Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.

Supporting Information for

High-Throughput Screening of Amorphous Polymers with High Intrinsic Thermal Conductivity via Automated Physical Feature Engineering

Xiang Huang[a], Shengluo Ma[a], Yunwen Wu[b,*], Chaoying Wan[c], C. Y. Zhao[a], Hong Wang[b], and Shenghong Ju[a, b, *]

[a] China-UK Low Carbon College, Shanghai Jiao Tong University, Shanghai, China

[b] Materials Genome Initiative Center, School of Material Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[c] International Institute for Nanocomposites Manufacturing (IINM), WMG, University of Warwick, Coventry, UK

* **Corresponding Email:** shenghong.ju@sjtu.edu.cn (S. Ju), tlwuyunwen@sjtu.edu.cn (Y. Wu)

This PDF file includes:

Supporting Information Text

A. Visualization of polymer backbones in the 2D plane

The distribution of polymer chemical structures in different datasets was visualized in the 2D plane by uniform manifold approximation and projection (UMAP) [1]. The chemical structures in the benchmark dataset (Figure S1a) are able to well cover those in PoLyInfo and Polymides datasets without significant selection bias.

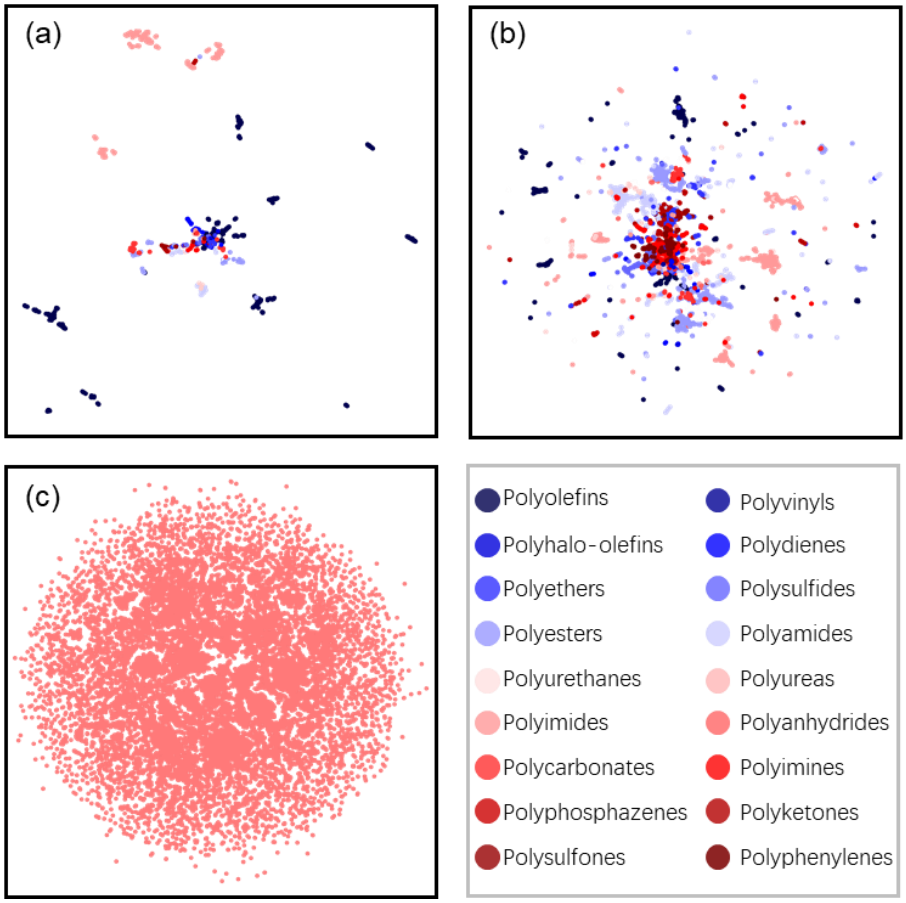

Fig. S1. Visualization of polymer backbones using UMAP. (a), (b) and (c) are the datasets A, B and C, described as benchmark, PoLyInfo and Polymides datasets, respectively. Datasets A and B are sourced from the PoLyInfo database, and datasets C consists of hypothetical polyimides formed by dianhydride and diamine/diisocyanate from PubChem. All polymers are sorted into 18 classes, color-coded according to the definition of PoLyInfo.

B. Initial descriptors generation and down-selection

The 325 initial descriptors were collected in this work, which consists of 294 Mordred-based descriptors and 31 MD-based descriptors, as listed in Table S1 and S2. The meaning of the Mordred-based descriptors can be found at https://mordred-descriptor.github.io/documentation/master/descriptors.html (accessed Dec. 30, 2022).

Table S1. List of Mordred descriptors.

ABC	nAcid	nBase	SpMax_A	SpMAD_A
VE1_A	VR1_A	VR3_A	nAromAtom	nAtom
nSpiro	nBridgehead	nHetero	nH	nN
nO	nS	nP	nF	nCl
nBr	nI	ATS0Z	AATS0dv	AATS0d
AATS2d	AATS0Z	AATS1Z	ATSC1dv	ATSC2dv
ATSC3dv	ATSC4dv	ATSC5dv	ATSC6dv	ATSC7dv
ATSC8dv	ATSC1d	ATSC2d	ATSC3d	ATSC4d
ATSC5d	ATSC6d	ATSC7d	ATSC8d	ATSC0Z
ATSC1Z	ATSC2Z	ATSC3Z	ATSC4Z	ATSC5Z
ATSC6Z	ATSC7Z	ATSC8Z	AATSC0dv	AATSC1dv
AATSC2dv	AATSC3dv	AATSC0d	AATSC1d	AATSC2d
AATSC3d	AATSC0Z	AATSC1Z	AATSC2Z	AATSC3Z
MATS1dv	MATS2dv	MATS3dv	MATS1d	MATS2d
MATS3d	MATS1Z	MATS2Z	MATS3Z	GATS1dv
GATS2dv	GATS3dv	GATS1d	GATS2d	GATS3d
GATS1Z	GATS2Z	GATS3Z	BCUTdv-1h	BCUTdv-1l
BCUTd-1h	BCUTd-1l	BCUTZ-1h	BCUTZ-1l	BalabanJ
nBondsS	nBondsD	nBondsT	C2SP1	C1SP2
C3SP2	C1SP3	C2SP3	C3SP3	C4SP3
HybRatio	Xch-3d	Xch-4d	Xch-5d	Xch-6d
Xc-3d	Xc-4d	Xc-5d	Xc-6d	AXp-2d
AXp-3d	NsCH3	NdCH2	NtCH	NdsCH
NsssCH	NdssC	NaaaC	NssssC	NsNH2
NssNH	NaaNH	NtN	NdsN	NaaN
NsssN	NddsN	NaasN	NsOH	NssO
NaaO	NssS	NaaS	NdssS	NddssS
SsssCH	SdssC	SaasC	SaaaC	SsssN
SddsN	AETA_beta_s	ETA_beta_ns_d	AETA_beta_ns_d	AETA_eta_RL
ETA_eta_B	AETA_eta_B	ETA_epsilon_3	ETA_dBeta	fMF
GeomShapeIndex	nHBDon	IC0	IC1	IC2
IC3	SIC0	SIC1	SIC2	SIC3
SIC4	BIC3	BIC4	CIC0	CIC1
MIC0	MIC1	MIC2	Kier2	Kier3
Lipinski	GhoseFilter	Mor02	Mor03	Mor04

Mor06	Mor08	Mor09	Mor10	Mor11
Mor12	Mor13	Mor14	Mor15	Mor16
Mor17	Mor18	Mor19	Mor21	Mor22
Mor23	Mor24	Mor25	Mor26	Mor27
Mor28	Mor29	Mor30	Mor31	Mor32
PEOE_VSA1	PEOE_VSA2	PEOE_VSA3	PEOE_VSA4	PEOE_VSA6
PEOE_VSA7	PEOE_VSA8	PEOE_VSA9	PEOE_VSA10	PEOE_VSA11
PEOE_VSA12	PEOE_VSA13	SMR_VSA1	SMR_VSA3	SMR_VSA4
SMR_VSA5	SMR_VSA6	SMR_VSA9	SlogP_VSA1	SlogP_VSA2
SlogP_VSA3	SlogP_VSA4	SlogP_VSA5	SlogP_VSA7	SlogP_VSA8
SlogP_VSA10	SlogP_VSA11	EState_VSA1	EState_VSA2	EState_VSA3
EState_VSA4	EState_VSA5	EState_VSA6	EState_VSA7	EState_VSA8
EState_VSA9	EState_VSA10	VSA_EState3	VSA_EState4	VSA_EState5
VSA_EState7	VSA_EState8	VSA_EState9	AMID_h	AMID_C
AMID_N	AMID_O	AMID_X	MOMI-Z	PBF
n4Ring	n12Ring	nHRing	n6HRing	n5aRing
naHRing	n6aHRing	nARing	n5ARing	n6ARing
nAHRing	n6AHRing	nFRing	n7FRing	n8FRing
n9FRing	n10FRing	n12FRing	nG12FRing	n8FHRing
n10FHRing	nFaRing	nG12FaRing	nFaHRing	n10FARing
n12FARing	nG12FARing	n10FAHRing	nG12FAHRing	nRot
RotRatio	SLogP	TopoPSA(NO)	JGI2	JGI3
JGI4	JGI5	JGI6	JGI7	JGI8
JGI9	JGI10	JGT10	TopoShapeIndex	

Table S2. Description of MD-based descriptors.

Name	Description
Charge_max	Maximum value of the atom chargaes
Charge_min	Minimum value of the atom chargaes
Charge_ave	Average value of the atom chargaes
Epsilon_max	Maximum value of the depth of the energy potential (Lennard–Jones parameter)
Epsilon_min	Minimum value of the depth of the energy potential (Lennard–Jones parameter)
Epsilon_ave	Average value of the depth of the energy potential (Lennard–Jones parameter)
Sigma_max	Maximum value of the equilibrium distance (Lennard–Jones parameter)
Sigma_min	Minimum value of the equilibrium distance (Lennard–Jones parameter)
Sigma_ave	Average value of the equilibrium distance (Lennard–Jones parameter)
K_bond_max	Maximum value of force constants of the bond
K_bond_min	Minimum value of force constants of the bond
K_bond_ave	Average value of force constants of the bond
R0_max	Maximum value of equilibration structural parameters of the bond
R0_min	Minimum value of equilibration structural parameters of the bond
R0_ave	Average value of equilibration structural parameters of the bond
K_ang_max	Maximum value of force constants of the bond angle
K_ang_min	Minimum value of force constants of the bond angle
K_ang_ave	Average value of force constants of the bond angle
Theta0_max	Maximum value of equilibration structural parameters of the bond angle
Theta0_min	Minimum value of equilibration structural parameters of the bond angle
Theta0_ave	Average value of equilibration structural parameters of the bond angle
K_dih_max	Maximum value of force constants of the dihedral angle
K_dih_min	Minimum value of force constants of the dihedral angle
K_dih_ave	Average value of force constants of the dihedral angle
Mass_max	Maximum atomic mass
Mass_min	Minimum atomic mass
Mass_ave	Average of atomic masses
VDW	Van der Waals volume of the monomer
MW	Molecular weight of monomer
Monomer_length	Monomer length after optimization using MMFF94 force field in RDKit
MW_ratio	The ratio of the molecular weight of the main chain to that of the monomer

Figure S2 illustrates the statistical correlation coefficients of Pearson, Spearman, Distance and maximum information (MIC) for 53 descriptors (Cor. descriptors). The thresholds of 0.050, 0.050, 0.213 and 0.186 for Pearson, Spearman, Distance and MIC. The descriptors that correspond to the numbers of the horizontal coordinate are listed in Table S3.

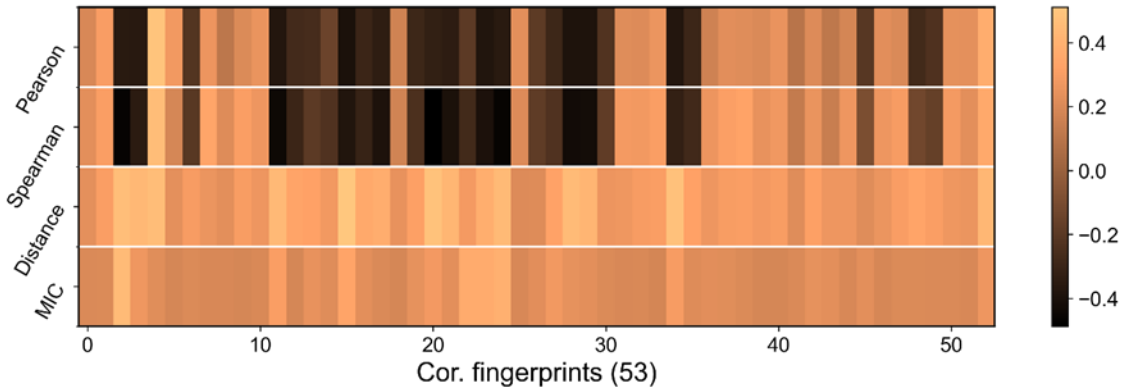

Fig. S2. The statistical correlation coefficients of Pearson, Spearman, Distance and maximum information (MIC) for 53 descriptors (Cor. descriptors).

Table S3. List of descriptors after filtering by statistical correlation coefficients.

No.	Name	Description	Source
0	Monomer_length	Monomer length after optimization using MMFF94 force field in RDKit	MD-based
1	MW_ratio	The ratio of the molecular weight of the main chain to that of the monomer	
2	Mass_max	Maximum atomic mass	
3	Mass_ave	Average of atomic masses	
4	K_bond_min	Minimum value of force constants of the bond	
5	K_bond_ave	Average value of force constants of the bond	
6	K_ang_ave	Average value of force constants of the bond angle	
7	Theta0_min	Minimum value of equilibration structural parameters of the bond angle	
8	nAtom	number of all atoms	
9	nH	number of H atoms	
10	nN	number of N atoms	
11	nF	number of F atoms	
12	nCl	number of Cl atoms	
13	AATS0dv	averaged moreau-broto autocorrelation of lag 0 weighted by valence electrons	Mordred
14	AATS0d	averaged moreau-broto autocorrelation of lag 0 weighted by sigma electrons	
15	AATS0Z	averaged moreau-broto autocorrelation of lag 0 weighted by atomic number	
16	AATS1Z	averaged moreau-broto autocorrelation of lag 1 weighted by atomic number	
17	ATSC2Z	centered moreau-broto autocorrelation of lag 2 weighted by atomic number	
18	ATSC5Z	centered moreau-broto autocorrelation of lag 5 weighted by atomic number	

No.	Name	Description	Source
19	AATSC0dv	averaged and centered moreau-broto autocorrelation of lag 0 weighted by valence electrons	
20	AATSC0Z	averaged and centered moreau-broto autocorrelation of lag 0 weighted by atomic number	
21	AATSC2Z	averaged and centered moreau-broto autocorrelation of lag 2 weighted by atomic number	
22	BCUTdv-1h	first heighest eigenvalue of Burden matrix weighted by valence electrons	
23	BCUTd-1h	first heighest eigenvalue of Burden matrix weighted by sigma electrons	
24	BCUTZ-1h	first heighest eigenvalue of Burden matrix weighted by atomic number	
25	C1SP2	SP2 carbon bound to 1 other carbon	
26	Xc-3d	3-ordered Chi cluster weighted by sigma electrons	
27	Xc-5d	5-ordered Chi cluster weighted by sigma electrons	
28	Xc-6d	6-ordered Chi cluster weighted by sigma electrons	
29	NssssC	number of ssssC	
30	ETA_eta_B	ETA branching index	
31	nHBDon	number of hydrogen bond donor	
32	CIC0	0-ordered complementary information content	
33	CIC1	1-ordered complementary information content	
34	MIC0	0-ordered modified information content	
35	MIC1	1-ordered modified information content	
36	Kier2	kappa shape index 2	
37	Kier3	kappa shape index 3	
38	Mor02	3D-MoRSE (distance = 2)	
39	Mor04	3D-MoRSE (distance = 4)	
40	Mor08	3D-MoRSE (distance = 8)	
41	Mor13	3D-MoRSE (distance = 13)	
42	Mor14	3D-MoRSE (distance = 14)	
43	Mor19	3D-MoRSE (distance = 19)	
44	Mor31	3D-MoRSE (distance = 31)	
45	SMR_VSA1	MOE MR VSA Descriptor 1 (-inf < x < 1.29)	
46	SMR_VSA3	MOE MR VSA Descriptor 3 (1.82 <= x < 2.24)	
47	SlogP_VSA5	MOE logP VSA Descriptor 5 (0.10 <= x < 0.15)	
48	SlogP_VSA10	MOE logP VSA Descriptor 10 (0.40 <= x < 0.50)	
49	EState_VSA1	EState VSA Descriptor 1 (-inf < x < -0.39)	
50	EState_VSA5	EState VSA Descriptor 5 (1.17 <= x < 1.54)	
51	VSA_EState4	VSA EState Descriptor 4 (5.41 <= x < 5.74)	
52	VSA_EState7	VSA EState Descriptor 7 (6.07 <= x < 6.45)	

The accuracies of RF models during recursive feature elimination are shown in Figure S3. Combining the variations of R2 and mean-square error (MSE) curves, we identified the 25 optimized descriptors.

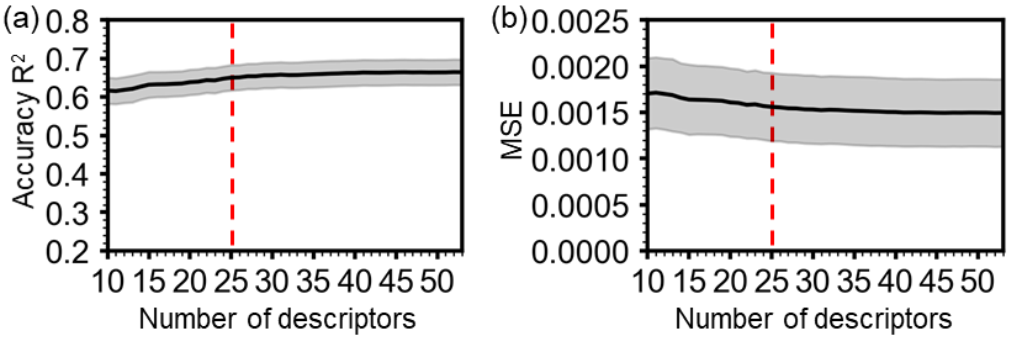

Fig. S3. Accuracies of RF models during recursive feature elimination. (a) and (b) R2 and mean-square error (MSE). The solid black line indicates the average precision value of 10 cross-validations, and the gray area indicates a standard deviation.

Table S4. Description of optimized descriptors.

No.	Labels	Description	Source
1	BCUTZ-1h	first heighest eigenvalue of Burden matrix weighted by atomic number	Moedred
2	AATS0d	averaged moreau-broto autocorrelation of lag 0 weighted by sigma electrons	Moedred
3	MW_ratio	Ratio of mainchain molecular weight to monomer molecular weight	MD
4	K_bond_ave	Average of different bond force constants in monomer	MD
5	BCUTd-1h	first heighest eigenvalue of Burden matrix weighted by sigma electrons	Moedred
6	AATS0Z	averaged moreau-broto autocorrelation of lag 0 weighted by atomic number	Moedred
7	Mass_max	Maximum atomic mass in a monomer	MD
8	Monomer_length	Monomer length after relaxation	MD
9	Mor02	3D-MoRSE (distance = 2)	Moedred
10	ATSC5Z	centered moreau-broto autocorrelation of lag 5 weighted by atomic number	Moedred
11	nHBDon	number of hydrogen bond donor	Moedred
12	Mor19	3D-MoRSE (distance = 19)	Moedred
13	Kier3	kappa shape index 3	Moedred
14	ATSC2Z	centered moreau-broto autocorrelation of lag 2 weighted by atomic number	Moedred
15	Mor14	3D-MoRSE (distance = 14)	Moedred
16	Mass_ave	Average atomic mass in a monomer	MD
17	AATSC2Z	averaged and centered moreau-broto autocorrelation of lag 2 weighted by atomic number	Moedred
18	K_ang_ave	Average of different bond angle force constants in monomer	MD
19	AATSC0Z	averaged and centered moreau-broto autocorrelation of lag 0 weighted by atomic number	Moedred
20	SMR_VSA3	MOE MR VSA Descriptor 3 (1.82 <= x < 2.24)	Moedred
21	MIC0	0-ordered modified information content	Moedred
22	SMR_VSA1	MOE MR VSA Descriptor 1 (-inf < x < 1.29)	Moedred
23	VSA_EState4	VSA EState Descriptor 4 (5.41 <= x < 5.74)	Moedred
24	MIC1	1-ordered modified information content	Moedred
25	nH	number of H atoms	Moedred

To generate statistically meaningful results, 20 evaluations were conducted for each ML model with optimized descriptors. During these evaluations, the training and test data were randomly sampled from a total of 1051 benchmark data at a ratio of 80/20%. The performance of each model was evaluated using the R^2 metric, as shown in Fig. S4a. The mean values of test R^2 from 20 outcomes for RF, KRR and MLP are 0.72, 0.71 and 0.77, respectively. Additionally, the average of the 20 predicted TC for 1051 polymers from each model versus the MD-calculated TC are visualized in Fig. S4b-d. Overall, the performance of the three models is comparable.

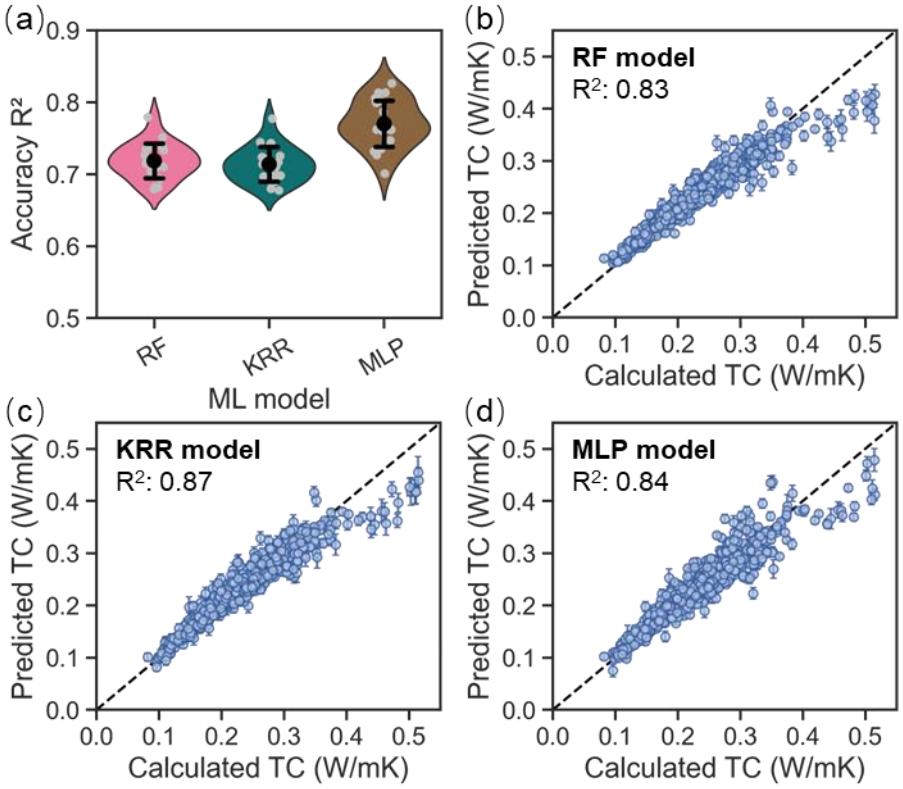

Fig. S4. Accuracy of different ML models trained with optimized descriptors. (a) Testing R^2 for three models trained with optimized descriptors. Each model was repeated 20 runs with randomly split of train/test data by 80/20%. Violins represent the distributions of the values; individual subsamples are shown in gray, and mean and standard deviation of R^2 in black. (b), (c) and (d) MD calculated TC versus predicted TC from RF, KRR and MLP. Each point is the average of the 20 prediction results, and the error bar is a standard deviation.

C. Principal component analysis for descriptors dimensionality reduction and comparison of the performance of different ML models

Principal component analysis (PCA) in Scikit-learn [2] was additionally performed for comparison with RF-based RFE. The PCA with 26 principal components of 98.0% variance was picked in Figure S5.

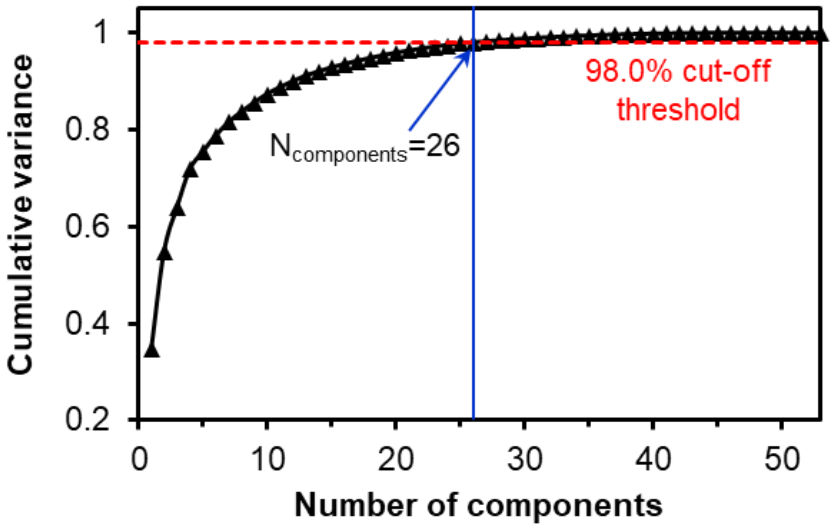

Fig. S5. Number of components versus cumulative variance in principal component analysis.

We additionally performed support vector machine (SVM) and Gaussian process (GP) regression with different descriptors in Scikit-learn [3] and Gpytorch [4] toolkits. Each combination of descriptors and ML model was evaluated by 20 repeated runs using randomly split candidates from the benchmark dataset according to a train/test ratio of 80/20%. As shown in Fig. S6a, the results suggest that these two ML models are more sensitive to high-dimensional redundant information, resulting in improved accuracy of models during down-selection stages. Also, the performance of graph descriptors in these two types of models is unsatisfactory, with the obtained test R^2 is much lower than that trained with optimized descriptors (Fig. S6b).

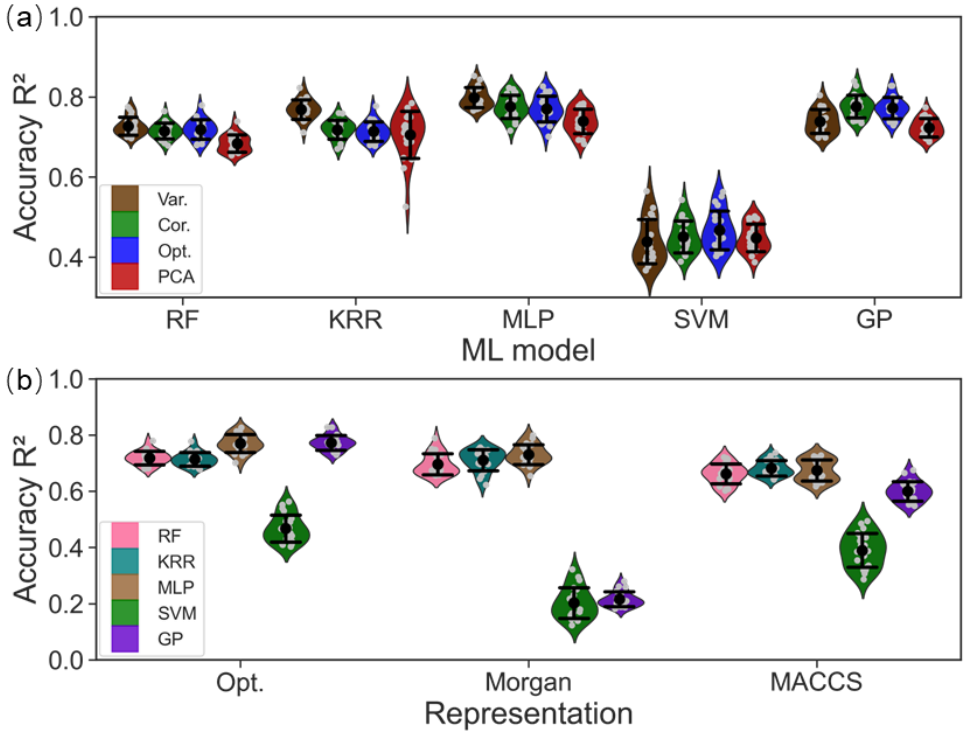

Fig. S6. Comparison of different combinations of descriptors and ML models, to predict TC. (a) R^2 for RF, KRR, MLP, SVM and GP models trained with each set of descriptors. (b) R^2 for five types of ML models trained with Optimized descriptors or graph descriptors. R^2 values were computed from 20 stratified subsampling repeats of the training data set. Violins represent the distributions of the subsampling results, mean and standard deviation of R^2 are shown in black, and individual subsample results are in gray.

D. Identification of polymers with high thermal conductivity in this work

In this work, a total of 40 amorphous polymers with TC > 0.400 W/mK were verified by MD simulations, displayed in Table S5 and Figure S7.

Table S5. MD-confirmed polymers with TC > 0.400 W/mK.

ID	SMILES	TC	SA
AP1	[*]C#CC=C[*]	1.072	7.105
AP2	[*]c1ccc2cc(-c3nc4cc5nc([*])[nH]c5cc4[nH]3)ccc2c1	0.974	3.552
AP3	[*]C=Cc1ccc(C=Cc2nc3cc4nc([*])[nH]c4cc3[nH]2)cc1	0.893	3.690
AP4	[*]c1ccc(-c2ccc(-c3ccc(-n4c(=O)c5cc6c(=O)n([*])c(=O)c6cc5c4=O)cc3)cc2)cc1	0.810	3.044
AP5	[*]c1ccc(-c2nc3cc4nc([*])[nH]c4cc3[nH]2)cc1	0.800	3.726
AP6	[*]c1ccc(-c2ccc(-c3nc4cc5nc([*])oc5cc4o3)cc2)cc1	0.729	3.322
AP7	[*]c1ccc(-n2c(=O)c3cc4c(=O)n([*])c(=O)c4cc3c2=O)cc1	0.725	3.550
AP8	[*]c1ccc(-c2ccc(-n3c(=O)c4cc5c(=O)n([*])c(=O)c5cc4c3=O)cc2)cc1	0.677	3.266
AP9	[*]c1ccc2[nH]c([*])nc2c1	0.619	4.647
AP10	[*]c1ccc(-c2nc3cc4nc([*])oc4cc3o2)c(O)c1	0.618	3.880
AP11	[*]NC(=O)C=CC(=O)Nc1nc([*])nc(N)n1	0.606	4.097
AP12	[*]Nc1ccc(C#Cc2ccc(NC(=O)c3ccc(C([*])=O)cc3)cc2)cc1	0.588	2.849
AP13	[*]c1ccc(N2C(=O)c3cc4cc5c(cc4cc3C2=O)Cc2cc3cc4c(cc3cc2C5)C(=O)N([*])C4=O)nc1	0.582	3.748
AP14	[*]C=CC=CN1C(=O)c2cc3c(cc2C1=O)Oc1cc2c(cc1C3)C(=O)N([*])C2=O	0.580	4.092
AP15	[*]c1ccc2c(c1)Cc1cc(-n3c(=O)c4cc5c(=O)n([*])c(=O)c5cc4c3=O)ccc1-2	0.566	3.563
AP16	[*]c1ccc(N2C(=O)c3cc4cc5c(cc4cc3C2=O)Cc2cc3cc4c(cc3cc2C5)C(=O)N([*])C4=O)cc1	0.564	3.551
AP17	[*]c1ccc2cc(-c3nc4ccc(-c5ccc6nc([*])[nH]c6c5)cc4[nH]3)ccc2c1	0.553	3.284
AP18	[*]c1nc2cc3nc(-c4ccc([*])o4)[nH]c3cc2[nH]1	0.547	4.023
AP19	[*]Nc1ccc(NC(=O)c2ccc(C([*])=O)cc2)cc1	0.545	2.758
AP20	[*]c1ccc(-c2ccc(-c3nc4ccc(-c5ccc6nc([*])oc6c5)cc4o3)cc2)cc1	0.542	3.187
AP21	[*]C=CC=CN1C(=O)c2cc3cc4c(cc3cc2C1=O)Cc1cc2cc3c(cc2cc1C4)C(=O)N([*])C3=O	0.533	3.964
AP22	[*]Nc1ccc(C([*])=O)cc1	0.527	3.783
AP23	[*]c1ccc([*])[nH]1	0.517	5.890
AP24	[*]NNC(=O)C([*])=O	0.504	5.361
AP25	[*]c1ccc(-c2ccc(N3C(=O)c4cc5cc6c(cc5cc4C3=O)Cc3cc4cc5c(cc4cc3C6)C(=O)N([*])C5=O)cc2)cc1	0.502	3.485
AP26	[*]C=CN1C(=O)c2cc3cc4c(cc3cc2C1=O)Cc1cc2cc3c(cc2cc1C4)C(=O)N([*])C3=O	0.494	3.790
AP27	[*]NC(=O)c1ccc(C(=O)Nc2cnc([*])nc2)cc1	0.491	3.173
AP28	[*]c1ccc(N2C(=O)c3cc4cc5c(cc4cc3C2=O)Cc2cc3cc4c(cc3cc2C5)C(=O)N([*])C4=O)c2cccnc12	0.485	3.751

ID	SMILES	TC	SA
AP29	[*]Nc1ccc(NC(=O)C=CC([*])=O)cc1	0.479	3.401
AP30	[*]c1ccc(-c2nc3cc(-n4c(=O)c5cc6c(=O)n([*])c(=O)c6cc5c4=O)ccc3[nH]2)cc1	0.472	3.513
AP31	[*]CNC(=O)N[*]	0.470	5.778
AP32	[*]c1ccc(NC(=O)c2ccc(C(=O)Nc3ccc(-n4c(=O)c5cc6c(=O)n([*])c(=O)c6cc5c4=O)cc3)cc2)cc1	0.468	3.125
AP33	[*]NNC(=O)C=CC(=O)Nc1ccc(NC(=O)C=CC([*])=O)cc1	0.460	3.279
AP34	[*]Nc1nnc(Nc2n[nH]c([*])n2)[nH]1	0.451	5.063
AP35	[*]c1ccc(N2C(=O)c3cc4cc5c(cc4cc3C2=O)Cc2cc3cc4c(cc3cc2C5)C(=O)N([*])C4=O)c2c1CC2	0.432	3.777
AP36	[*]NNC(=O)c1ccc(-c2ccc(C(=O)NNC(=O)c3cccc(C([*])=O)c3)cc2)cc1	0.419	2.778
AP37	[*]c1ccc(Nc2ccc(-n3c(=O)c4cc5c(=O)n([*])c(=O)c5cc4c3=O)cc2)cc1	0.419	3.314
AP38	[*]C(=O)NNC(=O)c1ccc([*])nc1	0.415	3.882
AP39	[*]c1ccc(-c2nc3cc(-c4ccc5oc([*])nc5c4)ccc3o2)cc1	0.410	3.311
AP40	[*]c1ccc(N2C(=O)c3cc4cc5c(cc4cc3C2=O)Cc2cc3cc4c(cc3cc2C5)C(=O)N([*])C4=O)c(C)c1	0.410	3.682

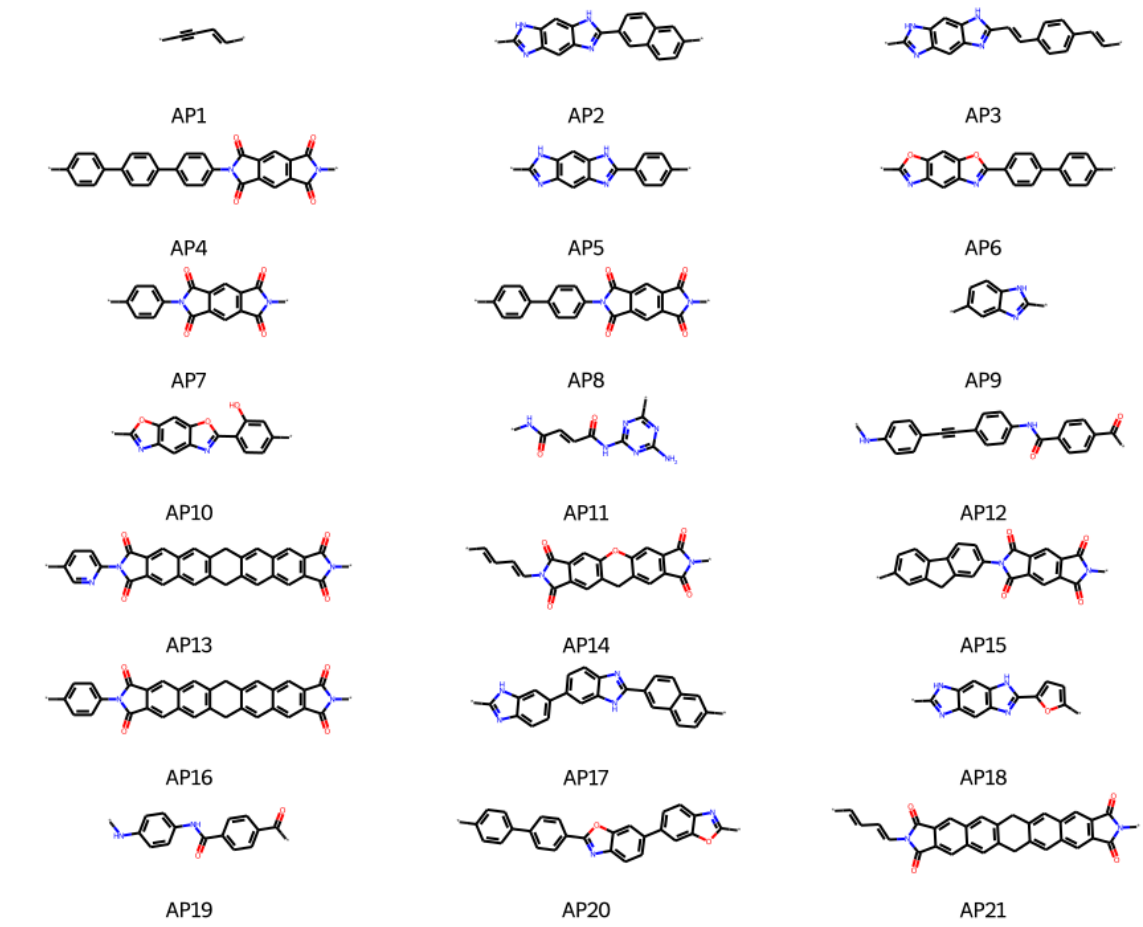

Fig. S7. Continued.

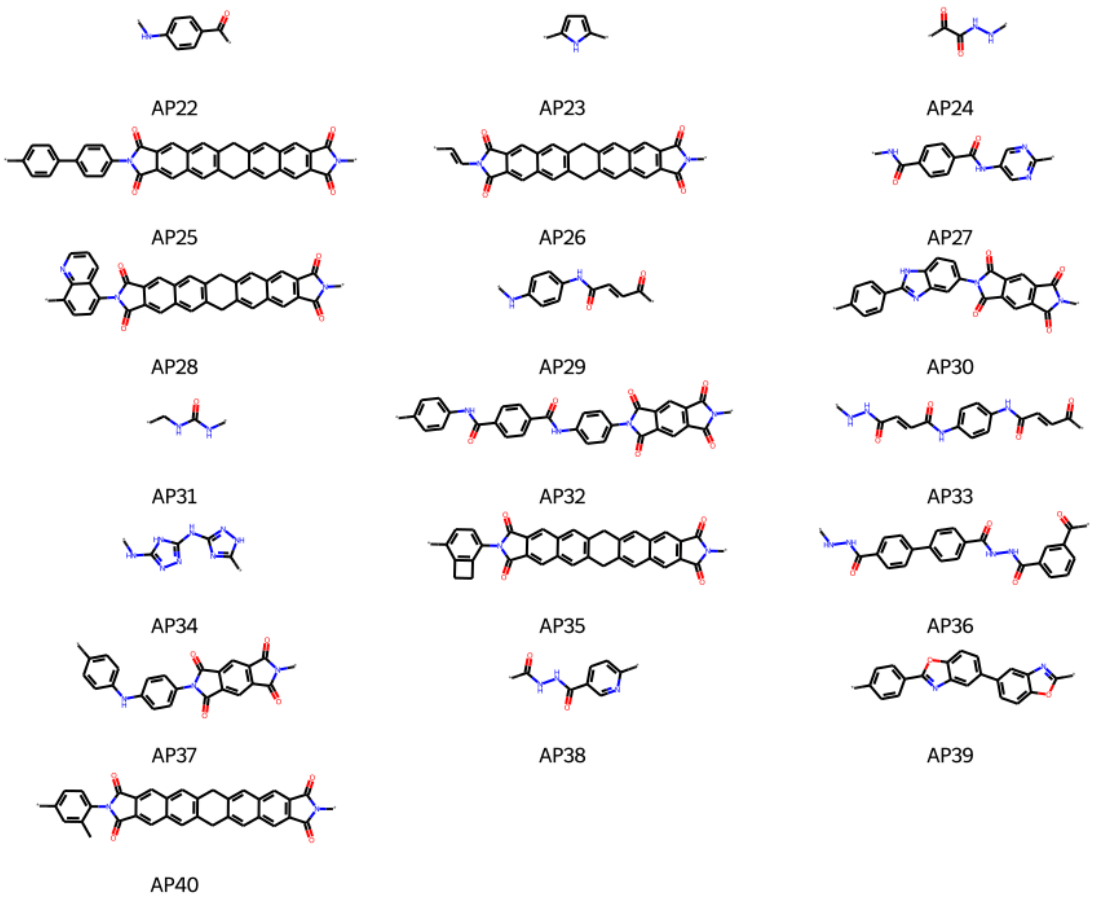

Fig. S7. MD-validated polymer monomers with TC > 0.400 W/mK.

E. Analysis of the feature importance

Figure S8 reflects the strong positive correlation between BCUTZ-1h and maximum atomic mass (Mass_max) in the monomer (Figure S8a). Typically, the presence of large masses of atoms in the system suppresses lattice vibrations, resulting in small phonon group velocities and low TC (Figure S8b).

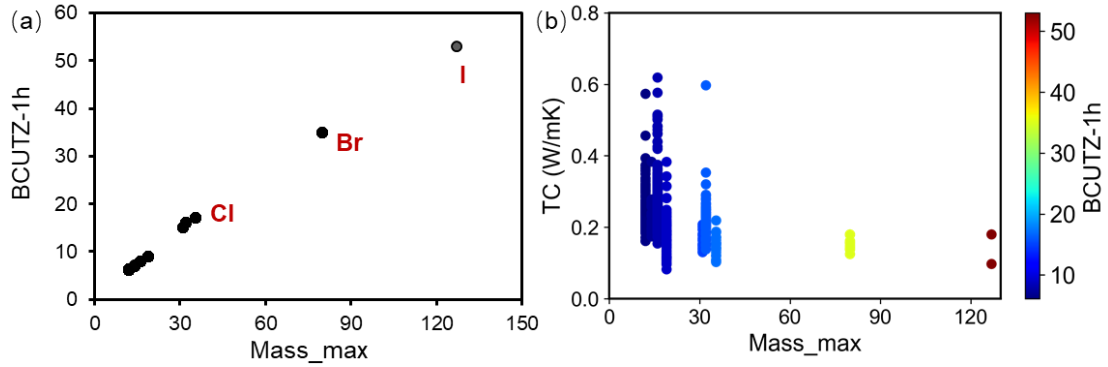

Fig. S8. Feature impact analysis. (a) Descriptor BCUTZ-1h versus maximum atomic mass (Mass_max) in the monomer. (b) Mass_max versus TC. The value of BCUTZ-1h is color coded to the data points.

F. Symbolic regression for the construction of analytical models

The symbolic regression with 11 parameters was applied for construction of analytical models. Table S6 listed the 11 parameters, which includes the amorphous system properties density, number density, radius of gyration, persistence length, specific heat capacity at constant pressure and at constant volume after equilibrium, in addition to the first six MD-based parameters. The relationships between these polymer properties and TC and its decomposition terms of convection, intra chain and inter chain are shown in Figure S9.

Table S6. Setting of parameters in symbolic regression.

No.	Description	Symbol
x_0	MW_ratio	M
x_1	K_bond_ave	k_{bavg}
x_2	K_ang_ave	k_{aavg}
x_3	Mass_max	m_{max}
x_4	nHBDon	n_H
x_5	density	ρ
x_6	number density	n
x_7	radius of gyration	R_g
x_8	persistence length	ξ
x_9	specific heat capacity at constant pressure	C_P
x_{10}	specific heat capacity at constant volume	C_V

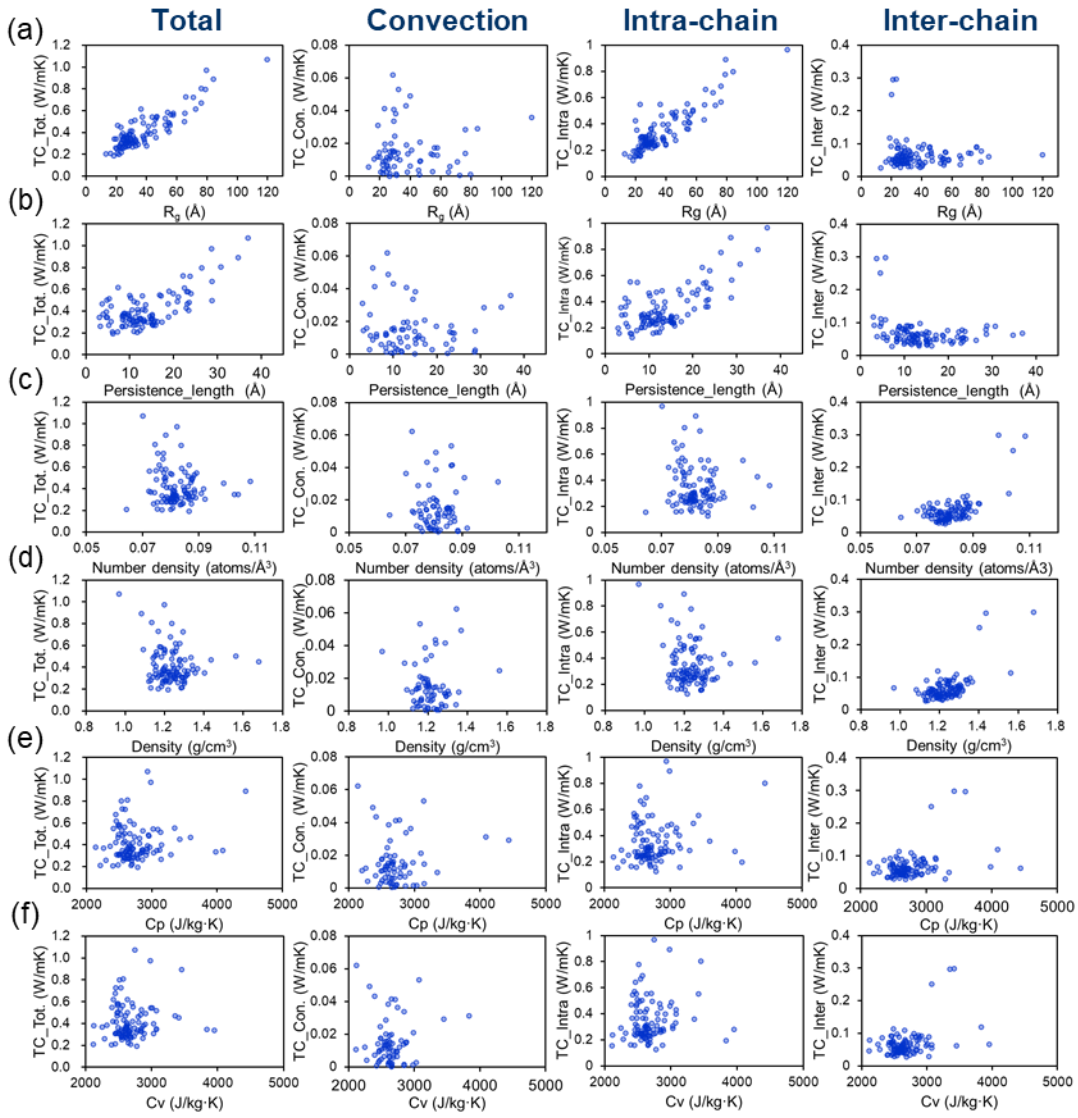

Fig. S9. The relationships between the total thermal conductivity (TC_Tot.), $\boldsymbol{\lambda}$-convection (TC-Con.), λ-intra chain (TC-Intra), λ-inter chain (TC-Inter) and (a) radius of gyration ($\boldsymbol{R_g}$), (b) persistence length, (c) number density, (d) density, (e) and (f) specific heat capacity at constant pressure ($\boldsymbol{C_P}$) and at constant volume ($\boldsymbol{C_V}$), respectively.

Figure S10 illustrates the fitting results for the thermal conductivity of the six formulas at the Pareto front, which have good agreement with the TC calculated by MD simulations.

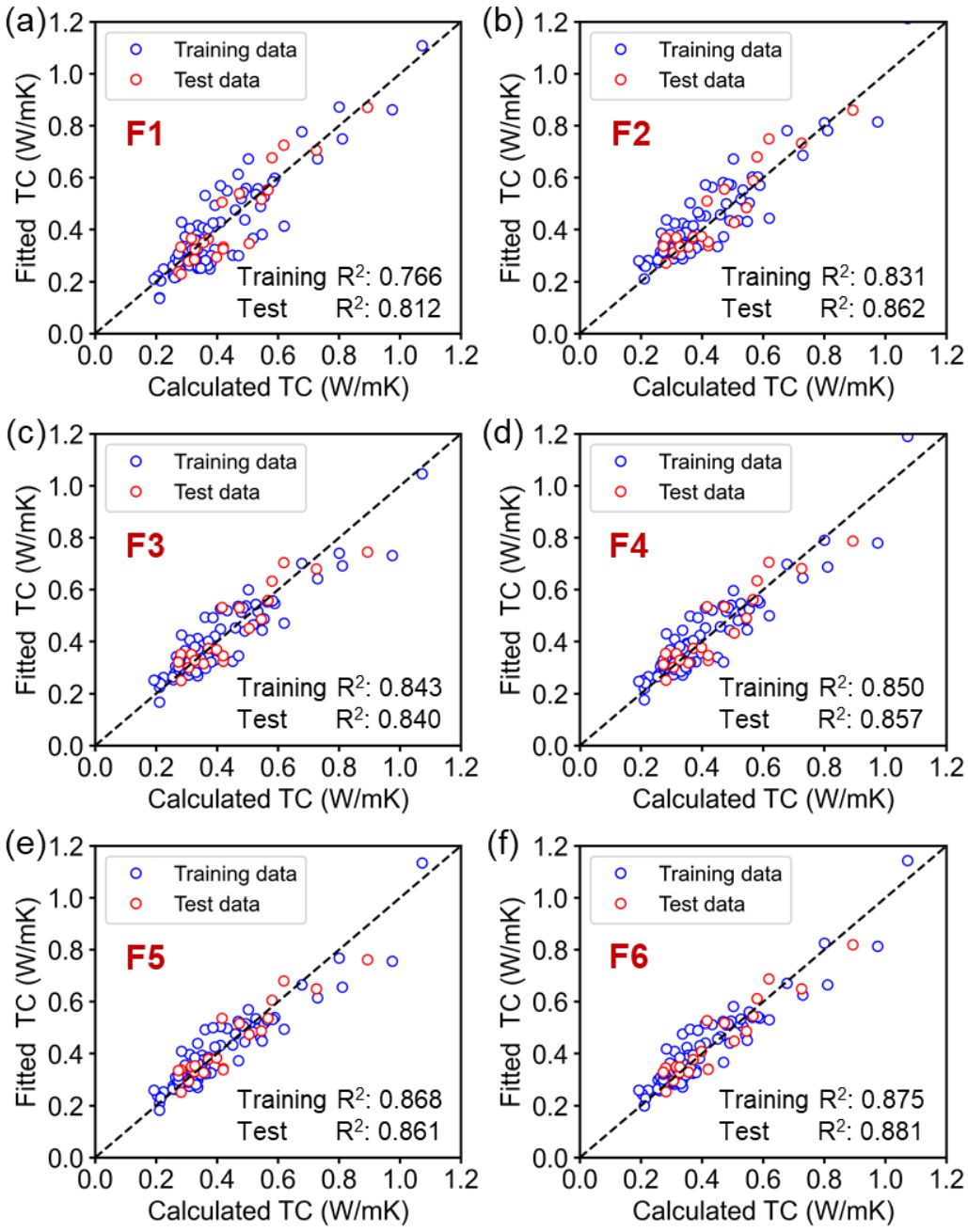

Fig. S10. Six formulas F1~F6 at Pareto front in Fig 6b of maintext.

G. Amorphous polymer generation and thermal conductivity calculation

Figure S11 illustrates the procedure for amorphous polymer generation and thermal conductivity calculation. A polymer monomer was loaded into RDKit software [5] in the form of simplified molecular input line entry system (SMILES) [6] and then polymerized by a self-avoiding random walk algorithm to form a ~1000 atom polymer chain. Next, a single polymer chain is randomly arranged and rotated for replication to form a ~10000 atom simulation cell. Next, 10 polymer chains were randomly arranged and rotated to prevent overlap with each other and to form a ~10,000 atoms simulation cell. After that, the simulation cell was initialized and annealed at 700 K to form an amorphous cell. Moreover, the amorphous unit was equilibrated by the 21-steps equilibration scheme [7].

The equilibrated amorphous unit was replicated 3 times in the x-direction and the reverse NEMD simulations proposed by Müller-Plathe [8] was performed to obtain the temperature gradient. Ultimately, the thermal conductivity of the final amorphous polymer is calculated by Fourier's law. The whole simulation process was implemented in Radonpy, more details can be found elsewhere [9].

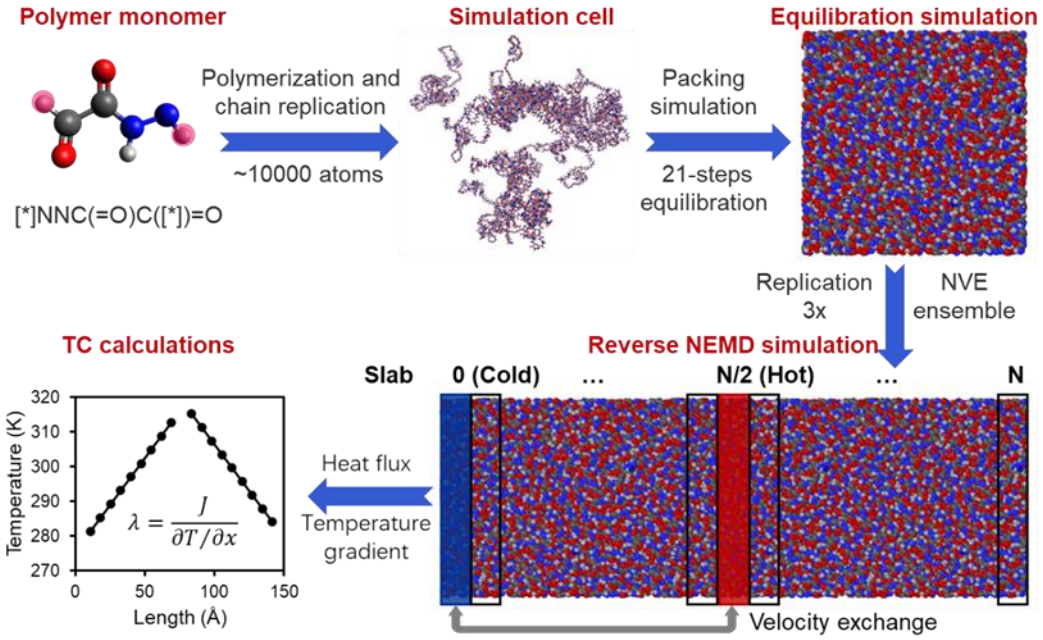

Fig. S11. Procedure of amorphous polymer generation and thermal conductivity calculation through MD simulations**.**

SI References

1. L. McInnes, J. Healy and J. Melville, *arXiv preprint arXiv:1802.03426*, 2018.
2. E. Bisong, in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, ed. E. Bisong, Apress, Berkeley, CA, 2019, DOI: 10.1007/978-1-4842-4470-8_24, pp. 287-308.
3. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *the Journal of machine Learning research*, 2011, **12**, 2825-2830.
4. J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel and A. G. Wilson, *Advances in neural information processing systems*, 2018, **31**.
5. G. Landrum, *Greg Landrum*, 2013.
6. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31-36.
7. G. S. Larsen, P. Lin, K. E. Hart and C. M. Colina, *Macromolecules*, 2011, **44**, 6944-6951.

8. F. Müller-Plathe, *The Journal of Chemical Physics*, 1997, **106**, 6082-6085.
9. Y. Hayashi, J. Shiomi, J. Morikawa and R. Yoshida, *npj Computational Materials*, 2022, **8**, 222.