Supplementary Materials for

Self-Improving Photosensitizer Discovery System via Bayesian Search with First-Principle Simulations

Shidang Xu[‡1], Jiali Li[‡1], Pengfei Cai[3], Xiaoli Liu[1], Bin Liu[1,2]* and Xiaonan Wang[1,+]*

[1]Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585, Singapore

[2]Joint School of National University of Singapore and Tianjin University, International Campus of Tianjin University, Binhai New City, Fuzhou 350207, China

[3]Department of Materials Science and Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore 117575, Singapore

[‡]Both authors contributed equally to the work.

[+]Current address: Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

*Corresponding author. E-mail: cheliub@nus.edu.sg; chewxia@nus.edu.sg

General

All starting materials are commercially available and were used as supplied unless otherwise indicated. All experiments were conducted in air unless otherwise noted. 2-Bromoanthracene-9,10-dione, 1,4-dibromo-2,3-dihydronaphthalene-2,3-diamine, 2-chlorobenzoic acid, 3,6-dibromophenanthrene-9,10-dione, and other chemicals and reagents for the synthesis were purchased from Sigma-Aldrich and Tee Hai Chem Ltd. and used as received without any further purification. Compounds **1-4** and related intermediates were synthesized and characterized according to the methods described in the supporting information.

NMR spectra were recorded on a Bruker ARX 400 NMR spectrometer. Chemical shifts were recorded in parts per million referenced according to residual solvent ($CDCl_3$ = 7.26 ppm) in ^1H NMR and ($CDCl_3$ = 77.0 ppm) in ^{13}C NMR. Mass spectra were reported on the AmaZon X LC-MS for ESI. Data were measured using omega and phi scans of 0.5° per frame. UV-vis absorption spectra were obtained on a Shimadzu Model UV-1700 spectrometer. Photoluminescence (PL) spectra were measured on a Perkin-Elmer LS 55 spectrofluorometer. All UV and PL spectra were collected at 24 ± 1 °C.

Hyperparameter Optimization with Gaussian Process

Hyperparameters are external configurations of a model that are used in the training process to estimate the model parameters and it is necessary to tune the hyperparameters finely to obtain an accurate prediction model. Hyperparameters for both initial DA and DAD models were optimized with a Bayesian optimization-based gaussian search process. The optimization process was done with scikit-optimize (https://scikit-optimize.github.io/#skopt.gp_minimize) to minimize MAE till 50 trials have been done. In each cycle, a fixed training (80%) and test set (20%) were used for DA and DAD models. The following hyperparameters were tuned:

- *Graph convolutional layers:* A list of graph convolutional layers with each value representing the number of nodes in each layer.
- *Dense layers:* A list of dense fully connected layers with each value representing the number of nodes in each layer.
- *Dropout:* Probability (between 0 and 1) that neurons in the hidden layers are ignored; dropout is added to prevent overfitting.
- *Learning rate:* The multiplier for gradient descent and determines how fast the parameter changes.
- *Epochs:* Number of complete passes through the training dataset by the model
- *Batch size:* Number of training samples used in each epoch.

It is noted that it is impossible to determine the best hyperparameters for a specific problem. Thus, the table below shows the final hyperparameters that are used in all models in the initial model training and across all active learning cycles, in which they are considered to produce accurate enough model predictions.

Table S1. Hyperparameters for both DA and DAD models

	DA Model	**DAD Model**
Structures in the training set	DA	DA and DAD
Graph convolutional layers	295, 295, 295, 295, 295, 295	512, 512, 512, 512
Dense layers	382, 382, 382, 382	128, 128, 128
Dropout	0.00874	0.01
Learning rate	0.0001	0.001

Batch size	10	10

Modified Equation for Expected Improvement used in Bayesian Optimization

The expected improvement (EI) equation was adapted to optimize a minimized value of ΔE_{ST}, and the following equation is used in our BO algorithm for each active learning cycle.

$$EI(\Delta E_{ST}) = \begin{cases} (\Delta E_{ST}{}^{-} - \mu(\Delta E_{ST}) - \xi)\Phi(\mathrm{Z}) + \sigma(\Delta E_{ST})\phi(Z) & if\ \sigma(\Delta E_{ST}) > 0 \\ 0 & if\ \sigma(\Delta E_{ST}) = 0 \end{cases}, \quad (\mathbf{Eq.\ S1})$$

where

$$Z = \begin{cases} \dfrac{\Delta E_{ST}{}^{-} - \mu(\Delta E_{ST}) - \xi}{\sigma(\Delta E_{ST})} & if\ \sigma(\Delta E_{ST}) > 0 \\ 0 & if\ \sigma(\Delta E_{ST}) = 0 \end{cases}$$

Here, $\Delta E_{ST}{}^{-}$ represents the smallest ΔE_{ST} observed in the labeled dataset so far, $\mu(\Delta E_{ST})$ and $\sigma(\Delta E_{ST})$ represent the mean and standard deviation of each predicted ΔE_{ST} value in the screened dataset by the current surrogate navigation models, Φ and ϕ represent the cumulative distribution function and the probability density function, respectively, and a trade-off value of $\xi = 0.01$ is used. In our case, each prediction's mean and standard deviation is derived from navigation models with different dropouts. As seen from the EI score calculation, there are two counteracting terms in the equation, each focusing on exploitation or exploration of the molecular search space. The first term, $(\Delta E_{ST}{}^{-} - \mu(\Delta E_{ST}) - \xi)\Phi(\mathrm{Z})$, is the exploitation term and it contributes a high value when the molecule is predicted to have a small ΔE_{ST}. The second term, $\sigma(\Delta E_{ST})\phi(Z)$, is the exploration term and it contributes a high value when the molecule is predicted with high uncertainty by the surrogate models.

Detailed Active Learning Data Progression Breakdown

The specific breakdown in labeled data used for model training in each cycle is summarized below.

Table S2. Breakdown of the number of labeled structures used for model training in each active learning cycle before a screening of the unlabeled space

Cycle (N)	DA Model			DAD Model			
	DA			DA	DAD		
	Training	Suggestions added for Cycle N+1	Screening Space	Training	Training	Suggestions added for Cycle N+1	Screening Space
0 (Initial)	7101	-	-	7691	4914	-	-
1	7101	119	123239		4914	119	200000
2	7220	112	123120		5033	120	200000
3 *	7332	120	25102		5153	120	17730
4	7452	119	24982		5273	120	17610
5 *	7571	120	24862		5393	120	18621
6	7691	120	24742		5513	120	18501

7	7811	120	24622		5633	120	18381
8	7931	120	24502		5753	120	18261
9 #	8051	-	24382		5873	120	18141
10					5993	120	18021
11 #					6113	-	17901
Total	**8051**	**950**			**6113**	**1199**	

* It is noted that initially, the search space is formed by the combinations of 96 donors, 98 acceptors, and 14 bridges (including single bond). From cycle 3 onwards, 13 new donors, 5 new acceptors, and 9 new bridges were added to the substructure list, and from cycle 5 onwards, 3 new acceptors were added for DAD.

In these cycles, the DA model is trained on 8051 PSs and the DAD model is trained on 13804 PSs before these final models are used for the predictions on the remaining unlabeled dataset for final recommendations.

To evaluate the model performances, the model in every cycle was validated on a fixed test set of 806 DA and 612 DAD structures, respectively. The fixed test set includes random structures from initial and all active learning cycles. A mean MAE was obtained by running 5 repetitions of model training on the same training set for each cycle.

Prediction of HOMO-LUMO (H-L) Energy Gap

To evaluate the absorption onset of PS candidates, prediction models for both DA and DAD form PSs are trained to predict the H-L gaps as well. Note that, instead of the S1 level, the H-L gap is being used because DFT-B3LYP is less sensitive to issues with charge transfer than TD-B3LYP. Figure S1 shows the initial prediction performance for H-L gap. Similar to the prediction of ΔE_{ST} (Figure 2b-c), the predicted values of H-L gap are very close to the quantum calculated values. The H-L gap prediction performance for the DA model is also better than that of the DAD model, due to significantly larger design space for DAD and larger molecules for DAD compared to DA PSs.

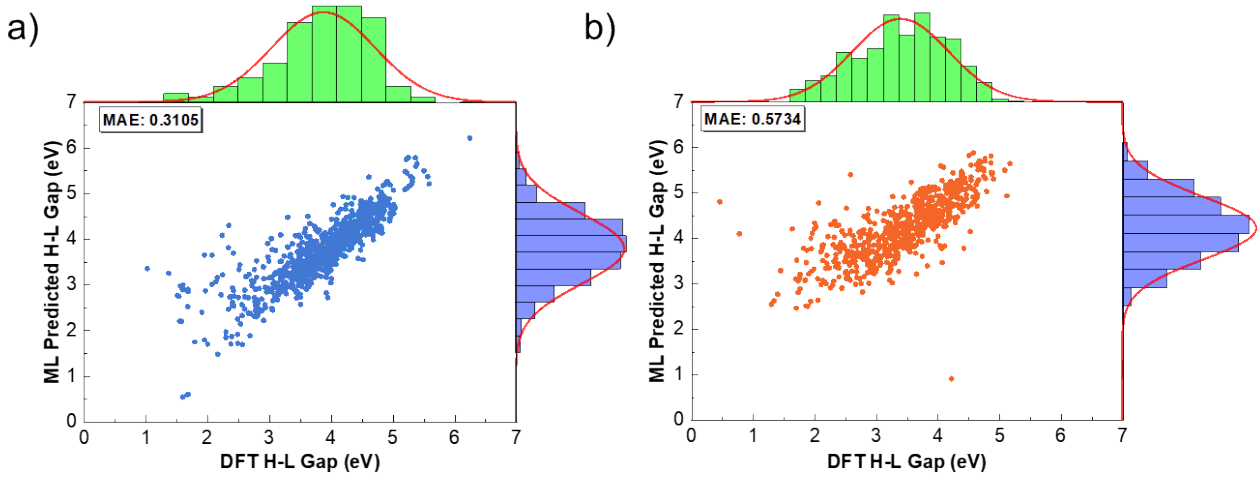


Figure S1. Prediction of H-L gap by initial models (a) MAE and distribution of H-L gap predictions on a fixed test set by an initial model against calculation by DFT for DA form PSs and (b) DAD form PSs.

t-Distributed Stochastic Neighbor Embedding from Neural Fingerprints

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised machine learning algorithm. It is often used for clustering and visualization of high-dimensional data such as the molecular fingerprints in this study. The algorithm starts by calculating the conditional probability of similarity between high-dimensional data points (i.e., the high-dimensional molecular fingerprints) and also between their low-dimensional counterparts (i.e., two-dimensional vectors that can be visualized) by the Euclidean distances of data points. A cost function, which is defined as a single Kullback-Leibler (KL) divergence between joint probability distributions in the high-dimensional space and the low-dimensional space is then minimized. By minimizing the cost function, t-SNE can ensure the points that are similar in high-dimensional space are close to each other in the low-dimensional space. KL divergence measures the distance between two random distributions. When two random distributions are the same, their KL divergence is equal to zero. When the difference between two random distributions increases, their KL divergence also increases.

To visualize the molecular space of the DA and DAD datasets through active learning progression, the neural fingerprint of every structure was predicted by the final DA model from the last round of active learning. The predicted neural fingerprints were fitted in a t-SNE model with 2 components, perplexity of 50, a learning rate of 200 and optimized for 1000 iterations to reduce KL divergence. The final 2-dimensional embedded features were plotted for structures in the unlabeled space, initial training set, predictions by each active learning cycle, and the final recommended 4 structures. In this work, neural fingerprints were predicted with help of the DeepChem package (https://github.com/deepchem/deepchem), and t-SNE model was done with sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html)

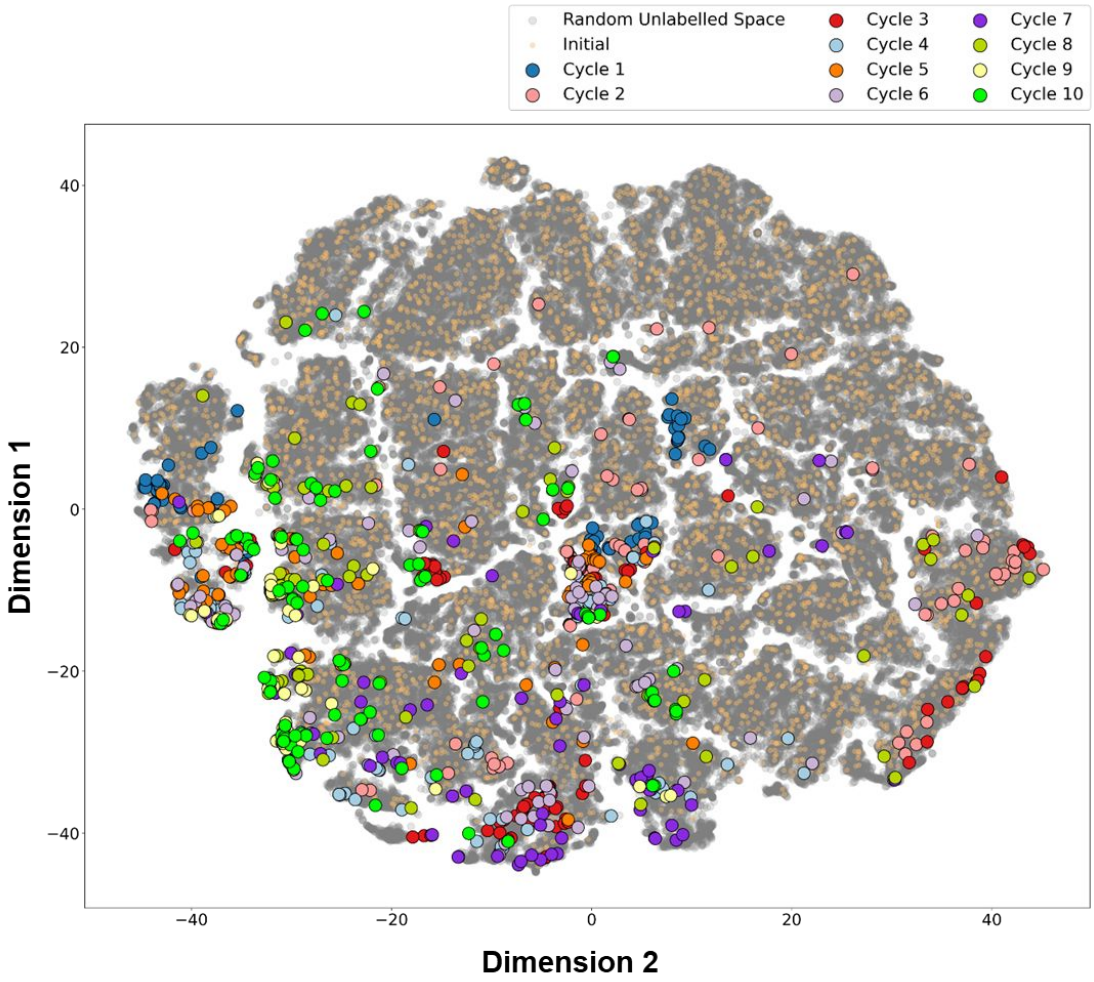


Figure S2. t-SNE for DAD Sample Space and with Active Learning Progression

For the visualizations of DA space (in **Figure 3e**), the final DA model from the last active learning cycle was used to predict the neural fingerprints of the full unlabeled DA molecular space, initial 7101 random DA structures used for initial model training, and structures added in every active learning cycle from cycle 1 to cycle 8. For the visualizations of DAD space (in **Figure S2**), the final DAD model from the last active learning cycle was used to predict the neural fingerprints of a random subset of the initial unlabeled space (> 110000 DAD), initial 4914 random DAD structures used for initial model training, and structures added in every active learning cycle from cycle 1 to cycle 10. From the visualizations of DAD space, a similar conclusion of the effectiveness of the active learning strategies can be derived. For the visualizations of the combined DA and DAD space along with the 4 recommended structures (in **Figure 4i**), the final DA model from the last active learning cycle was used to predict the neural fingerprints of the initial 7101 DA and 4914 DAD structures, and the final 4 selected DA and DAD structures.

Comparing Molecular Fingerprint Methods

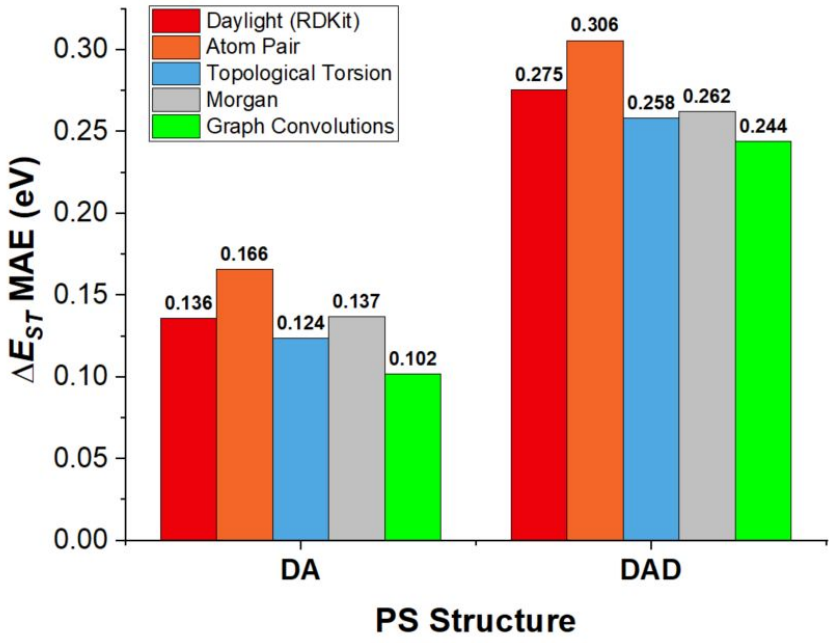


Figure S3. Comparison of average model performances (MAE) between graph-based deep learning method and traditional molecular fingerprints methods. For all methods, the respective model is trained on the initial training set and evaluated on the fixed test set and an average of 5 performances was obtained. As shown, for both DA and DAD form PSs, the graph-based deep learning method has shown better performance than traditional molecular fingerprint methods. This result is aligned with signature references as graphs are the most suitable representations of molecules and self-learned features are usually more efficient.

Daylight

Daylight fingerprint captures the patterns of molecular features such as atoms, the nearest neighbours of atoms, and so on. Then the information will be hashed into bit strings and all bit strings will be linearly combined to form a final binary fingerprint.[1]

Atom Pair

The atom pair fingerprint is defined in terms of the atomic environments of, and shortest path separations between, all pairs of atoms in the topological representation of a chemical structure.[2]

Topological Torsion

Topological torsion consists of four consecutively bonded non-hydrogen atoms along with the number of non-hydrogen branches. It is essentially a topological analog of the basic conformational element, the torsion angle.[3]

Morgan

Morgan fingerprints are one kind of topological fingerprint for molecular characterization.[4-6] It contains substructure information according to the different radius of a molecule and can represent novel structural classes.

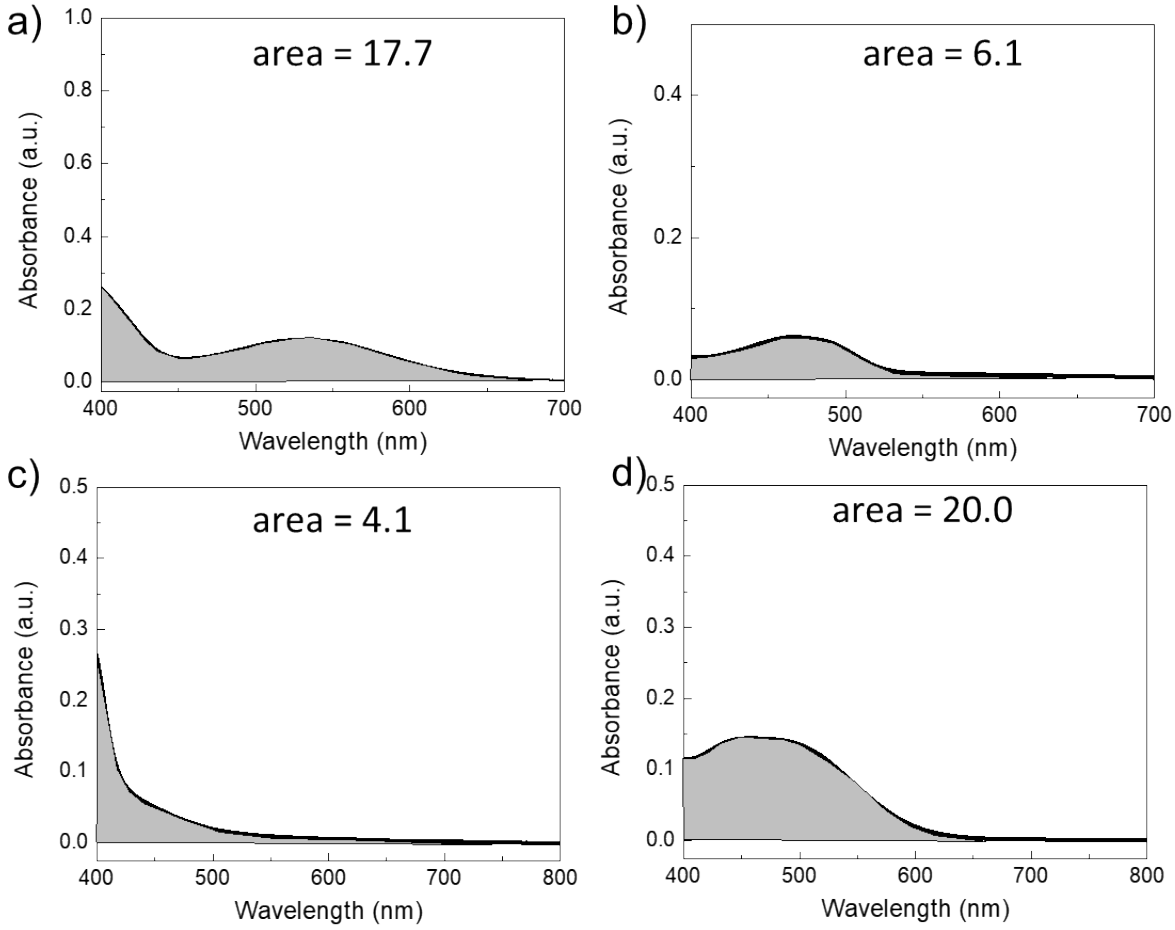


Figure S4. The absorption peak areas of **1-4** (a-d).

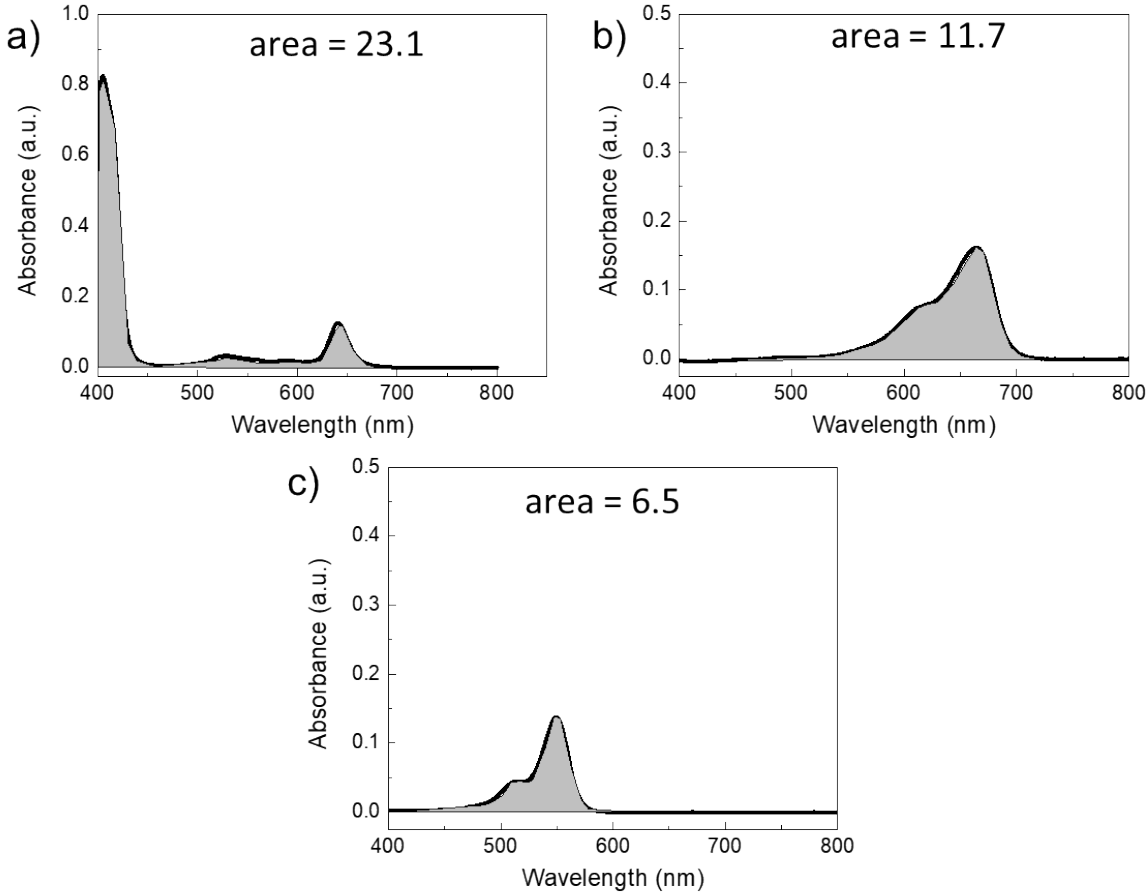


Figure S5. The absorption peak areas of Ce6, MB, and RB (a-c).

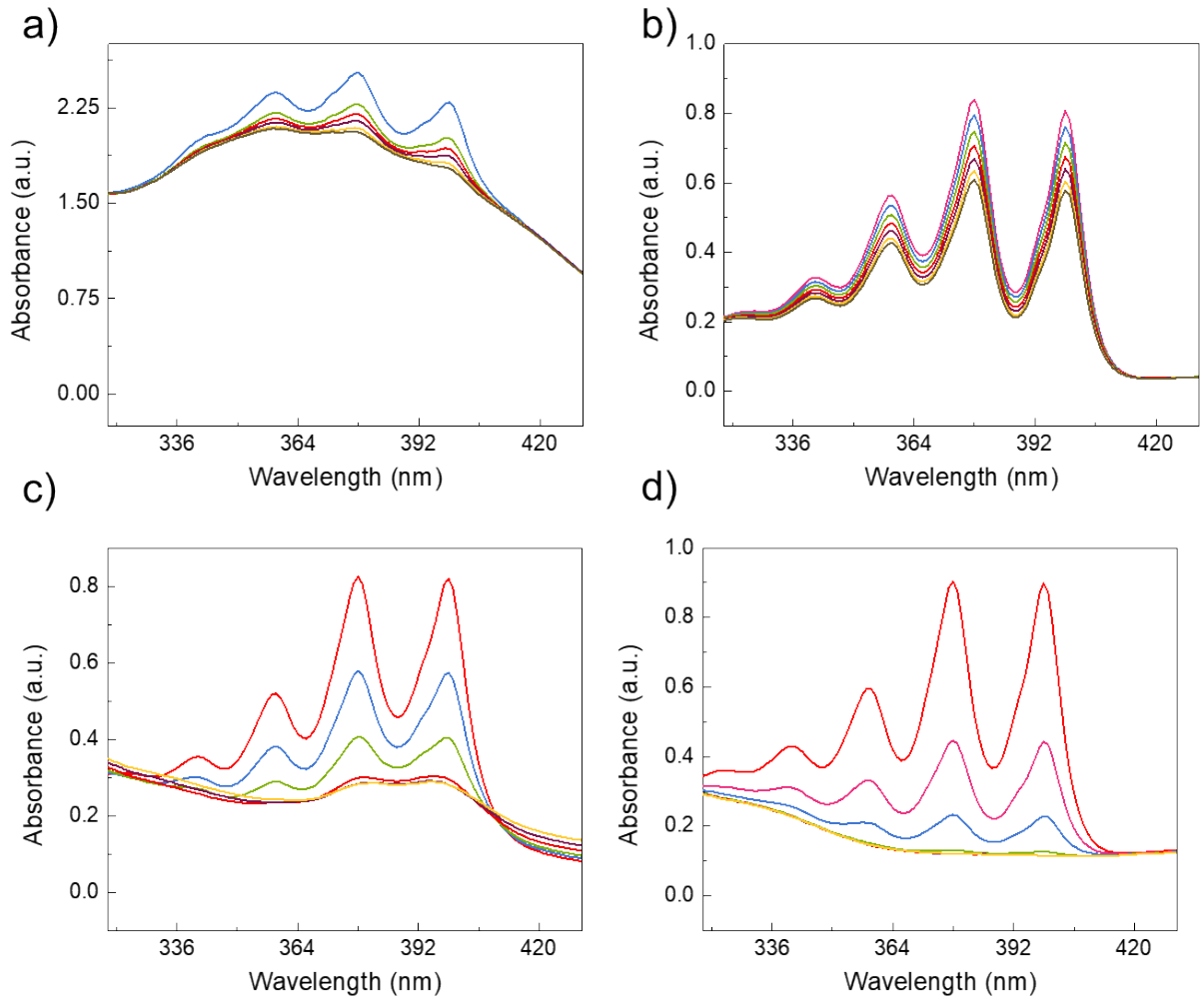


Figure S6. Photo-degradation of ABDA with **1-4** (a-d) in DMSO/water (v/v = 1/99) in five minutes upon white light irradiation in five minutes, concentration of PSs: 5×10^{-6} M, power density of light: 50 mW/cm^2.

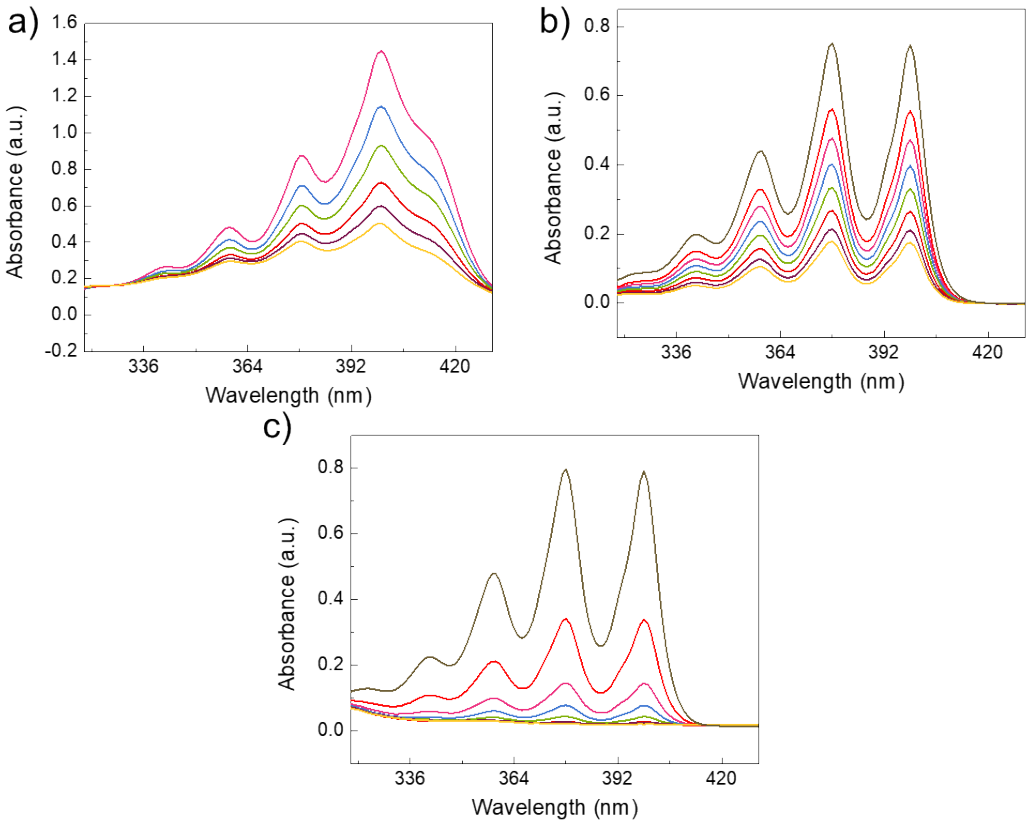
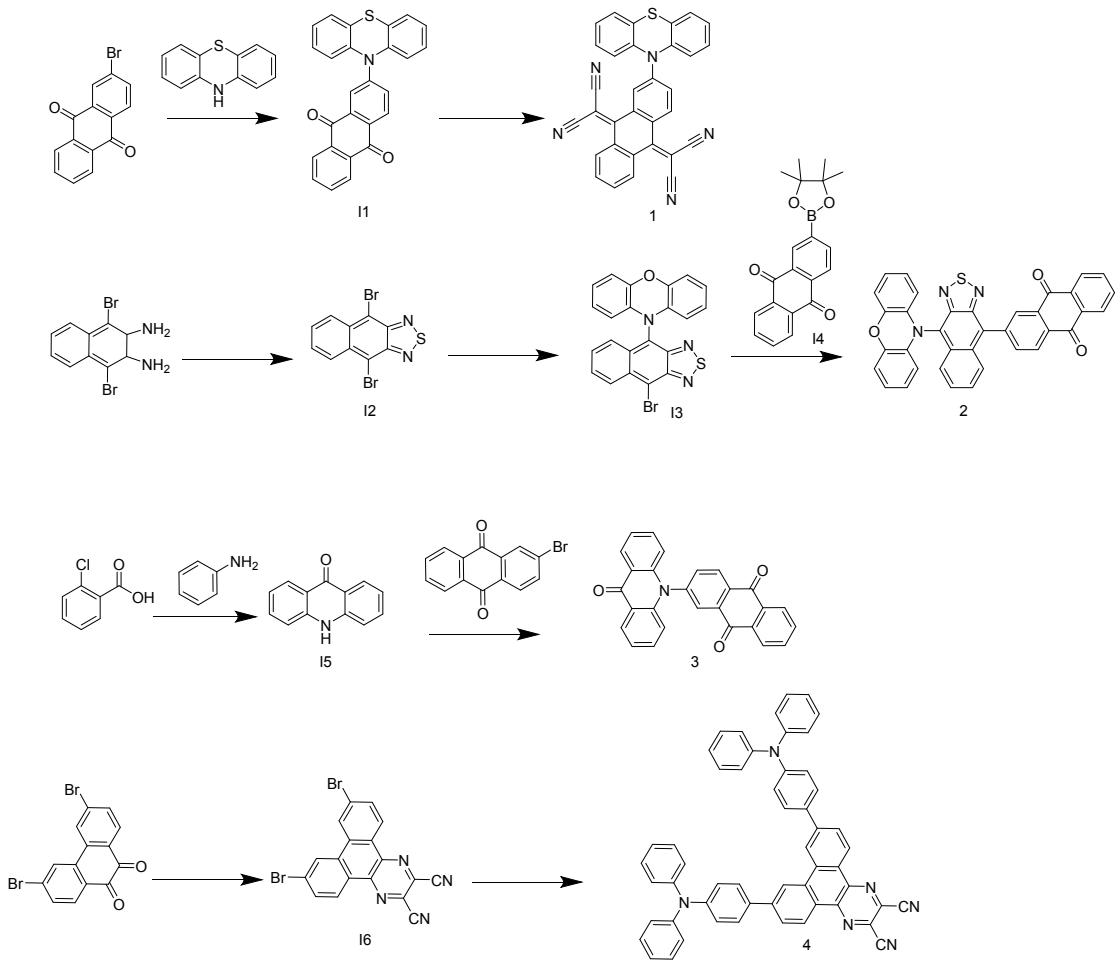
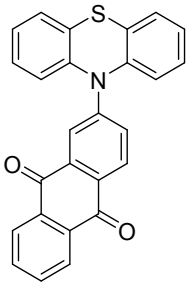


Figure S7. Photo-degradation of ABDA with Ce6, MB, and RB (a-c) in DMSO/water (v/v = 1/99) upon white light irradiation in five minutes, concentration of PSs: 5×10^{-6} M, power density of light: 50 mW/cm^2.

Synthesis of 1-4.



Scheme S1. The synthetic route towards compounds 1-4.



Synthesis of I1. A 100 mL round-bottom flask equipped with a magnetic stir bar was charged with 2-bromoanthracene-9,10-dione (375 mg, 1.29 mmol), phenothiazine (283 mg, 1.42 mmol), cesium carbonate (535 mg, 3.87 mmol) and toluene (15 mL). The solution was stirred at room temperature. After 10 min, a solution of palladium(II) acetate (8.67 mg, 0.04 mmol) and tri-tert-butyl phosphine (29 mg, 0.14 mmol) in toluene (5 mL) was added dropwise over 5 min. The reaction mixture was stirred and heated to 120 °C under reflux for 24 h. After cooling to room temperature, the resulting mixture was treated with water (40 mL) and extracted with chloroform (20 mL × 3). The organic phase was separated, washed twice with brine, dried over anhydrous $MgSO_4$. Then the solution was concentrated under reduced pressure, and the residue was purified by column chromatography on silica gel (hexane/chloroform = 10/1) to afford **I1** (355 mg, 70% yield) as a light-yellow solid. ^{1}H NMR (400

MHz, $CDCl_3$) δ 8.29 (d, *J* = 7.5 Hz, 1H), 8.23 (d, *J* = 7.2 Hz, 1H), 8.16 (d, *J* = 8.8 Hz, 1H), 7.90 (d, *J* = 2.7 Hz, 1H), 7.82 – 7.66 (m, 2H), 7.54 – 7.46 (m, 4H), 7.45 – 7.35 (m, 3H), 7.28 (t, *J* = 8.3 Hz, 2H).

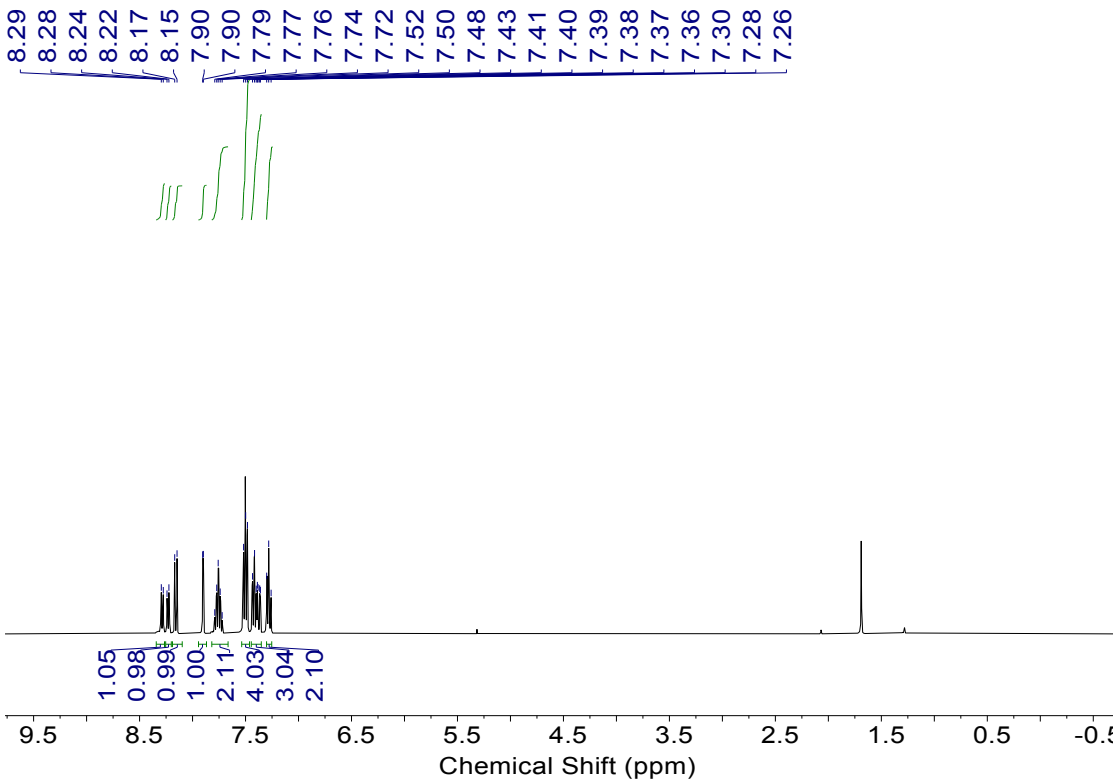
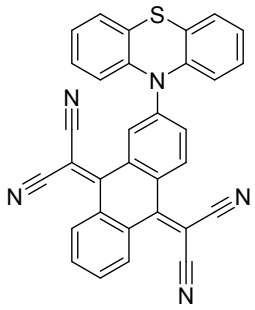


Figure S8. The ^{1}H NMR spectrum of I**1** in $CDCl_3$.



Synthesis of 1. To the solution of compound **I1** (40.5 mg, 0.10 mmol) and malononitrile (39.6 mg, 0.60 mmol) in dichloromethane (10 mL) was added titanium tetrachloride (0.08 mL, 0.7 mmol) slowly at 0 °C. After the reaction mixture was stirred for 30 min, pyridine (0.06 mL, 0.7 mmol) was injected and stirred for another 30 min. Then the mixture was heated at 45 °C for 48 h. After the mixture was cooled down to room temperature, the reaction was quenched by water (30 mL) and the mixture was extracted with dichloromethane. The collected organic layer was washed by brine, dried over Na_2SO_4 and concentrated under reduced pressure. The desired residue was purified by column chromatography using n-hexane/dichloromethane (1/5, v/v) as eluent to give the desired product **1** as a dark red solid (32 mg, 64% yield).^{1}H NMR (400 MHz, $CDCl_3$) δ 8.14 (d, *J* = 7.3 Hz, 1H), 8.02 (d, *J* = 7.3 Hz, 1H), 7.94 (d, *J* = 9.0 Hz, 1H), 7.63 – 7.54 (m, 5H), 7.48 – 7.40 (m, 4H), 7.26 (t, *J* = 7.0 Hz, 2H), 7.16 (dd, *J* = 9.0, 2.6 Hz, 1H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 161.15, 159.48, 148.49, 139.40, 135.23, 132.35,

132.11, 131.81, 130.87, 130.01, 129.59, 129.32, 128.07, 127.94, 127.41, 127.27, 119.76, 115.45, 114.15, 113.92, 113.18, 111.34, 83.10, 79.03. APCI-MS, m/z: calcd 501.1048, found 501.1057.

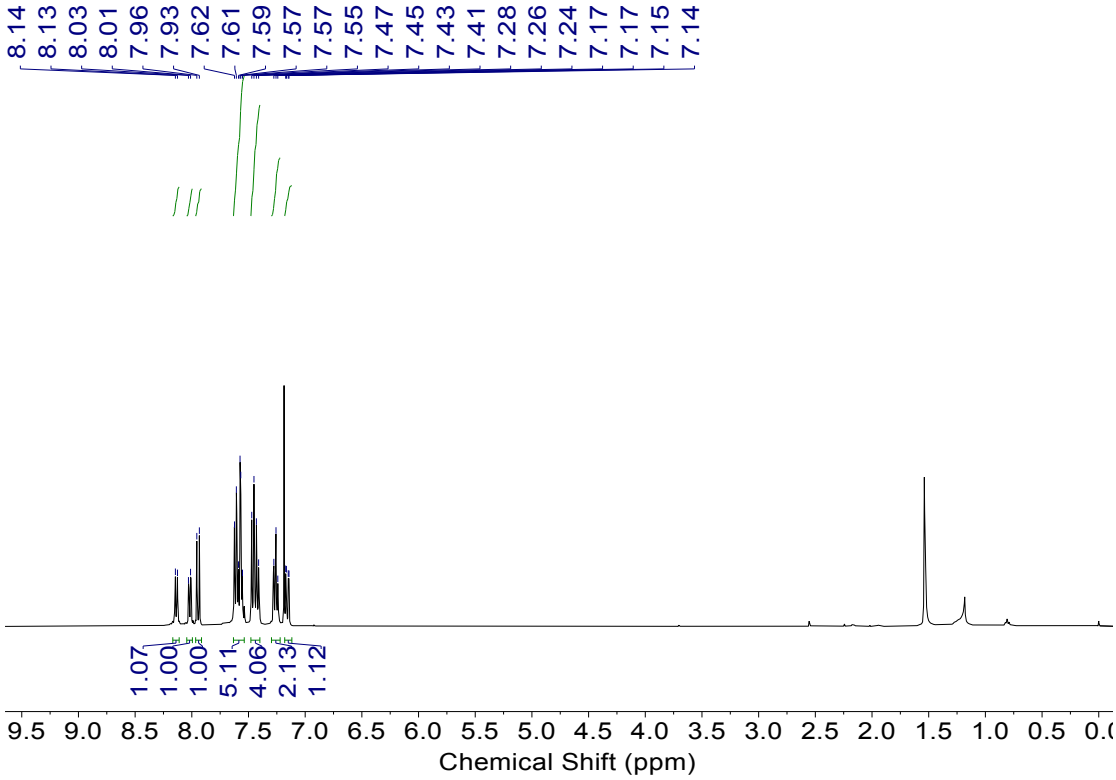


Figure S9. The ^{1}H NMR spectrum of **1** in $CDCl_3$.

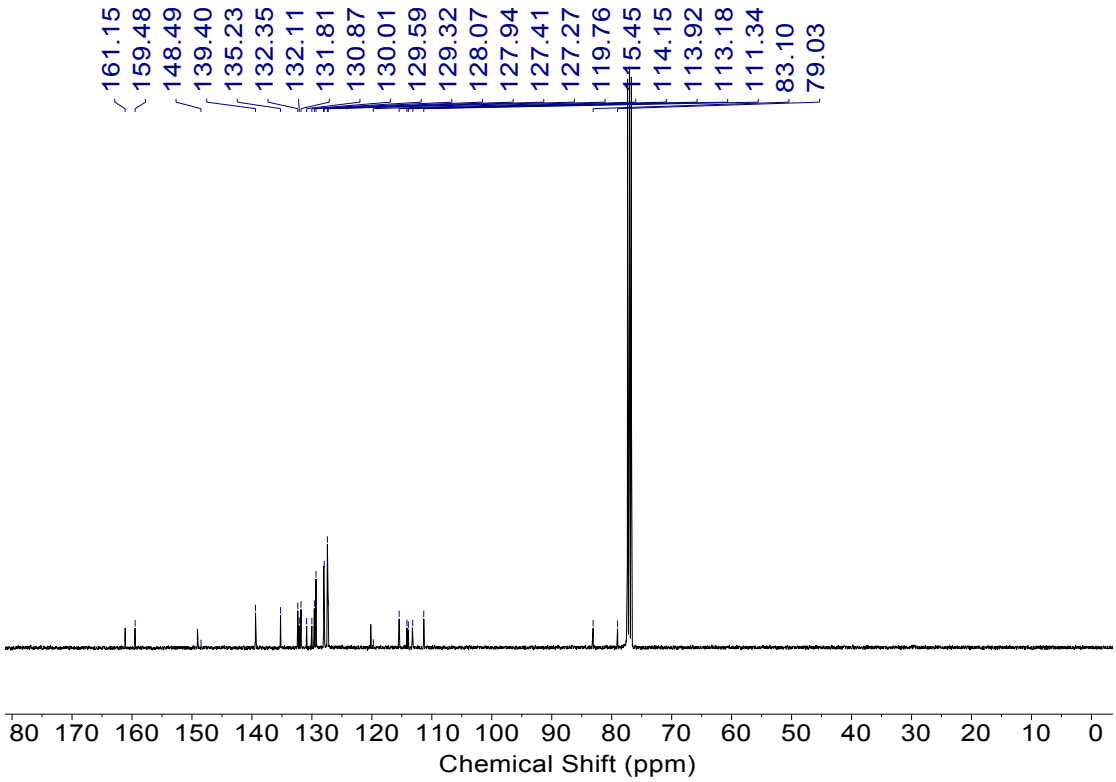
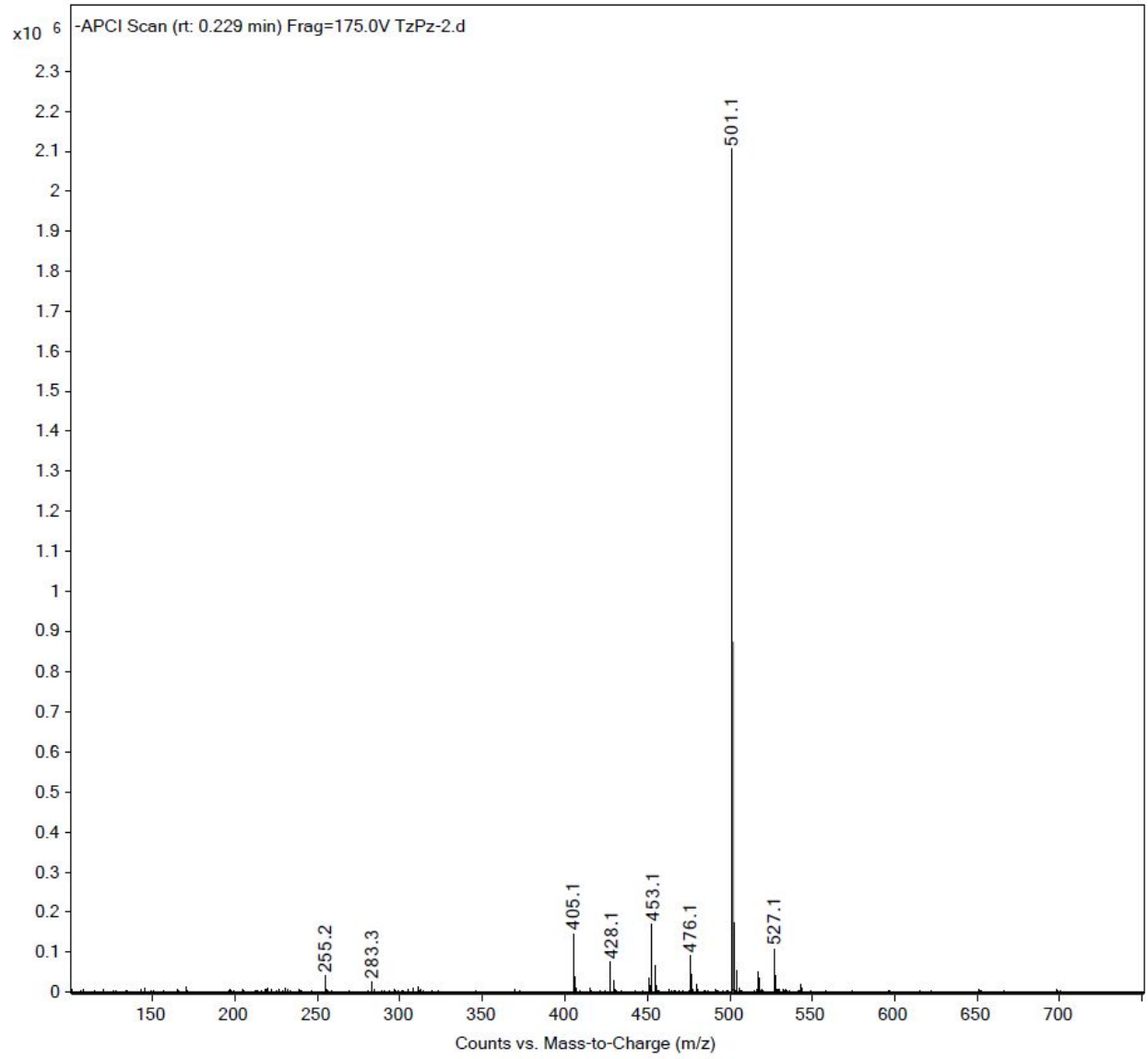


Figure S10. The ^{13}C NMR spectrum of **1** in $CDCl_3$.



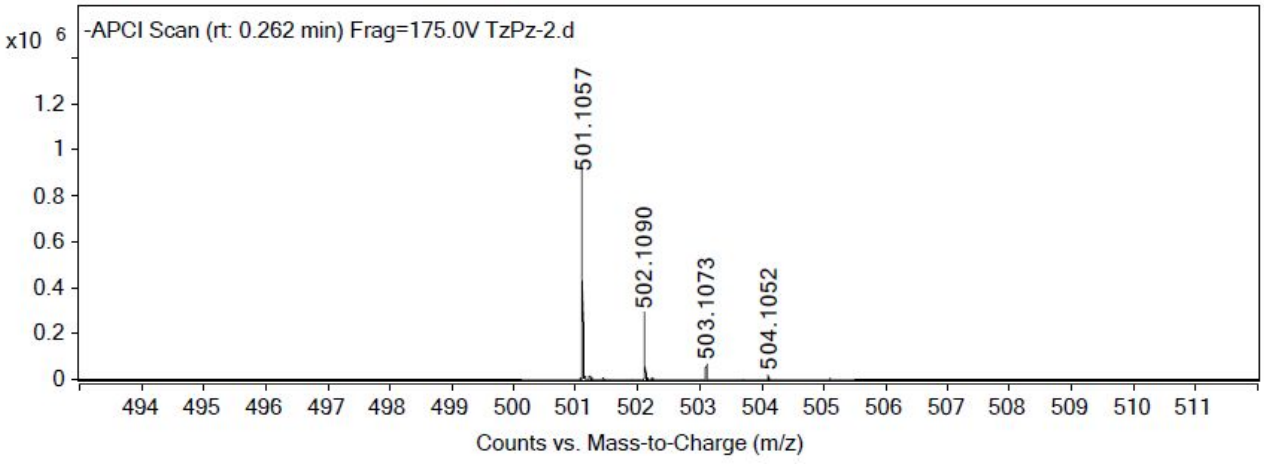
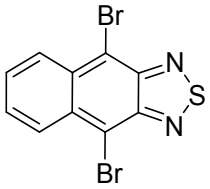


Figure S11. The HRMS spectrum of **1**.



Synthesis of I2. In a 100 mL flask,1,4-dibromo-2,3-diaminonaphthalene (1.8 g, 5.5 mmol) was dissolved in 23 mL of anhydrous pyridine and then 1.31 mL (16.1 mmol) of thionylaniline and 7.0 mL (55 mmol) of chlorotrimethylsilane were added. The reaction was heated at 80 °C overnight with stirring. After the reaction was cooled to room temperature, 20 mL of ethanol was added to the mixture. The precipitate was filtered, washed with ethanol, and then recrystallized from a mixture of ethanol and chloroform to give **I2** (1.5 g, 70% yield) as orange needles. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.43 (dd, *J* = 7.0, 3.2 Hz, 2H), 7.61 (dd, *J* = 7.0, 3.2 Hz, 2H).

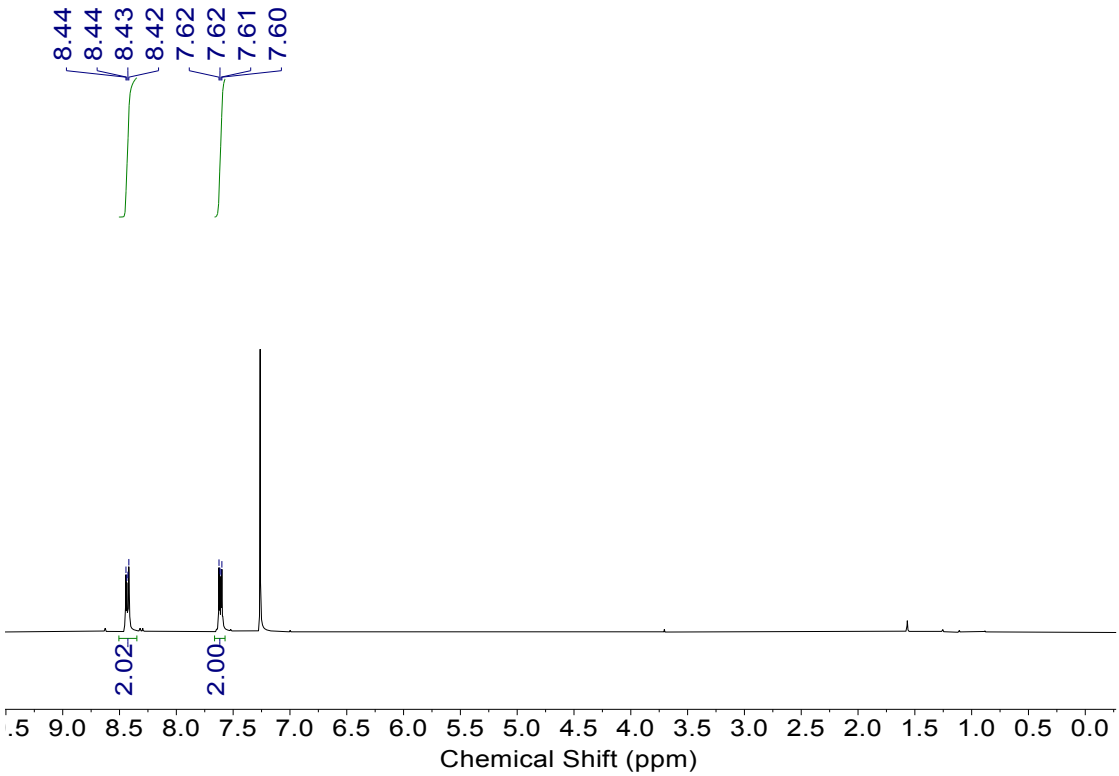
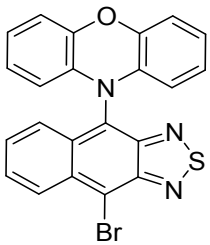


Figure S12. The ^{1}H NMR spectrum of **I2** in $CDCl_3$.



Synthesis of I3. A mixture of phenoxazine (0.5 g, 2.6 mmol), **I2** (400 mg, 2.4 mmol), palladium acetate (90 mg, 0.4 mmol), [(*t*-Bu)$_3$P]HBF$_4$ (348 mg, 1.2 mmol), and sodium *tert*-butoxide (0.92 g, 9.6 mmol) in 25 mL anhydrous toluene was stirred and reflux at 110 °C under argon atmosphere for 72 h. After cooling down to room temperature, the reaction mixture was poured into saturated brine and extracted with dichloromethane. Then, the organic phase was dried over anhydrous Na_2SO_4. After

solvent removal, the crude product was purified by column chromatography (silica, hexane/dichloromethane (v/v) = 1:10) to give I3 (273 mg, 45% yield) as a purple powder. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.38 – 8.23 (m, 1H), 8.05 – 7.88 (m, 1H), 7.32 – 7.16 (m, 2H), 6.60 (d, *J* = 9.4 Hz, 2H), 6.45 (t, *J* = 7.6 Hz, 2H), 6.32 – 6.17 (m, 2H), 5.28 (d, *J* = 8.1 Hz, 2H).

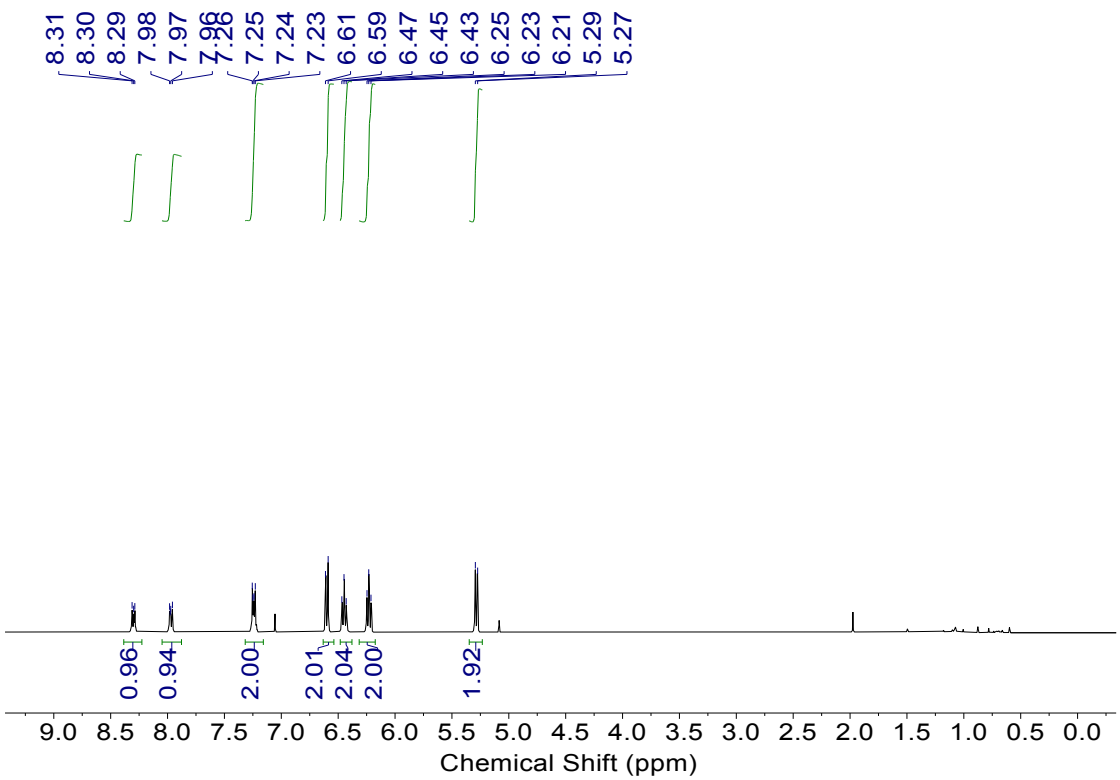
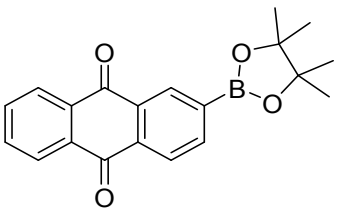


Figure S13. The ^{1}H NMR spectrum of **I3** in $CDCl_3$.



Synthesis of I4. 2-Bromoanthracene-9,10-dione (574 mg, 2.0 mmol), bis(pinacolato)diborane (1.02 g, 4.0 mmol), potassium acetate (687 mg, 7.0 mmol), Pd(dppf)Cl_2 (73 mg, 0.23 mmol, dppf = 1,1'-bis (diphenylphosphanyl)ferrocene) and dioxane (20 mL) were mixed together in a 250 mL flask. After degassing, the reaction mixture was kept at 100 °C for 2 days, and then cooled down to room temperature. The organic solvent was distilled out, and the residual solid was dissolved in dichloromethane and washed with water. After solvent removal, the crude product was purified on a silica gel column using n-hexane/ethyl acetate (20:1, v/v) as the eluent to afford compound **I4** (504 mg, 75% yield) as a very viscous liquid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.75 (s, 1H), 8.34 – 8.27 (m, 3H), 8.20 (d, *J* = 9.0 Hz, 1H), 7.82 – 7.76 (m, 2H), 1.26 (s, 12H).

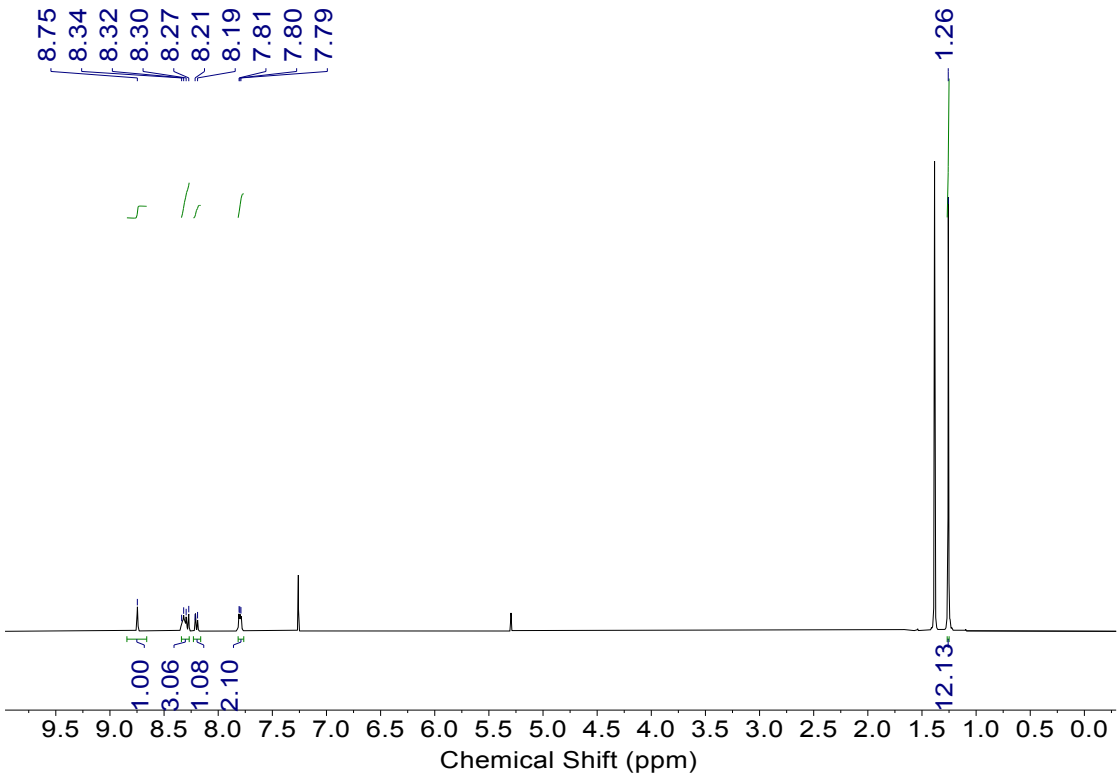
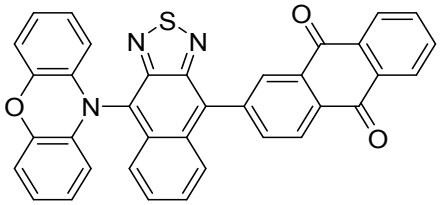


Figure S14. The ^{1}H NMR spectrum of **I4** in $CDCl_3$.



Synthesis of 2. Compound **I4** (270 mg, 0.60 mmol), compound **I3** (130 mg, 0.20 mmol), potassium carbonate (552 mg, 4.0 mmol), THF (12 mL)/water (4 mL), and $Pd(PPh_3)_4$ (15 %) were carefully degassed and charged with nitrogen. The reaction mixture was then stirred at 60 °C for 12 h. After cooling down the reaction mixture to ambient temperature, it was extracted with DCM and washed with water. The DCM layer was separated and dried over $MgSO_4$. After evaporation of the solvent, the crude product was purified by column chromatography on silica gel by using n-hexane/dichloromethane (1/2 ~ 0/1, v/v) as the eluent to afford a dark blue solid **2** (112 mg, 52% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.61 (s, 1H), 8.56 (d, *J* = 7.9 Hz, 1H), 8.36 – 8.28 (m, 3H), 8.12 (d, *J* = 9.9 Hz, 1H), 7.97 (d, *J* = 9.0 Hz, 1H), 7.82 – 7.78 (m, 2H), 7.48 – 7.40 (m, 2H), 6.78 (d, *J* = 7.8 Hz, 2H), 6.64 (t, *J* = 7.6 Hz, 2H), 6.42 (t, *J* = 7.6 Hz, 2H), 5.52 (d, *J* = 8.1 Hz, 2H). APCI-MS, m/z: calcd 573.1147, found 573.1157.

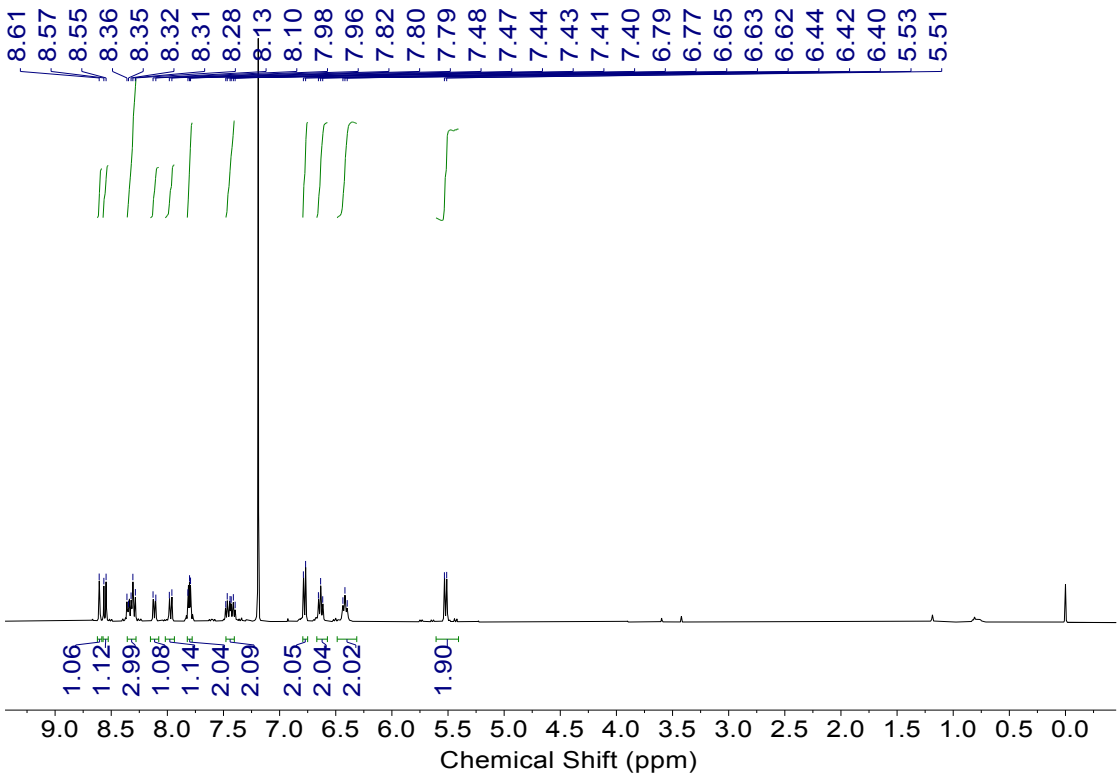


Figure S15. The ^{1}H NMR spectrum of **2** in $CDCl_3$.

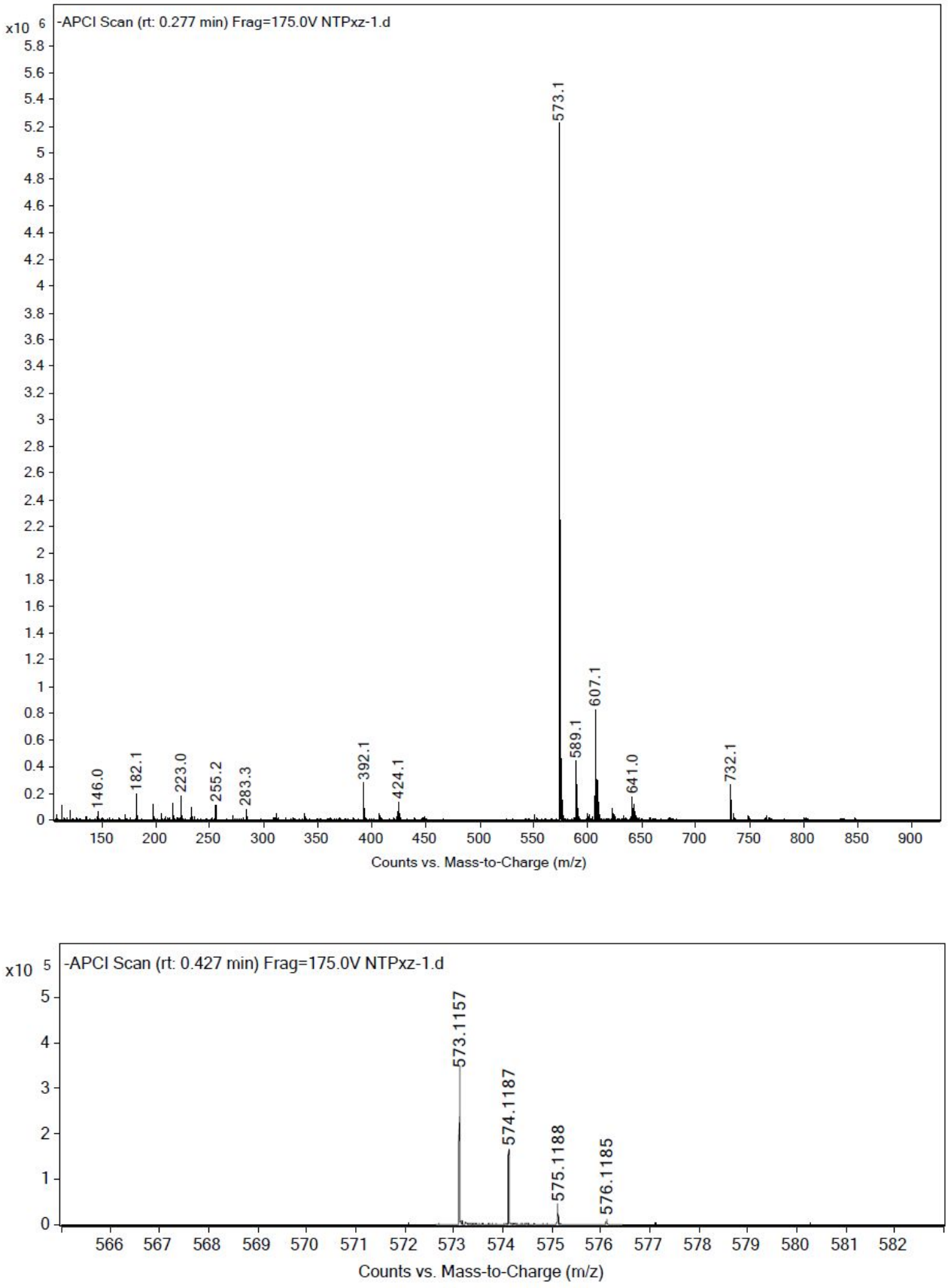
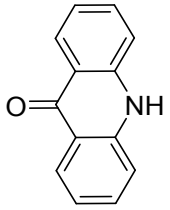
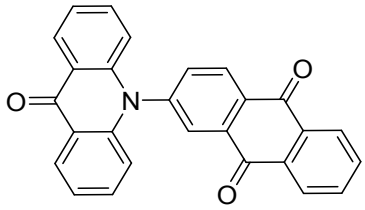


Figure S16. The HRMS spectrum of **2**.



Synthesis of I5. *N*-phenylanthranilic acid (6.00 g, 27.8 mmol) was suspended in polyphosphoric acid (60 g) and heated to 120 °C in a round-bottom flask. which was equipped with a strong magnetic stirring bar. The dark green viscous mixture was also occasionally mixed thoroughly with a glass rod. After about 3.5 h, the *N*-phenylanthranilic acid was completely dissolved and the reaction mixture was held at this temperature for additional 0.5 h and then carefully poured into a beaker of ice/water (100 mL). The greenish yellow suspension was brought to pH 7 by slow addition of NaOH solution. The solid material was filtered off by suction filtration and washed with hot water (3×100 mL). The greenish yellow solid was dried in air overnight at 120 °C to give crude **I5** (5.4 g, 95% yield), which was further used without purification.



Synthesis of 3. A mixture of **I5** (256 mg, 2.6 mmol), 2-bromoanthracene-9,10-dione (700 mg, 2.4 mmol), palladium acetate (90 mg, 0.4 mmol), [(*t*-Bu)$_3$P]HBF$_4$ (348 mg, 1.2 mmol), and sodium *tert*-butoxide (0.92 g, 9.6 mmol) in 25 mL anhydrous toluene was stirred and reflux at 110 °C under argon atmosphere for 72 h. After cooling down to room temperature, the reaction mixture was poured into saturated brine and extracted with dichloromethane. Then, the organic phase was dried over anhydrous Na_2SO_4. After solvent removal, the crude product was purified by column chromatography (silica, hexane/dichloromethane (v/v) = 5:1) to give **I5** (480 mg, 50% yield) as a purple power. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.70 (dd, *J* = 29.4, 7.5 Hz, 3H), 8.44 (q, *J* = 7.2, 6.4 Hz, 3H), 7.92 (d, *J* = 13.0 Hz, 3H), 7.56 (d, *J* = 8.2 Hz, 2H), 7.37 (d, *J* = 8.3 Hz, 2H), 6.76 (d, *J* = 8.8 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 182.15, 181.95, 178.06, 144.44, 142.53, 136.75, 136.17, 134.80, 134.10, 133.68, 133.37, 130.72, 129.44, 127.76, 122.78, 116.24. APCI-MS, m/z: calcd 401.1052, found 401.1060.

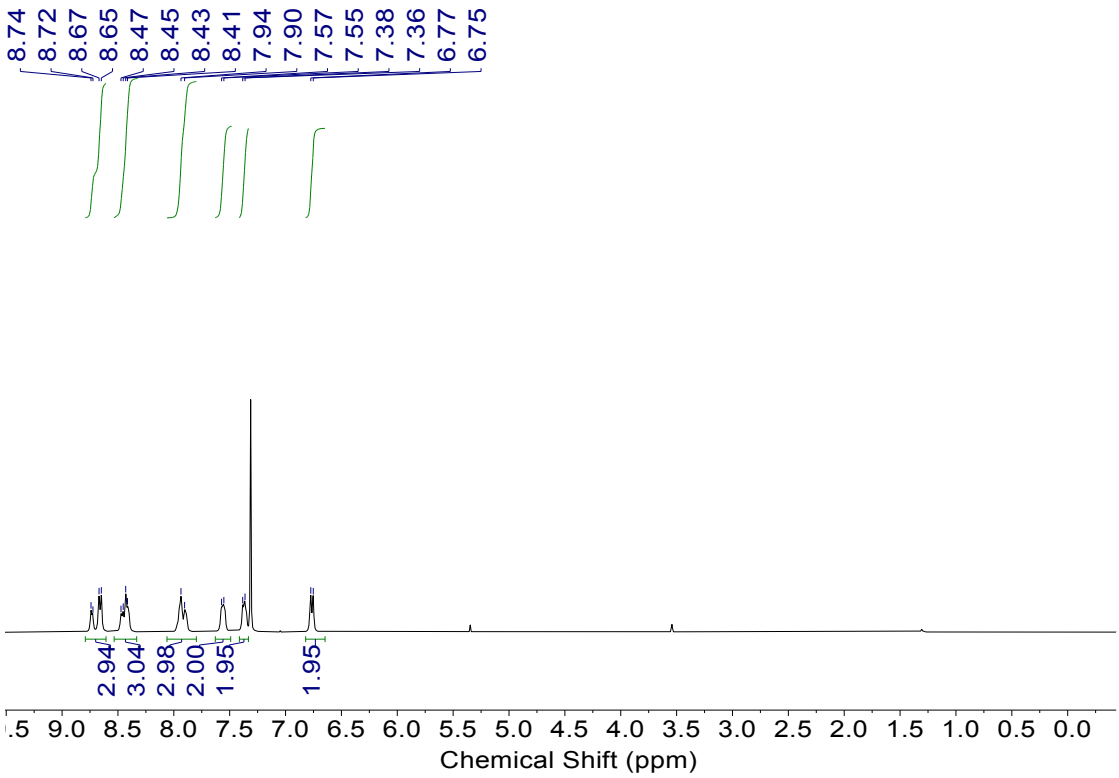
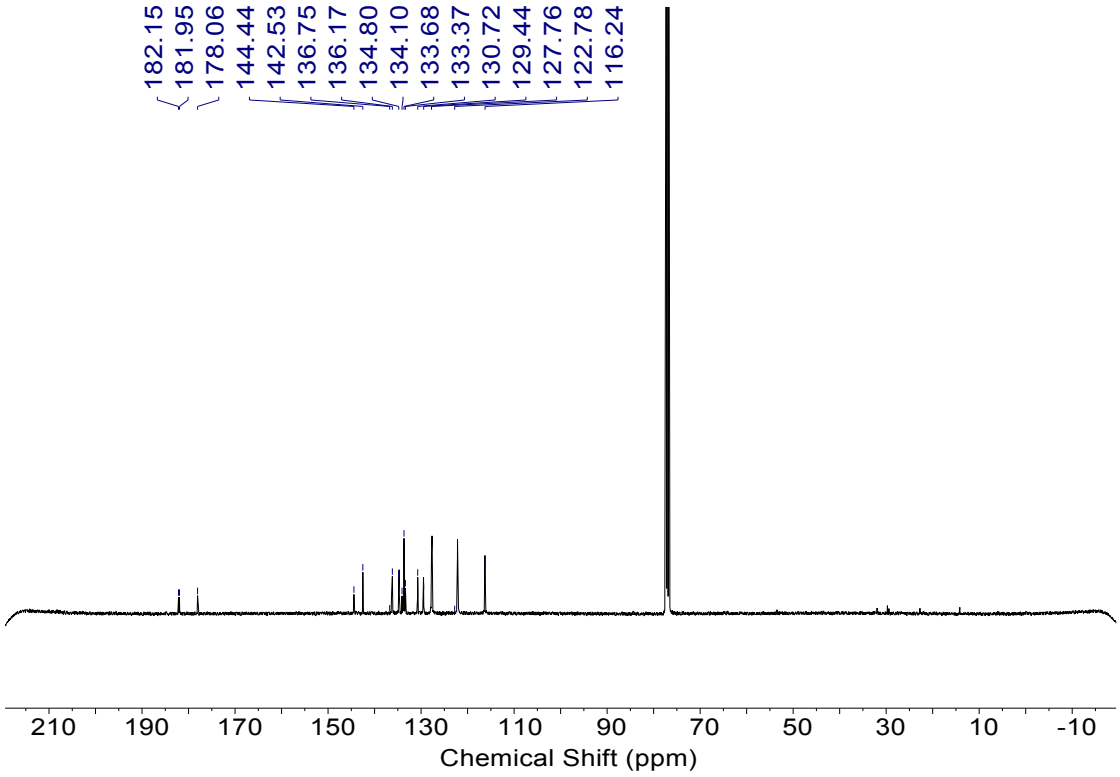
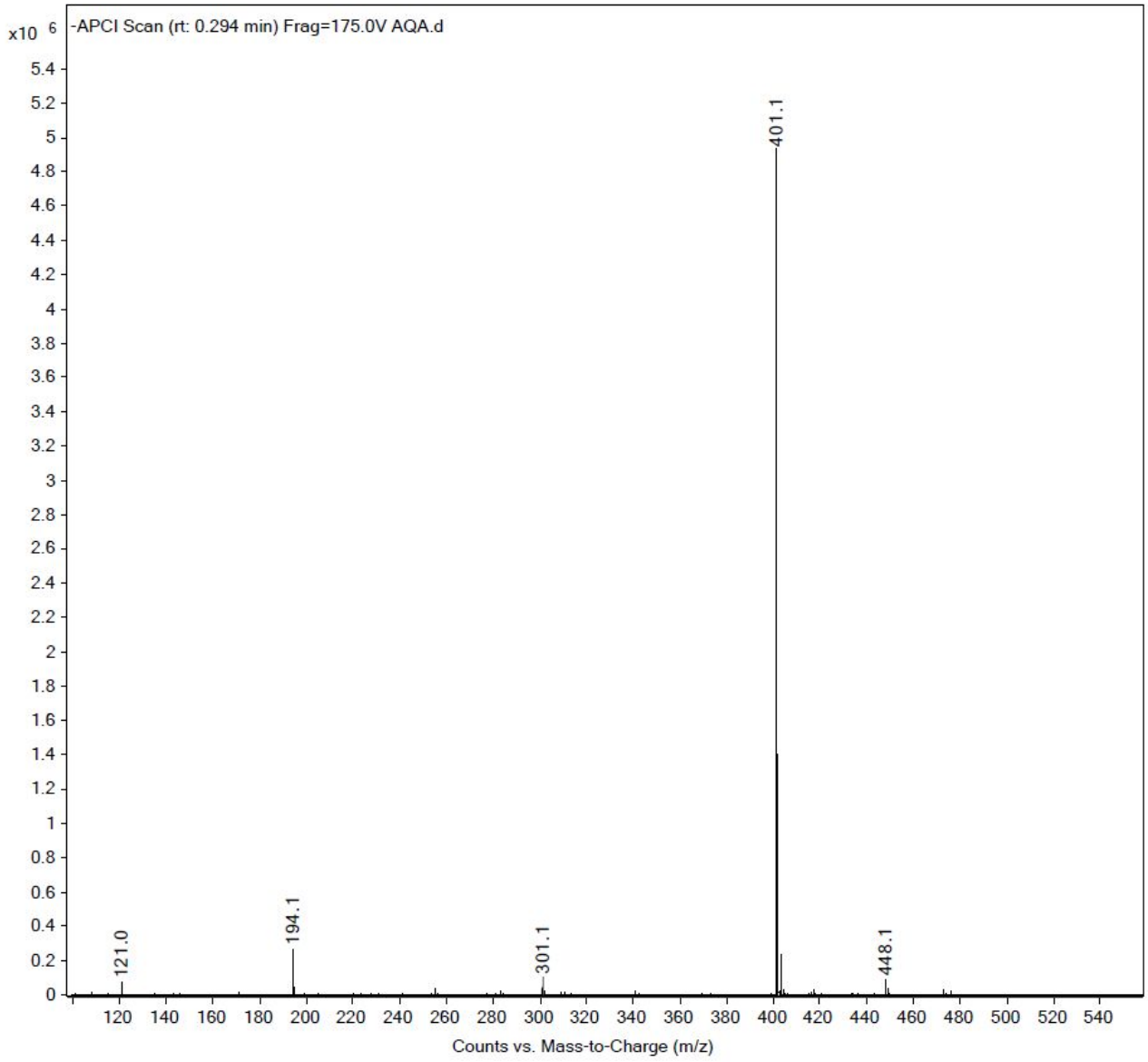


Figure S17. The ^{1}H NMR spectrum of **3** in $CDCl_3$.




-APCI Scan (rt: 0.294 min) Frag=175.0V AQA.d
x10 6
401.1
194.1
121.0
301.1
448.1
Counts vs. Mass-to-Charge (m/z)

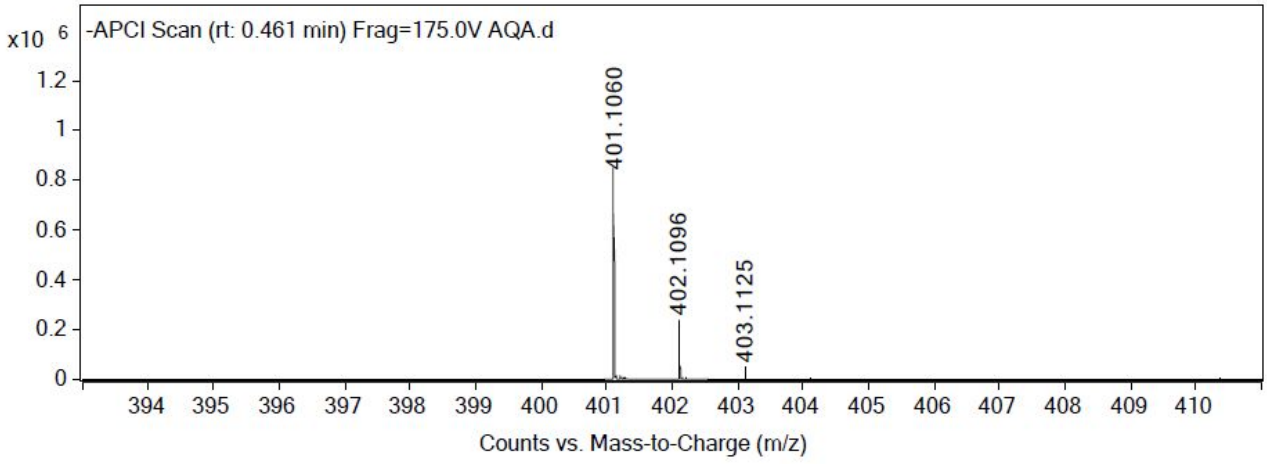
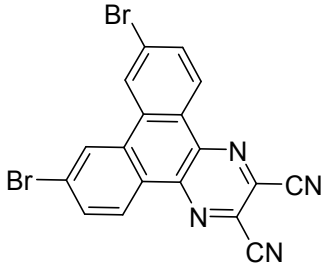
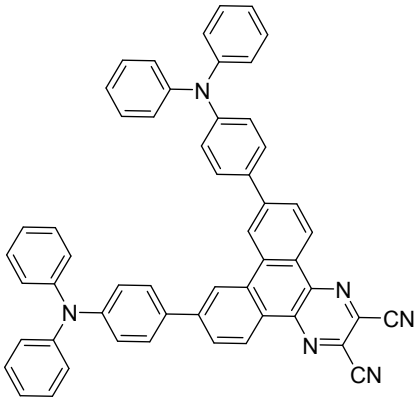

-APCI Scan (rt: 0.461 min) Frag=175.0V AQA.d
x10 6
401.1060
402.1096
403.1125
Counts vs. Mass-to-Charge (m/z)

Figure S18. The HRMS spectrum of **3**.



Synthesis of I6. A suspension of 3,6-dibromophenanthrene-9,10-dione (1.10 g, 3 mmol) and diaminomaleonitrile (0.32 g, 3 mmol) in acetic acid (10 mL) was heated to reflux for 8 hours. After cooling to room temperature, the resulting mixture was poured into ice water (100 mL) and then filtered. The solid was washed with water several times. The crude product was purified by column chromatography on silica gel (eluent: dichloromethane) and dried under vacuum to give **I6** (1.05 g, 80% yield) as a light-yellow solid.



Synthesis of 4. (4-(Diphenylamino)phenyl)boronic acid (180 mg, 0.62 mmol), compound **I6** (87 mg, 0.20 mmol), potassium carbonate (552 mg, 4.0 mmol), THF (12 mL)/water (4 mL), and $Pd(PPh_3)_4$ (15 %) were carefully degassed and charged with nitrogen. The reaction mixture was then stirred at 60 °C for 12 h. After cooling the reaction mixture to ambient temperature, it was extracted with DCM and washed with water. The DCM layer was separated and dried over $MgSO_4$. After evaporation of the solvent, the crude product was purified by column chromatography on silica gel by using n-hexane/dichloromethane (1/5, v/v) as the eluent to afford **4** (38 mg, 52% yield) as a dark blue solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.99 – 8.70 (m, 1H), 8.56 – 8.30 (m, 1H), 7.90 – 7.75 (m, 1H), 7.68 – 7.50 (m, 2H), 7.35 (t, *J* = 7.9 Hz, 4H), 7.23 (d, *J* = 7.8 Hz, 6H), 7.13 (t, *J* = 7.3 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 148.83, 146.61, 144.47, 141.69, 133.03, 132.18, 129.45, 128.18, 127.37, 127.05, 125.42, 125.16, 123.74, 122.90, 119.92, 113.84. APCI-MS, m/z: calcd 766.2845, found 766.2854.

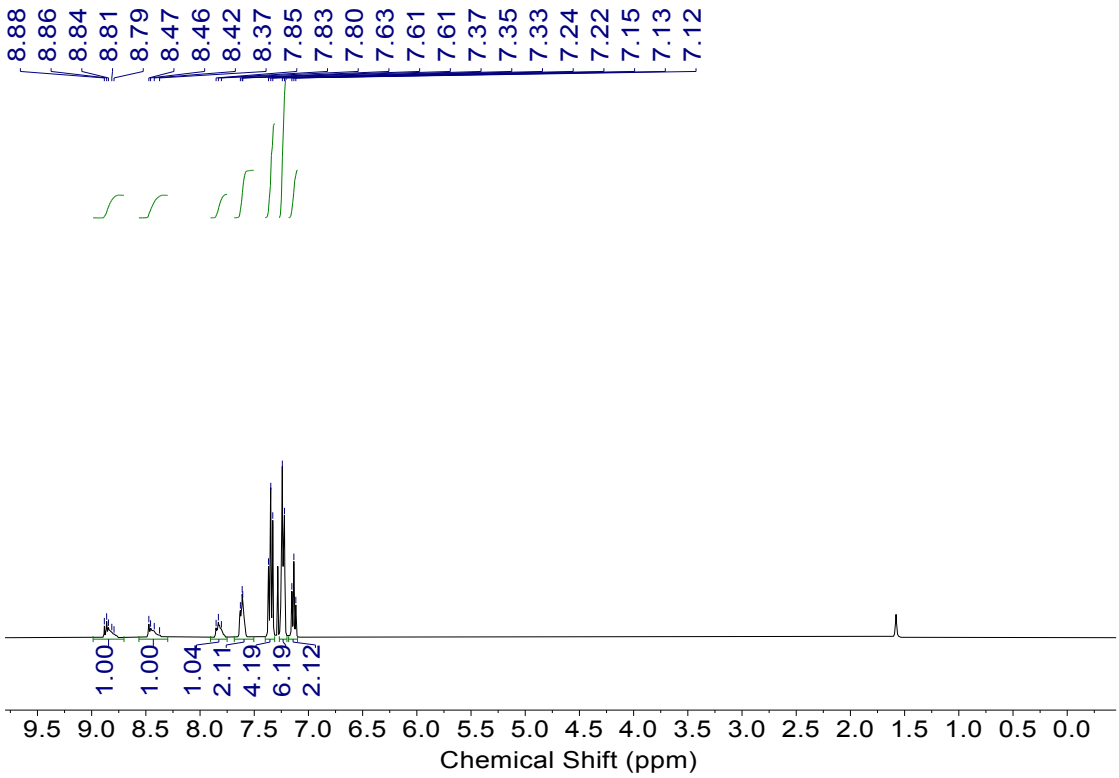


Figure S19. The ^{1}H NMR spectrum of **4** in $CDCl_3$.

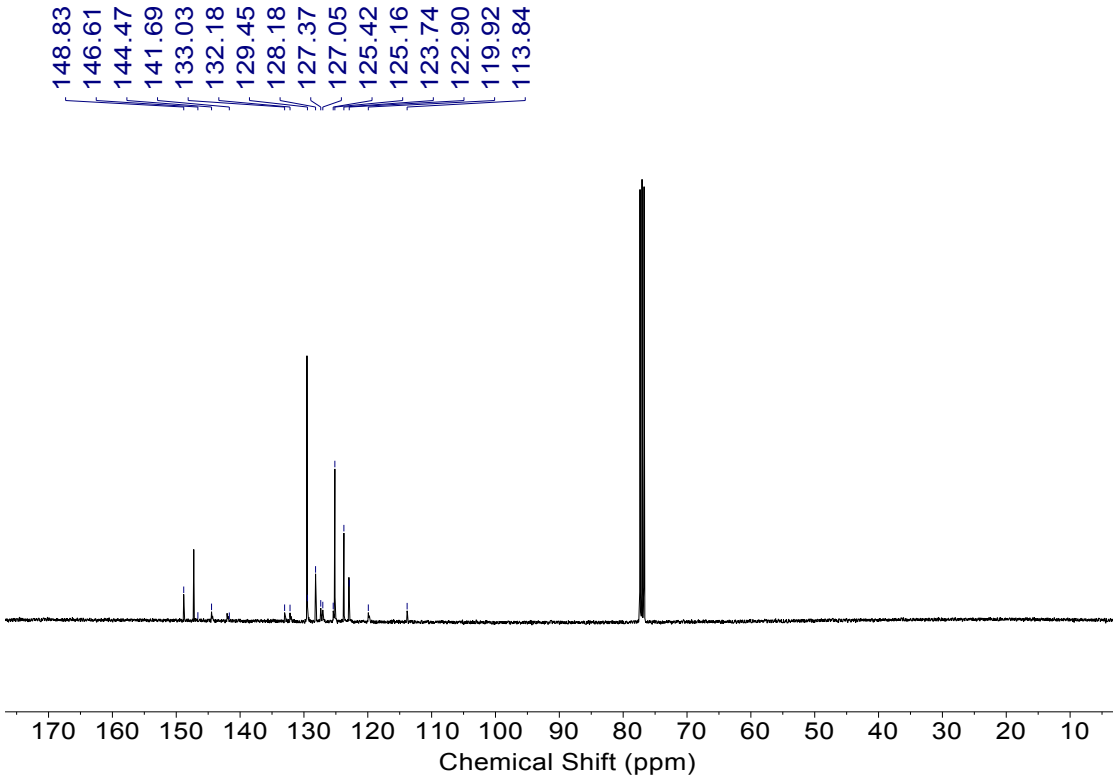
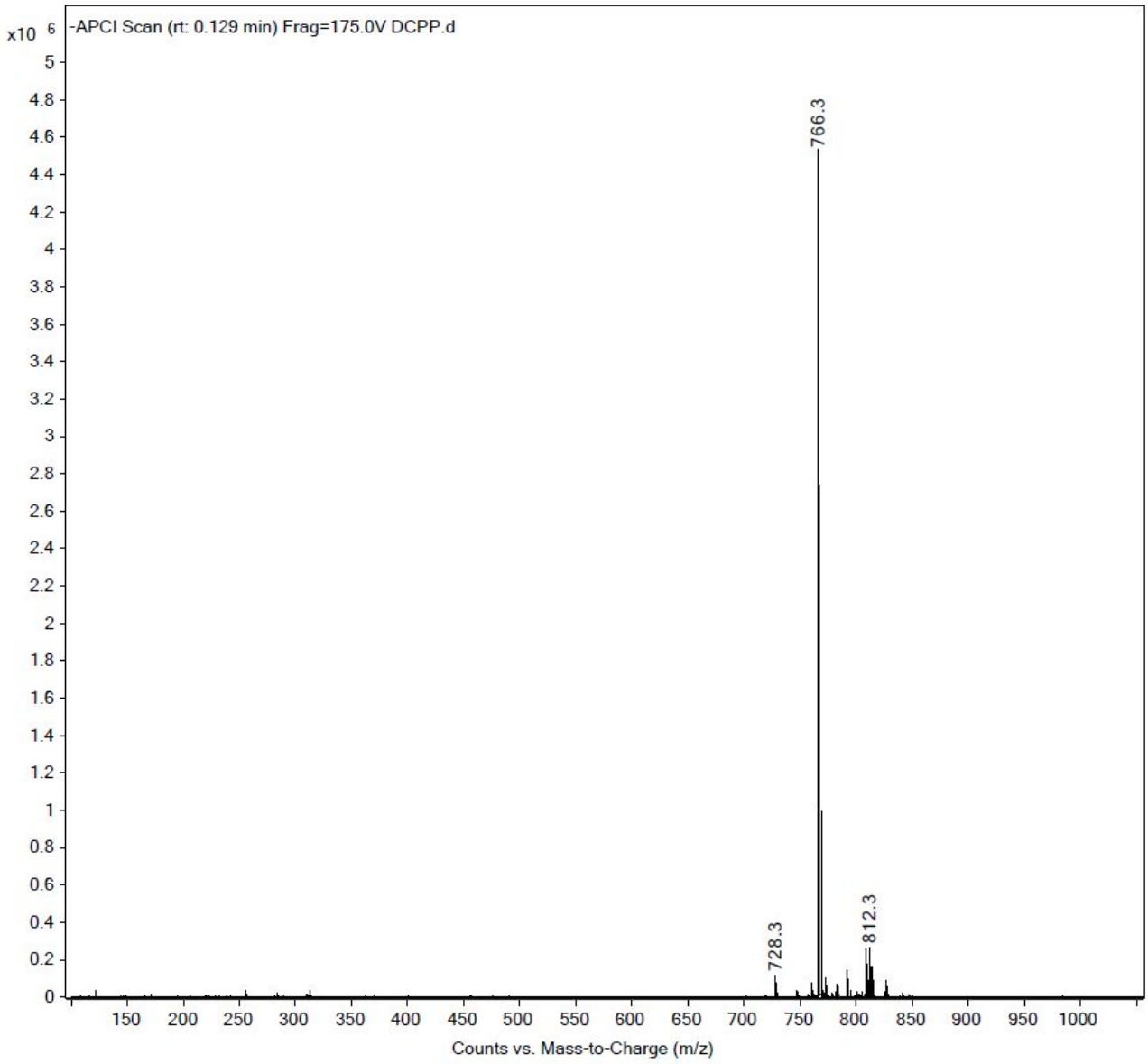


Figure S20. The ^{13}C NMR spectrum of **4** in $CDCl_3$.


-APCI Scan (rt: 0.129 min) Frag=175.0V DCPP.d
x10 6
766.3
728.3
812.3
Counts vs. Mass-to-Charge (m/z)

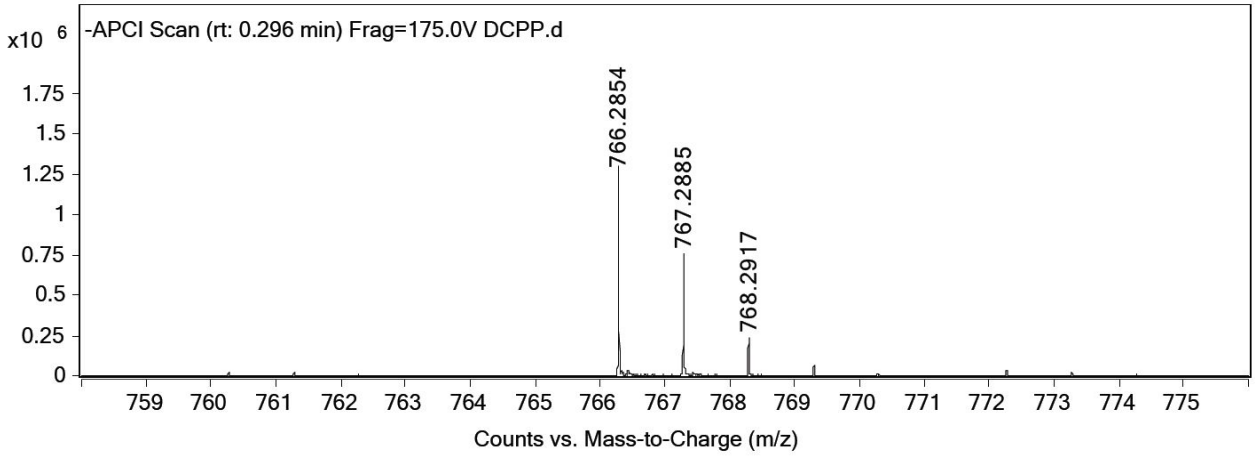

-APCI Scan (rt: 0.296 min) Frag=175.0V DCPP.d
x10 6
766.2854
767.2885
768.2917
Counts vs. Mass-to-Charge (m/z)

Figure S21. The HRMS spectrum of **4**.

REFERENCES

1. Muegge, I. & Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **11**, 137–148 (2016).
2. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73 (1985).
3. Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27**, 82–85 (1987).
4. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
5. Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A., Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule* **1**, 857-870. (2017).
6. Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A., Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25,** 6495-6502 (2015).