

Supplementary Information for

Discovery of thermosetting polymers with low hygroscopicity, low thermal expansivity, and high modulus by machine learning

Xinyao Xu, Wenlin Zhao, Yaxi Hu, Liquan Wang, Jiaping Lin*, Huimin Qi, and Lei Du*

Shanghai Key Laboratory of Advanced Polymeric Materials, Key Laboratory for Ultrafine Materials of Ministry of Education, Frontiers Science Center for Materiobiology and Dynamic Chemistry, School of Materials Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China

Contents

S1.	Details of structural representation learning	SI 3
S2.	All-atomic molecular simulations	SI 6
S3.	Procedure for curing of polycyanurates in MD simulations	SI 7
S4.	Calculation of hygroscopicity using MC simulations	SI 9
S5.	Fitting model of coefficient of linear thermal expansion (CLTE)	SI 10
S6.	Calculation of tensile modulus using MD simulations	SI 11
S7.	Polycyanurates and their experimental data	SI 12
S8.	Detailed comparisons between experimental and calculated results.....	SI 14
S9.	Property distributions of sampling space	SI 16
S10.	Details of multi-fidelity surrogate model	SI 17
S11.	Implementation of gene definition, cleaning, and combination	SI 21
S12.	Detailed structures of promising and existing CE monomers	SI 23
S13.	Preparation of the FCE monomer.....	SI 26
S14.	Structural characterization and the purity of the FCE	SI 27
S15.	The thermal property and processability of the FCE	SI 29
S16.	Specimen preparation for testing.....	SI 31
S17.	A statistical method for quantifying the importance of molecular fragments.....	SI 32
	References.....	SI 34

S1. Details of structural representation learning

There is no standard way to represent the chemical structure in digital form. Since the chemical structures are typically non-Euclidean space graphs, graph neural network (GNN) models can help develop chemical and materials informatics.^{S1-S5} Inspired by the work of Hatakeyama-Sato *et al.*,^{S5} we attempted to learn the low-dimensional representation of CE monomers based on the graph self-supervised learning framework (see Figure S1a). Some chemical properties which can be calculated freely by the chemical toolkit were used as graph-level auxiliary properties. Moreover, we used several state-of-the-art GNNs, including the graph convolutional network (GCN),^{S6} graph attention network (GAT),^{S7} and gated attention network (GaAN),^{S8} to convert CE monomers to latent vectors.

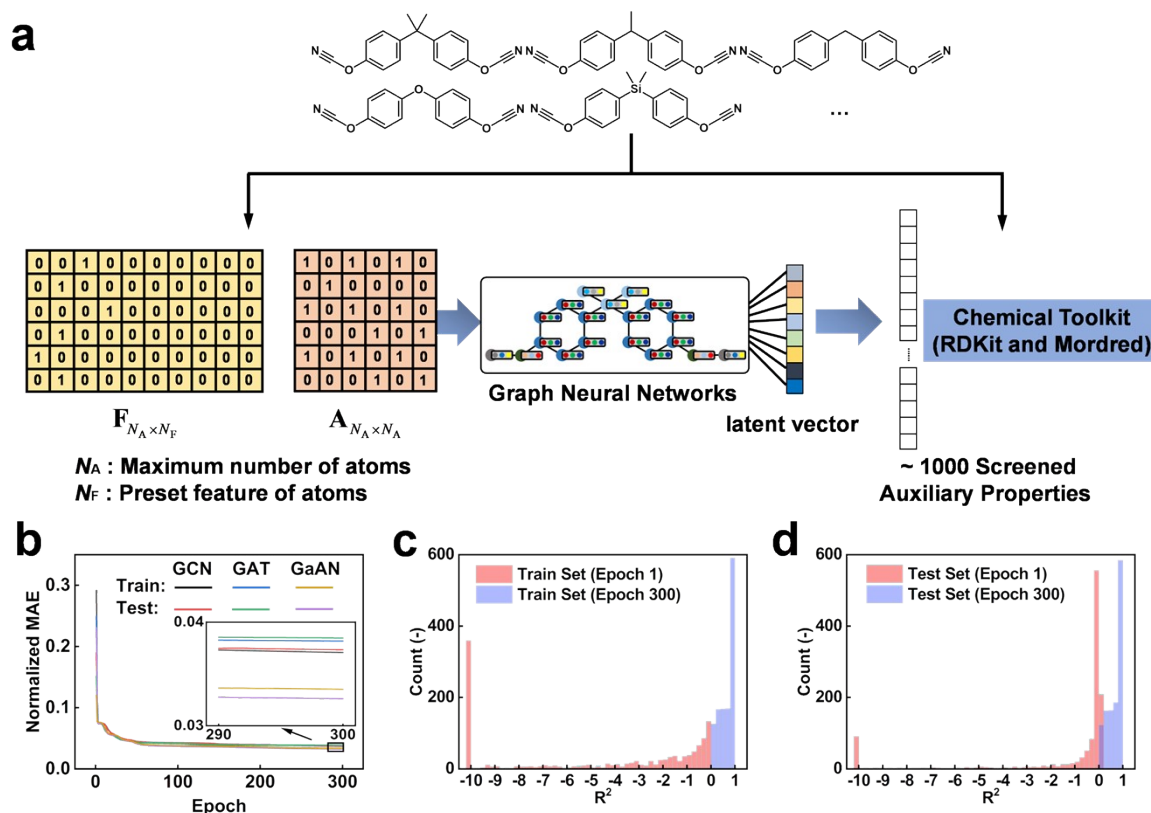


Figure S1. A framework of representation learning to express CE monomers. a) The implementation of graph self-supervised learning in this work. b) Learning process of the GNN models. Distribution of R^2 for each auxiliary property of the GaAN model at the beginning (Epoch 1) and the end (Epoch 300) of the learning process on c) train set and d) test set, respectively.

On the one hand, the CE monomers can be represented by a non-Euclidean graph, which is the input of the framework. The CE monomer can be defined as a graph (G) that consists of a tuple (F, A). Here, F is a matrix of atom features, and A is the adjacency matrix that reflects the atom connection in the monomer. We used one-hot encoding to generate a feature matrix of the CE monomers. The maximum number of atoms (N_A) in a molecule is set to 100 (Implicit H). If the number is smaller than 100, the matrix is completed with 0 to ensure consistency of the representation. The preset atom features (N_F) include atom type {B, C, N, O, F, Si, P, S, Cl, I}, number of bonded atoms {0, 1, 2, 3, 4, 5, 6}, number of bonded hydrogens {0, 1, 2, 3}, atomic implicit valence {0, 1, 2, 3, 4, 5}, and aromaticity indicator {True, False}. If the feature of an atom is not in the preset feature set of atom type, number of bonded hydrogens, and atomic implicit valence, it is recorded as 'others', respectively. As such, each atom can be represented by a 32-dimensional one-hot vector. The adjacency matrix reflects only the connectivity of atom pairs since the bond information (such as single bond, double bond, triple bond, and aromatic bond) can be inferred from the atom feature.

On the other hand, descriptors (graph-level auxiliary properties) calculated using RDKit and Mordred packages can effectively describe chemical features and are the output of the framework here.^{S9,S10} The steps are as follows. 1) All CE monomers are optimized by MMFF force field using the RDKit package. 2) The descriptors of CE monomers are calculated by the Mordred package. Here, all default molecular descriptors implemented by the Mordred package (including 1613 2D descriptors and 213 3D descriptors) are calculated. 3) Data preprocessing was performed on the calculated descriptors, the descriptors with outliers (such as calculation errors and values containing characters) and the descriptors with repetition ratio greater than 0.5 were removed. As a result, 1224 descriptors were obtained. 4) Each screened descriptor is normalized.

We constructed several GNN models and connected them to a multilayer perceptron (MLP), as shown in Figure S1a. The hyper-parameters of the GNNs in this work are presented

as follows. A 3-layers GNN model with the output dimension of 64, 64, and 64 was used to update atom states. Afterward, a single-layer MLP with an output dimension of 32 was used for the readout function. A 4-layers MLP with 16, 8, 128, and 1224 units was then connected with the model above. Here, 1224 is the number of the screened descriptors, and the latent vectors of chemical structures (8-dimensional vectors) were obtained from the layer with eight units. In addition, the optimizer, loss function, learning rate, epoch, and batch size are Adam, mean-square error (MSE), 0.001, 300, and 32, respectively. The dataset was split into train and test sets with a ratio of 7:3. The GNN models are implemented by TensorFlow 2.3.0.^{S11} The model learns how to represent CE monomers based on the virtual CE dataset. And the characteristics of CE monomers are represented by low-dimensional vectors using a well-trained GNN model.

After learning 300 epochs, each model achieved convergence. Figure S1b shows the change in the normalized mean absolute error (MAE) on the train and test set during the learning process. After convergence, the GaAN has the lowest value of the normalized MAE among these three GNN models, and thus we used the well-trained GaAN model for digitizing chemical structures. Furthermore, we calculated the distribution of the coefficient of determination (R^2) for each screened auxiliary property at the beginning (Epoch 1) and end (Epoch 300) of the learning process to show the ability of the GaAN model to represent molecules, which are presented in Figure S1c (on the train set) and Figure S1d (on the test set). When only one epoch was trained, the model could not capture the molecular characteristics well (R^2 is almost less than 0). The R^2 is significantly improved at the end of the training, indicating that the model can better describe molecules.

S2. All-atomic molecular simulations

In this work, we used all-atomic molecular simulations to obtain the properties data (calc.) of polycyanurates. The calculations include: simulating the curing process of polycyanurates by all-atomic molecular dynamics (MD) simulations (see Section S3), calculating the ultimate hygroscopicity (RT, RH60%) of cured polycyanurates by Monte Carlo (MC) simulations (see Section S4), recording the volume of cured polycyanurates simulation box at different temperatures (below glass transition temperature) and fitting coefficient of thermal expansion (CTE) by a formula (see Section S5), and calculating the tensile modulus by MD simulations (see Section S6).

Unless explicitly stated, the simulation was performed in the box with periodic boundary conditions based on the COMPASS force field, Nose thermostat, and Andersen barostat. The default pressure and temperature are 101.325 kPa and 473.15 K, respectively. The Particle-Particle-Mesh (PPPM) summation method was used to sum the long-range electrostatic interaction terms. The direct atom-based method was chosen for the short-range van der Waals interactions. The time step is 1fs.

S3. Procedure for curing of polycyanurates in MD simulations

We designed a crosslinking scheme based on a cut-off radius criterion and a multi-step relaxation to construct the cured polycyanurates networks,^{S12-S14} which is shown in Figure S2. Firstly, cyanate ester (CE) monomers were placed in a simulation box using an Amorphous Cell module. To keep consistency, the total number of cyanate groups in the simulation box is 300. Then, the system was subjected to geometry optimization, and the MD simulation was successively performed under *NPT*, *NVT*, and *NPT* ensembles with 100 ps, 100 ps, and 200 ps, respectively.

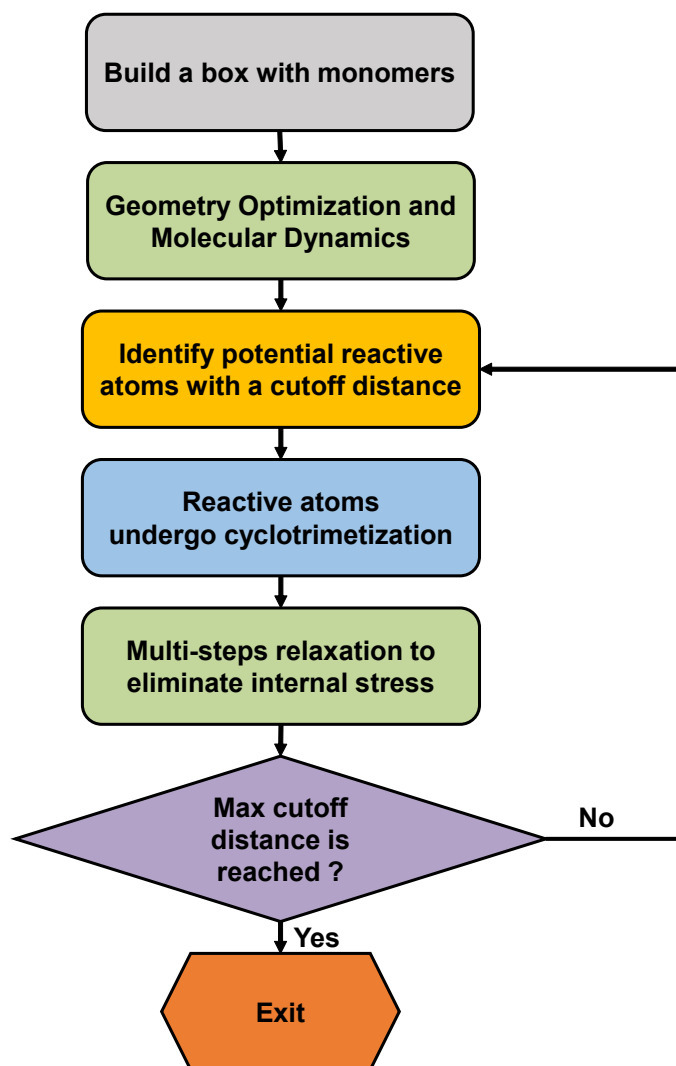


Figure S2. Crosslinking flowchart of polycyanurates in the MD simulations.

Afterward, the equilibrium configuration was used for the following crosslinking scheme. Only the cyclotrimerization of cyanate groups was considered in this procedure, as shown in Figure S3a (the reactive cyanate group and triazine ring) and Figure S3b (the topological structure of the cured matrix in the simulation box). Pairs of reactive atoms (C atom in one cyanate group and N atom in another cyanate group) that meet the cut-off distance criterion can be connected by forming new bonds. Bonds and neighboring atoms for the reacted atoms were adjusted, and the total system was subjected to *NPT-NVT-NPT* simulations with 100-100-100 ps to eliminate internal stress stemming from the crosslinking. The above simulations were repeated until no new pair of reactive atoms were generated within the maximum cut-off distance. To achieve the ultimate degree of crosslinking, we increased the cut-off distance from 3 Å to 13 Å, where the step is 0.1 Å. Under a certain cut-off distance, only one cyclotrimerization occurs per cycle. When there is no cyclotrimerization occurs, the following cut-off distance is entered. The entire process was implemented through Perl scripts.

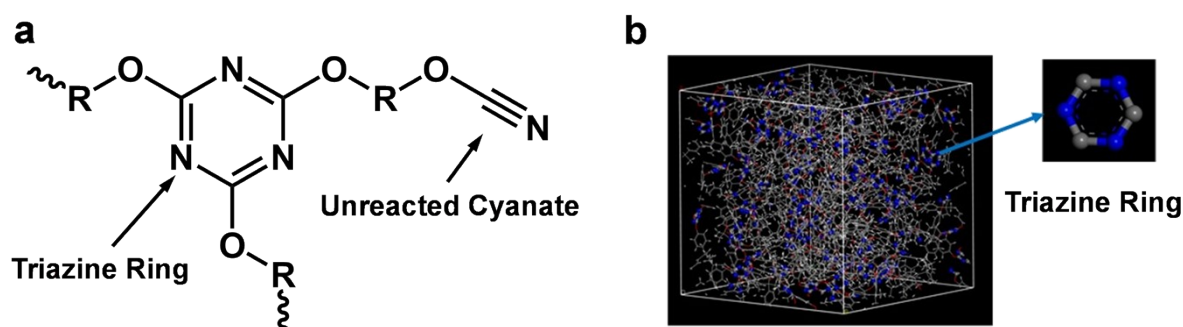


Figure S3. a) The reactive group (cyano group) and the product (triazine ring). b) The topological structure of a highly crosslinked polycyanurate in the simulation box.

S4. Calculation of hygroscopicity using MC simulations

Metropolis Monte Carlo (MC) method, realized through the fixed pressure task in the Sorption module,^{S15} was used to simulate water molecules absorbed in cured polycyanurates. The configurations are sampled from a grand canonical ensemble, and the probability of a configuration, m , is given by

$$\rho_m = CF(\{B\}_m) \exp[-E_m/k_B T] \quad (\text{S-1})$$

where C is an arbitrary normalization constant, E_m is the total energy of configuration m , k_B is the Boltzmann constant, and T is the temperature.

The adsorption system simulated an RH60% atmospheric environment at room temperature (25 °C). The Ewald summation method has been used to sum the long-range electrostatic interaction terms.^{S16} The number of equilibration steps was 20,000,000, and the number of production steps was set to 40,000,000. All unoccupied volumes in the simulation box were considered.

S5. Fitting model of coefficient of linear thermal expansion (CLTE)

We used MD simulations to obtain the coefficient of thermal expansion (CTE) of polycyanurates below the glass transition temperature.^{S17} The CTE refers to the regular coefficient that the geometric properties of a substance change with the temperature change. The coefficient β of volume thermal expansion is

$$\beta = \frac{1}{V} \frac{dV}{dT} \quad (\text{S-2})$$

The coefficient α of linear thermal expansion can be given as $\alpha = \beta / 3$ for isotropic systems.

Herein, we recorded the density of the polycyanurates in the simulation box at various temperatures. The temperature was increased by 10K from 293K to 473K, with the *NPT* dynamics of 200 ps performed at each temperature.

S6. Calculation of tensile modulus using MD simulations

We used a constant strain method in Forcite to calculate the tensile modulus of cured polycyanurates.^{S15} This approach estimates the constant elastic matrix by a series of finite difference approximations. Some strains are applied to the simulation box, and the metric tensor, \mathbf{G} , is given by

$$\mathbf{G} = \mathbf{H}_0^T [2\boldsymbol{\varepsilon} + \mathbf{I}] \mathbf{H}_0 \quad (\text{S-3})$$

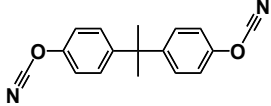
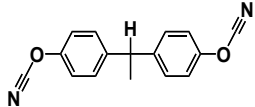
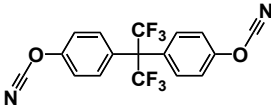
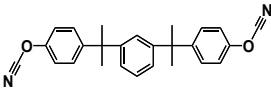
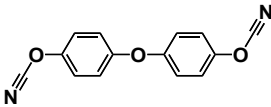
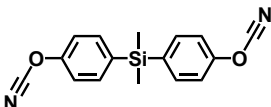
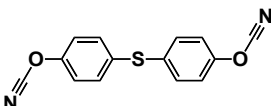
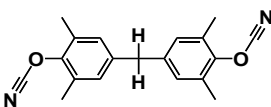
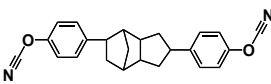
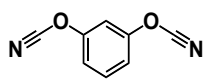
where \mathbf{H}_0 is formed from the lattice vectors, \mathbf{I} is the identity matrix, and \mathbf{H}_0^T is the transpose of \mathbf{H}_0 .

The new lattice parameters can be derived from the metric tensor \mathbf{G} and are used to transform the cell parameters. Then, the strained structure is optimized for internal relaxation, and the stress is calculated. A stiffness matrix is obtained from a linear fit between the applied strain and the resulting stress. In our simulations, the Ewald summation method has been used to sum the long-range electrostatic interaction terms. The number of strains was 15, and the maximum strain was 0.003.

S7. Polycyanurates and their experimental data

We collected existing CE resins with experimental properties of hygroscopicity (RT, RH60%), CLTE (below T_g), and tensile modulus (RT). Table S1 shows the properties of several cured neat polycyanurates, including CLTE and tensile modulus. The data were collected from existing literature given in Refs. S18-S20.

Table S1. Structure and properties of neat polycyanurates*

Structure	Abbreviation	Hygroscopicity [%]	CLTE [ppm/°C]	Tensile Modulus [Gpa]
	BADCy	0.90	64	3.17
	BECy	\	64	2.90
	HFBACy	\	54	3.11
	MBCy	\	70	3.16
	OXOCy	\	65	3.35
	SiMCy	\	69	2.80
	THIOBCy	\	68	2.76
	TMBFCy	0.53	71	2.97
	DOCy	\	68	2.89
	RECORCy	\	49	4.76

*Hygroscopicity is tested under the environment of RH60%, RT in our lab.

In addition, we blended the TMBFCy with DOCy with various ratios. The properties of these samples, including hygroscopicity and CLTE, were characterized. The results are shown in Table S2.

Table S2. Composition and properties of cured polycyanurates*.

Sample Index of TMBFCy/DOCy	Molar Ratio of TMBFCy/DOCy	Hygroscopicity [%]	CLTE [ppm/°C]
1	1:9	0.318	61.44
2	2:8	0.395	60.28
3	3:7	0.427	59.44
4	4:6	0.445	58.44
5	5:5	0.452	62.21
6	1:0	0.530	71.00

*Hygroscopicity is tested under the environment of RH60 %, RT. CLTE is tested below T_g .

S8. Detailed comparisons between experimental and calculated results

To verify the reliability of the simulation scheme in Sections S2-S6, we first collected the experimental data of neat CE resins (nCEs) and TMBFCy/DOCy blend resins (bCEs) (see Table S1 and TableS2) and then computed the properties of nCEs and bCEs using this simulation scheme to obtain calculation data. The comparisons of hygroscopicity, CLTE, and tensile modulus between calculated and experimental results are depicted in Figure S4.

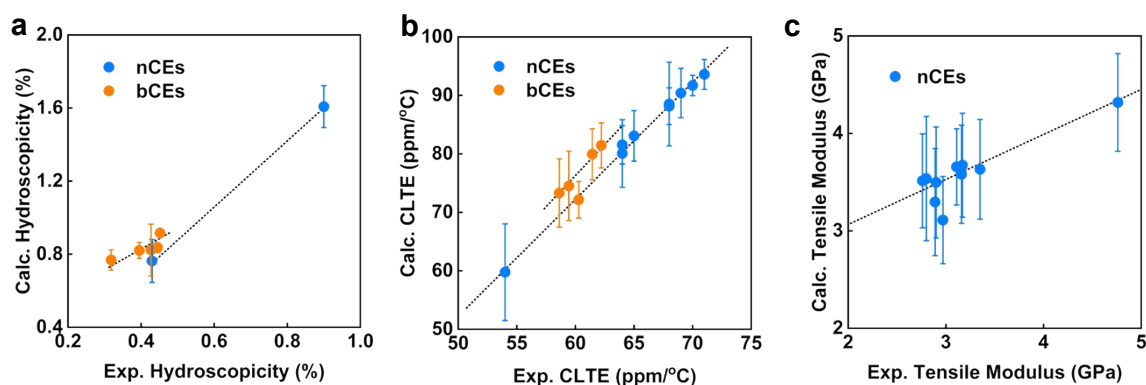


Figure S4. Verifying the reliability of the simulation scheme. The comparison between the calculated and experimental data of a) hygroscopicity, b) CLTE, and c) tensile modulus of existing neat CE resins (nCEs) and TMBFCy/DOCy blend resins (bCEs).

The figures show that the calculated and experimental properties show similar trends. The calculated hygroscopicity is close to once larger than the experimental one (Figure S4a). The reasons may be as follows. First, the calculated property is the extreme moisture uptake, and the time tends to infinity, while the experimental result is the plateau value of the moisture curve. Second, all the unoccupied volume in the simulation box is considered, but in the actual situation, water molecules cannot enter some closed pores in the material. For the CLTEs shown in Figure S4b, there is an excellent linear correlation between the calculated and experimental results of the control group, indicating that the computations can estimate the materials reasonably in terms of CLTEs. One can see from Figure S4c that the average value of the calculated tensile modulus is basically consistent with the experimental one. However, some

deviations exist, such as the fluctuation of calculated tensile moduli. This is because, in the calculation process, we applied a small strain to the simulation box and then performed relaxation to calculate the resulting stress. The internal stress in the simulation box due to the simulated crosslinking reaction may affect the calculated results and lead to fluctuation.

Although there is a gap between the calculated and the experimental results, the consistency of the trend indicates that this simulation scheme can be used to generate low-fidelity data required by multi-fidelity learning.

S9. Property distributions of sampling space

The distributions of calculated properties of hygroscopicity, CLTE, and tensile modulus for the low-fidelity dataset are depicted in Figures S5a-c, respectively. As shown, the body distribution of various properties is consistent with our scientific intuition, indicating that ML models have the potential to learn general rules. Furthermore, several extremely low (or high) property values exist, which benefits the ML models in learning some marginal features.

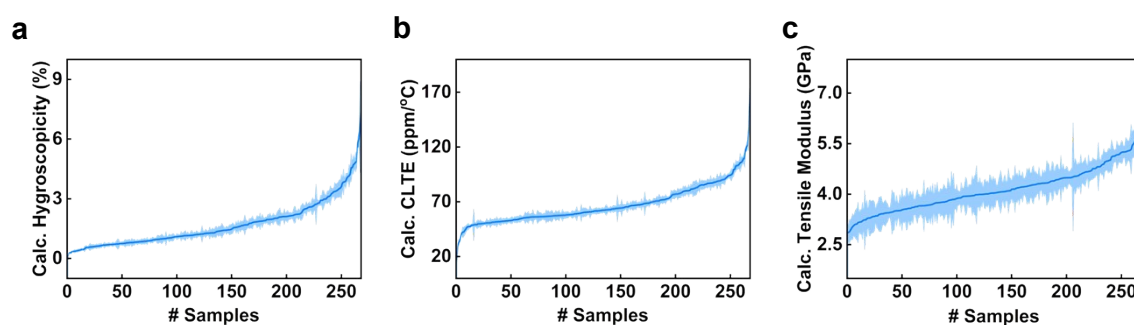


Figure S5. Preparing the low-fidelity calculated dataset. The calculated a) hygroscopicity, b) CLTE, and c) tensile modulus of polycyanurates and their distributions.

S10. Details of multi-fidelity surrogate model

We utilized a multi-fidelity surrogate model to learn polymer properties using their latent vector.^{S21} A surrogate function $Y_H(\mathbf{X})$ can be represented as

$$Y_H(\mathbf{X}) = \rho Y_L(\mathbf{X}) + Y_D(\mathbf{X}) \quad (\text{S-4})$$

Here, $Y_H(\mathbf{X})$ and $Y_L(\mathbf{X})$ represent the surrogate function of the high- and low-fidelity data, respectively; $Y_D(\mathbf{X})$ represents the difference between $Y_H(\mathbf{X})$ and $Y_L(\mathbf{X})$; and ρ is a constant scaling factor.

The Gaussian process regression (GPR) with four different kernel functions was used to learn $Y_L(\mathbf{X})$ and $Y_D(\mathbf{X})$.^{S22} The GPR is within the Bayesian framework, wherein a Gaussian process is used to obtain the functional mapping $f(x) \rightarrow y$ based on the available training set and the Bayesian prior, incorporated using the covariance (or kernel) function. The process was implemented using the Scikit-Learn package.^{S23}

The selected kernel functions ($k_{\#1} \sim k_{\#4}$) are given as Equations (S-5~S-8):

$$k_{\#1}(x_i, x_j) = \exp\left[-\frac{1}{l} d(x_i, x_j)\right] + k_0(x_i, x_j) \quad (\text{S-5})$$

$$k_{\#2}(x_i, x_j) = \left[1 + \frac{\sqrt{3}}{l} d(x_i, x_j)\right] \exp\left[-\frac{\sqrt{3}}{l} d(x_i, x_j)\right] + k_0(x_i, x_j) \quad (\text{S-6})$$

$$k_{\#3}(x_i, x_j) = \left[1 + \frac{\sqrt{5}}{l} d(x_i, x_j) + \frac{\sqrt{5}}{3l} d(x_i, x_j)^2\right] \exp\left[-\frac{\sqrt{5}}{l} d(x_i, x_j)\right] + k_0(x_i, x_j) \quad (\text{S-7})$$

$$k_{\#4}(x_i, x_j) = \exp\left[-\frac{1}{2l^2} d(x_i, x_j)^2\right] + k_0(x_i, x_j) \quad (\text{S-8})$$

Here, $d(x_i, x_j)$ is the Euclidean distance, and the white kernel $k_0(x_i, x_j)$ satisfies

$$k_0(x_i, x_j) = \begin{cases} \text{noise_level} & \text{if } x_i = x_j \\ 0 & \text{else} \end{cases} \quad (\text{S-9})$$

For the task of learning the surrogate function $Y_L(\mathbf{X})$ of the low-fidelity data, the dataset was randomly divided into train, validation, and test set with a ratio of 7:2:1. The average performance of the models was obtained by adjusting the prior distribution and randomly dividing the data set 500 times. Figure S6 shows the distribution of MAE, MSE, and R^2 on these three properties. The models with good performance on the train and validation set were selected.

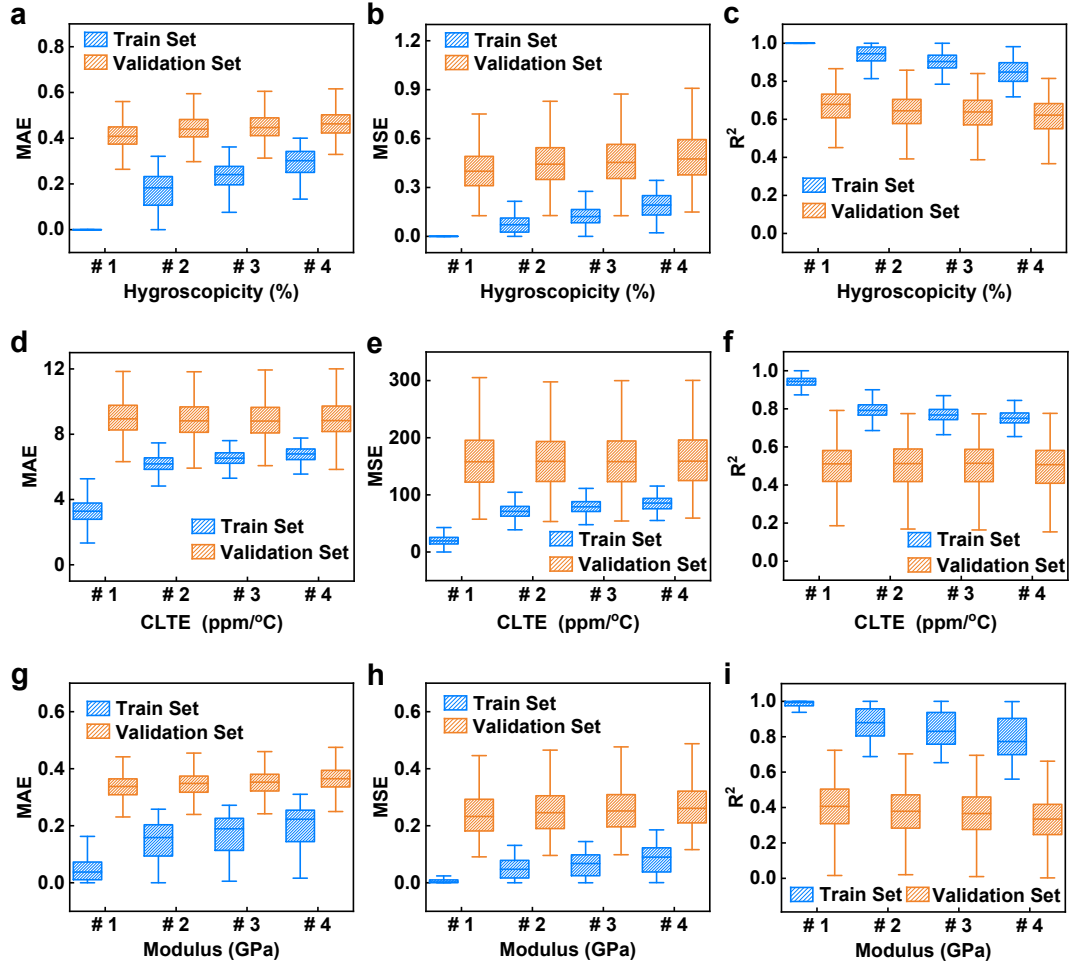


Figure S6. Comparison of different kernel functions by box plots. a) MAE, b) MSE, and c) R^2 of hygroscopicity for the train and validation set. d) MAE, e) MSE, and f) R^2 of CLTE for the train and validation set. g) MAE, h) MSE, and i) R^2 of tensile modulus for the train and validation set.

For the task of learning the difference function $Y_D(\mathbf{X})$, all available experiment data were used to map the relationship. The leave-one-out (LOO) cross-validation was used to utilize the

limited experimental data and avoid over-fitting efficiently. The $Y_H(\mathbf{X})$ model consists of the well-trained $Y_L(\mathbf{X})$ and the well-trained $Y_D(\mathbf{X})$. We used the $Y_H(\mathbf{X})$ model to carry out the high-throughput prediction.

We also trained ML-based models using only experimental data for comparisons to evaluate the advantages of multi-fidelity learning in addressing the "small data" challenge. All the steps used to train ML-based models are the same as for training $Y_D(\mathbf{X})$, except that the data used here are experimental properties rather than deviations. The model trained here is marked as $Y_E(\mathbf{X})$, where the subscript E means the experimental data only.

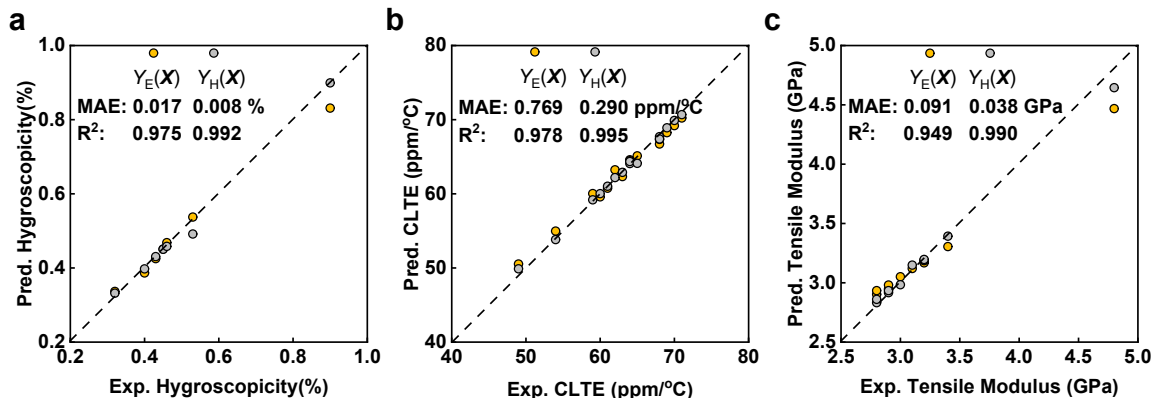


Figure S7. Comparisons of the established $Y_H(\mathbf{X})$ and $Y_E(\mathbf{X})$ models. ML-predicted properties of a) hygroscopicity, b) CLTE, and c) tensile modulus compared with the experimental ones. The performances of the ML-based models were evaluated by two metrics of R^2 and MAE.

For quantitative comparison, we also calculated the R^2 and MAE of $Y_E(\mathbf{X})$ models on experimental data. The R^2 - MAE pairs of hygroscopicity, CLTE, and tensile modulus are 0.975 - 0.017 %, 0.979 - 0.769 ppm/°C, and 0.949 - 0.091 GPa, respectively. As shown in Figure S7, we can see that $Y_H(\mathbf{X})$ models have better performance than $Y_E(\mathbf{X})$ models. The results demonstrated that through multi-fidelity learning, the limited experimental data could be well utilized with the assistance of all-atomic simulation data for establishing robust ML-based QSPR models.

In the experiments, we have prepared a resin of synthetically accessible structure of 9,9-bis(4-cyanatophenyl) fluorene (for details, see Sections S13-S16). Figure S8 shows the experimental results of this resin and the predicted results from the $Y_E(\mathbf{X})$ and $Y_H(\mathbf{X})$ models. Remarkably, the predicted results of the $Y_H(\mathbf{X})$ models are closer to the experimental results than those of the $Y_E(\mathbf{X})$ models. The results further confirm that the $Y_H(\mathbf{X})$ models have higher prediction accuracy and better generalization ability than $Y_E(\mathbf{X})$ models.

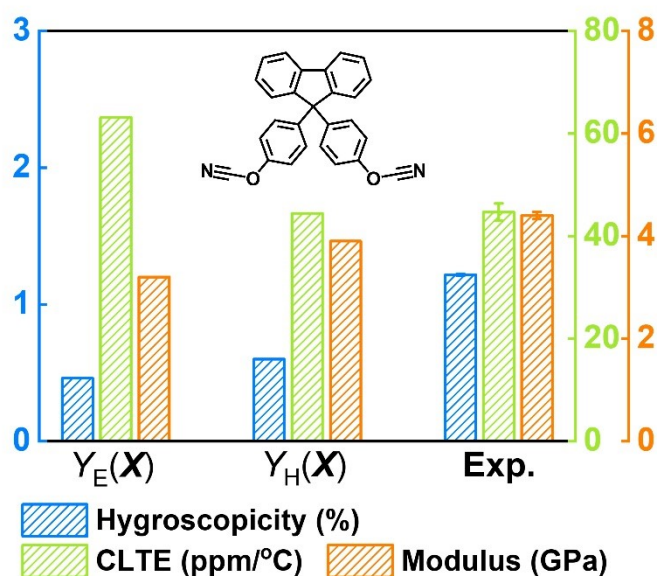


Figure S8. Comparison between the experimental results and the predicted results from the $Y_E(\mathbf{X})$ and $Y_H(\mathbf{X})$ models of one synthetically accessible resin of 9,9-bis(4-cyanatophenyl) fluorene. The blue, green, and orange columns are the hygroscopicity, CLTE, and tensile modulus of this resin, respectively.

S11. Implementation of gene definition, cleaning, and combination

In general, chemical units such as -F, -CH₃, -NH₂, -CO-, -O- can be combined arbitrarily to create a vast chemical space. However, random combinations inevitably create numerous unreasonable chemical structures. Herein, we first extracted a structural template based on di-functional aromatic CE monomers, as shown in Figure S9. Then, we labeled the para-, meta-, and ortho-carbon of the cyanate group on the template as #1, #2, and #3. The groups which bonded to #1, #2, and #3 carbon were defined as #A, #B, and #C 'genes', respectively.

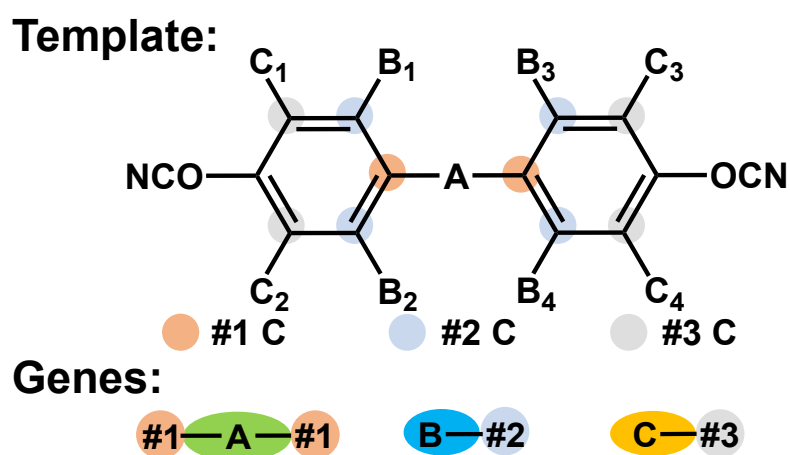


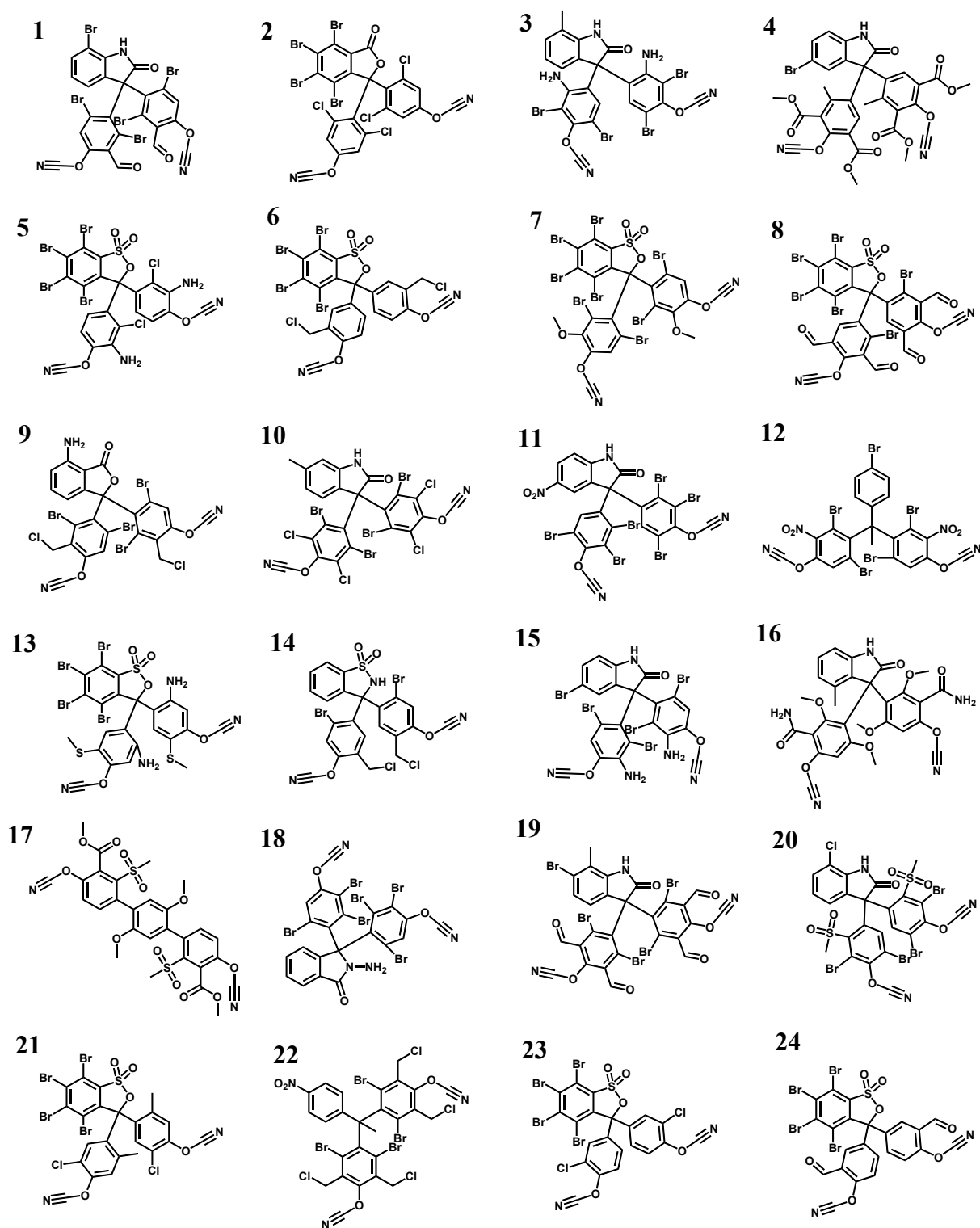
Figure S9. The template of CE monomer and the defined genes.

We need to obtain a diverse and reasonable 'gene' pool to create a reliable candidate space. Due to the limited number of existing CE monomers, we first collected a library of 5838 polyfunctional phenols from the existing chemical databases and literature, then converted them to corresponding virtual CE monomers based on the synthesis route of cyanation reaction.^{S24} Then, we screened the 5838 virtual CE monomers according to several rules: 1) the number of cyanate groups was limited to 2; 2) the atom types were limited to H, B, C, N, O, F, Si, P, S, Cl, Br; and 3) the structure of CE monomer satisfied the defined template. As a consequence, we obtained 2613 virtual CE monomers. According to the defined template and 'genes', we split these 2613 virtual CE monomers to obtain an initial 'gene' pool through the RDKit toolkit and obtained 1242 #A, 23 #B, and 110 #C 'genes'.^{S9} To keep the candidate space reliable, we

screened the 'gene' pool and removed these 'genes' containing the groups such as $\text{C}\equiv\text{C}$ triple bond, $\text{C}\equiv\text{N}$ triple bond (not the part of $-\text{OC}\equiv\text{N}$), $\text{C}=\text{C}$ double bond, $\text{C}=\text{N}$ double bond, epoxy group, a long chain containing more than five aliphatic atoms. Upon the above treatments, we obtained 256 #A, 17 #B, and 46 #C 'genes'. Finally, we created a library of 573591 candidate CE monomers.

S12. Detailed structures of promising and existing CE monomers

We screened the virtual candidates with the criteria of the top 1% and obtained 574 promising structures. Then, we verified the 50 structures randomly selected from 574 promising ones by theoretical simulation. Figure S10 shows the detailed structures of these 50 CE monomers in descending order of the overall score.



(continued from the previous page)

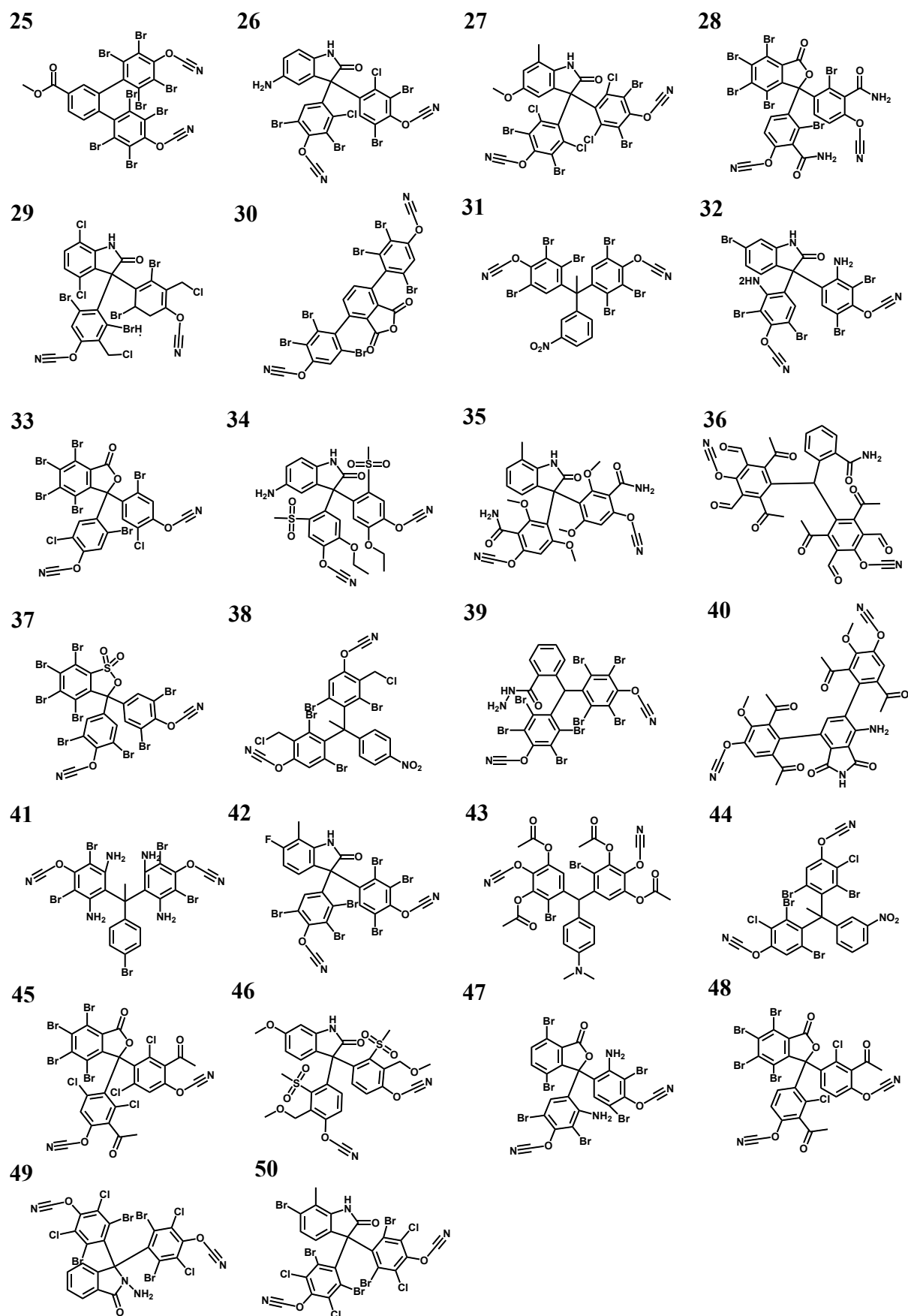


Figure S10. The structures of CE monomers are arranged in descending order of overall score.

We collected a library of existing CE monomers from the SciFinder database. For comparison, we then screened these structures that satisfy the template defined in Section S11. The detailed structures of these screened existing CE monomers (other than the ones listed in Table S1) and their chemical abstracts service (CAS) numbers are shown in Figure S11.

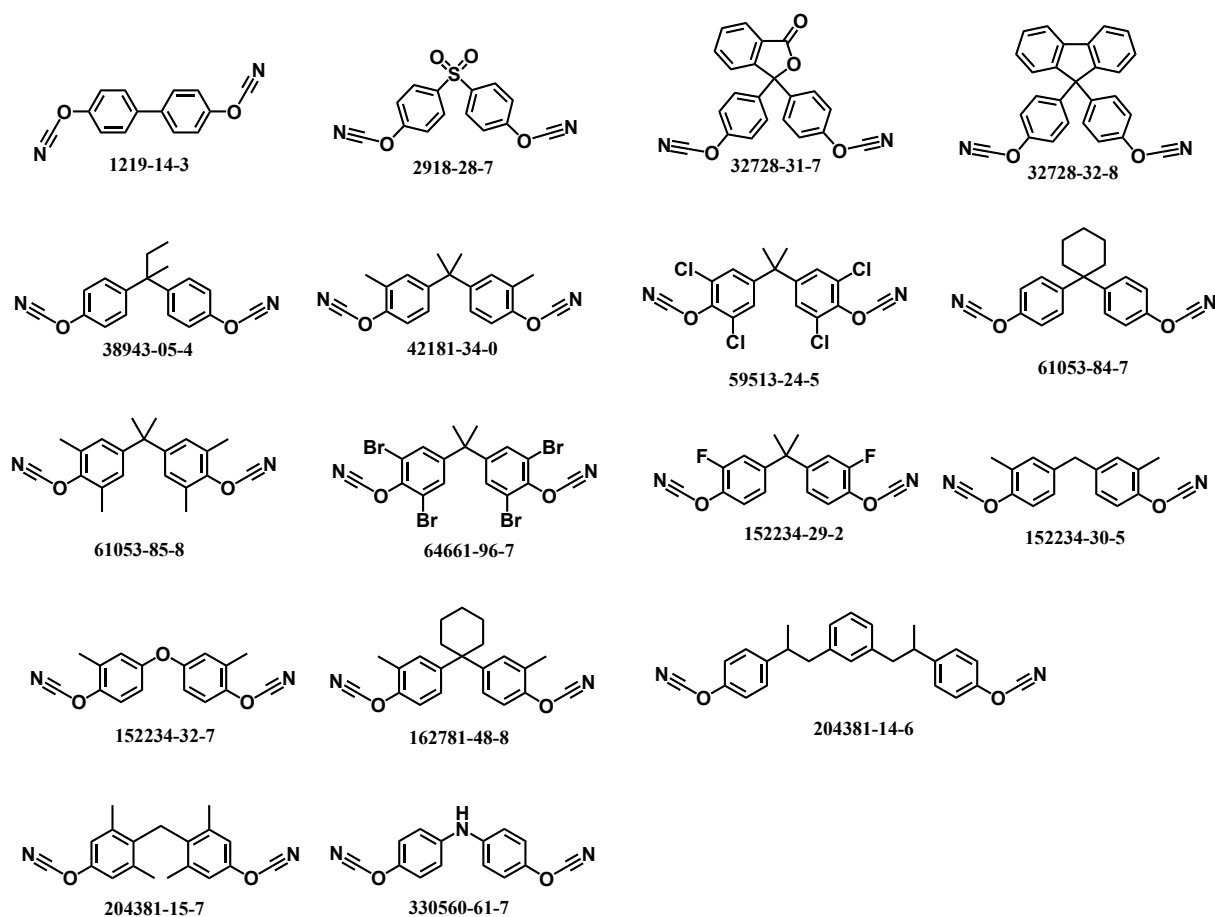


Figure S11. The detailed structures of several existing CE monomers and their chemical abstracts service (CAS) number.

S13.Preparation of the FCE monomer

The FCE monomer was synthesized by the reaction of bisphenols with cyanogen bromide in the presence of triethylamine. Figure S12 shows the synthesis route of 9,9-bis(4-cyanatophenyl) fluorene (FCE).

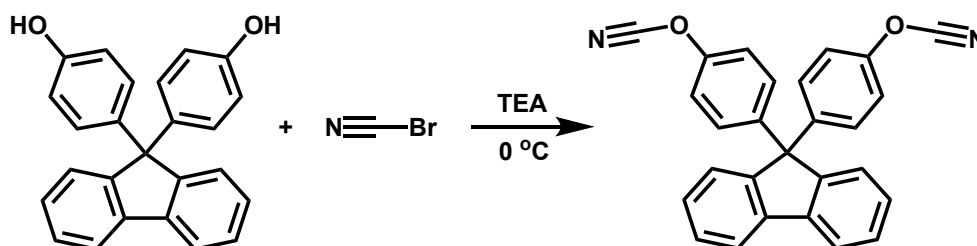


Figure S12. Synthesis route of 9,9-bis(4-cyanatophenyl) fluorene

9,9-bis(4-hydroxyphenyl)fluorene (175 g, 0.5 mol) was dissolved in acetone (500 mL) in a flask and cooled to -10 °C under a dry nitrogen atmosphere. 105 g (1 mol) of cyanogen bromide was added, with the solution stirring at 0 °C. Then 101 g (1 mol) of triethylamine was added dropwise into the mixture. After keeping the reaction for another 20 min, the mixture was poured into 2 L of deionized water (0 °C) with moderate stirring, and the 9,9-bis(4-cyanatophenyl) fluorene was precipitated. We washed the filter residue with deionized water until neutral and dried it in a vacuum oven at 30 °C. After purifying by recrystallization from petroleum ether and drying, 176.52 g of the final product (white powder) was obtained. The yield is 88.3 %.

S14. Structural characterization and the purity of the FCE

The structure of the 9,9-bis(4-cyanatophenyl) fluorene (FCE) was characterized by ^1H -NMR, FT-IR, and mass spectrum. The purity was measured by high-performance liquid chromatography (HPLC). Figure S13 shows the ^1H -NMR spectrum of the FCE. The protons and their corresponding chemistry shifts are marked in the figure. The peaks in the range of 7.80-7.79, 7.43-7.39, and 7.31-7.30 ppm represent the protons on the fluorene. The peaks in the range of 7.25-7.24 and 7.19-7.17 ppm correspond to the benzene rings connected to the oxygen. Figure S14 shows the FT-IR spectrum of the FCE. The peaks at 2272-2235 cm^{-1} refer to the $\text{C}\equiv\text{N}$ absorption. The $\text{C}-\text{C}$ and $\text{C}=\text{C}$ aromatic stretching vibrations appear at 1598, 1497, and 1448 cm^{-1} , respectively. The absorptions between 1167 cm^{-1} and 1188 cm^{-1} are derived from $\text{C}-\text{O}-\text{C}$. Figure S15 shows the mass spectrum of the FCE. The characteristic peak of FCE is at $m/z = 400$. Figure S16 shows the result of the HPLC of the FCE. The purity of FCE is 97.0 %.

The above results indicate that the target product was successfully synthesized.

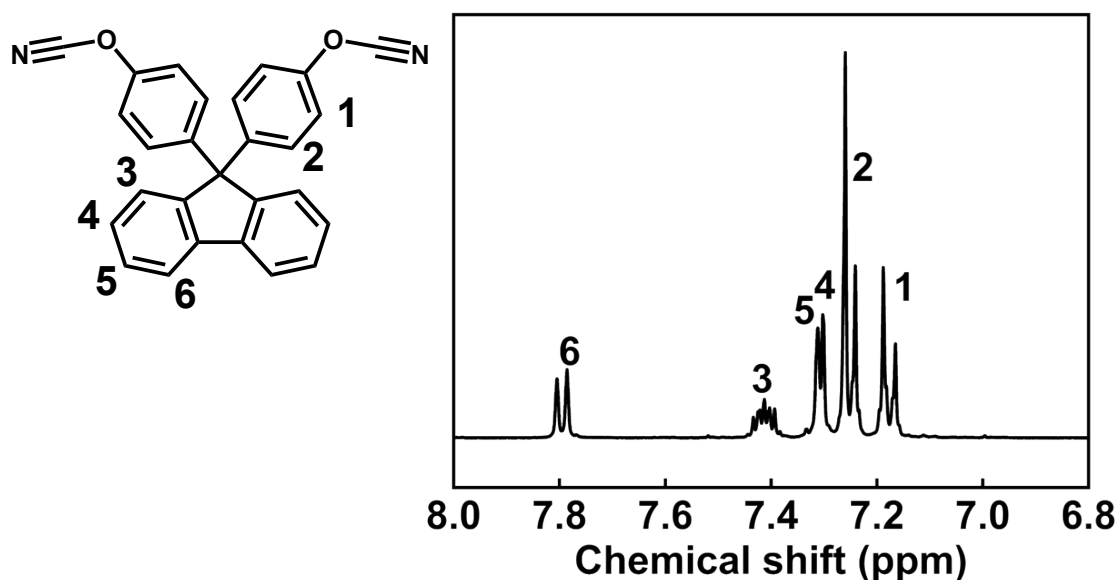


Figure S13. The ^1H -NMR spectrum of the FCE.

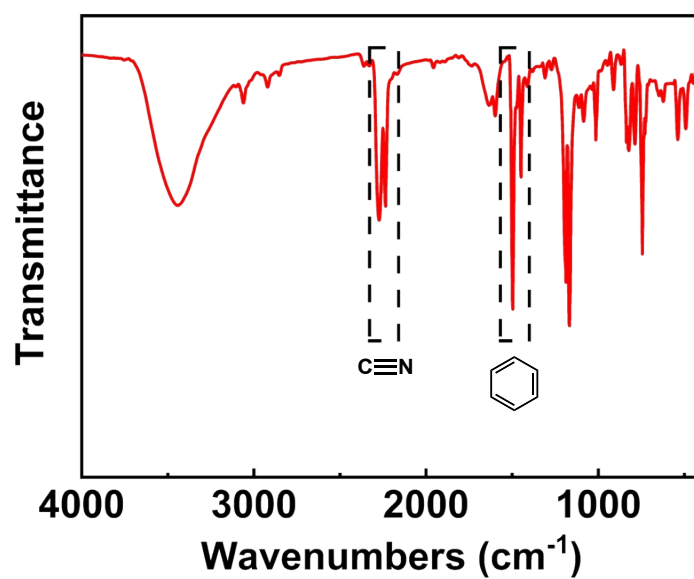


Figure S14. FT-IR spectrum of the FCE.

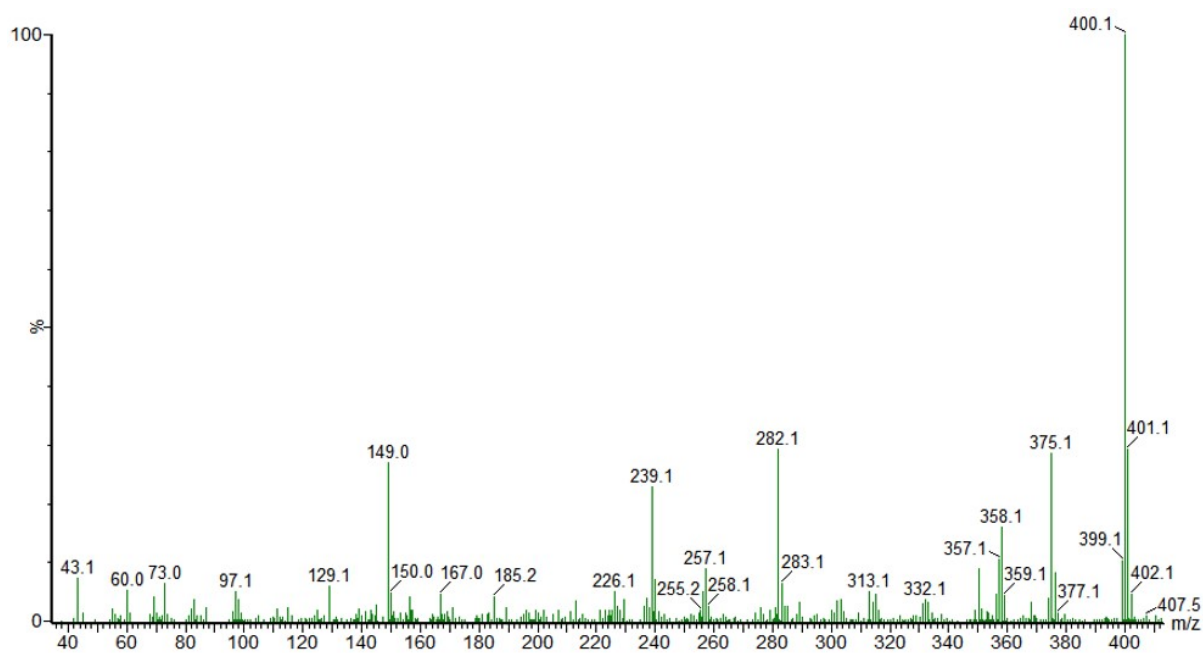


Figure S15. Mass spectrum of the FCE.

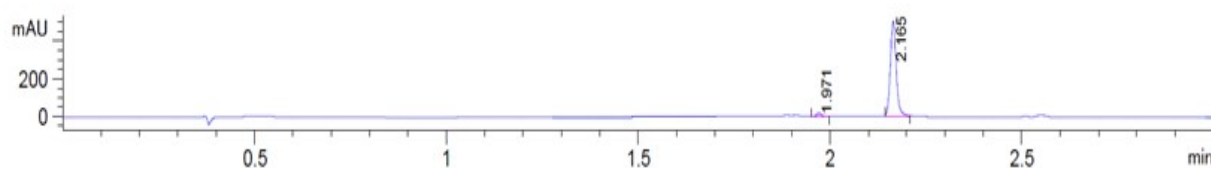


Figure S16. High-performance liquid chromatography of the FCE.

S15. The thermal property and processability of the FCE

The thermal stability and processing performance of the FCE were characterized by the differential scanning calorimeter (DSC), thermo-gravimetry analysis (TGA), and rheometer. Figure S17 shows the DSC curve of the FCE. DSC analysis was performed on a TA DSC 250 instrument from 25 °C to 400 °C at a heating rate of 10 °C/min under a nitrogen atmosphere. The FCE appears as white powders at room temperature. The endothermic peak at 167 °C indicates the melting point of the FCE. There is no other thermal effect until it begins curing at 315 °C. The exothermic peak temperature of FCE is 332 °C. The exothermic enthalpy during the curing process is 473 J/g.

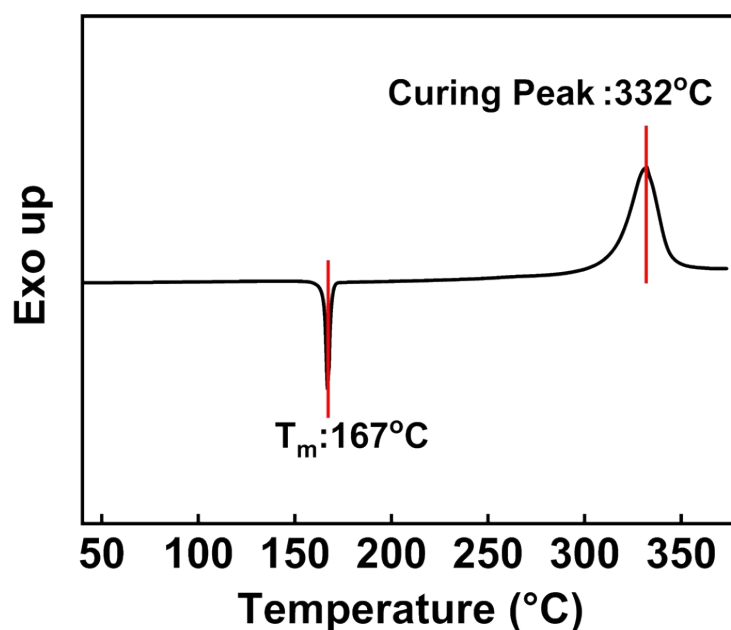


Figure S17. DSC curve of initial FCE.

Figure S18 shows the TGA curve of cured FCE. The thermal decomposition of cured FCE was analyzed using Netzsch TG 209Fi Libra instrument in flowing nitrogen (30 mL/min) at a heating rate of 10 °C/min. The initial decomposition temperature, 5 % decomposition temperature (T_{d5}), and char yield at 800 °C ($Y_{800^{\circ}\text{C}}$) are 418 °C, 434 °C, and 66.62 %, respectively.

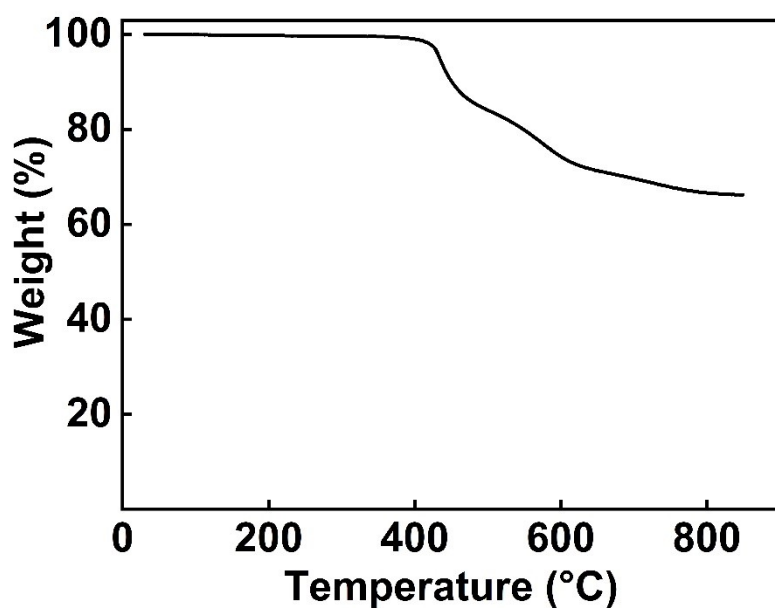


Figure S18. TGA curve of cured FCE.

The viscosity-temperature curve of the FCE is shown in Figure S19. The viscosity is below 2 Pa·s between the melting point and initial curing point. It rapidly increases when the temperature reaches the gel point at 270 °C. The processing window is as wide as 103 °C, which shows good processability.

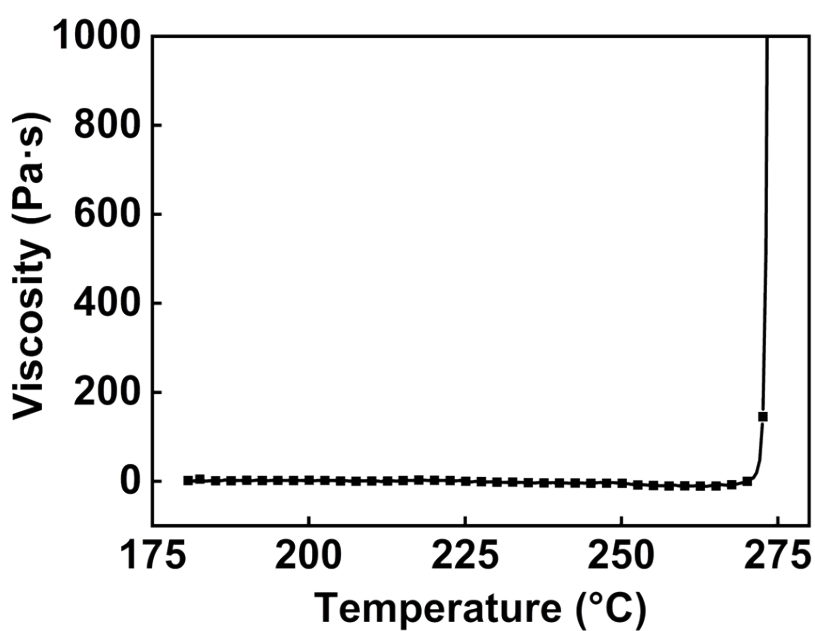


Figure S19. The viscosity-temperature curve of FCE.

S16. Specimen preparation for testing

The FCE monomer was added to a 250 mL eggplant flask with magnetic stirring and was heated at 180 °C until all the powder melted. The flowing liquid was quickly poured into a steel mold preheated at 180 °C and precoated with a release agent. After all the bubbles were thoroughly removed at 180 °C under vacuum, the resin was transferred to a blast drying oven for curing following a procedure of 230 °C/3 h, 260 °C/3 h, 290 °C/3 h, and 320 °C/3 h.

The cured sample was cut into specific dimensions for hygroscopicity, CLTE, and tensile tests. The tests were conducted according to ISO 62: 2008, ISO 11359-2: 1999, and ASTM D638-14, respectively. The experimental results of the three properties of the cured FCE resin are shown in Figure S20, in which detailed results for each sample are plotted (three samples for hygroscopicity tests, five samples for CLTE tests, and five samples for tensile tests).

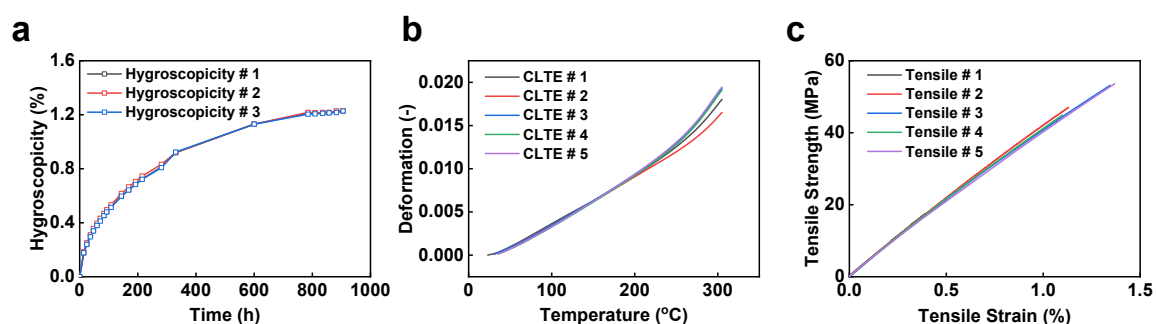


Figure S20. a) Hygroscopicity of the cured FCE resin over time under a certain condition of 25 °C and 60 % humidity. b) TMA plot of the cured FCE resin (25 - 300 °C, 5 °C/min). c) Stress-strain curve (tensile property) of the cured FCE resin (gauge length 25 mm, tensile rate 2 mm/min, 25 °C)

S17.A statistical method for quantifying the importance of molecular fragments

The underlying physics in the black-box function of the structure-property relationship can provide helpful insights and inspiration for the rational design of polymers with property objectives. Several methods, such as Shapley additive explanations (SHAP) analysis^{S25-S27}, have been widely used by researchers since these methods can be used to analyze the contribution of input variables to output variables in the black-box model. However, such a method is unsuitable for our work since the input of our model represents a whole molecule. In our case, an 8-dimensional latent vector (**X**) of a chemical structure can be obtained through a well-trained GNN model. Then, the relationship between **X** and the property (**Y**) is mapped using Gaussian process regression. Although the latent vector of **X** has eight dimensions, it represents a whole molecule; that is, the values of each dimension have no physical meaning. Thus, it is not very sensible to use methods like SHAP to analyze the contribution of **X** to **Y**. For this reason, we proposed a statistical method to obtain insights into structural design.

Numbers of factors determine the macroscopic properties of materials. A chemical structure may contain multiple fragments acting in opposite directions simultaneously. For example, methyl and methoxyl groups may have opposite effects on moisture uptake.^{S28,S29} A CE resin containing both ortho-methyl and ortho-methoxyl groups near the cyanate group may exhibit moderate moisture uptake. Therefore, statistics only on local regions of the polymer space (such as the promising space) may miss some essential features.

The proposed statistical method quantifies the contribution of a fragment to a property by estimating its effect on property distribution (i.e., the cumulative distribution function of a property, CDF). The CDF of a real-valued random variable X is the function given by

$$F_X(x) = P(X \leq x) \quad (\text{S-10})$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x .

For instance, as depicted in Figure S21, the black line is the property distribution of all virtual candidates, and the red, blue, and green line shows the property distribution of the virtual candidates that contain # 1, # 2, and #3 fragment, respectively. We can see that the red and blue lines move to a low property region compared with the black line, while the green line shifts to a high property region. Based on the results, we can deem that the presence of # 1 fragment and # 2 fragment could lead to a low property value, and the presence of # 3 fragment could cause a high property value. Furthermore, the contribution of each fragment can be estimated by the area between the property distribution affected by the fragment and the baseline (the property distribution of all virtual candidates), as shown in the light gray area of Figure S21.

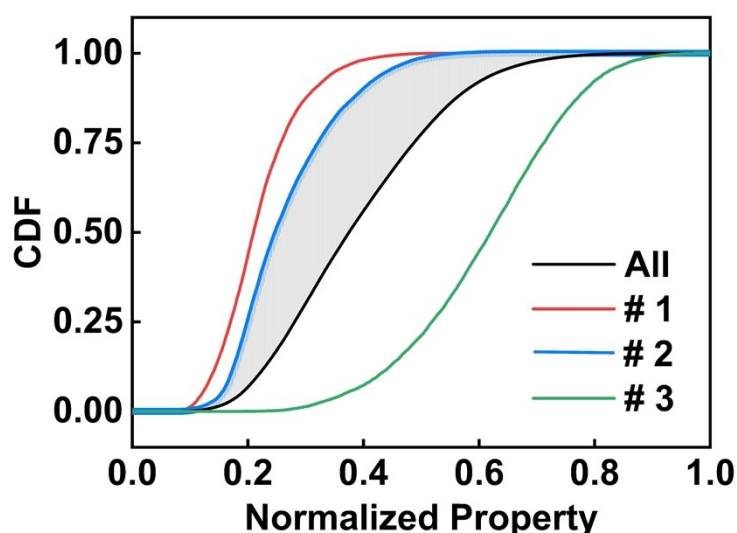


Figure S21. A graphic illustration of calculating the contribution of a fragment to a property. The property distribution of all candidates (black line) and candidates containing # 1 fragment (red line), # 2 fragment (blue line), and # 3 fragment (green line) are counted by the cumulative distribution function (CDF). The contribution of each fragment to a property is estimated by the area between the CDF of the fragment and the black line (*i.e.*, the gray area in the figure).

References

- S1 V. Fung, J. Zhang, E. Juarez, B. G. Sumpter, *npj Comput. Mater.*, 2021, **7**, 84.
- S2 S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wong, Y. Song, F. Liu, J. Hu, (Preprint) arXiv:2003.13379, v1, submitted: Mar, 2020.
- S3 Z. Li, G. P. Wellawatte, M. Chakraborty, H. A. Gandhi, C. Xu, A. D. White, *Chem. Sci.*, 2020, **11**, 9524.
- S4 S. Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu, J. Hu, *Phys. Chem. Chem. Phys.*, 2020, **22**, 18141.
- S5 K. Hatakeyama-Sato, T. Tezuka, M. Umeki, K. Oyaizu, *J. Am. Chem. Soc.*, 2020, **142**, 3301.
- S6 J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, (Preprint) arXiv:1312.6203, v3, submitted: May, 2014.
- S7 P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, (Preprint) arXiv:1710.10903, v3, submitted: Feb, 2018.
- S8 J. Zhang, X. Shi, J. Xie, H. Ma, I. King, D.-Y. Yeung, (Preprint) arXiv:1803.07294, v1, submitted: Mar, 2018.
- S9 RDKit: Open Source Toolkit for Cheminformatics, <http://www.rdkit.org/>, accessed: Nov, 2022.
- S10 H. Moriwaki, Y. S. Tian, N. Kawashita, T. Takagi, *J. Cheminform.*, 2018, **10**, 4.
- S11 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, (Preprint) arXiv:1603.04467, v1, submitted: Mar, 2016.
- S12 J. Zhu, M. Chu, Z. Chen, L. Wang, J. Lin, L. Du, *Chem. Mater.*, 2020, **32**, 4527.
- S13 M. S. Radue, V. Varshney, J. W. Baur, A. K. Roy, G. M. Odegard, *Macromolecules*, 2018, **51**, 1830.
- S14 V. Varshney, S. S. Patnaik, A. K. Roy, B. L. Farmer, *Macromolecules*, 2008, **41**, 6837.
- S15 BIOVIA Materials Studio, <https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/biovia-materials-studio/>, accessed: Nov, 2022.
- S16 M. P. Tosi, *Solid State Physics*, 1964, **16**, 107.
- S17 L. M. J. Moore, N. D. Redeker, A. R. Browning, J. M. Sanders, K. B. Ghiassi, *Macromolecules*, 2021, **54**, 6275.

- S18 T. Fang, D. A. Shimp, *Prog. Polym. Sci.*, 1995, **20**, 61.
- S19 A. J. Guenthner, G. R. Yandek, M. E. Wright, B. J. Petteys, R. Quintana, D. Connor, R. D. Gilardi, D. Marchant, *Macromolecules*, 2006, **39**, 6046.
- S20 A. J. Guenthner, K. R. Lamison, V. Vij, J. T. Reams, G. R. Yandek, J. M. Mabry, *Macromolecules*, 2012, **45**, 211.
- S21 A. I. J. Forrester, A. Söbester, A. J. Keane, *Proc. R. Soc. A*, 2007, **463**, 3251.
- S22 C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, London 2006.
- S23 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *J. Mach. Learn. Res.*, 2011, **12**, 2825.
- S24 E. Grigat, R. Pütter, *Angew. Chem. Int. Ed.*, 1967, **6**, 206.
- S25 Z. -Y. Yang, J. Dong, Z. -J. Yang, A. -P. Lu, T. -J. Hou, D. -S. Cao, *J. Chem. Inf. Model.*, 2020, **60**, 2031.
- S26 W. -L. Ye, C. Shen, G. -L. Xiong, J. -J. Ding, A. -P. Lu, T. -J. Hou, D. -S. Cao, *J. Chem. Inf. Model.*, 2020, **60**, 4126.
- S27 Y. R. Xie, D. C. Castro, S. E. Bell, S. S. Rubakhin, J. V. Sweedler, *Anal. Chem.*, 2020, **92**, 9338.
- S28 A. J. Guenthner, M. E. Wright, A. P. Chafin, J. T. Reams, K. R. Lamison, M. D. Ford, S. P. J. Kirby, J. J. Zavala, J. M. Mabry, *Macromolecules*, 2014, **47**, 7691.
- S29 B. G. Harvey, A. J. Guenthner, W. W. Lai, H. A. Meylemans, M. C. Davis, L. R. Cambrea, J. T. Reams, K. R. Lamison, *Macromolecules*, 2015, **48**, 3173.