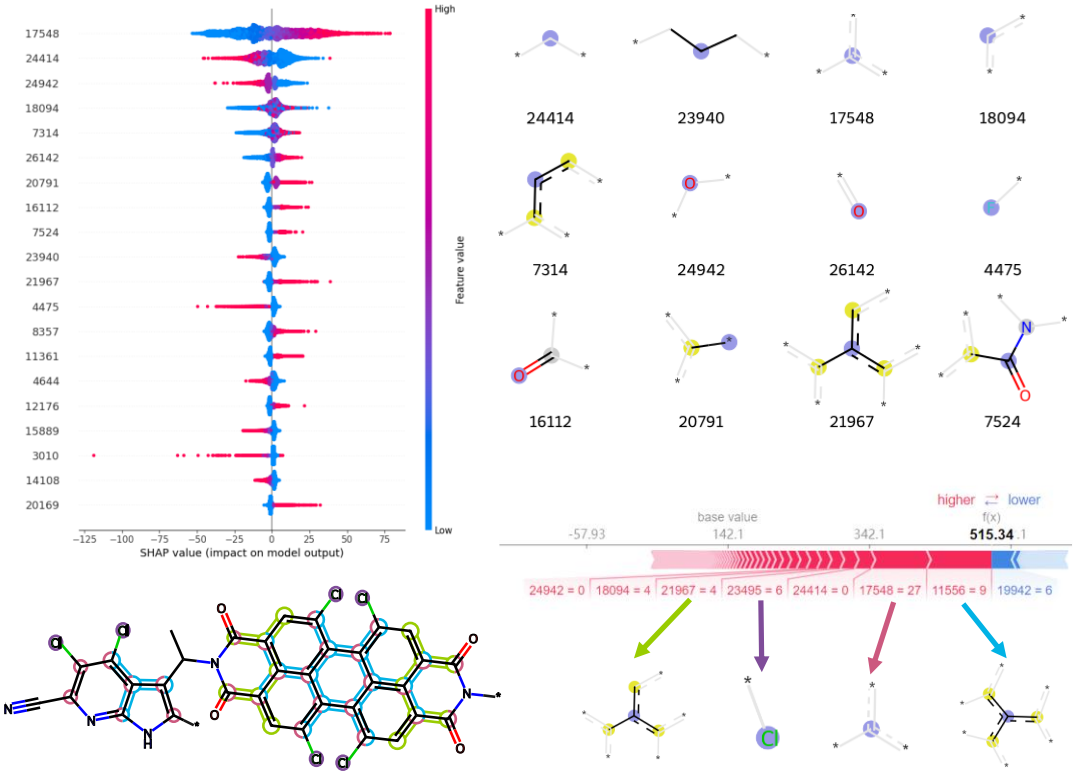


Figure S5. The substructure importance plot for T_g displays the most significant substructures in descending order, with each dot representing the impact from a specific sample in the training set. The plot highlights the 12 most important substructures associated with T_g according to SHAP values. In the plot, the central atom of each substructure is marked in blue, aromatic atoms are highlighted in yellow, and the connectivity of the atoms is shown in light gray. Below this, the individual SHAP value plot for the promising hypothetical polyimide structure is presented. Red and blue arrows indicate the positive and negative contributions of each substructure, respectively. The feature value of a substructure can be "0," indicating its absence in the polymer structure, but its feature importance remains valid as indicated by the length of the arrow. The top substructures in this polyimide are highlighted in different colors.

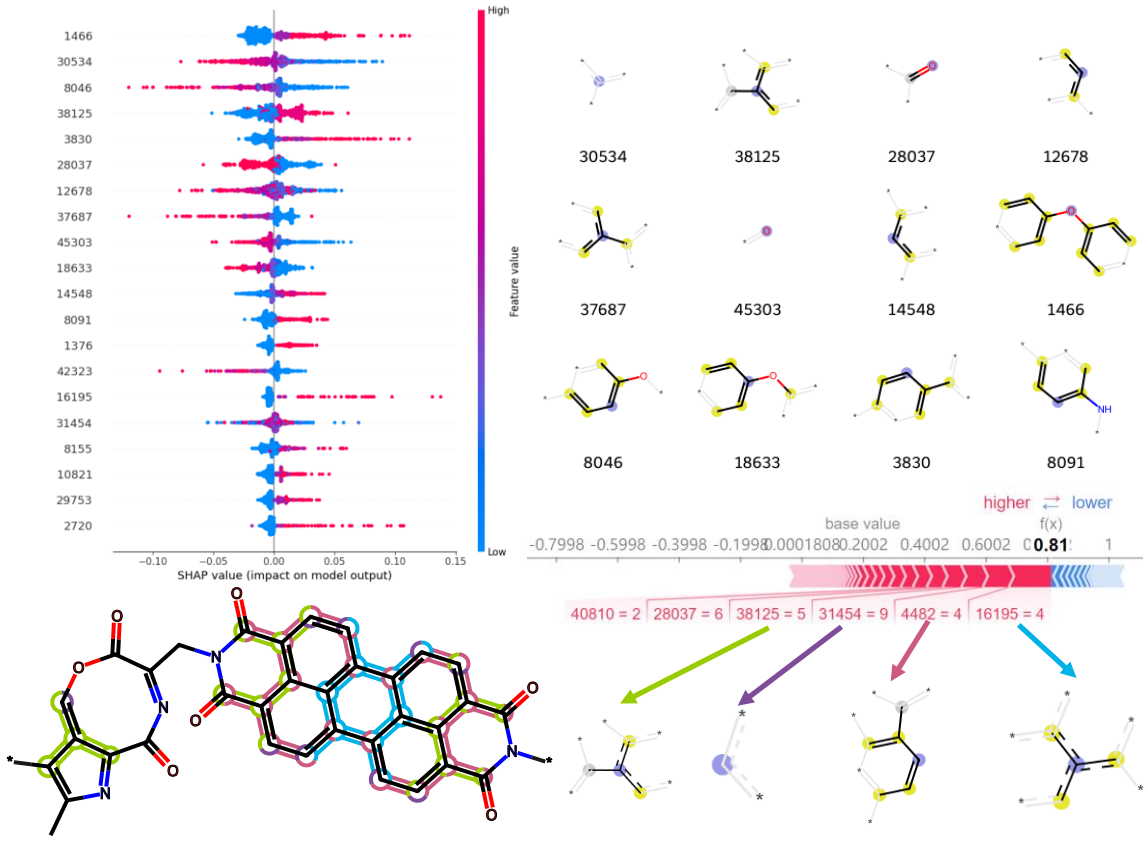


Figure S6. The substructure importance plot for E displays the most significant substructures in descending order, with each dot representing the impact from a specific sample in the training set. The plot highlights the 12 most important substructures associated with E according to SHAP values. In the plot, the central atom of each substructure is marked in blue, aromatic atoms are highlighted in yellow, and the connectivity of the atoms is shown in light gray. Below this, the individual SHAP value plot for the promising hypothetical polyimide structure is presented. Red and blue arrows indicate the positive and negative contributions of each substructure, respectively. The feature value of a substructure can be "0," indicating its absence in the polymer, but its feature importance remains valid as indicated by the length of the arrow. The top substructures in this polyimide are highlighted in different colors.

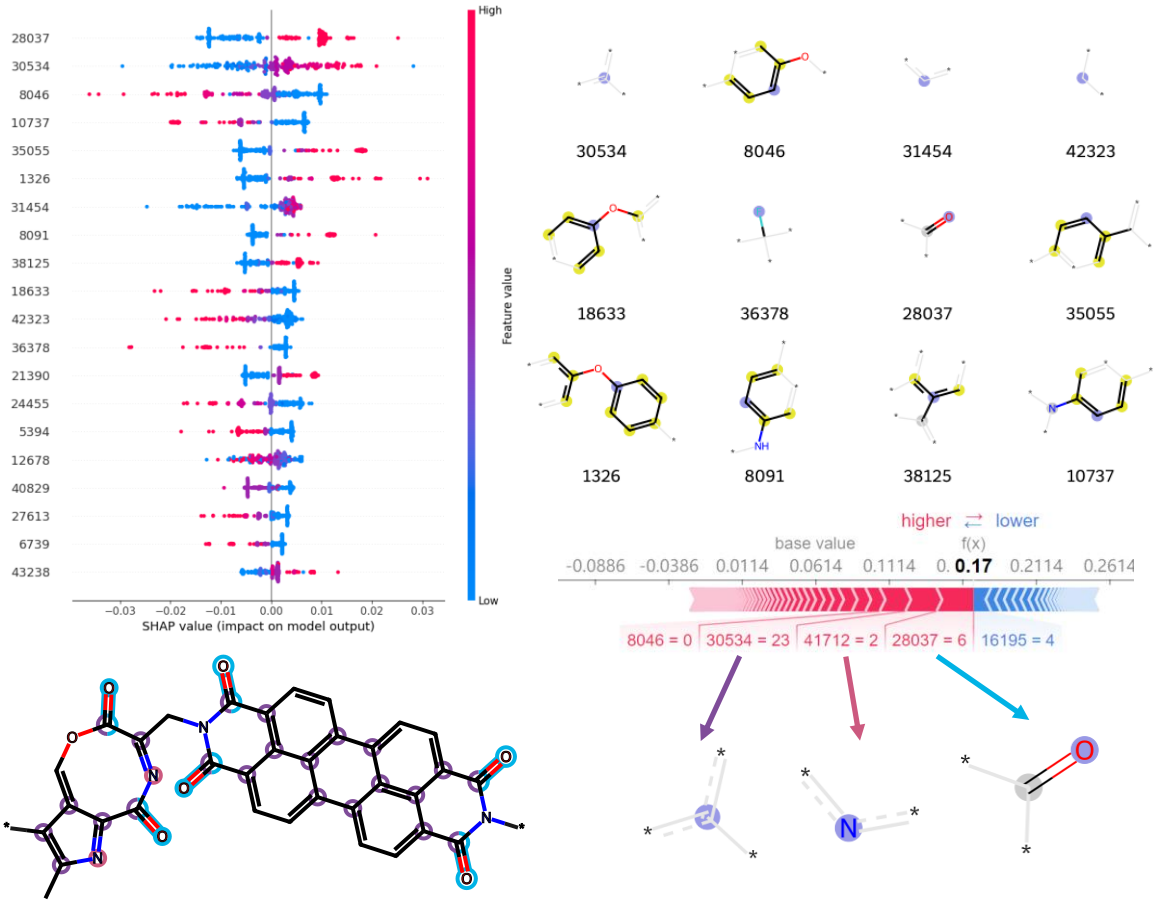


Figure S7. The substructure importance plot for σ_y displays the most significant substructures in descending order, with each dot representing the impact from a specific sample in the training set. The plot highlights the 12 most important substructures associated with σ_y according to SHAP values. In the plot, the central atom of each substructure is marked in blue, aromatic atoms are highlighted in yellow, and the connectivity of the atoms is shown in light gray. Below this, the individual SHAP value plot for the promising hypothetical polyimide structure is presented. Red and blue arrows indicate the positive and negative contributions of each substructure, respectively. The feature value of a substructure can be "0," indicating its absence in the polymer, but its feature importance remains valid as indicated by the length of the arrow. The top substructures in this polyimide are highlighted in different colors.

Details of Network Training and Dataset

The largest database, PoLyInfo[1], contains over 18,000 reported polymers, including 12,854 homopolymers with their chemical structures and around 100 types of properties. This homopolymer dataset is suitable for training neural networks in our study. Within this dataset, 6,906 homopolymers have reported values for T_g, 3,633 for T_m, 5,237 for T_d, 923 for E, 230 for σ_y, and 983 for σ_b. Using these reported property values and the corresponding monomer

structures, machine learning models can be trained to establish a composition-property mapping for polymers.

For the gas permeability models training, the dataset includes 778 homopolymers (representing 353 unique polymer chemistries), each associated with at least one reported gas permeability value for He, H_2, O_2, N_2, CO_2, and CH_4.

The model for T_g was trained using 90% of the data points with reported experimental values, with the remaining 10% used as a test set. The model achieved an R^2 of 0.96 for training and 0.89 for validation. For T_d, 90% of the data points were pseudorandomly selected for the training set, and the remaining 10% were used for testing. Using the same training process as the tensile modulus model, it achieved an R^2 of 0.99 for training and 0.75 for validation. Similarly, for T_m, 90% of the data points were pseudorandomly selected for training, and the other 10% were used for testing. The model obtained an R^2 of 1.00 for training and 0.84 for validation.

The model for E was trained using 95% of the data points with reported experimental values, while the remaining 5% were used as a test set. The model achieved an R^2 of 0.96 for training and 0.84 for validation. For σ_y, 90% of the data points were pseudorandomly selected for the training set, and the remaining 10% were used for testing. Following the same training process as for the tensile modulus model, the model achieved an R^2 of 0.95 for training and 0.80 for validation. Similarly, for σ_b, 90% of the data points were selected pseudorandomly for training, and the other 10% were used for testing. Using the same training process, the model achieved an R^2 of 0.93 for training and 0.72 for validation.

For the gas permeability models training, the metric of interest is the R^2 correlation between the predicted and actual permeabilities on both the training and test sets. The data was split randomly, with 80% used for training and the remaining 20% reserved for testing, as summarized in **Table S5**.

Table S5. Summary of the performances of supervised ML models as scored by the R^2 value between the predicted and actual permeabilities.

He		H_2		O_2		N_2		CO_2		CH_4	
Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.88	0.91	0.88	0.90	0.90	0.92	0.90	0.91	0.89	0.90	0.89	0.88

Details of molecular dynamics verification

To validate the performance of the ML model, we conducted all-atom molecular dynamics simulations on nine hypothetical polymer structures randomly selected from the PolyUniverse database. The SMILES representations of their repeat units are shown in **Table S6**.

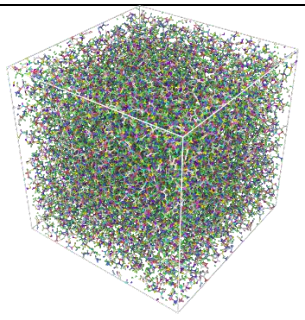
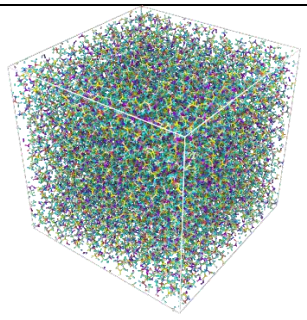
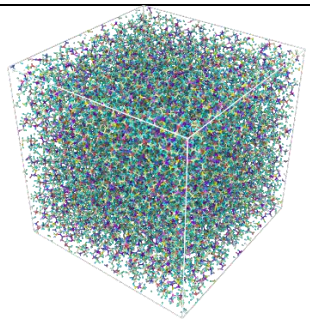
Table S6. SMILES of repeat units for nine hypothetical polymers used for MD simulations

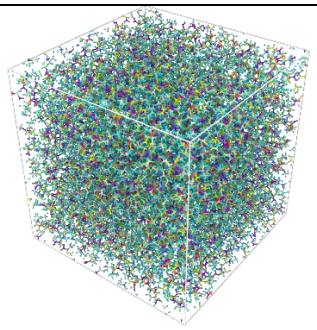
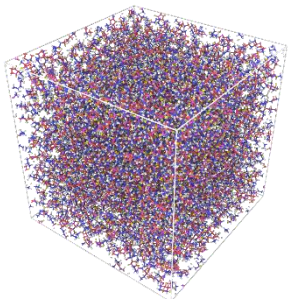
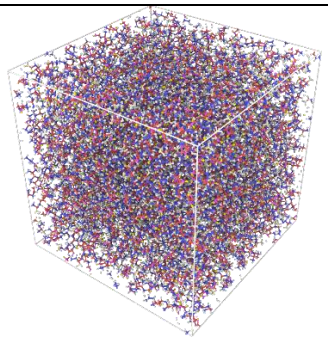
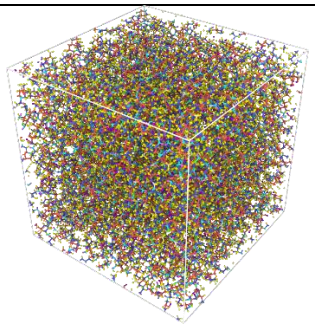
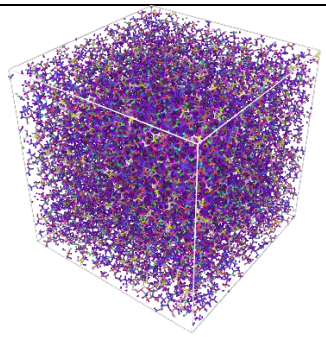
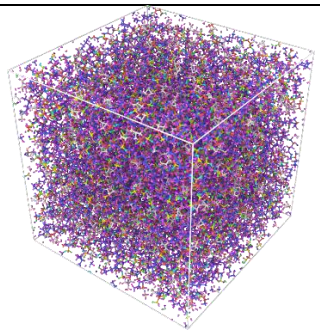
Model	SMILES
1	*Nc1oncc1C1(C)CCC=C1C(*)=O
2	*NCC1CC(=NO)CC1(C)NCCC(*)=O
3	*NC(C=O)CCC1(C)CCCC1CC(*)=O
4	*OCCC1(C(*)=O)C=CC(CC)CC1
5	*CCCNCCCC1=Nc2ccc(-c3ccc4c(c3)N=C(*)[N]4)cc2[N]1
6	*CCCCC(C)CC1=Nc2ccc(-c3ccc4c(c3)N=C(*)[N]4)cc2[N]1
7	*CC(O)c1ccccc1C(O)CCNCC1c2oncc2C(C)CN1CCN*
8	*CC(O)CCOCCCCC(O)COC(=O)C1CC2(C)CC1C2C(=O)O*
9	*CC(O)CCCCCCCC(O)COC(=O)C1C2CC(C2C)C1C(=O)O*

We selected T_g as the target property and conducted all-atomic MD simulations to validate the ML-predicted T_g of these selected polymers. Among the properties used to validate ML models, T_g is often chosen, as all-atom MD simulations can, in most cases, provide reasonable estimates of Tg for crystalline, semi-crystalline, and amorphous polymers by analyzing simulated density vs. temperature curves.

Each polymer model contains approximately 40,000 atoms, with a box side length of about 75 Å, as shown in **Table S7**. Periodic boundary conditions were applied in all three dimensions. The polymer consistent force field (PCFF) was used to define interatomic interactions. PCFF, a second-generation force field, is parameterized for organic compounds containing H, C, N, O, S, P, halogens, and ions.[2-5] It offers broad coverage for calculating cohesive energies, mechanical properties, compressibilities, heat capacities, and elastic constants of organic polymers. The LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) package was used for the MD simulations. Prior to simulating the properties, these polymers were equilibrated using a 21-step MD equilibration protocol, as employed in our previous work.[6]

Table S7. Initial configuration of nine selected polymers for molecular dynamics simulation

		
#1 42680 atoms 40chains	#2 39760 atoms 40chains	#3 39520 atoms

75.25Å×75.25Å×75.25Å	72.83Å×72.83Å×72.83Å	40chains 72.56Å×72.56Å×72.56Å
 #4 39520 atoms 40chains 70.9Å×70.9Å×70.9Å	 #5 39760 atoms 40chains 73.05Å×73.05Å×73.05Å	 #6 37880 atoms 40chains 701.9Å×71.9Å×71.9Å
 #7 37600 atoms 40chains 74.28Å×74.28Å×74.28Å	 #8 38480 atoms 40chains 71.72Å×71.72Å×71.72Å	 #9 39760 atoms 40chains 72.59Å×72.59Å×72.59Å

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 density vs. temperature curves is shown in **Figure S8.** A comparison of the calibrated MD results and the corresponding ML predictions can be found in **Table S8** and **Figure S9**. It is worth noting that the timescale of MD simulation is in the nanosecond range, resulting in a much faster cooling rate than in experiments. In our previous work[6], we compared these results with experimental data, demonstrating that this MD simulation workflow reasonably agrees with experimental values. Therefore, we believe the T_g results from our MD simulations can be used to validate the ML-predicted T_g. Our results show that the ML predictions reasonably agree with the MD results, confirming that our model reliably predicts T_g across different polymers in the dataset.

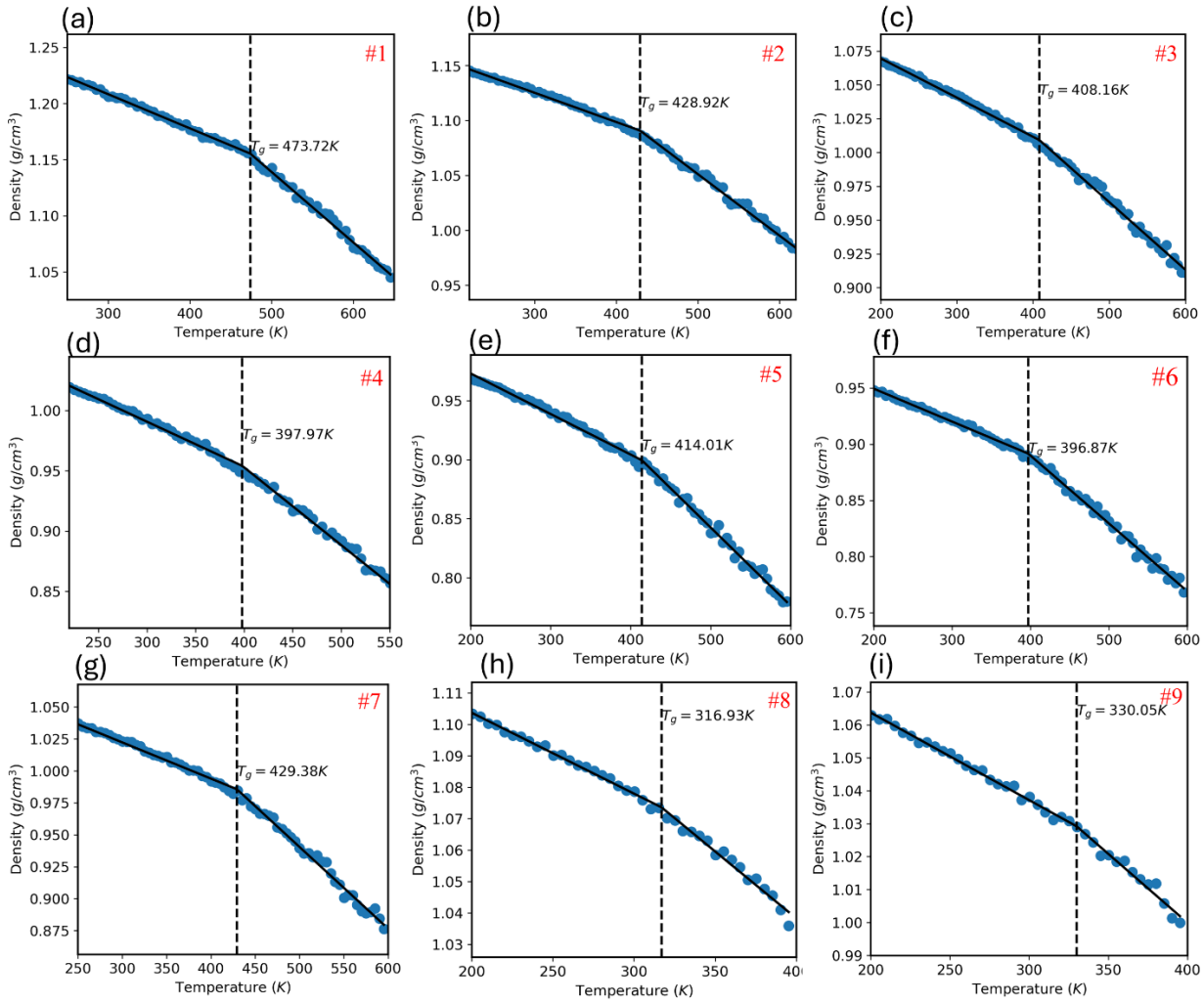


Figure S8. Glass transition temperature for the nine selected polymers based on MD simulations.

Table S8. Comparison of ML predictions and MD simulations of the nine selected polymers

Model	Tg (K)		
	MD	ML	Error[a]
#1	473.56	411.1933	15.18%
#2	428.92	403.0879	6.41%
#3	408.16	383.5832	6.41%
#4	397.97	374.0717	6.39%
#5	414.01	429.9912	3.72%
#6	396.87	418.2139	5.10%
#7	429.38	384.8127	11.58%
#8	316.93	293.2459	8.08%
#9	330.05	314.2498	5.03%

[a]: Error=(ML-MD/MD)×100%

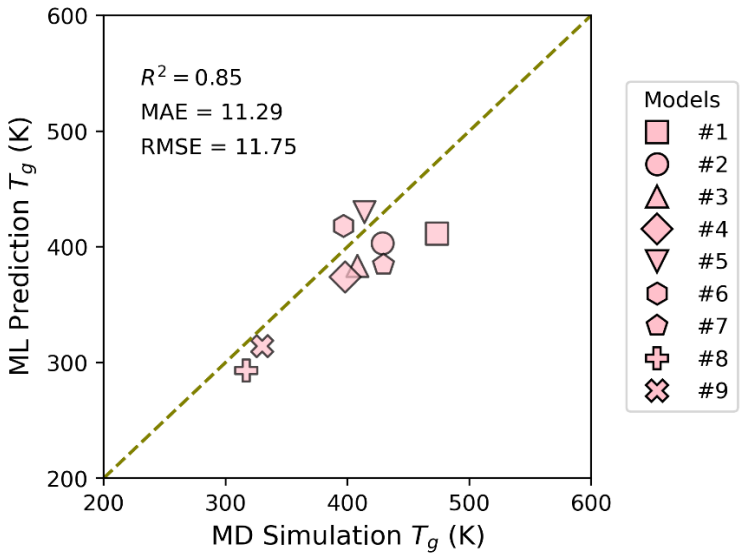


Figure S9. Glass transition temperature for nine hypothetical polymers based on MD simulations.

Furthermore, we conducted additional simulations on both the real polymer #R1 and the hypothetical polymer #ML1. The simulated specific volume vs. temperature curves is shown in **Figure S9.** A comparison of the calibrated MD results and the corresponding ML predictions can be found in **Table S9.** The MD results for three sets of the real polymer compared with the experimental T_g values show that the MD-predicted T_g values are close to the experimental ones, with errors of 4.85%, 6.79%, and 3.65%. This demonstrates that MD can serve as a reliable reference for the ML-predicted T_g values. For the #ML1 hypothetical polymer, the MD-predicted T_g values show reasonable agreement with our ML predictions, with errors of 0.66%, 2.52%, and 7.66%. This suggests that the #ML1 hypothetical polymer indeed may exhibit a relatively high T_g.

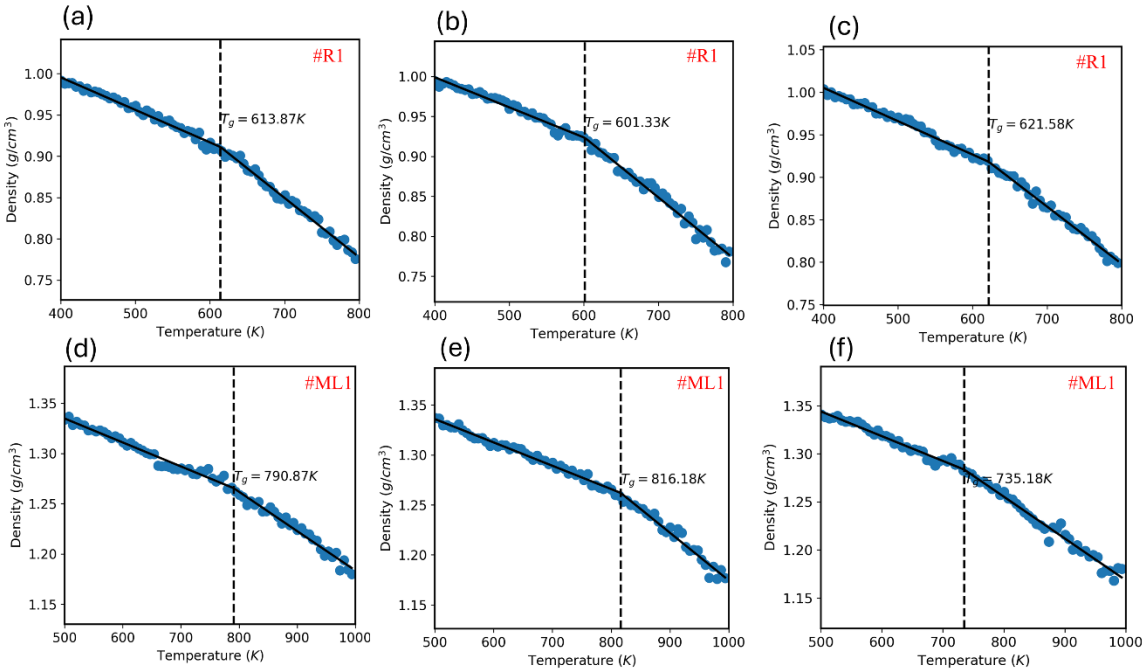
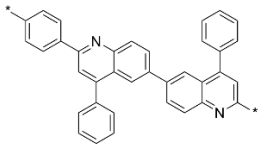
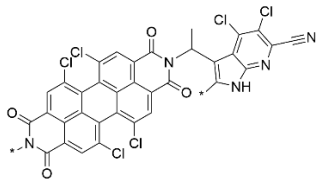


Figure S9. Glass transition temperature for the real and hypothetical polymers based on MD simulations.

Table S9. Comparison of ML predictions and MD simulations of the real and hypothetical polymers.

Model	Repeat Unit	Exp T_g (K)	ML T_g (K)	MD T_g (K)	Error
#R1		645.15	-	613.87	4.85%
				601.33	6.79%
				621.58	3.65%
#ML1		-	796.15	790.87	0.66%
				816.18	2.52%
				735.18	7.66%

[a]: Error=|(ML-MD)/MD)|×100% or Error=|(MD-Exp)/Exp)|×100%

References

(1) Ishii, M.; Ito, T.; Sado, H.; Kuwajima, I. NIMS polymer database PoLyInfo (I): an overarching view of half a million data points. *Science and Technology of Advanced Materials: Methods* **2024**, (just-accepted), 2354649.

(2) Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T. An ab initio CFF93 all-atom force field for polycarbonates. *Journal of the American Chemical society* **1994**, *116* (7), 2978-2987.

(3) Sun, H.; Ren, P.; Fried, J. The COMPASS force field: parameterization and validation for phosphazenes. *Computational and Theoretical Polymer Science* **1998**, *8* (1-2), 229-246.
(4) Sun, H. Ab initio calculations and force field development for computer simulation of polysilanes. *Macromolecules* **1995**, *28* (3), 701-712.
(5) 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.
(6) Tao, L.; He, J.; Munyaneza, N. E.; Varshney, V.; Chen, W.; Liu, G.; Li, Y. Discovery of multi-functional polyimides through high-throughput screening using explainable machine learning. *Chemical Engineering Journal* **2023**, *465*, 142949.