Supporting Information

Interpretable Prediction of Aggregation-Induced Emission Molecules Based on Graph Neural Network

Shi-Chen Zhang,[a] Jun Zhu,[b] Yi Zeng,[a] Hua-Qi Mai,[a] Dong Wang*[b] and Xiao-Yan Zheng*[a]

a. Key Laboratory of Cluster Science of Ministry of Education, Key Laboratory of Medicinal Molecule Science and Pharmaceutics Engineering of Ministry of Industry and Information Technology, Beijing Key Laboratory of Photoelectronic/ Electro-photonic Conversion Materials, School of Chemistry and Chemical Engineering, Beijing Institute of Technology, Beijing 100081, P. R. China.
E-mail: xiaoyanzheng@bit.edu.cn
b. Center for AIE Research, Shenzhen Key Laboratory of Polymer Science and Technology, Guangdong Research Center for Interfacial Engineering of Functional Materials, College of Materials Science and Engineering, Shenzhen University, Shenzhen 518060, P. R. China
E-mail: wangd@szu.edu.cn

Table of Contents

1. Supplementary Notes

1.1 Data collection and cleaning

The molecules in the dataset were collected through three channels: (1) 816 AIEgens and 14 ACQ molecules from the ASBase database[1]; (2) 134 AIEgens and 222 ACQ molecules from the open-source dataset by Liu *et al.*[2]; (3) 58 AIEgens and 41 ACQ molecules obtained by searching novel literatures. The process of collecting molecules from the literature was entirely manual. All collected molecules were converted to standard simplified molecular-input line-entry system (SMILES) strings, and duplicates were removed. This resulted in the construction of a dataset containing 934 AIEgens and 255 ACQ molecules. The dataset was subjected to t-SNE analysis using the Scikit-learn package (Fig.2b).

1.2 t-SNE

In this study, we employed the t-SNE algorithm[3] to reduce the high-dimensional chemical space into a two-dimensional representation for visualization purposes. T-SNE is a nonlinear dimensional reduction technique that is particularly effective in embedding high-dimensional data into a space of lower dimensions (typically two or three dimensions) while retaining the local structure of the data points. The algorithm is based on the probability distribution of the similarities between data points, with the goal of transforming the high-dimensional similarities into low-dimensional probabilities that can be visualized in a scatter plot.

The t-SNE implementation utilized in this work is derived from the sklearn.manifold module in the scikit-learn library, a widely recognized machine learning library in Python. Our implementation begins with the computation of Morgan fingerprints for each molecule in the dataset, which serve as the input features for the t-SNE algorithm. These fingerprints are then transformed into a two-dimensional space using the t-SNE model with the Jaccard similarity metric, which is suitable for binary fingerprint data.

The parameters for the t-SNE model were carefully chosen to optimize the visualization: n_components=2 for the two-dimensional output, init='pca' for the initialization method, and random_state=0 to ensure reproducibility of the results.

1.3 Graph Neural Network

The GNN utilized in this work was constructed based on the Chemprop package[4]. For each molecule, RDKit package was employed to generate graph-based molecular representations from the SMILES strings of the compounds. Feature vectors for each atom and bond in the molecule were generated based on the following computable features: atomic features include atom type (type of atom (ex. C, N, O), by atomic number), bonds (number of bonds the atom is involved in), formal charge (integer electronic charge assigned to atom), chirality (unspecified, tetrahedral CW/CCW, or other), Hs (number of bonded hydrogen atoms), hybridization (sp, sp2, sp3, sp3d, or sp3d2), aromaticity (whether this atom is part of an aromatic system), and atomic mass (mass of the atom, divided by 100); bond features include bond type (single, double, triple, or aromatic), conjugated (whether the bond is conjugated), in ring (whether the bond is part of a ring), and stereo (none, any, E/Z or cis/trans)[5]. Based on such initial graph data, the GNN performs directed message-passing steps, updated by summing messages from adjacent bonds, concatenating the current bond's message with the sum, and then applying a single neural network layer with a nonlinear activation function (Fig.2c). After a fixed number of message-passing steps, messages across the entire molecule are summed to generate the final message representing the molecule. This message is then passed through a feedforward neural network, which outputs a prediction of the AIE/ACQ properties of the compound.

1.4 Extreme gradient boosting

Extreme gradient boosting (XGBoost) is a powerful and widely-used ensemble learning algorithm that excels in classification and regression tasks. It operates on the principle of boosting, where weak learners, typically decision trees, are combined to form a strong predictive model. The core idea is to sequentially add new trees that correct the errors made by the previously trained trees, thereby improving accuracy.

In this study, the XGBoost model was implemented using the xgboost Python package, which offers a comprehensive set of tools for gradient boosting. The XGBoost algorithm can be summarized in several key steps: 1) Initialize the model with a constant value; 2) For each iteration, compute the pseudo-residuals, which are the differences between the true values and the predicted values; 3) Fit a new decision tree to these pseudo-residuals; 4) Update the model by adding the predictions from this new tree, scaled by a learning rate; 5) Repeat the process until a specified number of trees is reached or no further improvement is observed.

Mathematically, the update for the predicted value can be expressed as: $\hat{y}_i = \hat{y}_i + \eta f(x_i)$, where $\hat{y}_i$ is the predicted value for the i^{th} instance, η is the learning rate, and $f(x_i)$ is the output of the newly added tree for the i^{th} instance.

1.5 Support vector machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It operates on the principle of finding the optimal hyperplane that separates data points from different classes in a high-dimensional space. The goal of SVM is to maximize the margin between the closest points of each class, known as support vectors.

In this study, the SVM model was implemented using the sklearn.svm library, which provides a robust and efficient interface for SVM classification. The SVM process can be described in several key steps: 1) Transform the input data into a higher-dimensional space using a kernel function if necessary; 2) Identify the support vectors that are closest to the decision boundary; 3) Calculate the optimal hyperplane that maximizes the margin between the classes; 4) Classify new data points based on which side of the hyperplane they fall.

Mathematically, the decision function can be formulated as: $f(x) = sign(\omega^T x + b)$, where ω is the weight vector, x is the input feature vector, b is the bias term, and signsign determines the class label based on the position relative to the hyperplane.

1.6 Random forest

Random forest is a powerful ensemble learning algorithm widely used for classification and regression tasks. This algorithm builds multiple decision trees and aggregates their results to enhance prediction accuracy and stability. The key aspect of Random Forest lies in the word "random": during the construction of each tree, a random subset of features and samples is selected, ensuring that each tree maintains a degree of independence.

In this study, the RF model was implemented using the sklearn.ensemble library, which provides a comprehensive and efficient interface for ensemble methods. The process can be described as follows: 1) From the training dataset, randomly sample N data points with replacement to create a new training set; 2) Build a decision tree using this training set, typically employing the CART (Classification and Regression Trees) algorithm; 3) Repeat steps 1 and 2 to construct multiple decision trees, forming a forest; 4) For classification tasks, use majority voting to aggregate the predictions of all trees, returning the class with the highest frequency as the final prediction; for regression tasks, take the average of all tree predictions as the final output.

The prediction for a given input can be expressed with the following formula:

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

where *T* is the total number of trees, and $f_t(x)$ is the prediction of the *t*-th tree for the sample *x*.

1.7 K-nearest neighbor

KNN is a relatively mature pattern recognition algorithm and one of the simplest classification algorithms. Considering the k closest samples of a data point, if most of the samples belong to a certain category, the data point also belongs to this category. Two key factors that affect KNN were the number of neighbors k and the calculation of distance. k was usually an integer not greater than 20 and the distance was generally using Euclidean distance. The Euclidean distance is defined as $d = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$, where n is the number of samples.

The KNN model was implemented using the sklearn.neighbors library, which provides a comprehensive and efficient interface for KNN-based methods. The neighbors selected in the KNN algorithm were all objects that have been correctly classified. This method determined the category to which the sample to be classified belongs only based on the category of the nearest sample or samples. Therefore, the KNN algorithm process could be described as: 1) calculate the distance between test data and each training data; 2) sort by increasing distance; 3) select the K points with the smallest distance; 4) determine the occurrence frequency of the category of the first K points; 5) return the category with the highest frequency in the first K points as the predicted classification of test data.

1.8 Multilayer perceptron

Multilayer Perceptron (MLP) is a fundamental neural network architecture widely used for various classification and regression tasks. It consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. Each neuron applies a weighted sum of its inputs followed by a non-linear activation function, allowing the network to learn complex patterns in the data.

The MLP model was implemented using the sklearn.neural_network library, which provides a robust and efficient interface for neural network-based methods. The MLP process can be outlined in several key steps: 1) Initialize the weights and biases of the network; 2) Perform a forward pass, where input data is fed through the network, and each layer computes its output; 3) Calculate the loss by comparing the predicted output to the actual target values; 4) Implement backpropagation to update the weights and biases based on the gradient of the loss function; 5) Repeat the process for multiple epochs until convergence is achieved.

Mathematically, the output of a neuron can be expressed as: $y = f(\sum_{i=1}^{n} \omega_i x_i + b)$, where y is the output, ω_i are the weights, x_i are the inputs, b is the bias, and f is the activation function (such as sigmoid, ReLU, or tanh).

1.9 Model optimization and evaluation

Three model optimization strategies were employed to enhance model performance. First, 200 kinds of additional molecular-level features calculated by RDKit package (Table S1) were added to the graph-based representations of each molecule. This step was performed to provide additional information on the global properties of each molecule that local message-passing methods might not encapsulate. Second, we utilized hyperparameter optimization to select the best-performing hyperparameters for the model. A limited grid search combined with ten-fold cross-validation was used to find and evaluate hyperparameters, resulting in good performance, with the parameter search ranges represented in Table S2. The XGBoost, SVM, RF, KNN, and MLP constructed in this work also underwent hyperparameter optimization, with the parameter search ranges represented in Tables S4-S8. Third, an ensemble model strategy was employed, combining five GNN models in an attempt to enhance performance.

According to the ratio of AIE/ACQ molecules, 10% of the molecules (94 AIE molecules, 26 ACQ molecules, a total of 120 molecules) were retained as the test set. Evaluate model performance using metrics such as accuracy, AUROC, AUPRC, F1 score, and MCC coefficient.

1.10 Monte Carlo tree search for key substructures

Inspired by empirical rules derived from a multitude of experiments, we posit that certain key substructures present within molecules largely confer AIE properties to the entire molecule. The Monte Carlo search method is employed to identify molecular substructures with high predictive scores, facilitated by the built-in "interpret" function of Chemprop package. Specifically, the root of the search tree is a complete AIEgens, and each state within the search tree represents a subgraph derived from sequences of bond or ring omissions. To ensure that each state is chemically valid and maintains connectivity, we only permit the removal of one peripheral bond or ring from each state. If the molecule remains connected after deletion, it is referred to as a bond or ring. We set the minimum predictive value for key substructures at 0.5, with the number of non-hydrogen atoms ranging from 15 to 50.

1.11 Molecular fragment generation and docking

The BRICS functionality of the RDKit package was utilized to randomly fragment AIEgens from the database to obtain molecular fragments, and their SMILES structures were saved. The docking of molecular fragments was also accomplished using the RDKit package. Specifically, the SMILES of the molecular fragments were first converted to mol format, then a chemical site was randomly selected on each of the two molecular fragments, and finally a single bond was added to connect these two molecular fragments, resulting in the virtually generated molecule. Invalid structures were discarded.

1.12 Experiment

Four compounds: 2MeO-TPE-OH, 2MeO-TPE-NH_2, TPE-$COCH_3$, and TPE-PhCN, were synthesized to process experiment verification. Through the application of synthetic methods such as suzuki coupling and acid-catalyzed condensation, obtained products yield between 27% and 72%. The synthesized molecules were performed using nuclear magnetic resonance (NMR) and high-resolution mass spectrometry (HRMS) (Fig.S8-S19). Moreover, the optical properties of these compounds were measured by ultraviolet-visible (UV-vis) and photoluminescence (PL) spectroscopy (Fig.5c, S20). Additionally, AIE characteristics of these molecules were evaluated by examining their PL intensity in THF/water mixtures in different water volume fractions (f_w). The PL intensity of four compounds were weak due to active intramolecular motion that result rapidly energy dissipation from the excited state to ground state. The PL intensity have significantly increase when f_w increased from 0% to 100%, indicating that four compounds have strong AIE performance (Fig.5b).

1.12.1 Main Materials

The initial reagents included bis(4-methoxyphenyl)methanone, (4-hydroxyphenyl)(phenyl)methanone, (4-aminophenyl)(phenyl)methanone, (2-bromoethene-1,1,2-triyl)tribenzene, (4-acetylphenyl)boronic acid, (4-(1,2,2-triphenylvinyl)phenyl)boronic acid, 4-bromobenzonitrile, $TiCl_4$, pyridine, $Pd(PPh_3)_4$, Aliquat 336, zinc powder, K_2CO_3, Na_2SO_4 and various solvents, all of which were sourced from J&K Chemicals, Macklin Chemicals, or Aladdin Industrial Corporation. These materials were used as received without further purification. Anhydrous solvents were dehydrated using standard methods before application.

1.12.2 Instruments

Proton (^1H) and carbon (^{13}C) nuclear magnetic resonance (NMR) spectra were recorded using instruments operating at frequencies of 400, 500, and 600 MHz. The chemical shifts were expressed in parts per million (ppm) and referenced to either tetramethylsilane or the residual solvent peak as an internal standard. High-resolution mass spectral analysis (HRMS) was performed using a Finnigan MAT TSQ 7000 mass spectrometer. Optical absorption properties were measured with a PerkinElmer Lambda 950 spectrophotometer. Additionally, photoluminescence (PL) spectra were obtained using Edinburgh FS5 and FLS1000 spectrofluorometers to assess the emission characteristics.

1.12.3 Synthesis and Characterization

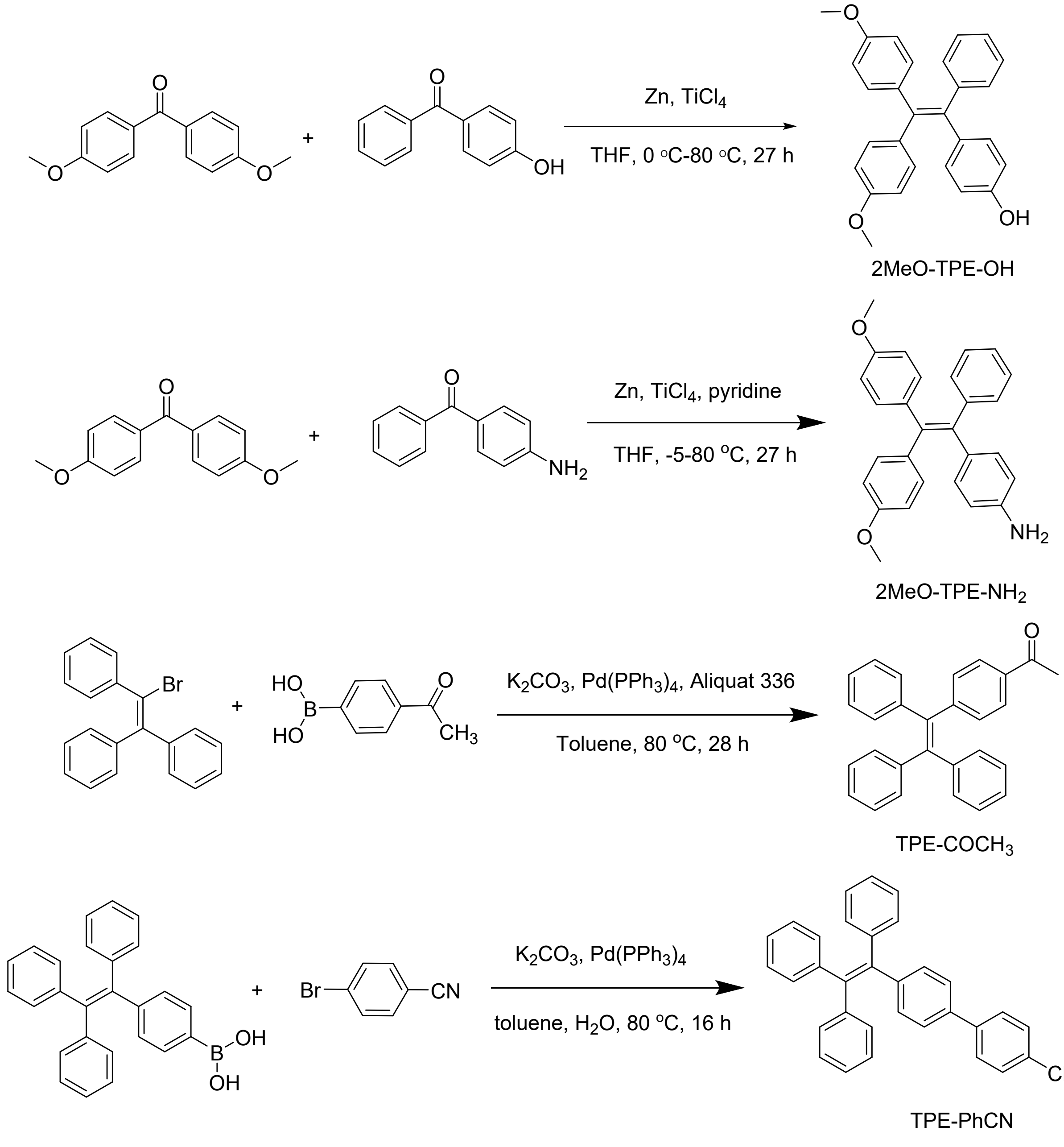


Fig.S1. The structures and synthetic routes of 2MeO-TPE-OH, 2MeO-TPE-NH_2, TPE-$COCH_3$ and TPE-PhCN.

1.12.4 General procedure for the synthesis of 2MeO-TPE-OH, 2MeO-TPE-NH$_2$, TPE-COCH$_3$ and TPE-PhCN

Synthesis of 2MeO-TPE-OH:

Under an inert atmosphere, bis(4-methoxyphenyl)methanone (484 mg, 2 mmol), (4-hydroxyphenyl)(phenyl)methanone (396 mg, 2 mmol) and zinc powder (520 mg, 8 mmol) were mixed in dry THF (20 mL) at 0 °C. Then TiCl$_4$ (440 uL, 4 mmol) were added to the reaction mixture slowly and stirred for 1 hours at 0 °C, then the mixture was stirred for 24 h at 80 °C. After the reaction was complete, the solution was adjusted to neutral by the addition to hydrochloric acid (16 mL, 1mol/L), subsequently, the resulting mixture was extracted with DCM (50 × 3 mL), and the combined organic phases were dried over Na$_2$SO$_4$. After the removal of the solvent under reduced pressure, the crude product was then purified by column chromatography on silica gel (PE/EA = 40:1) to give 2MeO-TPE-OH as white solid (217 mg, yield: 27%): ^{1}H NMR (CDCl$_3$, 500 MHz) δ (ppm) 7.11-6.87 (m, 11H), 6.66-6.56 (m, 6H), 3.75 (s, 3H), 3.73 (s, 3H). ^{13}C NMR (125 MHz, CDCl$_3$) δ 157.98, 153.91, 144.52, 139.38, 138.90, 136.95, 136.79, 136.68, 132.82, 132.68, 131.51, 127.76, 126.16, 114.74, 113.19, 113.11, 55.24, 55.22. HRMS (ESI) calculated for: $C_{28}H_{24}O_3$ [M]$^+$: 408.1725, found: 408.1731.

Synthesis of 2MeO-TPE-NH$_2$:

Under an inert atmosphere, zinc powder (1.6 g, 24 mmol) was stirred in dry THF (40 mL) at -5 °C, TiCl$_4$ (1.3 mL, 12 mmol) were added to the reaction mixture slowly and stirred for 0.5 hours at 0 °C. Then the mixture was stirred for 2.5 h at 80 °C. Pridine (0.5 mL, 6 mmol), bis(4-methoxyphenyl)methanone (1.74 g, 7.2 mmol) and (4-aminophenyl)(phenyl)methanone (1.18 g, 6 mmol) were solved in dry THF (15 mL), this solution were added to the solution of TiCl$_4$ and zinc powder, subsequently, the reaction solution were stirred for 27 h at 80 °C. After the reaction was complete, the react was quenched by 10% K$_2$CO$_3$ solution and extracted with DCM (50 × 3 mL), the combined organic phases were dried over Na$_2$SO$_4$. After the removal of the solvent under reduced pressure, the crude product was then purified by column chromatography on silica gel (PE/DCM = 2:1~1:10) to give 2MeO-TPE-NH$_2$ as white solid (870 mg, yield: 36%): ^{1}H NMR (CDCl$_3$, 400 MHz) δ (ppm) 7.11-7.02 (m, 5H), 6.97-6.90 (m, 4H), 6.84-6.83 (m, 2H), 6.66 (d, J = 6.0 Hz, 2H), 6.61 (d, J = 6.0 Hz, 2H), 6.55-6.53 (m, 2H), 3.75 (s, 3H), 3.73 (s, 3H). ^{13}C NMR (100 MHz, CDCl$_3$) δ 157.83, 144.71, 144.26, 139.26, 138.55, 136.99, 136.85, 134.82, 132.62, 132.57, 132.48, 131.51, 127.59, 125.95, 114.63, 114.60, 113.05, 112.96, 55.11, 55.09. HRMS (ESI) calculated for: $C_{28}H_{25}NO_2$ [M]$^+$: 408.1885, found: 408.1989.

Synthesis of TPE-COCH$_3$

(2-bromoethene-1,1,2-triyl)tribenzene (1 g, 2.89 mmol), (4-acetylphenyl)boronic acid (0.54 g, 3.28 mmol), K_2CO_3 (1.24 g, 9 mmol), and $Pd(PPh_3)_4$ (0.7 g, 0.6 mmol) were mixed in dry toluene (30 mL), and Aliquat 336 (5 drops) were added to the reaction solution. Then the reaction was heat up to 80°C for 28 h. After the reaction was completed, which extracted with DCM (50 × 3 mL), the combined organic phases were dried over Na_2SO_4. After the removal of the solvent under reduced pressure, the crude product was then purified by column chromatography on silica gel (PE/DCM = 2:1~1:1) to give TPE-COCH$_3$ as white solid (270 mg, yield: 72%): ^{1}H NMR (CDCl$_3$, 500 MHz) δ (ppm) 7.69 (d, *J* = 8.5 Hz, 2H), 7.13-7.10 (m, 11H), 7.04-7.00 (m, 6H), 2.53 (s, 3H). ^{13}C NMR (125 MHz, CDCl$_3$) δ 197.80, 149.12, 143.27, 143.21, 143.14, 142.67, 139.94, 135.01, 131.58, 131.36, 131.33, 127.97, 127.96, 127.89, 127.81, 127.01, 126.87, 126.85, 26.63. HRMS (ESI) calculated for: $C_{28}H_{23}O$ $[M]^+$: 375.1749, found: 375.1742.

Synthesis of TPE-PhCN

(4-(1,2,2-triphenylvinyl)phenyl)boronic acid (376 mg, 1 mmol), 4-bromobenzonitrile (182 mg, 1mmol), K_2CO_3 (0.65 g, 5 mmol), and $Pd(PPh_3)_4$ (100 mg, 0.09 mmol) were mixed in solution of dry toluene (50 mL), ethanol (50 mL) and water (5 mL). Then the reaction was heat up to 80°C for 16 h. After the reaction was completed, which extracted with DCM (50 × 3 mL), the combined organic phases were dried over Na_2SO_4. After the removal of the solvent under reduced pressure, the crude product was then purified by column chromatography on silica gel (PE/DCM = 3:1~1:2) to give TPE-PhCN as white solid (295 mg, yield: 68%): ^{1}H NMR (CDCl$_3$, 600 MHz) δ (ppm) 7.61-7.55 (m, 4H), 7.27 (d, *J* = 8.4 Hz, 2H), 7.06-7.03 (m, 11H), 7.00-6.95 (m, 6H). ^{13}C NMR (150 MHz, CDCl$_3$) δ 145.17, 144.46, 143.56, 143.54, 143.49, 141.83, 140.10, 136.77, 132.56, 132.15, 131.41, 131.37, 131.34, 127.89, 127.85, 127.75, 127.45, 126.74, 126.70, 126.66, 126.43, 119.05, 110.68. HRMS (ESI) calculated for: $C_{33}H_{23}N$ $[M]^+$: 433.1830, found: 433.1826.

2. Supplementary Figures

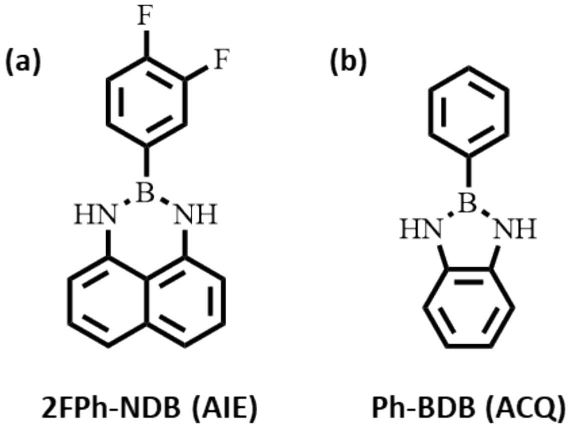


Fig.S2. Chemical Structure of 2FPh-NDB and Ph-BDB[6].

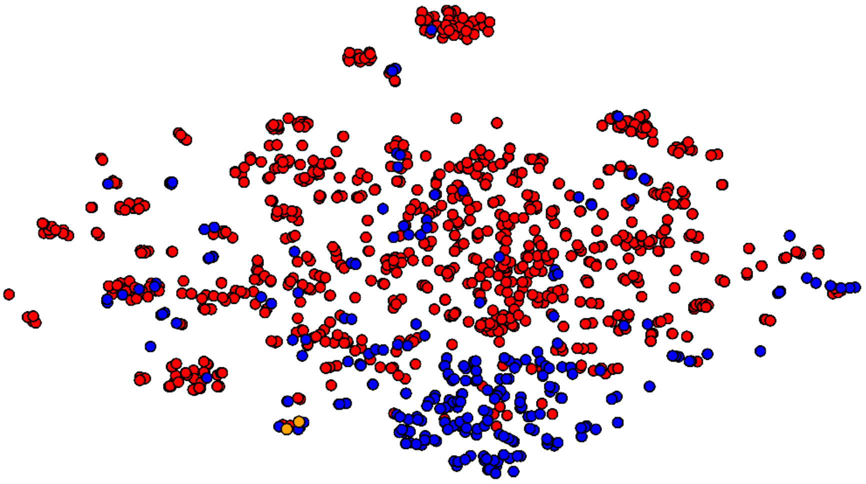
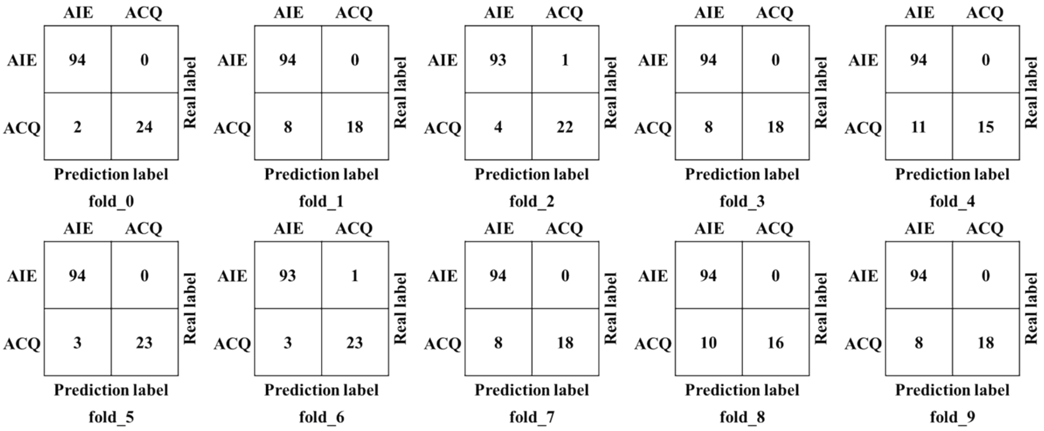


Fig.S3. T-SNE of the database, using Morgan fingerprint with jaccard similarity. In this figure, the red dots represent AIE molecules, the blue dots represent ACQ molecules, and the orange dots represent 2FPh-NDB and Ph-BDB.


Prediction label
Real label

fold_0

	AIE	ACQ
AIE	94	0
ACQ	2	24

fold_1

	AIE	ACQ
AIE	94	0
ACQ	8	18

fold_2

	AIE	ACQ
AIE	93	1
ACQ	4	22

fold_3

	AIE	ACQ
AIE	94	0
ACQ	8	18

fold_4

	AIE	ACQ
AIE	94	0
ACQ	11	15

fold_5

	AIE	ACQ
AIE	94	0
ACQ	3	23

fold_6

	AIE	ACQ
AIE	93	1
ACQ	3	23

fold_7

	AIE	ACQ
AIE	94	0
ACQ	8	18

fold_8

	AIE	ACQ
AIE	94	0
ACQ	10	16

fold_9

	AIE	ACQ
AIE	94	0
ACQ	8	18

Fig.S4 Confusion matrices for the 10-fold cross-validation of the GNN.

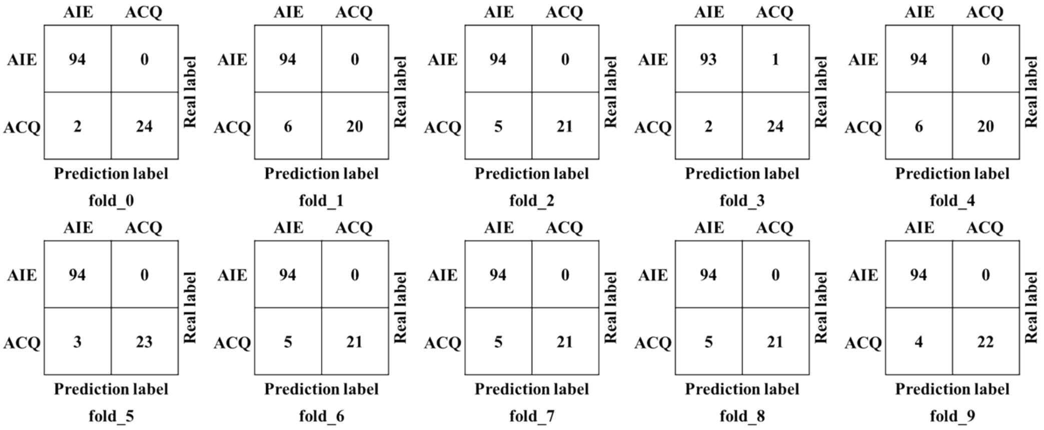


Fig.S5 Confusion matrices for the 10-fold cross-validation of the GNN (add RDkit feature).

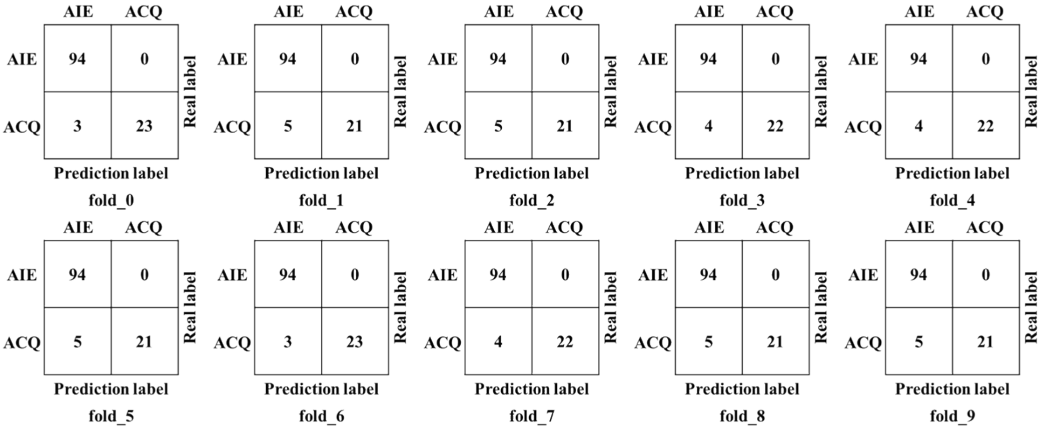


Fig.S6 Confusion matrices for the 10-fold cross-validation of the GNN (ensemble model, add RDkit feature).

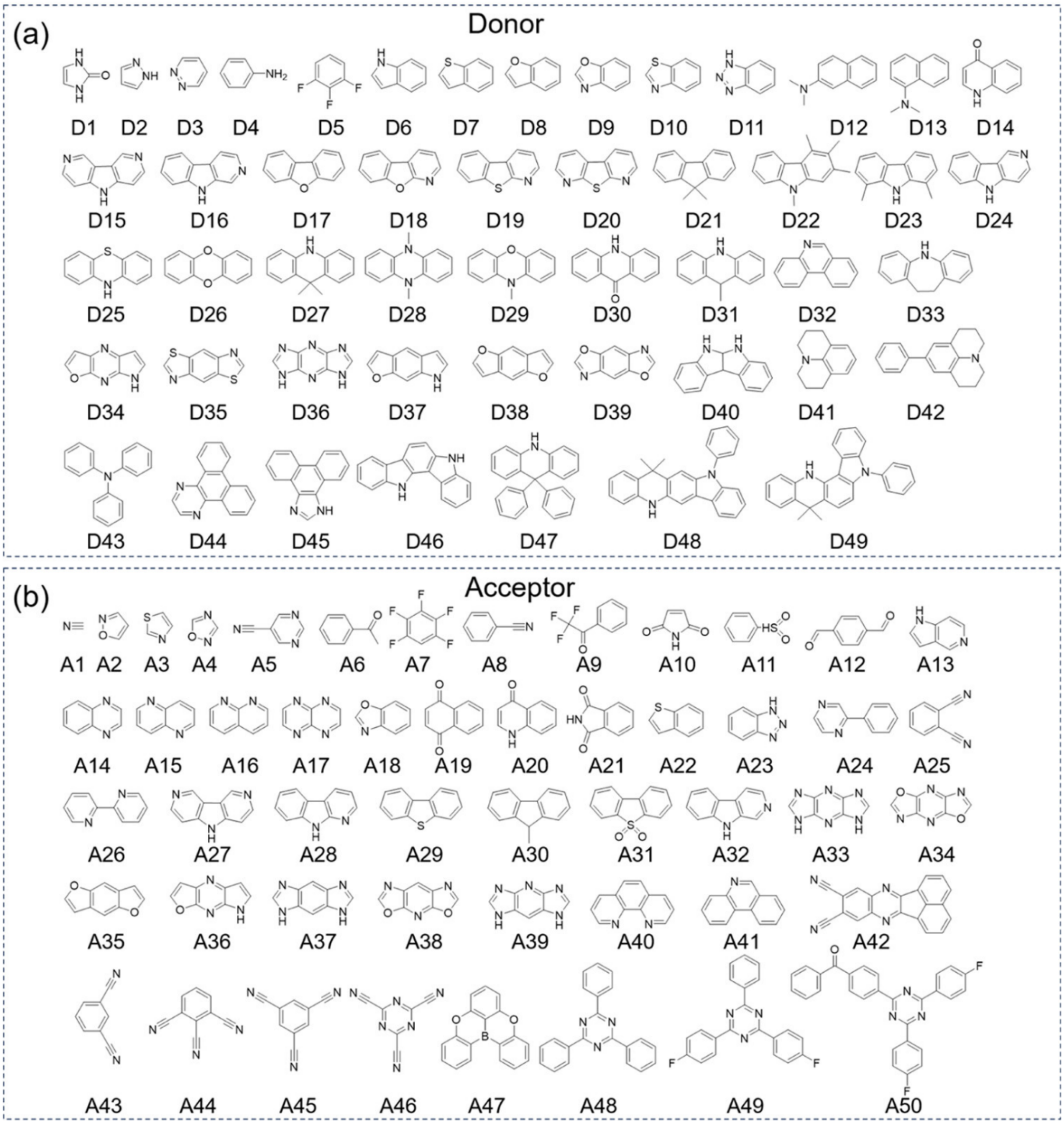

(a)
Donor
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14
D15 D16 D17 D18 D19 D20 D21 D22 D23 D24
D25 D26 D27 D28 D29 D30 D31 D32 D33
D34 D35 D36 D37 D38 D39 D40 D41 D42
D43 D44 D45 D46 D47 D48 D49
(b)
Acceptor
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11 A12 A13
A14 A15 A16 A17 A18 A19 A20 A21 A22 A23 A24 A25
A26 A27 A28 A29 A30 A31 A32 A33 A34
A35 A36 A37 A38 A39 A40 A41 A42
A43 A44 A45 A46 A47 A48 A49 A50

Fig.S7. Chemical Structure of 49 Donors and 50 Acceptors[7].

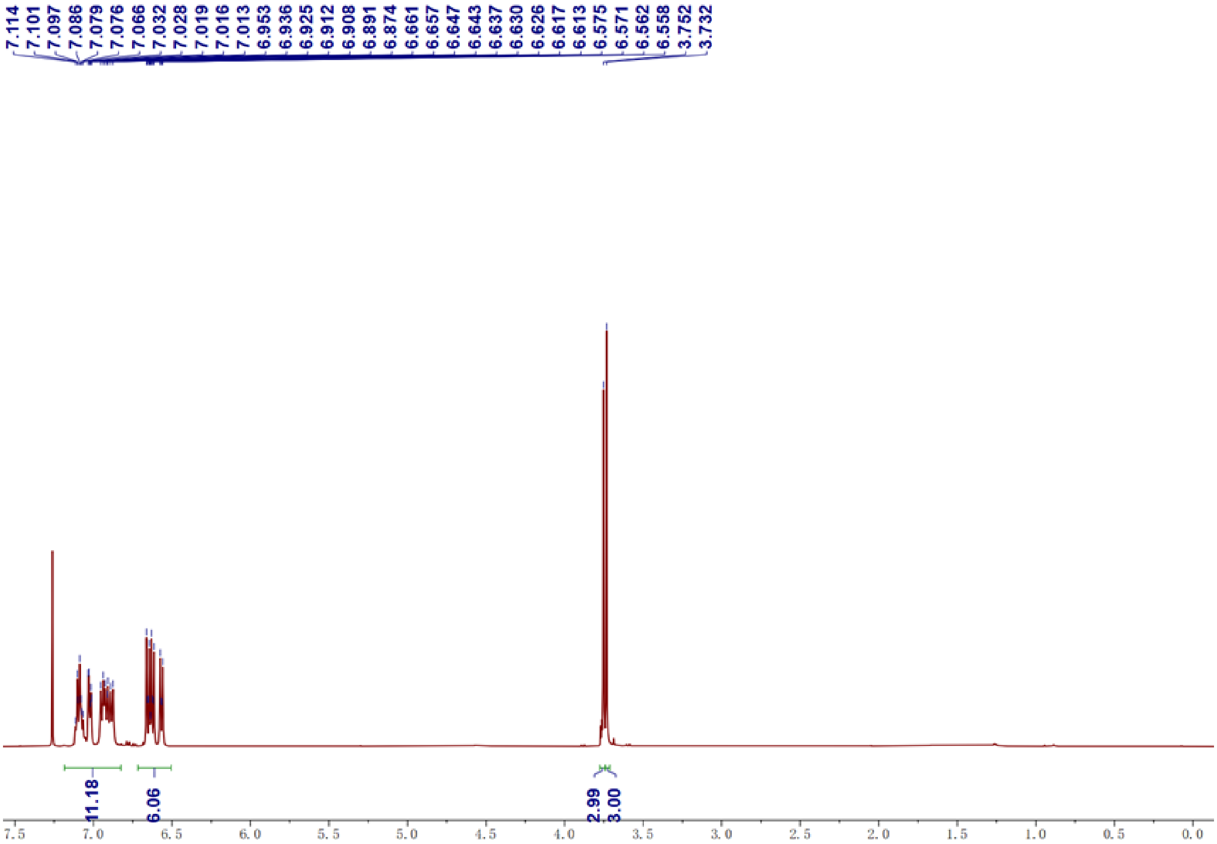


Fig.S8. ^{1}H NMR of 2MeO-TPE-OH.

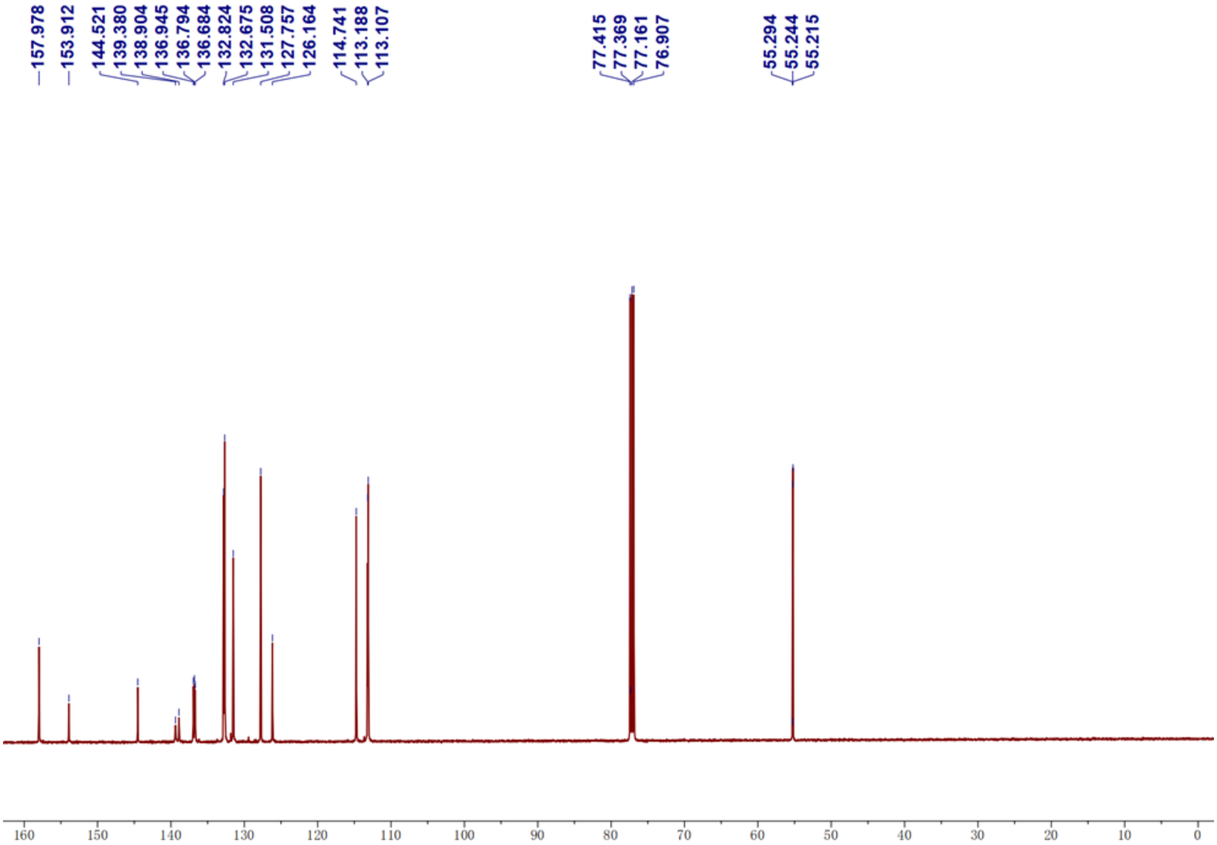


Fig.S9. ^{13}C NMR of 2MeO-TPE-OH.

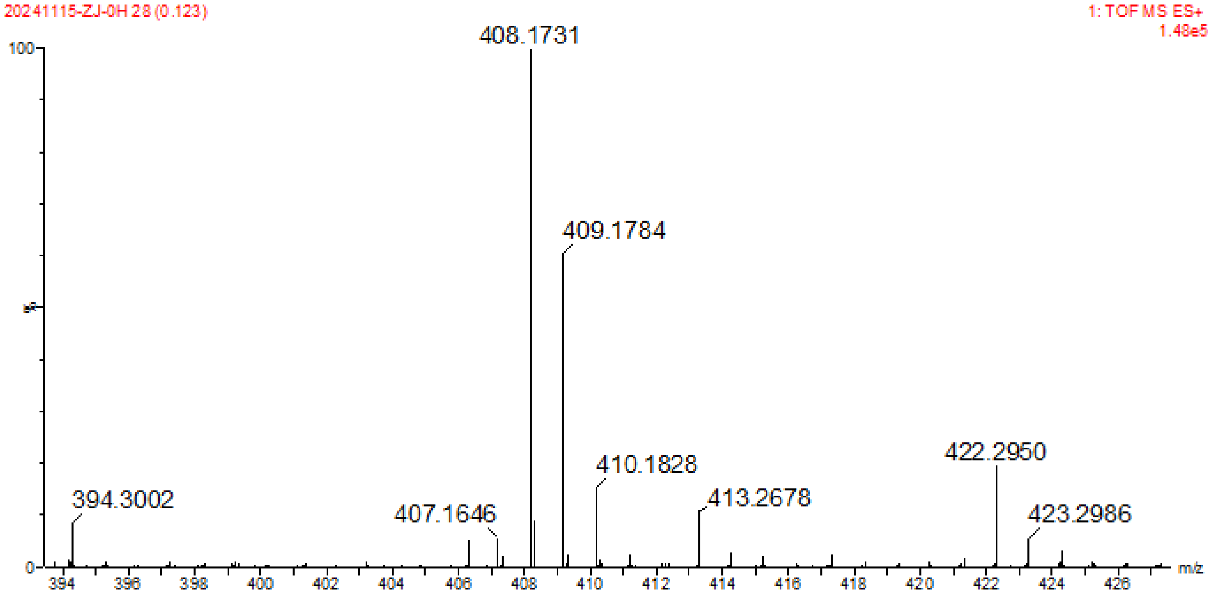


Fig.S10. Mass spectrum of 2MeO-TPE-OH.

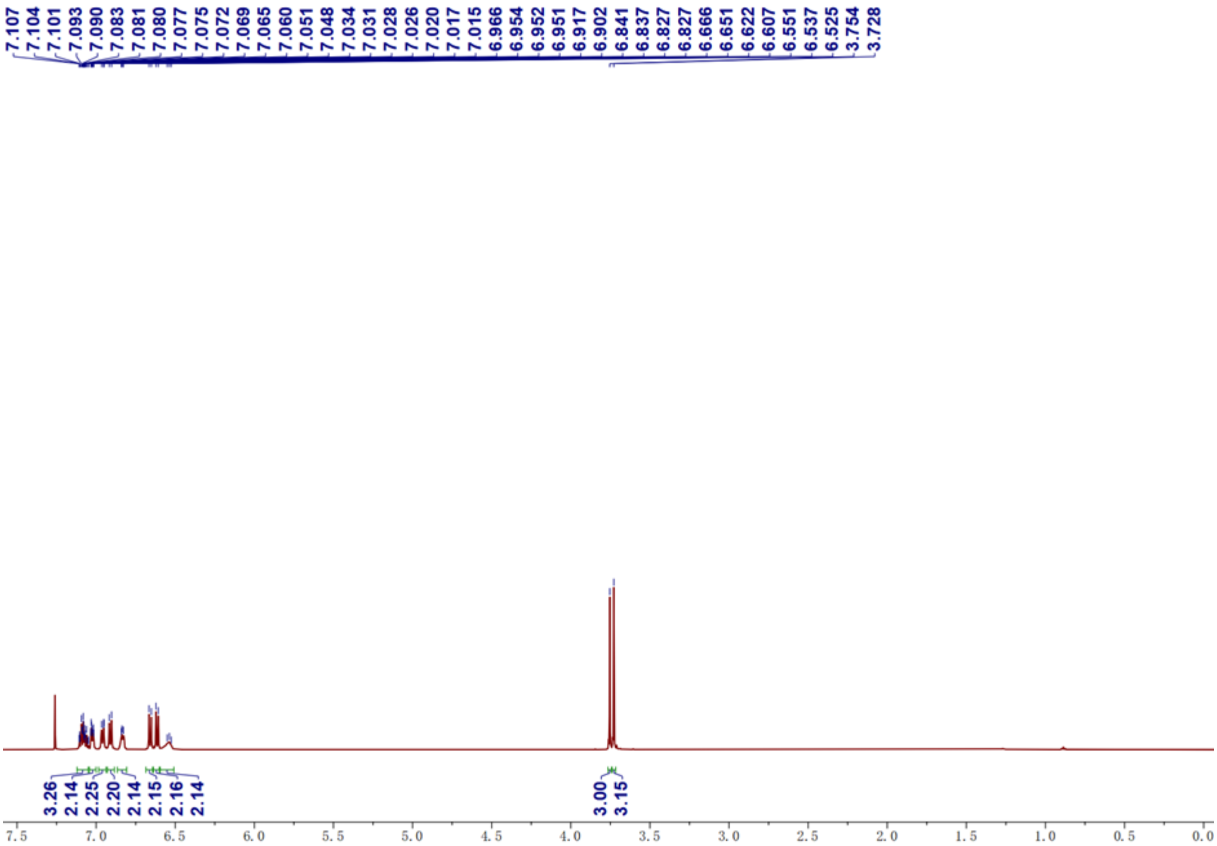


Fig.S11 ^{1}H NMR of 2MeO-TPE-NH_2.

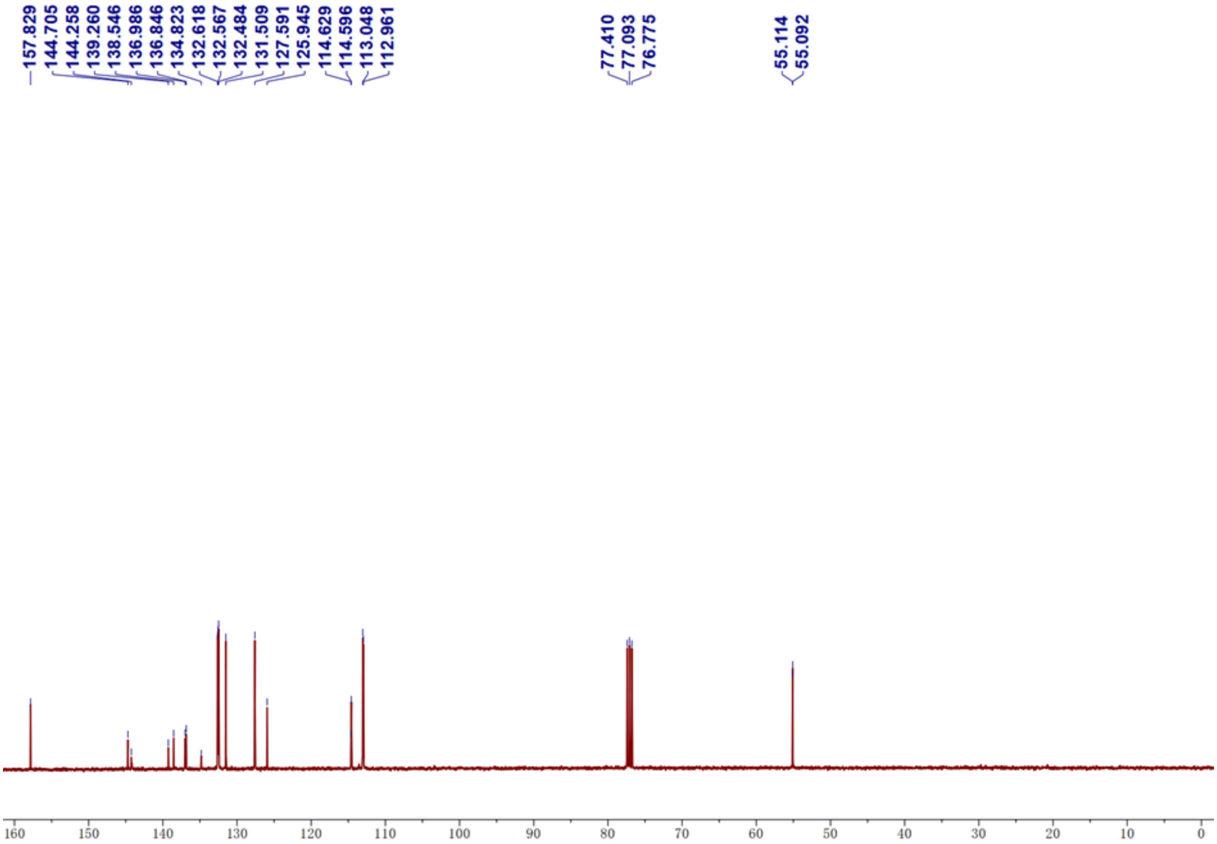


Fig.S12. ^{13}C NMR of 2MeO-TPE-NH$_2$.

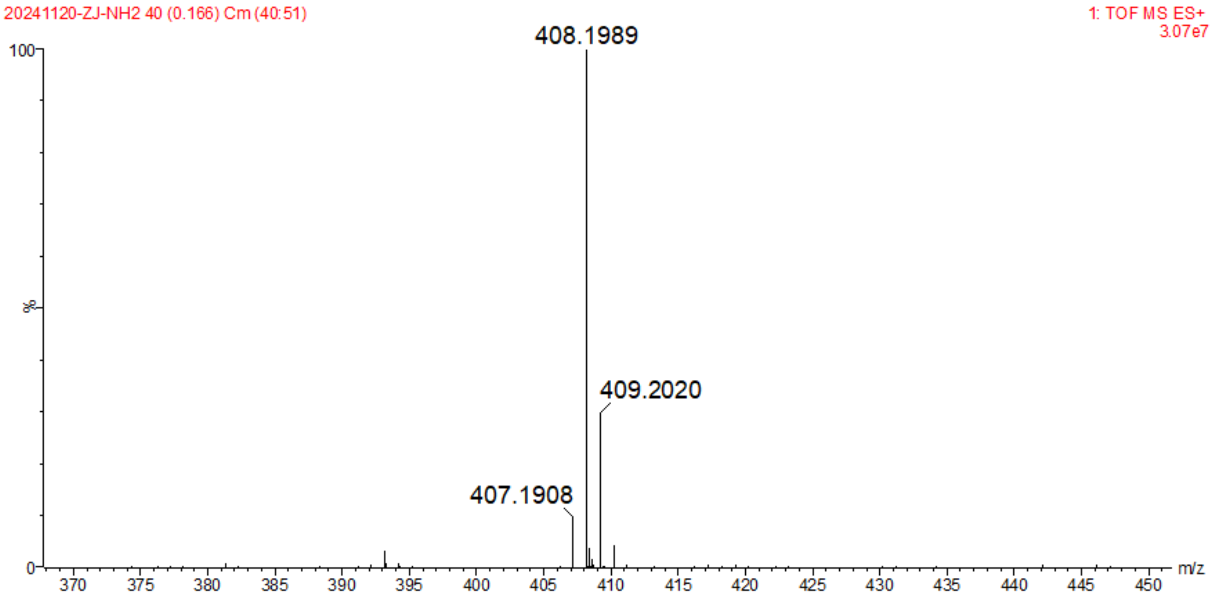


Fig.S13. Mass spectrum of 2MeO-TPE-NH$_2$.

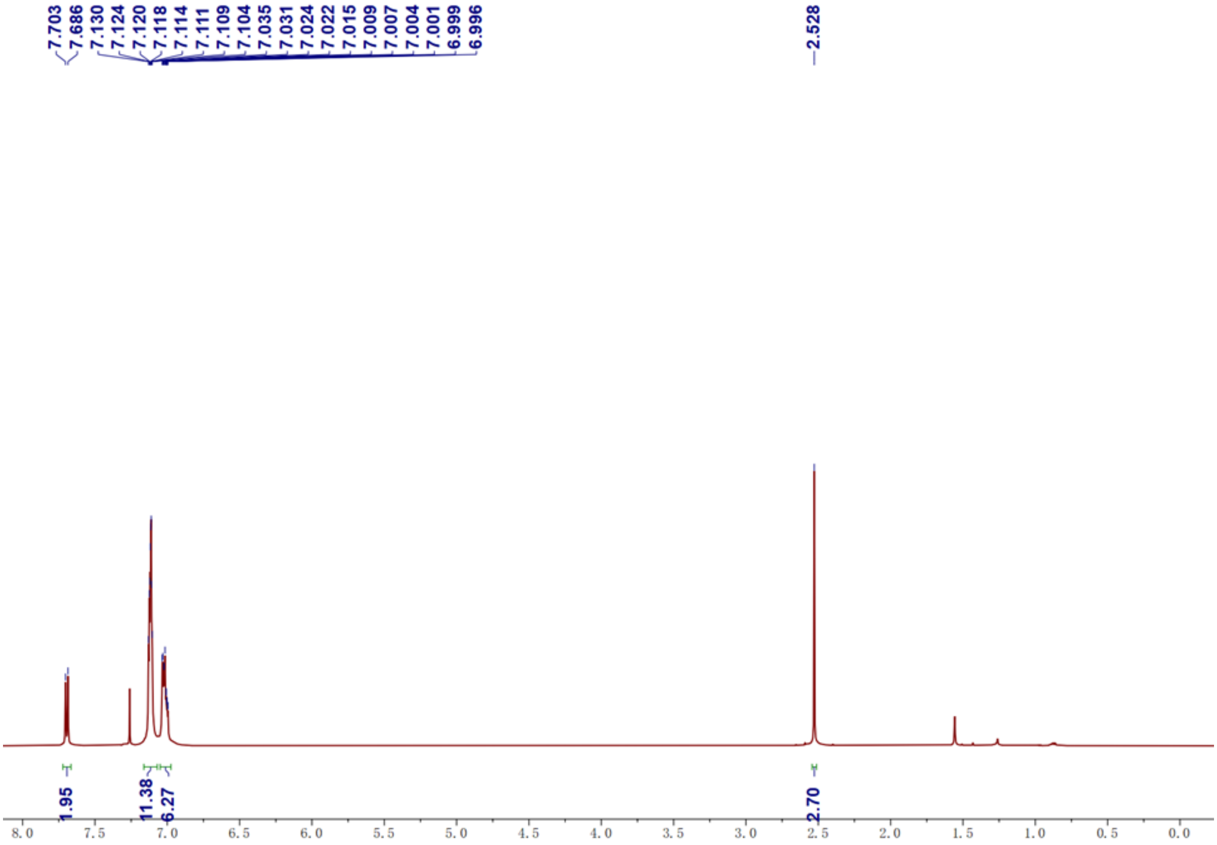


Fig.S14. ^{1}H NMR of TPE-COCH$_3$.

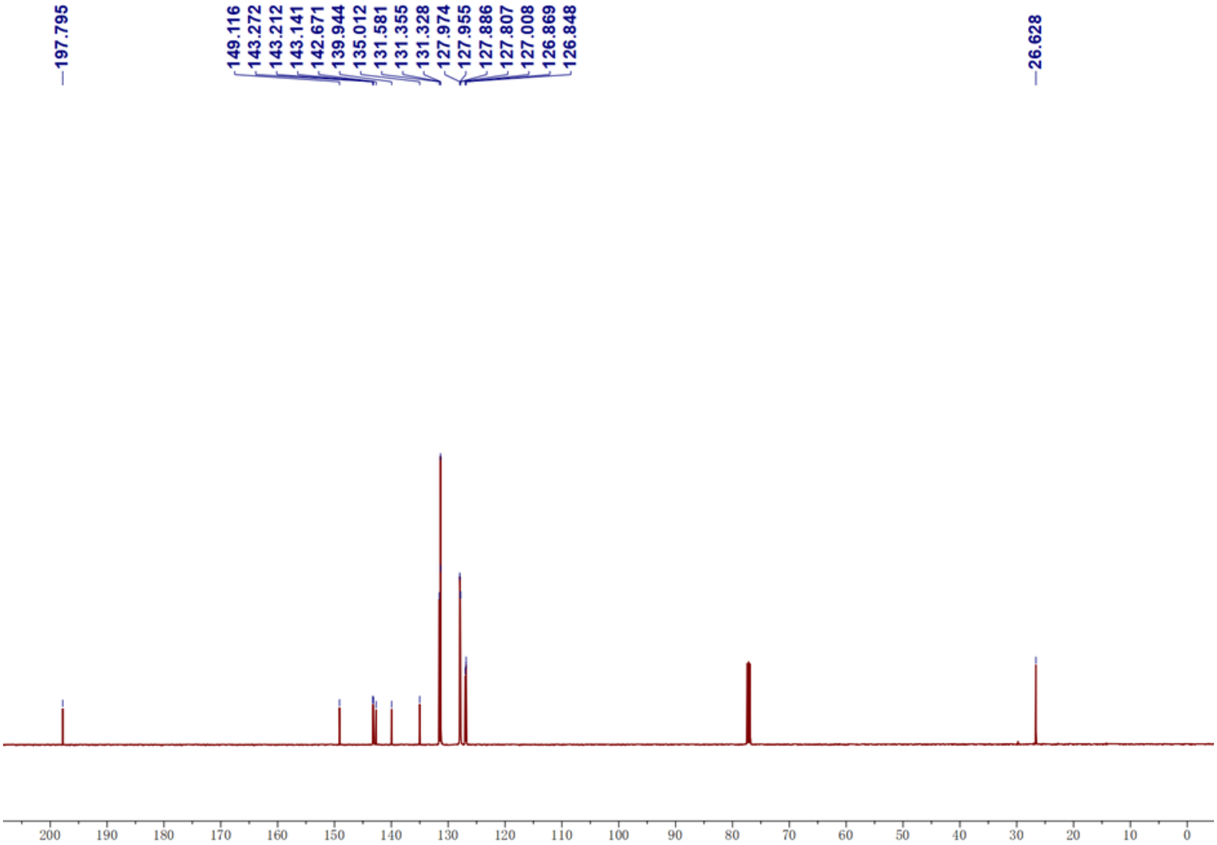


Fig.S15. ^{13}C NMR of TPE-COCH$_3$.

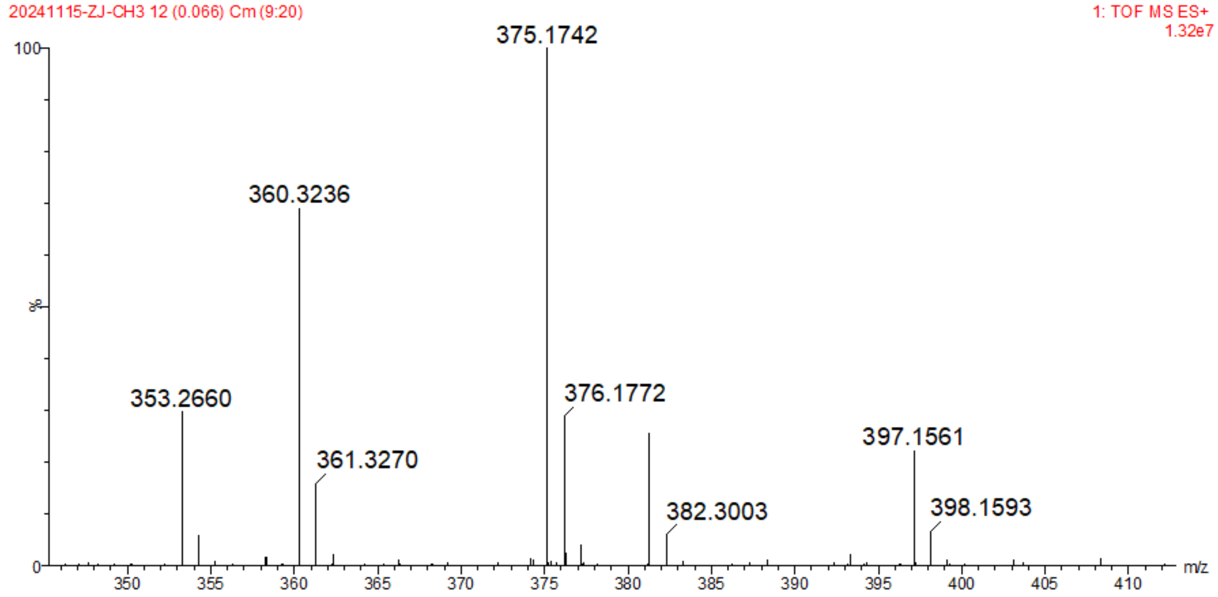


Fig.S16. Mass spectrum of 2MeO-TPE-OH.

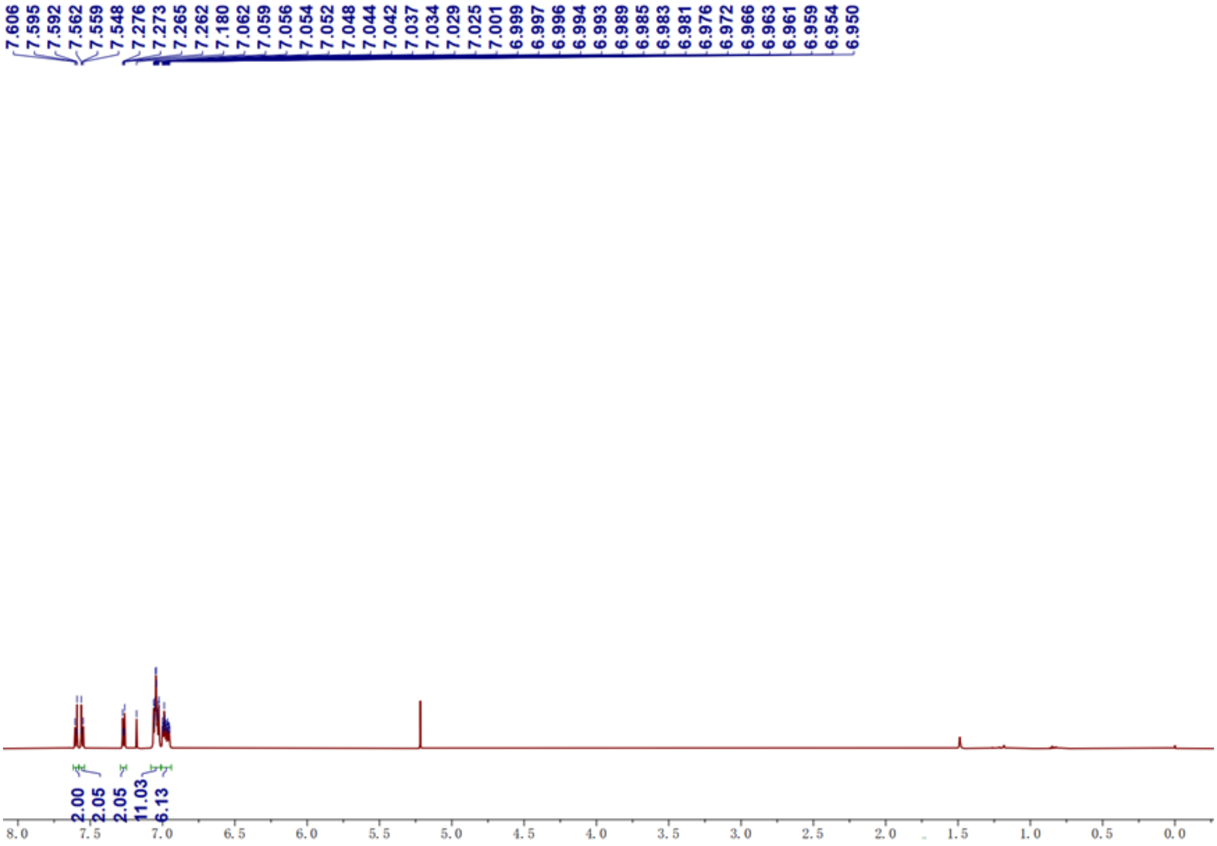


Fig.S17. ^{1}H NMR of TPE-PhCN.

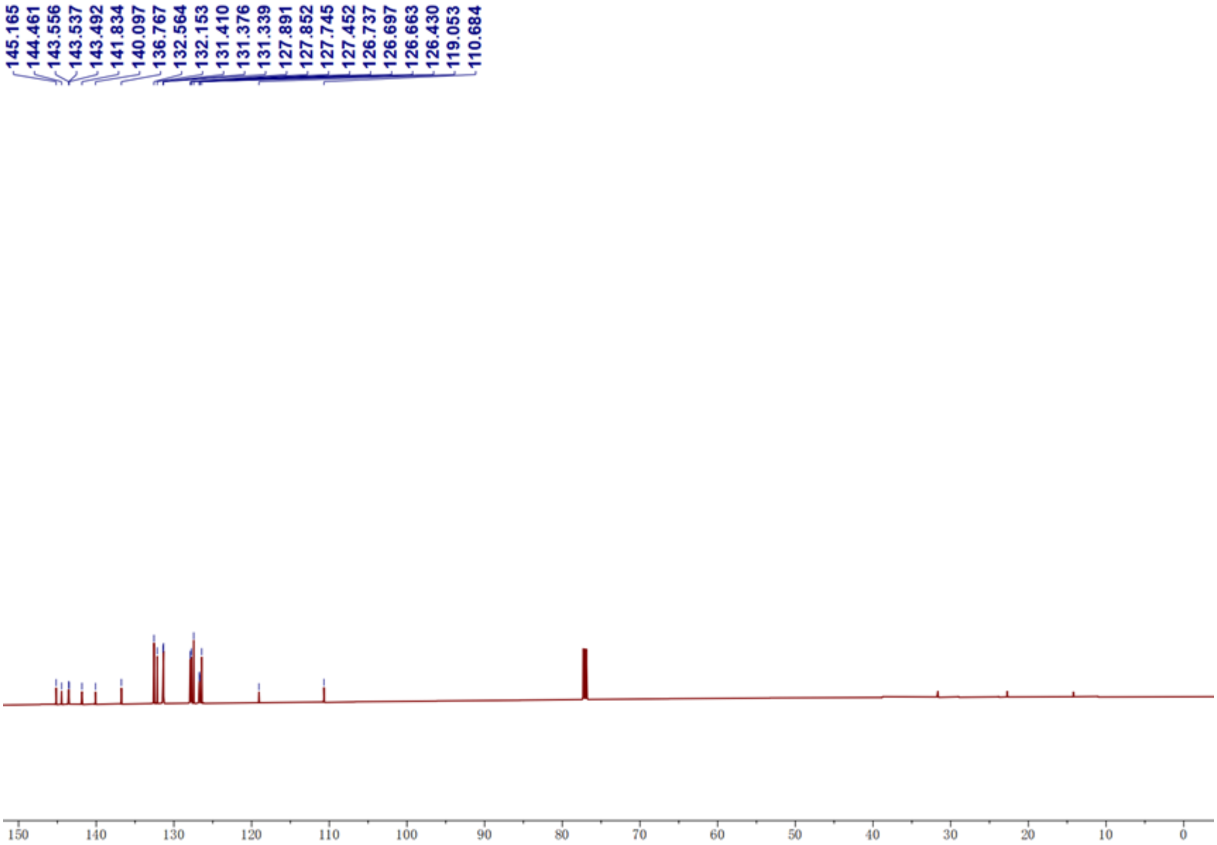


Fig.S18. ^{13}C NMR of TPE-PhCN.

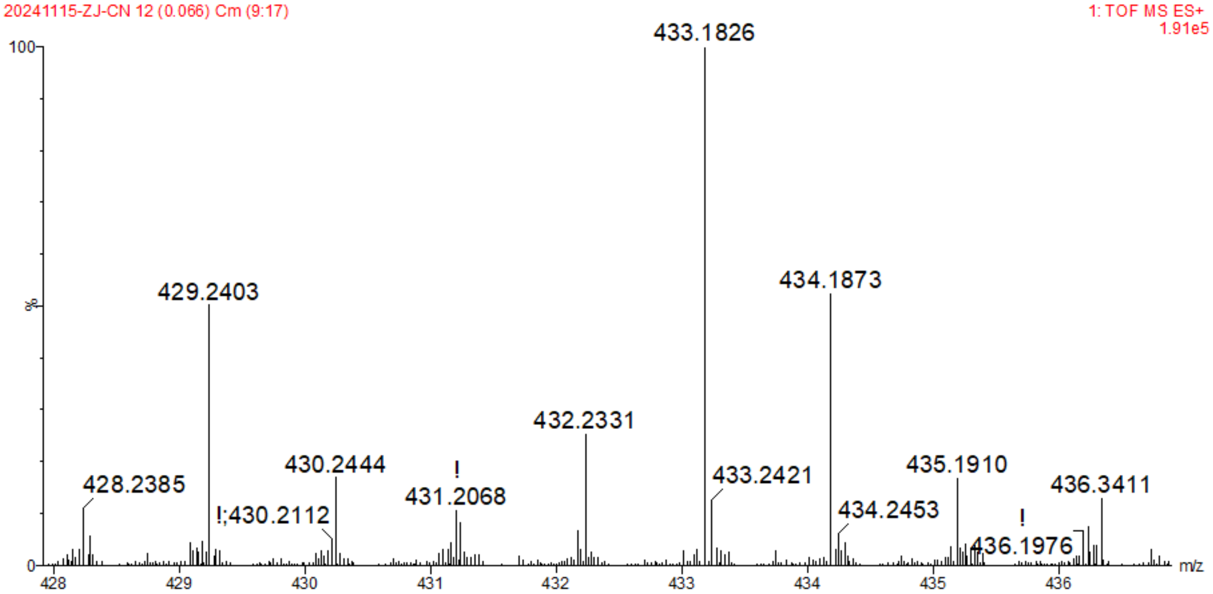


Fig.S19. Mass spectrum of TPE-PhCN.

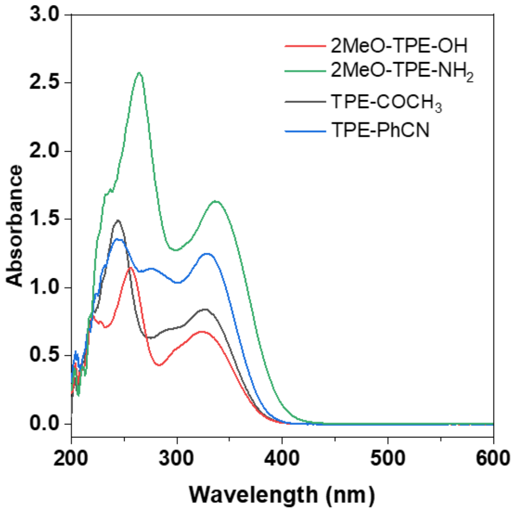


Fig.S20. UV visible absorption spectra of 2MeO-TPE-OH, 2MeO-TPE-NH$_2$, TPE-COCH$_3$, and TPE-PhCN in THF (5 x 10^{-5} M).

3. Supplementary Tables

Table S1 Molecular-level features calculated by RDKit package.

Features		
BalabanJ	BertzCT	Chi0
Chi0n	Chi0v	Chi1
Chi1n	Chi1v	Chi2n
Chi2v	Chi3n	Chi3v
Chi4n	Chi4v	EState_VSA1
EState_VSA10	EState_VSA11	EState_VSA2
EState_VSA3	EState_VSA4	EState_VSA5
EState_VSA6	EState_VSA7	EState_VSA8
EState_VSA9	ExactMolWt	FpDensityMorgan1
FpDensityMorgan2	FpDensityMorgan3	FractionCSP3
HallKierAlpha	HeavyAtomCount	HeavyAtomMolWt
Ipc	Kappa1	Kappa2
Kappa3	LabuteASA	MaxAbsEStateIndex
MaxAbsPartialCharge	MaxEStateIndex	MaxPartialCharge
MinAbsEStateIndex	MinAbsPartialCharge	MinEStateIndex
MinPartialCharge	MolLogP	MolMR
MolWt	NHOHCount	NOCount
NumAliphaticCarbocycles	NumAliphaticHeterocycles	NumAliphaticRings
NumAromaticCarbocycles	NumAromaticHeterocycles	NumAromaticRings
NumHAcceptors	NumHDonors	NumHeteroatoms
NumRadicalElectrons	NumRotatableBonds	NumSaturatedCarbocycles
NumSaturatedHeterocycles	NumSaturatedRings	NumValenceElectrons
PEOE_VSA1	PEOE_VSA10	PEOE_VSA11
PEOE_VSA12	PEOE_VSA13	PEOE_VSA14
PEOE_VSA2	PEOE_VSA3	PEOE_VSA4
PEOE_VSA5	PEOE_VSA6	PEOE_VSA7
PEOE_VSA8	PEOE_VSA9	RingCount
SMR_VSA1	SMR_VSA10	SMR_VSA2
SMR_VSA3	SMR_VSA4	SMR_VSA5
SMR_VSA6	SMR_VSA7	SMR_VSA8
SMR_VSA9	SlogP_VSA1	SlogP_VSA10
SlogP_VSA11	SlogP_VSA12	SlogP_VSA2
SlogP_VSA3	SlogP_VSA4	SlogP_VSA5
SlogP_VSA6	SlogP_VSA7	SlogP_VSA8
SlogP_VSA9	TPSA	VSA_EState1
VSA_EState10	VSA_EState2	VSA_EState3
VSA_EState4	VSA_EState5	VSA_EState6
VSA_EState7	VSA_EState8	VSA_EState9
fr_Al_COO	fr_Al_OH	fr_Al_OH_noTert
fr_ArN	fr_Ar_COO	fr_Ar_N
fr_Ar_NH	fr_Ar_OH	fr_COO
fr_COO2	fr_C_O	fr_C_O_noCOO
fr_C_S	fr_HOCCN	fr_Imine
fr_NH0	fr_NH1	fr_NH2
fr_N_O	fr_Ndealkylation1	fr_Ndealkylation2

fr_Nhpyrrole	fr_SH	fr_aldehyde
fr_alkyl_carbamate	fr_alkyl_halide	fr_allylic_oxid
fr_amide	fr_amidine	fr_aniline
fr_aryl_methyl	fr_azide	fr_azo
fr_barbitur	fr_benzene	fr_benzodiazepine
fr_bicyclic	fr_diazo	fr_dihydropyridine
fr_epoxide	fr_ester	fr_ether
fr_furan	fr_guanido	fr_halogen
fr_hdrzine	fr_hdrzone	fr_imidazole
fr_imide	fr_isocyan	fr_isothiocyan
fr_ketone	fr_ketone_Topliss	fr_lactam
fr_lactone	fr_methoxy	fr_morpholine
fr_nitrile	fr_nitro	fr_nitro_arom
fr_nitro_arom_nonortho	fr_nitroso	fr_oxazole
fr_oxime	fr_para_hydroxylation	fr_phenol
fr_phenol_noOrthoHbond	fr_phos_acid	fr_phos_ester
fr_piperdine	fr_piperzine	fr_priamide
fr_prisulfonamd	fr_pyridine	fr_quatN
fr_sulfide	fr_sulfonamd	fr_sulfone
fr_term_acetylene	fr_tetrazole	fr_thiazole
fr_thiocyan	fr_thiophene	fr_unbrch_alkane
fr_urea	qed	

Table S2 Hyperparameters used for GNN model.

Hyperparameter	Range	Value used
Number of MPNN Layers	[2, 6]	5
MPNN Hidden Layer Size	[100, 2400]	1300
Number of FFN Layers	[1, 3]	2
FFN Hidden Layer Size	[100, 2400]	1400
Dropout Rate	[0, 0.4]	0.25
Initial Learning Rate	[0.0001, 1]	0.0016475184132470895
Maximum Learning Rate	[0.000001, 1]	0.00026550826157806045
Final Learning Rate	[0.0001, 1]	0.00017424185057835588
Warmup Epochs	[1, 30]	4

Table S3 Hyperparameters used for XGBoost model.

Hyperparameter	Range	Value used
Number of Estimators	[50, 200]	50
Learning Rate	[0.01, 0.5]	0.5
Maximum Depth	[3, 7]	7
Minimum Loss Reduction	[0, 0.5]	0.1
Colsample by Tree	[0.5, 1.0]	0.5

Table S4 Hyperparameters used for SVM model.

Hyperparameter	Range	Value used
Kernel Type	Linear, Rbf	Rbf
Regularization Parameter C	[0.1, 100]	10
Gamma	Scale, Auto	Auto

Table S5 Hyperparameters used for RF model.

Hyperparameter	Range	Value used
Number of Estimators	[50, 200]	200
Maximum Depth	[0, 30]	0
Minimum Samples Split	[2, 10]	2
Minimum Samples Leaf	[1, 4]	1

Table S6 Hyperparameters used for KNN model.

Hyperparameter	Range	Value used
Number of Neighbors	[3, 9]	9
Weights	Uniform, Distance	Distance
Power Parameter	[1, 3]	1

Table S7 Hyperparameters used for MLP model.

Hyperparameter	Range	Value used
Hidden Layer Sizes	100, 200, (100, 100)	100
Activation Function	Tanh, ReLU	ReLU
Solver	Adam, Sgd	Sgd
Maximum Iterations	[500, 800]	500
Alpha	[0.0001, 0.001]	0.0001

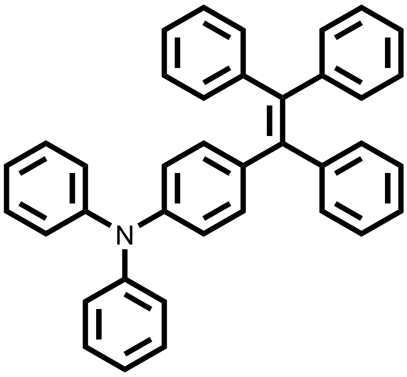
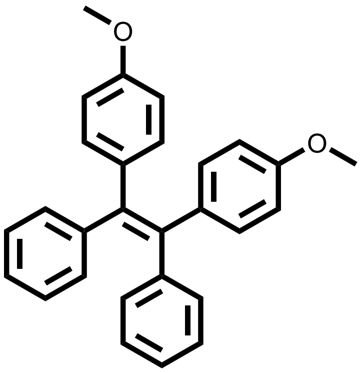
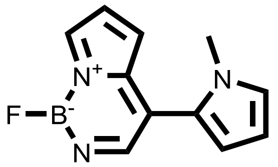
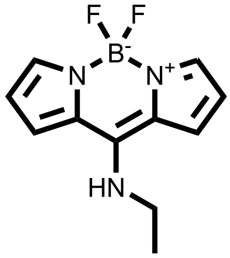
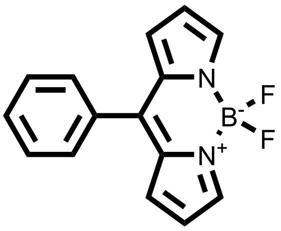
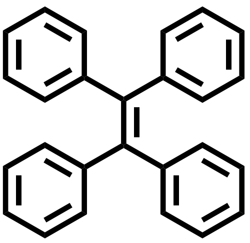
Table S8 Performance of XGBoost, SVM, RF, KNN, MLP, GNN, GNN* (add RDkit feature), and GNN** (ensemble model, add RDkit feature) on the test set.

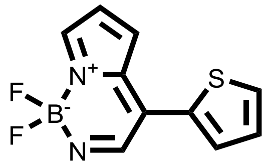
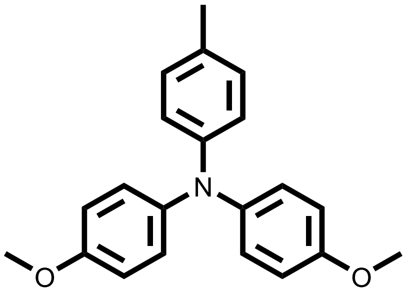
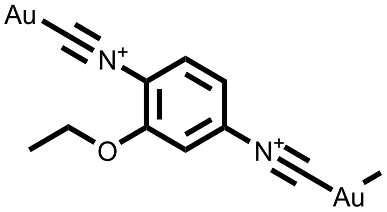
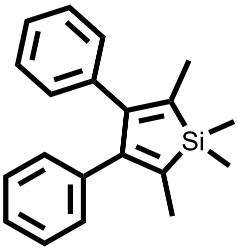
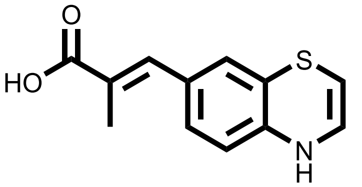
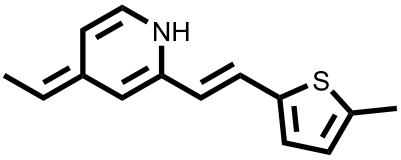
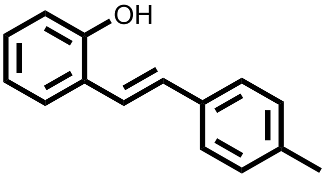
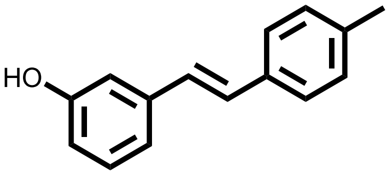
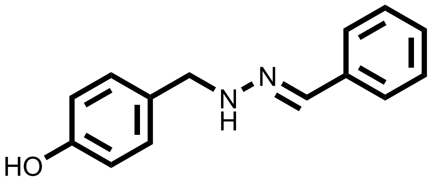
Model	Accuracy	AUPRC	AUROC	F1 Score	MCC	Precision	Recall
XGBoost	0.942	0.996	0.987	0.962	0.937	0.978	0.946
SVM	0.925	0.995	0.981	0.953	0.772	0.938	0.968
RF	0.933	0.992	0.975	0.957	0.804	0.957	0.957
KNN	0.875	0.990	0.962	0.922	0.610	0.899	0.947
MLP	0.917	0.996	0.984	0.947	0.748	0.978	0.947
GNN	0.944	0.994	0.980	0.966	0.831	0.936	0.998
GNN*	0.963	0.997	0.990	0.977	0.891	0.956	0.999
GNN**	0.964	0.997	0.990	0.978	0.893	0.956	1.000

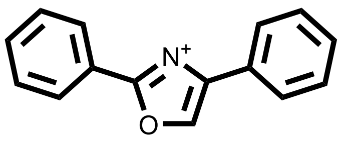
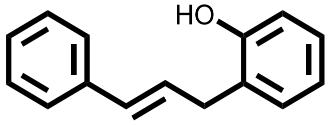
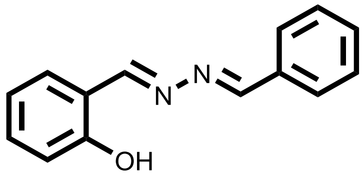
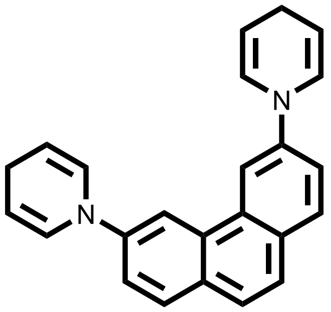
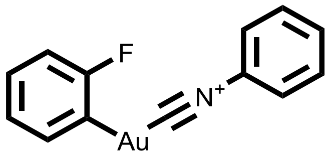
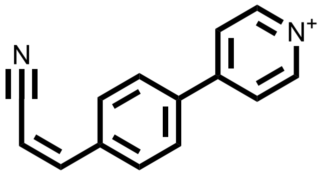
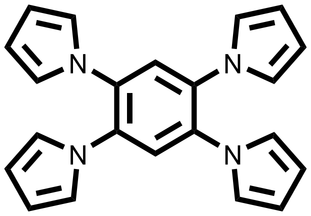
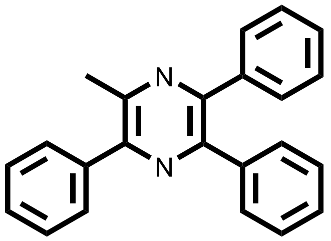
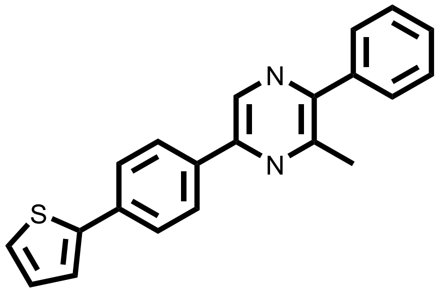
Table S9 Performance of XGBoost, SVM, RF, KNN, MLP, GNN, GNN* (add RDkit feature), and GNN** (ensemble model, add RDkit feature) on the training set.

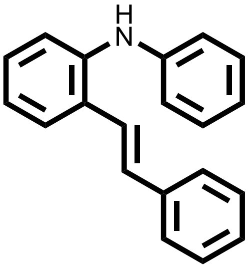
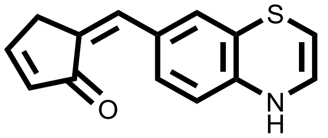
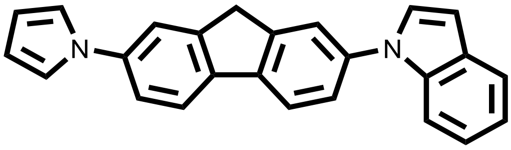
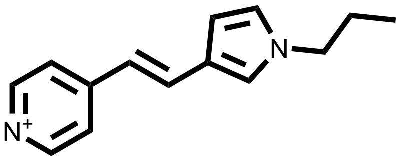
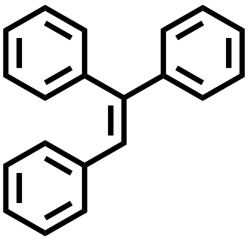
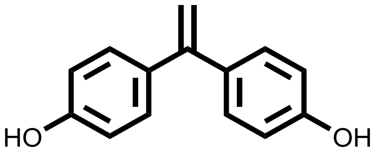
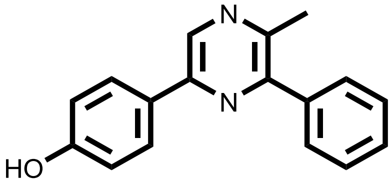
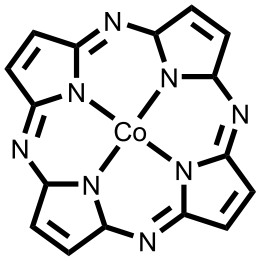
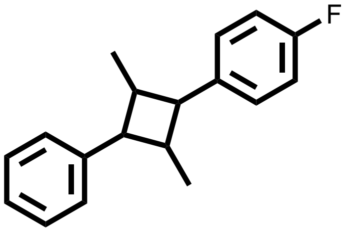
Model	Accuracy	AUPRC	AUROC	F1 Score	MCC	Precision	Recall
XGBoost	0.995	1.000	1.000	0.997	0.986	0.996	0.998
SVM	0.993	0.996	0.996	0.996	0.980	0.994	0.998
RF	1.000	1.000	1.000	1.000	1.000	1.000	1.000
KNN	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MLP	0.997	1.000	1.000	0.998	0.992	0.999	0.998
GNN	0.940	0.965	0.928	0.963	0.816	0.939	0.988
GNN*	0.946	0.981	0.951	0.966	0.834	0.941	0.993
GNN**	0.946	0.981	0.951	0.966	0.834	0.942	0.992

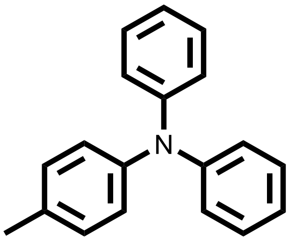
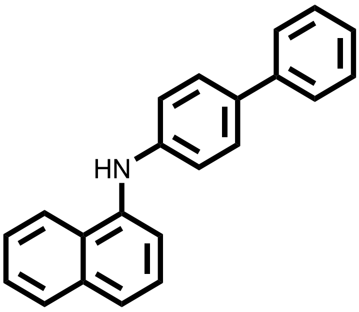
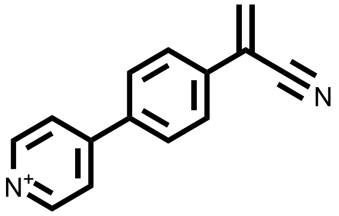
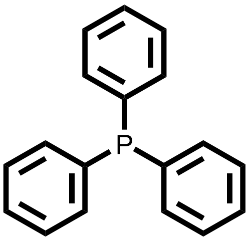
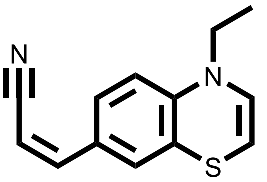
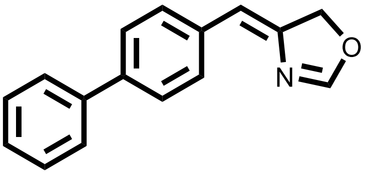
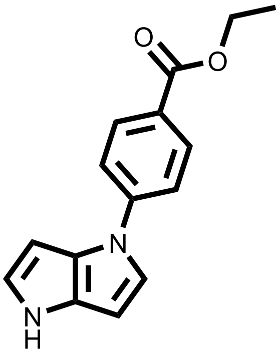
Table S10 AIE functional groups identified through Monte Carlo search methods.

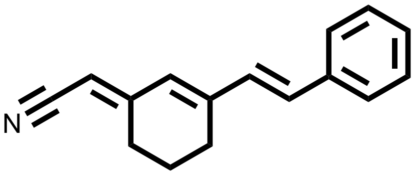
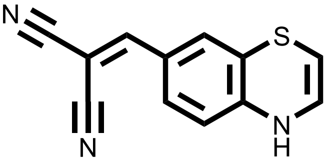
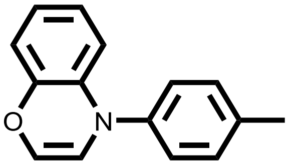
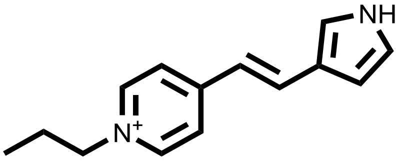
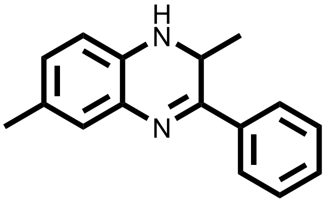
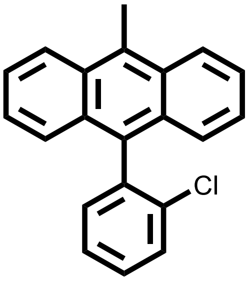
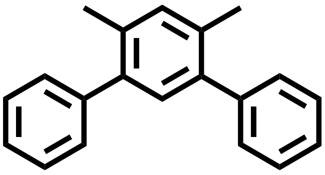
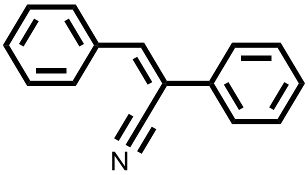
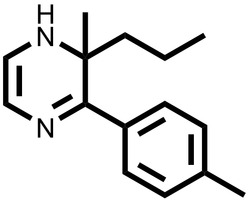
Structural formulas	Predicted score	Number of occurrences
	0.964	14
	0.961	51
	0.926	2
	0.894	3
	0.892	36
	0.890	217

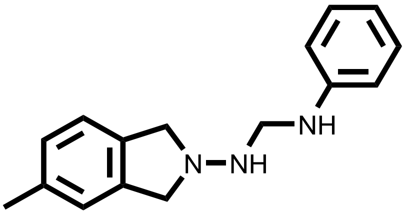
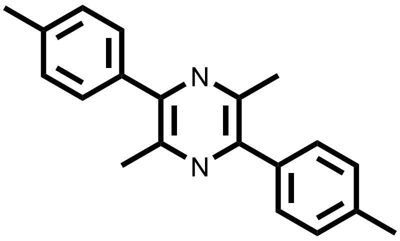
	0.882	2
	0.878	18
	0.860	2
	0.856	31
	0.825	6
	0.788	2
	0.777	3
	0.772	2
	0.730	2

	0.728	4
	0.719	3
	0.713	6
	0.705	3
	0.698	2
	0.686	2
	0.675	3
	0.674	2
	0.670	3

	0.667	2
	0.663	3
	0.650	2
	0.637	2
	0.627	29
	0.619	2
	0.617	4
	0.615	2
	0.609	2

	0.600	141
	0.596	2
	0.594	2
	0.585	2
	0.569	5
	0.567	3
	0.554	2

	0.549	2
	0.548	2
	0.541	2
	0.538	2
	0.532	2
	0.531	2
	0.518	4
	0.518	3
	0.508	2

	0.503	2
	0.501	5

4. References

1. J. Y. Gong, W. J. Gong, B. Wu, H. R. Wang, W. He, Z. Y. Dai, Y. Z. Li, Y. Liu, Z. M. Wang, X. J. Tuo, J. W. Y. Lam, Z. J. Qiu, Z. Zhao and B. Z. Tang, *Aggregate*, 2022, e263.
2. S. D. Xu, X. L. Liu, P. F. Cai, J. L. Li, X. N. Wang and B. Liu, *Adv. Sci.*, 2021, **9**, 2101074.
3. F. Wong, E. J. Zheng, J. A. Valeri, N. M. Donghia, M. N. Anahtar, S. Omori, A. Li, A. Cubillos-Ruiz, A. Krishnan, W. Jin, A. L. Manson, J. Friedrichs, R. Helbig, B. Hajian, D. K. Fiejtek, F. F. Wagner, H. H. Soutter, A. M. Earl, J. M. Stokes, L. D. Renner and J. J. Collins, *Nature*, 2024, **626**, 177-185.
4. E. Heid, K. P. Greenman, Y. Chung, S. C. Li, D. E. Graff, F. H. Vermeire, H. Y. Wu, W. H. Green and C. J. McGill, *J. Chem. Inf. Model.*, 2024, **64**, 9-17.
5. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, *J. Chem. Inf. Model.*, 2019, **59**, 3370-3388.
6. W. M. Wan, D. Tian, Y. N. Jing, X. Y. Zhang, W. Wu, H. Ren and H. L. Bao, *Angew. Chem. Int. Ed.*, 2018, **57**, 15510-15516.
7. Y. Bu and Q. Peng, *The Journal of Physical Chemistry C*, 2023, **127**, 23845-23851.