# A Self-Improving Photosensitizer Discovery System via Bayesian Optimization and Quantum Chemical Calculation

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

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

‡ These authors contributed equally to this work.

**Abstract** Artificial intelligence (AI) based self-learning or self-improving material discovery system is the holy grail of next-generation material discovery and materials science. Herein, we demonstrate how to combine accurate prediction of material performance via quantum chemical calculations and Bayesian optimization-based active learning to realize a self-improving discovery system for high-performance photosensitizers (PS). Through self-improving cycles, such a system can improve the model prediction accuracy (best mean average error of 0.09 eV for singlet-triplet spitting) and high-performance PS search ability, realizing the efficient discovery of PS. From a molecular space with more than 7 million molecules, 5950 potential high-performance PSs were discovered.

## Introduction

Material discovery is one of the most glorious duties for scientists and in a long history, it highly relies on the knowledge and experience of researchers. As such, the progress of material discovery is limited to the resource and manpower that are devoted to the field. The rapid development of artificial intelligence (AI), computational algorithms and hardware gradually changed the ways of material discovery through introducing technologies such as virtual screening, active learning, and neural network.[1–7] These technologies enable the rapid development of materials design and meanwhile lead to a huge demand for experimental data, which is limited by manpower and resource. As such, much effort has been devoted to robotics-based experimental data collection to accelerate material discovery with the desired property.[8-10] If the desired property of a material can be evaluated by computation (e.g. quantum chemical calculation), a self-improving material discovery system can be realized. However, it is challenging to accurately predict the material property simply by simulation.[11-13]

Photosensitizers (PSs) are molecules that can harvest light energy and generate singlet oxygen, reactive nitrogen species, and radicals for different applications.[14–17] Benefiting from the high reactivity of the photoproducts, PSs are highly desirable in many applications, including photodynamic therapy (PDT), antibacterial treatment, photocatalytic water treatment, and synthetic chemistry.[18–22] Owing to the limited PS structural library, it is difficult to find an optimized PS for a particular application, e.g., a PS with high efficiency and near-infrared (NIR) absorption for PDT, or high photostability and good solubility for water treatment. Therefore, it is of both research and application interest to discover new PS structures. [14,23]

Recently, based on the understanding of the role of intersystem crossing (ISC) in singlet oxygen generation, our lab proposed a new design principle for PS by tuning the singlet-triplet energy gap

($\Delta E_{ST}$).[24] Such a design strategy enables the precise design of PS with their efficiency optimized by molecular engineering on $\Delta E_{ST}$. As a result, series of high-efficiency PSs with molecular structures different from typical PSs have emerged.[23] Since $\Delta E_{ST}$ can be precisely calculated by Time-dependent Density Functional Theory (TD-DFT) method, pre-screening of promising PS candidates from $\Delta E_{ST}$ calculation candidates by quantum calculations before synthesis is becoming an efficient approach to PS discovery. However, the PS discovery is still limited to the time-consuming DFT and TD-DFT calculations and the construction of molecular candidates, which is experience-dependent, especially when multiple properties like both small $\Delta E_{ST}$ and particular absorption and emission wavelengths are required. [16,25] In this regard, the recent algorithm advancements in machine learning (ML) have shown that with sufficient training data and proper methods, instant prediction of molecular energy levels or molecular orbital information can be achieved with great accuracy.[26,27] As for PS, an accurate enough prediction system for energy levels will greatly reduce the time required for molecular screening. Moreover, active learning methods like Bayesian optimization can suggest molecules with a set goal, for example, in search for molecules with small $\Delta E_{ST}$. Such a recommendation process, together with DFT calculations, may be able to form a PS discovery loop system with self-improving capability for high throughput PS search and high accuracy of property prediction.

Herein, we describe a self-improving PS discovery system that is based on the combination of accurate prediction of PS properties via DFT calculation and Bayesian optimization-based active learning. First, a single figure of merit for high-efficiency PS design is identified from the Jablonski diagram-based understanding of singlet oxygen generation and the derivation of equations from the Fermi Golden Rule. Our design principles for the molecular generation algorithm and the task for first-principles calculations will then be based on this figure of merit. Next, with the molecular generation algorithm,

12015 randomly selected molecules were constructed and labeled with results from TD-DFT calculations. These structures form the initial dataset and are used for model training through a molecular graph-based forward prediction model. Then, a Bayesian optimization-based active learning method is implemented to form a loop with DFT calculation to improve the model's performance in our target property region. This will enable the framework to efficiently adapt to new chemical design space when there is a need to change the prior chemical knowledge. During the self-improving loops, the model prediction accuracy and PS search ability have been improved. Finally, choosing from the suggested promising candidates from our system, four new PSs were successfully synthesized with desired properties and performances comparable with or better than commercially available PSs.

**Results**

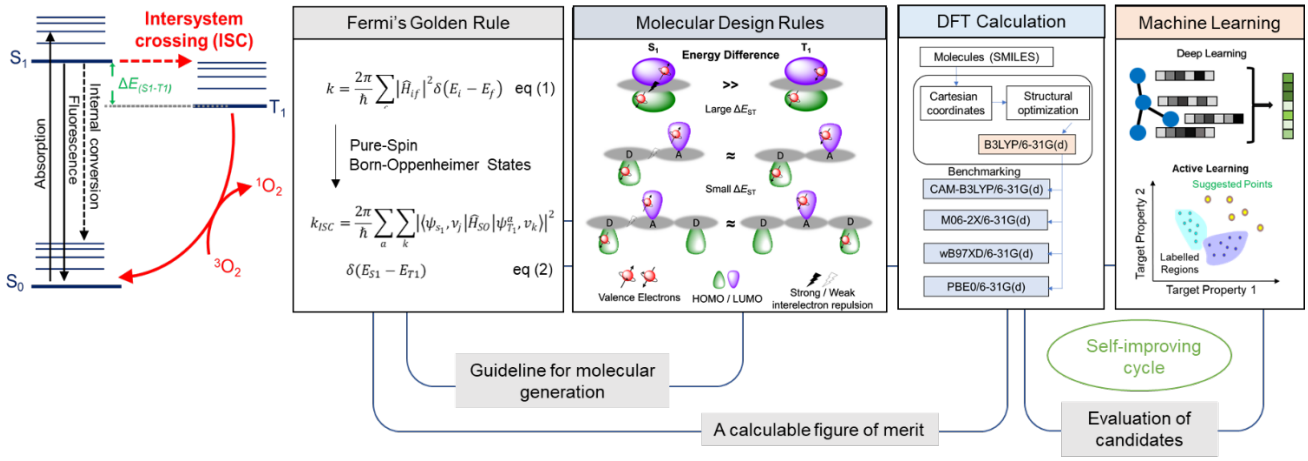



**Figure 1. Linking Jablonski diagram to machine learning.** A self-improving PS discovery system with quantum mechanism-based understanding and efficient molecular design rule for molecular generation algorithm, accurate calculation of figure of merit ($\Delta E_{ST}$), and machine learning.

**Linking Jablonski Diagram to Machine Learning**

As shown in the Jablonski diagram (**Figure 1**), the key process for singlet oxygen generation is intersystem crossing (ISC) from the singlet state to the triplet state. Thus, the design of high-performance PS is basically the design of PS with a high ISC rate ($k_{isc}$). According to the Fermi Golden Rule, the rate constant for the transition from one molecular state to another is shown in **Eq 1**, where in the ISC process, the initial state is $S_1$ and the end state is $T_1$. After pure spin Born-Oppenheimer (BO) approximation, the $k_{isc}$ can be expressed as **Eq 2**. Obviously, the ISC rate is inversely related to the $S_1$-$T_1$ gap. Therefore, reducing $\Delta E_{ST}$ is one of the key requirements to high efficiency in PSs, and this is also substantiated by the design of recently reported PSs.[23-24,28-29] According to these recent works, a molecule shows remarkable PS efficiency when its $\Delta E_{ST}$ is lower than 0.3 eV. With this design principle, we developed a molecular design recipe of donor-acceptor (DA) and donor-acceptor-donor (DAD), which are supposed to have relatively low $\Delta E_{ST}$ as compared with non-DA-based molecules. This is because the energy difference between $S_1$ and $T_1$ is attributed to the valence electrons of S1 with opposite spins, which causes electron repulsion to increase the total energy of the molecule. As such, reducing $\Delta E_{ST}$ is to mitigate the electron repulsion in the S1 state. As electron repulsion is inversely correlated to the distance between valence electrons, separating the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) distribution is the key to effectively reduce $\Delta E_{ST}$. Thus, our candidate libraries obey the following combination recipe: donor-pi-acceptor (DA) or donor-pi-acceptor-pi-donor (DAD). For the calculation of $\Delta E_{ST}$, the b3lyp/6-31g(d) hybrid functional and basis set were used after a benchmark comparison with other methods, the accuracy of the considered method is within 0.1 eV in most DA and DAD systems according to literature. As such, $\Delta E_{ST}$ calculation by TD-DFT can act as an evaluation tool in our

system, and molecules with small calculated $\Delta E_{ST}$ (< 0.2 eV) can be considered as good candidates. On the other hand, the HOMO-LUMO (H-L) gap of generated molecules can be used to evaluate their optical ranges and using the same DFT method, H-L gap values can be obtained with good accuracy as well. With such a first principles calculation-based evaluation tool, a self-improving loop with simultaneous improvement in molecular search ability and prediction accuracy in a particular molecular space (i.e. DA and DAD) can be formed with the assistance of neural networks and active learning.

**Molecular Space Generation and Initial Prediction Model**

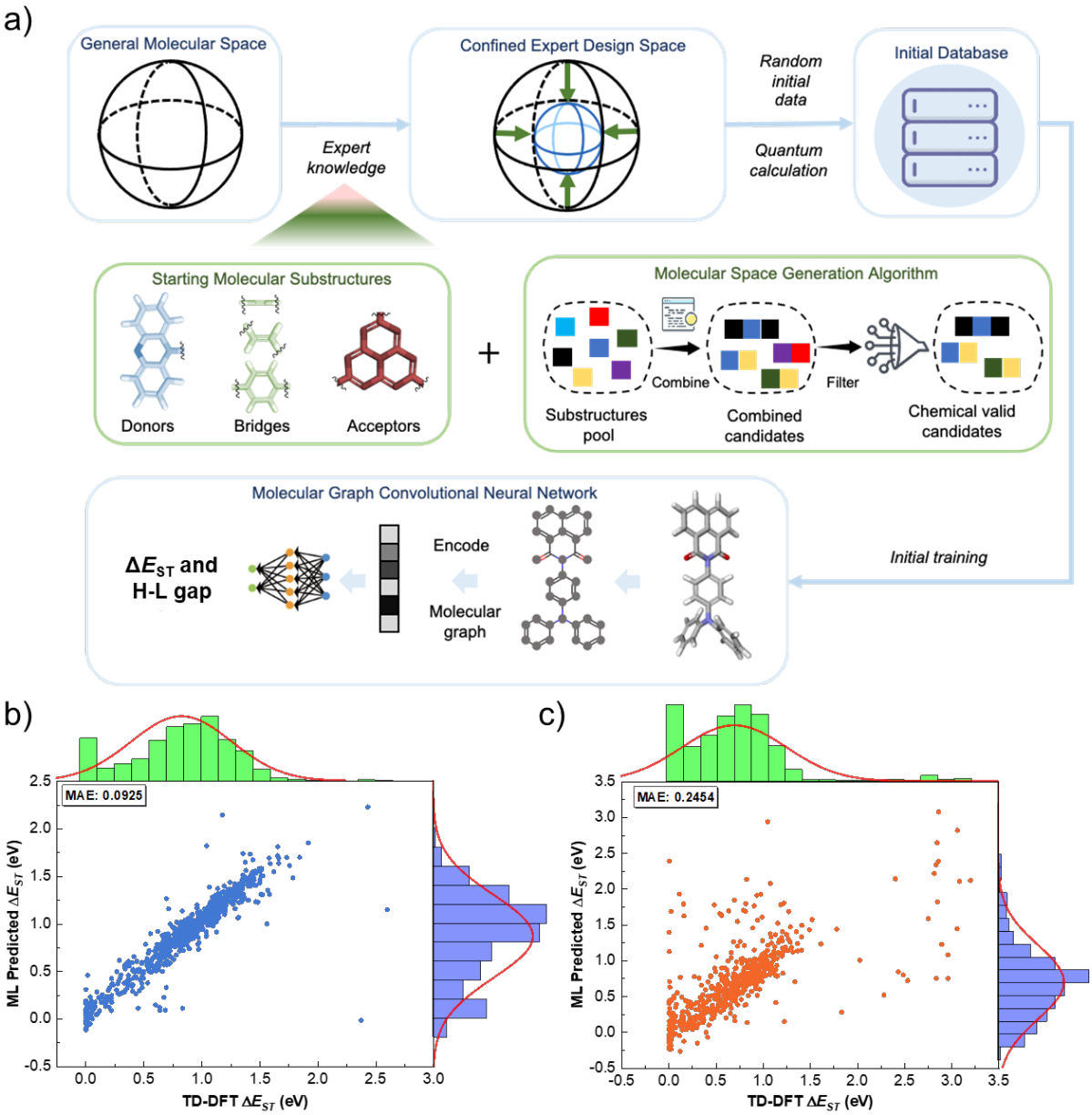



**Figure 2. Molecular Space Generation, Graph Convolutional Neural Network (GCNN), and Initial Model Prediction Performance.** (a) Schematic overview of molecular space generation algorithm for DA and DAD form PS and molecular graph convolutional neural network used for model training, (b) MAE and distribution of $\Delta E_{ST}$ values predicted by the initial model against calculations by TD-DFT for DA form PSs and (c) DAD form PSs

A general molecular design space can contain more than $10^{60}$ different molecules with different structures that are infeasible to screen with.[30] Using the expert design principles above and our developed algorithm for molecular generation[31] (discussed in detail in the Methods section), we could largely reduce the chemical space to the scale of $10^6$. After constructing the confined expert design space that is feasible to be screened with a fast deep learning model, the initial training databases were then created via the quantum calculations to label their $\Delta E_{ST}$ values and H-L gaps, respectively. To ensure broad coverage and diversity of the initial databases, the molecules labeled by quantum calculations were randomly picked from the corresponding molecule structural library. Next, a molecular graph neural network[32] was trained on the initial database to form an initial prediction model and this initial model was used as the navigation model in the first cycle of the active learning loop (Figure 2 and Figure S1). The hyperparameters used for the models were derived via hyperparameter optimization with a Gaussian process (Details in Table S1). An abstract structure of the molecular graph neural network is shown in Figure 2a, whereby a two-dimensional representation of a molecule is treated as the input. Useful information such as atomic number, element group, etc. was encoded in this representation. Several graph convolutional layers were then used to process the input into a fixed-length representative molecular fingerprint, which was further processed with several fully connected layers to obtain the final property prediction. Since graphs are the most natural representations of organic molecules, the graph-based methods have state-of-the-art performances and thus GCNN is chosen in our study. Some comparison studies with traditional molecular fingerprint methods are shown in Supplementary Note, Figure S3.

The $\Delta E_{ST}$ prediction performances for both the initial DA predictive model and DAD predictive model

are shown in Figure 3(b-c). These results are based on the models' performances on a fixed test set (770 for DA and 612 for DAD) that is independent of the training data. From both a single statistic indicator point of view (mean average error (MAE)) and based on the level of agreement between the distribution of ML predicted values and $\Delta E_{ST}$ computed by quantum calculations, the prediction performance of $\Delta E_{ST}$ is very close to the quantum calculated values.[33-35] In addition, the DA model performs much better than the DAD model. This could be due to two major reasons. First, the whole design space of DAD form PSs is much larger (about 70 times) than that of DA form PSs, and therefore, the labeled data for DAD is much sparser than the ones for DA. Second, since the DAD molecules are larger than DA molecules in terms of the number of atoms and molecular weight, this could make the information in molecules harder to be encoded, leading to more difficulties in learning their quantitative structure-property relationships.

The $\Delta E_{ST}$ prediction performances for both the initial DA predictive model and DAD predictive model are shown in Figure 3(b-c). These results are based on the models' performances on a fixed test set (770 for DA and 612 for DAD) that is independent of the training data. From both a single statistic indicator point of view, using mean average error (MAE), and the level of agreement between the distribution of ML predicted values and quantum calculated values point of view, the prediction performance of $\Delta E_{ST}$ is very close to the quantum calculated values.[33-35] In addition, the DA model performs much better than the DAD model. This could be due to two major reasons. First, the whole design space of DAD form PSs is much larger (about 70 times) than DA form PSs, and therefore, the labeled data for DAD is much sparser than the ones for DA. Second, since the DAD molecules are larger than DA molecules in terms of the number of atoms and molecular weight, this could make the information in molecules harder to be encoded, leading to more difficulties in learning their

quantitative structure-property relationships.

**Bayesian Optimization and Quantum Calculation Based Self-Improving Cycle**

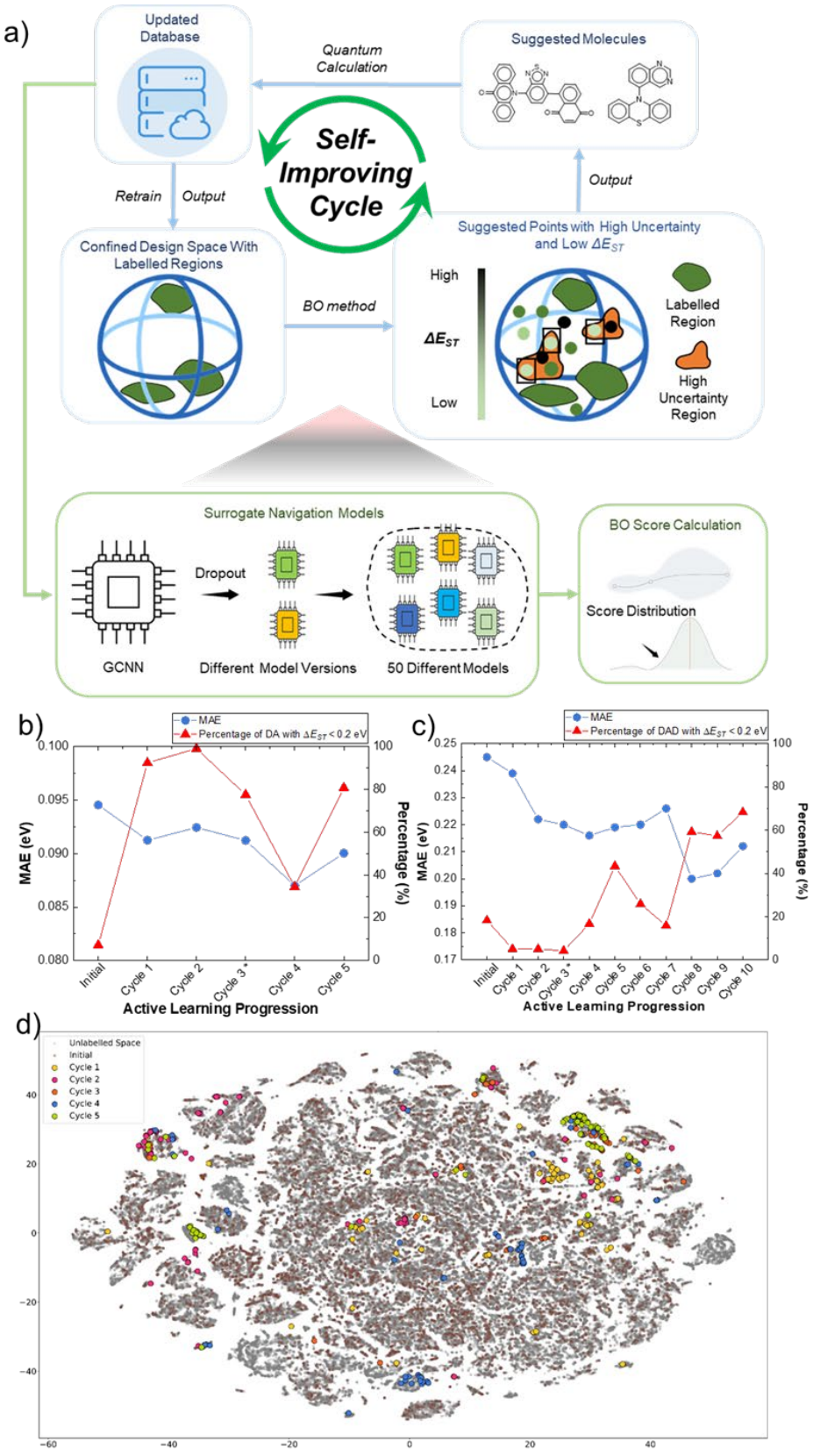


**Figure 3. Bayesian Optimization and Quantum Calculation Based Self-Improving Cycle.** (a) Schematic overview of Bayesian optimization-based active learning for both DA and DAD form PSs, (b-c) Performance of DA and DAD models with the progression of active learning based on the percentage of PSs found with $\Delta E_{ST} < 0.2$ eV in each active learning cycle and mean model performance on a fixed test set; * *There is a change in molecular design space from cycle 3 with the addition of new substructures,* (d) Visualization of the DA molecular space through active learning progression using t-SNE.

As illustrated in Figure 3a, the Bayesian optimization (BO) score, or expected improvement (EI), was calculated by utilizing the molecular graph-based prediction model as the surrogate navigation model. In each cycle of active learning, the surrogate navigation models were trained on the database provided from the previous cycle (i.e., for the first cycle of the active learning cycle, the training database is the initial database; then, the database is updated with newly suggested structures and used as the training database for the second cycle). By applying dropout methods, 50 different versions of molecular graph-based prediction models were provided from the same training database during each round. These navigation models were used to provide average predicted values of the $\Delta E_{ST}$ of all potential molecules that were unlabeled in the confined design spaces with labeled regions (i.e., the design space that is updated after adding labeled data in our database). The disagreement (i.e., standard deviation) between these navigation models was treated as the uncertainty part of the BO score calculation. The mean of all predicted $\Delta E_{ST}$ values was treated as the mean part of the BO score calculation. By considering the exploration-exploitation trade-off, whereby PSs with the highest EI (i.e. with both low predicted $\Delta E_{ST}$ and high prediction uncertainty), were recommended in each cycle. Adding such molecules with a high

EI score back to the training data will improve the model's performance in the region containing molecules with desired properties the most. The suggested molecules with the highest EI scores were then labeled by quantum calculations to update the training database for the next cycle.

Figures 3(b-c) show the change of MAE for the surrogate navigation models and the percentage of PSs with low $\Delta E_{ST}$ with the increasing cycles of active learning. Here, the navigation model performance was based on the model trained on the dataset updated after each cycle, and the percentage of PSs with low $\Delta E_{ST}$ were calculated out of the number of PSs selected in each cycle (i.e. for the initial dataset it was out of 7101 for DA and 4914 for DAD; for cycle 1, it was out of the number of PSs recommended to be added to the training set for the next cycle, with details in Table S2). By efficiently sampling from the design space with the active learning strategy, the MAEs for the models used for both DA and DAD generally decrease with the number of cycles. It is worthy to note that there were new PSs based on a new design space of donor, acceptor, and bridge substructures added from cycle 3 of active learning for synthetic accessibility reasons, and thus there was a decrease in model accuracy based on the test set MAE, for DA model especially. This indicates that even though the design space is changing, by efficiently sampling new data from the updated design space, the model could effectively adapt to the new space.

For DA, as the initial model was already quite accurate (with a low average MAE of 0.09455 in the prediction of $\Delta E_{ST}$), the number of PSs recommended in cycles 1 and 2 meeting the low $\Delta E_{ST}$ criteria were high (more than 90% of the recommended set). Therefore, the model would focus more on exploitation, compared to the exploration of the whole design space, and award a higher EI score to PSs with lower $\Delta E_{ST}$, while at the same time be able to predict these PSs with high accuracy in the recommendations. From cycle 3 to 4, there was a decrease in the percentage of low $\Delta E_{ST}$ PSs

recommended as there was a drastic change in the design space screened by the model. About 86% of the changed molecular space was made up of new DA PSs combined from new recipe donors, acceptors, and/or bridge substructures. Based on the balance between exploitation and exploration, the model would prioritize the exploration of the new design space in the recommendation of PSs with higher uncertainty instead. Nevertheless, towards the end of active learning, the improvement at cycle 5 is apparent with 80.8% of PSs meeting the low $\Delta E_{ST}$ criteria.

For DAD, on the other hand, the initial model's accuracy was low due to the vast molecular space of more than 7 million PSs while the model was only trained with 4914 DAD and 7691 DA PSs. Despite a larger training set with DA PSs included, the DAD structures were larger than those in DA form, and naturally, the model accuracy would not be as ideal. Thus, the model was focused more on the exploration of the whole design space and recommending DAD PSs with higher uncertainty instead of recommending PSs with low $\Delta E_{ST}$ accurately (for 4–5% of the recommendations meeting the criteria for cycles 1 to 3). With the changing design space from cycle 3 due to the addition of new donor, acceptor, and bridge substructures, fluctuating performances from cycles 3 to 7 are likely due to the changing trade-off between exploration and exploitation of the design space by the model. After several more cycles, there is a general improvement trend as seen from the decreasing average MAE and increasing percentage of PSs with low $\Delta E_{ST}$. By cycle 10, the percentage of PSs in the recommendations with low $\Delta E_{ST}$ has reached 68.3% and the MAE has decreased by 13% from the initial model. This signifies that our active learning strategy is efficient in finding the desired low $\Delta E_{ST}$ molecules with higher accuracy, despite adding only a smaller proportion of new training data.

Finally, through t-distributed stochastic neighbor embedding (t-SNE) (Details in Supplementary Note, Figure S2), a visualization of the whole active learning process was chosen to investigate further into

our self-improving system. The initial random training data spread evenly over the entire design space (Figure 4d and Figure S2), which ensured our model's generalizability for the whole design space. And from cycles 1 to 5 (for DA model in Figure 4d), it was clear that most of the suggested molecules formed several significant clusters for each round, and only a few points were explored far away from others. This phenomenon showed a good trade-off between exploitation in suggesting low $\Delta E_{ST}$ molecules to learn better over them and exploration in gaining more understandings of regions with considerable uncertainty. In addition, the proposed molecules were also spread over the whole design space across all the rounds of active learning. This indicated our active learning strategy could find new molecular structures with low $\Delta E_{ST}$ that were dissimilar with each other and ensured our model could predict well over several different low $\Delta E_{ST}$ molecular regions. The diversity in new structures was especially preferred in our new PSs structure discovery task.

**Discovery of High-Performance PSs**

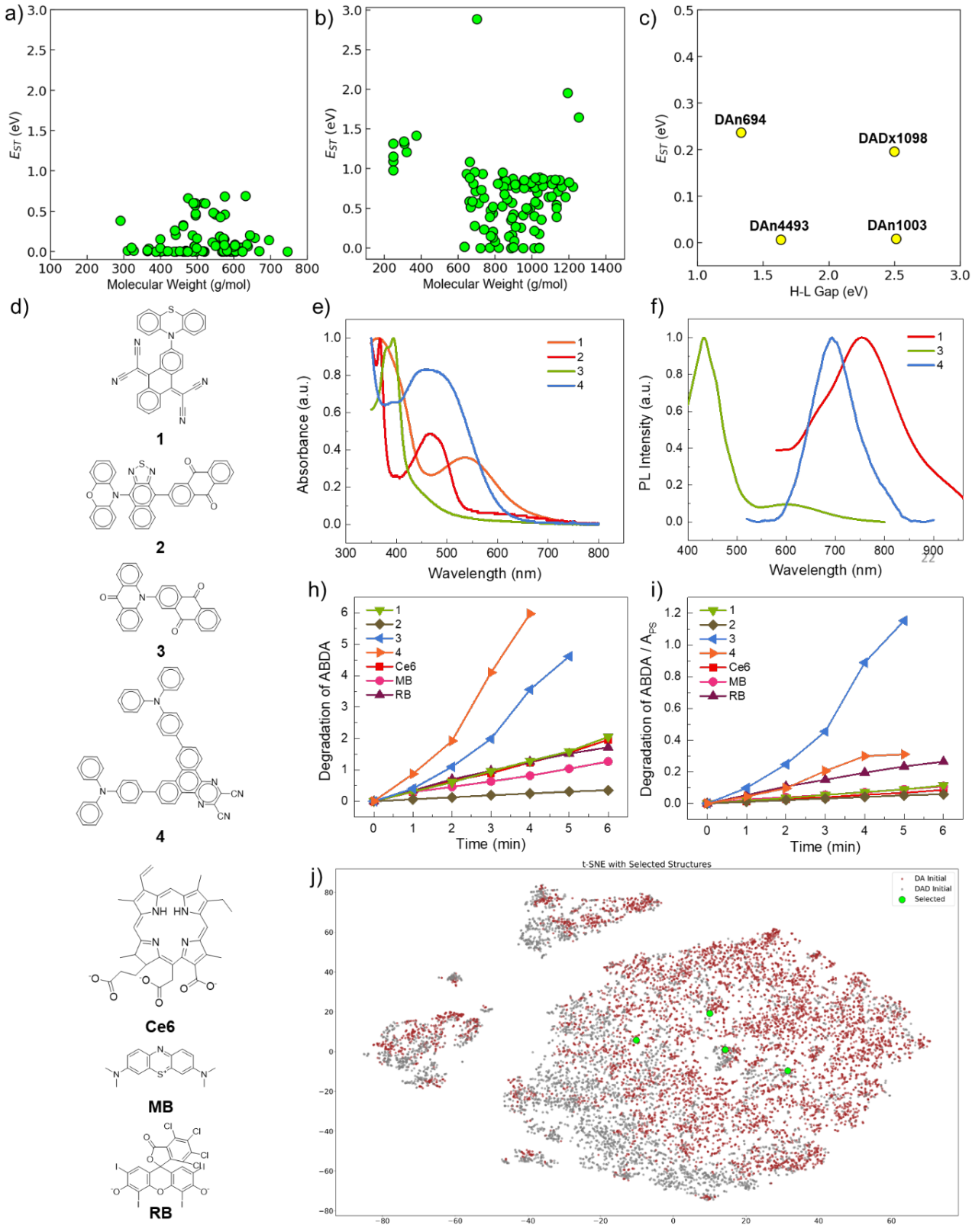



**Figure 4. Discovery of High-Performance PSs.** (a-c) S-T gap versus the molecular weight of DA and DAD suggested in the final active learning cycle and for the selected 4 PSs. (d) Molecular structures

of **1-4**, Ce6, MB, and RB. (e) UV-vis absorption spectra of **1-4** and (f) Photoluminescence spectra of 1,3 and 4 at $5\times10^{-6}$ M in DMSO/water mixture (v/v = 1/99). (h)-(i) The degradation rates, defined by ln ($a_0$/a) and ln ($a_0$/a)/$A_{PS}$ of by ABDA 1-4, Ce6, MB and RB; $A_0$ and A are the absorbance of ABDA in the presence of the PSs at 399 nm before and after irradiation, respectively. $a_0$ and a is the initial and real-time absorbance at 399 nm. $A_{PS}$ represents the light absorbed by the PSs, which is determined by the integration of areas under the absorption bands in the wavelength range of 400–800 nm. (j) Visualization of the selected 4 PSs among the DA and DAD molecular space by t-SNE.

After computation, the search space was reduced to human tractable decision batches from the final cycle of suggestions by active learning, and the candidates were then assessed by PS experts according to their predicted property, H-L gapgap, structural novelty, and synthetic accessibility. Four candidates with H-L gap ranging from 1.33 eV to 2.51 eV (red to blue) were synthesized and tested experimentally to confirm the discovery ability of our system (Figure 4c and 4d). The selected molecules contain phenothiazine, acridinone, phenoxazine and triphenylamine donors combined with 2,2'-(anthracene-9,10-diylidene)dimalononitrile, naphtho[2,3-c][1,2,5]thiadiazole, anthracene-9,10-dione and dibenzo[f,h]quinoxaline-2,3-dicarbonitrile acceptors. All these structures have rarely been reported from previous PS materials or even DA-based materials. We first characterized the optical properties of the four compounds by measuring their absorption and emission spectra in DMSO/water mixtures (Figure 4e and 4f). In general, the prediction from our system agreed with experiments within the known accuracy of TD-DFT calculation and the noise in experimental measurements. As expected, compounds **1-4** showed an absorption band that lies in the blue, green, yellow, and red regions, respectively. We next tested the PS efficiency of **1-4** by using a light-induced 9,10-anthracenediyl-

bis(methylene)dimalonic acid (ABDA) decomposition method (Figure S6, S7, Figure 4h, and 4i). In such a method, the faster ABDA absorption decrease represents the higher efficiency of singlet oxygen generation. Three commonly used commercial PSs, Chlorin e6 (Ce6), Methylene Blue (MB), and Rose Bengal (RB), were tested in the same conditions for comparison. Impressively, all four compounds showed a comparable ABDA degradation rate with three commercially available PSs. To exclude the influence of light absorption ability difference, we did another comparison by dividing the degradation rate with the integrated area of different compounds in the visible range (Figure 4i). Similarly, all compounds showed comparable efficiency with or higher efficiency than the three commercially available PSs. Compound **2** showed relatively lower PS efficiency, which should be attributed to its high planar structure and self-quenching effect in aggregate state. It is important to note that compound **3** showed the highest degradation rate, which is 5-fold, 3-fold, and 2.5-fold more effective than those of Ce6, MB, and RB, respectively. These results not only verified the search ability of our system in new molecular space (Figure 4j) but also proved the potential of our system in next-generation commercial PS discovery within a targeted optical range.

In summary, a self-improving PS discovery system with powerful molecular search ability and high accuracy of molecular property prediction was proposed in this study. To the best of our knowledge, this is the first time that self-learning or self-improving molecular material discovery has been realized. The idea to combine a calculable evaluation figure ($\Delta E_{ST}$) and active learning to form a self-improving loop could inspire material discovery in a wide range of fields. For the PS discovery system, the key to success lies in a sufficiently deep fundamental understanding of PS property. The quantum mechanism-based understanding of PS property not only enables the evaluation of PS efficiency by

simulation but also helps to construct a very promising molecular library for PS search. In total, a large database of 13,804 PSs have been labeled throughout this work and after self-improving, an error of 0.09 in MAE of $\Delta E_{ST}$ was achieved, a total of 2227 PSs with low $\Delta E_{ST}$ were labeled, and 3723 PSs covering the full visible range with predicted low $\Delta E_{ST}$ have been recommended. This makes our PS discovery system robust and useful, and the recommended PSs may find wide applications in different fields, such as photodynamic therapy, bacteria ablation, and photocatalytic water treatment.

## Methods

### *Molecular Space Generation Algorithm*

Given a list of donor, acceptor, and bridge molecular substructures, to generate the entire molecular space for both DA and DAD PSs, a unique generation algorithm following fixed bonding positions was used. Molecular structures were converted from simplified molecular-input line-entry system (SMILES) form to RDKIT[31] *MOL* object form for editing. For donor structures, they were given a single point of connection. For acceptor structures, some have more than one favorable point of connection. To create DA form PSs (Donor-Bridge-Acceptor), a bridge was used to connect a given donor at a specific given position with a given acceptor at one of its positions. To create DAD form PSs (Donor-Bridge-Acceptor-Bridge-Donor), an identical bridge structure was used to connect each donor with a central acceptor. In our database, not all acceptors have more than one connecting position. If both donor structures are the same, a symmetric DAD will be formed. If the donor structures are different, an asymmetric DAD will be formed. In our database, the naming convention is as follows: DA, DAD, and DxAD for DA, symmetric DAD, and asymmetric DAD PSs respectively. This process was repeated through all possible permutations according to our starting substructure list until all DA

and DAD PSs were formed. Chemical validity checks were done with RDKIT[31] and chemically invalid structures were removed from our database. After the structures were combined, a script was used to convert all SMILES into Gaussian input files (.gjf) for subsequent quantum calculations with Gaussian09.

***Density Functional Theory and Time-Dependent Density Functional Theory Calculations***

The ground states of all molecules were fully optimized by the hybrid B3LYP, in combination with 6-31G(d) basis set. The excited-state characteristics were calculated by the time-dependent density functional theory (TD-DFT) using optimized ground state geometries. TD-DFT in combination with the B3LYP hybrid functional method and the 6-31G (d) basis set has been shown to provide accurate energies for the excited-state of the DA molecular system with less than 0.15 eV error. In this work, Gaussian09 was used for all quantum calculation tasks.

***Molecular Graph Convolutional Neural Network***

Every molecule can be defined as an undirected graph $\mathcal{G} = (V, E)$ where nodes $V$ and edges $E$ represent atoms in the molecules and strong bonds (i.e., covalent bonds and ionic bonds) between atoms, respectively. Each node $V$ contains the property feature $x_i$ of the atom it represents and each edge $E$ represents the connectivity between the nodes. The topology of a molecular graph is contained in the adjacency matrix of itself. During each layer of the molecular graph convolutional neural network, a hidden feature $hi$ at a particular node will be updated with the information (hidden features, atom features) from the neighboring nodes and itself (the initial hidden features are initialized with zeros). During this information exchanging process, a graph convolutional weight parameter matrix will be learned from the data. This learned matrix can ensure the information is exchanged in the best way to encode the most representative features of the molecule. After several times of information

exchange (depending on the number of layers of graph neural network), the processed hidden features at each node will be gathered by a graph gathering layer (can be a sum or average operation) to form the final molecular vector (i.e., the molecular fingerprint). Finally, this molecular fingerprint will be passed through several fully connected layers (i.e., dense layers) to produce the final predictions of both the $\Delta E_{ST}$ value and H-L gap value of the molecule. All weight parameters in the graph convolutional layers and fully connected layers were trained and updated by gradient descent methods.

***Bayesian Optimization Method in Active Learning***

For Bayesian Optimization (BO), an expected improvement (EI) acquisition function that considers the trade-off between global search and local optimization, or in other words, exploration against exploitation, was used[7]. The EI function based on exploration-exploitation trade-off [36] was adapted to seek a minimized value of $\Delta E_{ST}$ instead of a maximum and it used a trade-off value $\xi = 0.01$. In our case, each prediction's mean and standard deviation was derived from 50 navigation models with different dropouts. The EI values were then ranked, and the leading 120 corresponding structures with the highest EI values were selected as the next points for labeling via DFT and TD-DFT before the next active learning cycle was repeated. It is noted that few molecules in the 120 set with invalid $\Delta E_{ST}$ values (i.e. 0) were removed. In this work, the surrogate model was based on the graph convolutional neural network built with the DeepChem package (https://github.com/deepchem/deepchem)[37]. The model uncertainty for every prediction was calculated with the built-in function which was based on 50 different navigation models with different dropout rates. After about 120 structures were recommended for labeling by DFT and TD-DFT, they were added back into the training set for the next cycle of model training before re-screening of the remaining unlabeled molecular space before another set of structures were recommended by the highest EI again. To evaluate the model

performances, the model in every cycle was validated on a fixed test set of 770 DA and 612 DAD structures, respectively. The fixed test set includes random structures from initial and all active learning cycles. An average MAE was obtained by running 5 repetitions of model training on the same training set for each cycle.

**Acknowledgement**

We acknowledge the Singapore RIE2020 Advanced Manufacturing and Engineering (AME) Programmatic grant "Accelerated Materials Development for Manufacturing" by the Agency for Science, Technology and Research under Grant No. A1898b0043. the Singapore NRF Investigatorship (grant no. R279-000-444-281), and the National University of Singapore (grant no. R279-000-482-133) for funding support.

**Author Contribution**

S.X. and P.C. performed all material modeling and quantum calculations. J.L., P.C., and X.L. performed all machine learning experiments. S.X performed the chemical synthesis. B.L. and X.W. supervised the project. S.X., J.L., P.C, B.L. and X.W. discussed the results and edited the manuscript.

**Competing financial interests**

The authors declare no competing financial interests.

**Data availability**

The data that support the finding of this work are available from the corresponding authors upon request.

**Code availability**

The code developed during the current study is available from the corresponding authors upon request.

**References**


1. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **119**, 10520–10594 (2019).
2. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015).
3. Cox, J. & Witten, I. B. Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.* **20**, 482–494 (2019).
4. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. DRUG Discov.* **18**, 463–477 (2019).
5. Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
6. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **5**, 1–36 (2019).
7. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, (2019).
8. Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237-241 (2020).
9. Roch, L. M. *et al.* ChemOS: Orchestrating autonomous experimentation. *Sci Robot* **3** (2018).
10. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **104**, 148-175,

doi:10.1109/jproc.2015.2494218 (2016).

11. Collins, C. *et al.* Accelerated discovery of two crystal structure types in a complex inorganic phase field. *Nature* **546**, 280-284 (2017).

12. Davies, D. W. *et al.* Computer-aided design of metal chalcohalide semiconductors: from chemical composition to crystal structure. *Chem Sci* **9**, 1022-1030 (2018).

13. Gomez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120-1127 (2016).

14. Zhao, J., Wu, W., Sun, J. & Guo, S. Triplet photosensitizers: from molecular design to applications. *Chem. Soc. Rev.* **42**, 5323–5351 (2013).

15. Liu, K. *et al.* Supramolecular Photosensitizers with Enhanced Antibacterial Efficiency. *Angew. Chem. Int. Ed.* **52**, 8285–8289 (2013).

16. Zhang, T. *et al.* In Situ Monitoring Apoptosis Process by a Self-Reporting Photosensitizer. *J. Am. Chem. Soc.* **141**, 5612–5616 (2019).

17. Dai, J. *et al.* Efficient Near-Infrared Photosensitizer with Aggregation-Induced Emission for Imaging-Guided Photodynamic Therapy in Multiple Xenograft Tumor Models. *ACS Nano* **14**, 854–866 (2020).

18. Jin, G. *et al.* Near-infrared light-regulated cancer theranostic nanoplatform based on aggregation-induced emission luminogen encapsulated upconversion nanoparticles. *Theranostics* **9**, 246–264 (2019).

19. Yuan, H. *et al.* Chemical Molecule-Induced Light-Activated System for Anticancer and Antifungal Activities. *J. Am. Chem. Soc.* **134**, 13184–13187 (2012).

20. Wang, B., Wang, M., Mikhailovsky, A., Wang, S. & Bazan, G. C. A Membrane-Intercalating Conjugated Oligoelectrolyte with High-Efficiency Photodynamic Antimicrobial Activity. *Angew. Chem. Int. Ed.* **56**, 5031–5034 (2017).

21. Liu, J. *et al.* Nanoscale metal-organic frameworks for combined photodynamic \& radiation therapy in cancer treatment. *Biomaterials* **97**, 1–9 (2016).

22. Li, X., Kwon, N., Guo, T., Liu, Z. & Yoon, J. Innovative Strategies for Hypoxic-Tumor Photodynamic Therapy. *Angew. Chem. Int. Ed.* **57**, 11522–11531 (2018).

23. Hu, F., Xu, S. & Liu, B. Photosensitizers with Aggregation-Induced Emission: Materials and Biomedical Applications. *Adv. Mater.* **30**, 1–29 (2018).

24. Xu, S. *et al.* Tuning the singlet-triplet energy gap: A unique approach to efficient photosensitizers with aggregation-induced emission (AIE) characteristics. *Chem. Sci.* **6**, 5824–5830 (2015).

25. Wu, W. *et al.* High performance photosensitizers with aggregation-induced emission for image-guided photodynamic anticancer therapy. *Mater. Horizons* **4**, 1110–1114 (2017).

26. Huang, S. *et al.* Computational prediction for singlet- and triplet-transition energies of charge-transfer compounds. *J. Chem. Theory Comput.* **9**, 3872–3877 (2013).

27. Hait, D., Zhu, T., McMahon, D. P. & Van Voorhis, T. Prediction of excited-state energies and singlet-triplet gaps of charge-transfer states using a restricted open-shell Kohn-Sham approach. *J. Chem. Theory Comput.* **12**, 3353–3359 (2016).

28. Wu, W. *et al.* A Highly Efficient and Photostable Photosensitizer with Near-Infrared Aggregation-Induced Emission for Image-Guided Photodynamic Anticancer Therapy. *Adv Mater* **29** (2017).

29. Wu, W. *et al.* Polymerization-Enhanced Photosensitization. *Chem* **4**, 1937-1951 (2018).

30. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).

31. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. (2013).

32. Duvenaud, D. *et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. 1–9 (2015) doi:10.1021/acs.jcim.5b00572.

33. Dong, Y. *et al.* Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. *NPJ Comput. Mater.* **5** (2019).

34. Mori-Sanchez, P., Cohen, A. J. & Yang, W. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Phys. Rev. Lett.* **100**, 146401 (2008).

35. Zheng, X., Cohen, A. J., Mori-Sanchez, P., Hu, X. & Yang, W. Improving band gap prediction in density functional theory from molecules to solids. *Phys. Rev. Lett.* **107**, 026403 (2011).

36. Brochu, E., Cora, V. M. & De Freitas, N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv Prepr. arXiv1012.2599* (2010).

37. Ramsundar, B., Eastman, P., Walters, P. & Pande, V. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. (‘ O’Reilly Media, Inc.’, 2019).

Supplementary Materials for

# A Self-Improving Photosensitizer Discovery System via Bayesian Optimization and Quantum Chemical Calculation

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

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

‡ These authors contributed equally to this work.

## 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 **1-12** 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 $^{1}H$ 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 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 |

**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 | Added Suggestions for Cycle N+1 | Training | Training | Added Suggestions for Cycle N+1 |
| 0 (Initial) | 7101 | | 7691 | 4914 | |
| 1 | 7101 | 119 | | 4914 | 119 |
| 2 | 7220 | 112 | | 5033 | 120 |
| 3 * | 7332 | 120 | | 5153 | 120 |
| 4 | 7452 | 119 | | 5273 | 120 |
| 5 | 7571 | 120 | | 5393 | 120 |
| 6 # | 7691 | | | 5513 | 120 |
| 7 | | | | 5633 | 120 |
| 8 | | | | 5753 | 120 |
| 9 | | | | 5873 | 120 |
| 10 | | | | 5993 | 120 |
| 11 # | | | | 6113 | |
| **Total** | **7691** | **590** | | **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.

# In these cycles, the DA model is trained on 7691 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 770 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 Energy Gap**

The prediction models for both DA and DAD form PSs are trained to predict the H-L gap as well. Figure S1 shows the initial prediction performance for H-L gap. Just like the prediction of $\Delta E_{ST}$ (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.

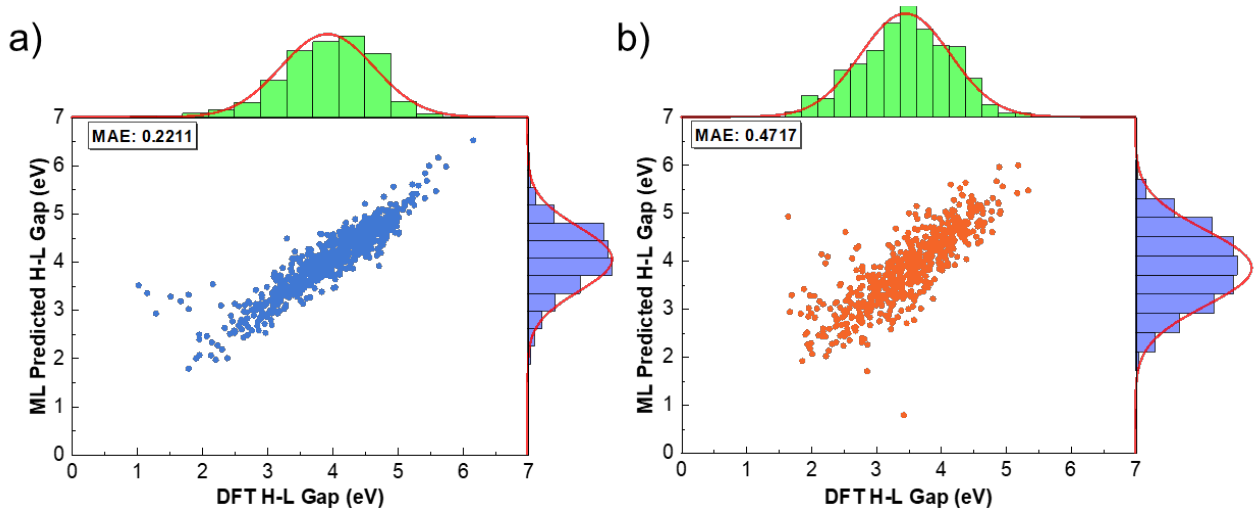



**Figure S1. Prediction of H-L gap by initial models** (a) MAE and distribution of H-L gap predicted by the initial model against calculation by TD-DFT for DA form PSs and (b) DAD form PSs.

**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)

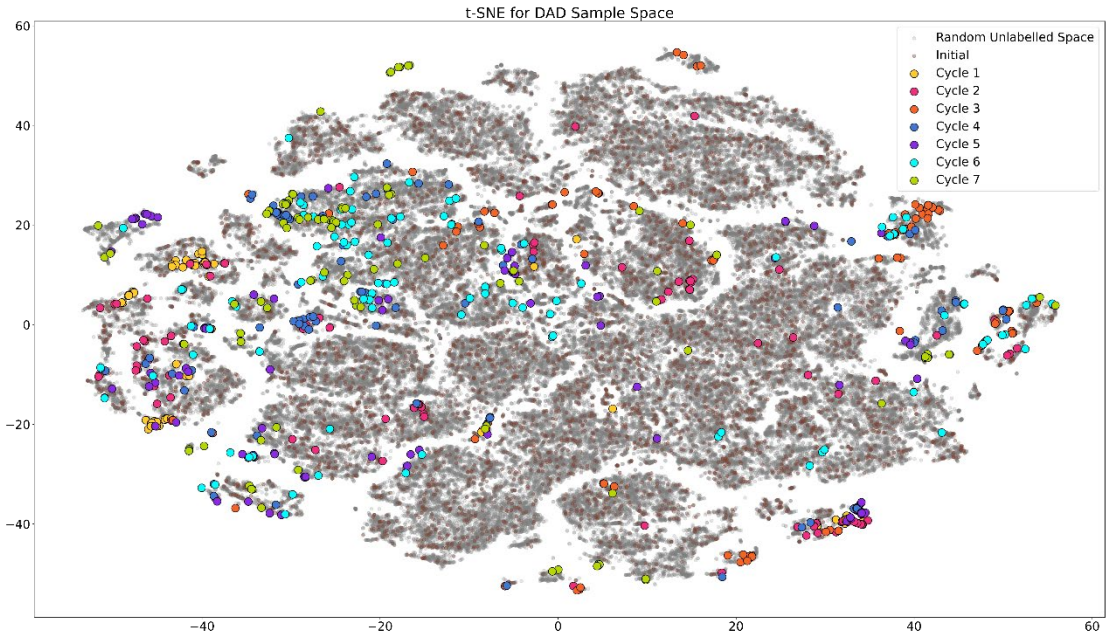



**Figure S2. t-SNE for DAD Sample Space and with Active Learning Progression**

For the visualizations of DA space (in **Figure 3d**), 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 5. 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 7. For the visualizations of the combined DA and DAD space along with the 4 recommended structures (in **Figure 4d**), 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**

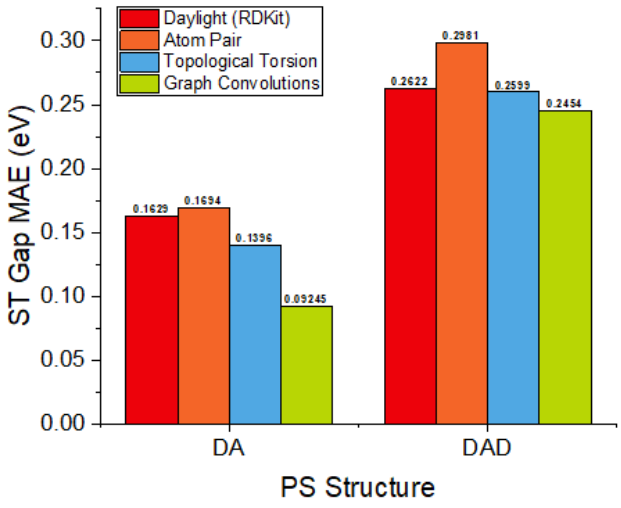



**Figure S3.** Comparison of model performances on initial dataset between graph-based deep learning method and traditional molecular fingerprints methods. As shown, for both DA and DAD form PSs, the graph-based deep learning method has shown better performance than traditional molecular fingerprints 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 neighbors 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]

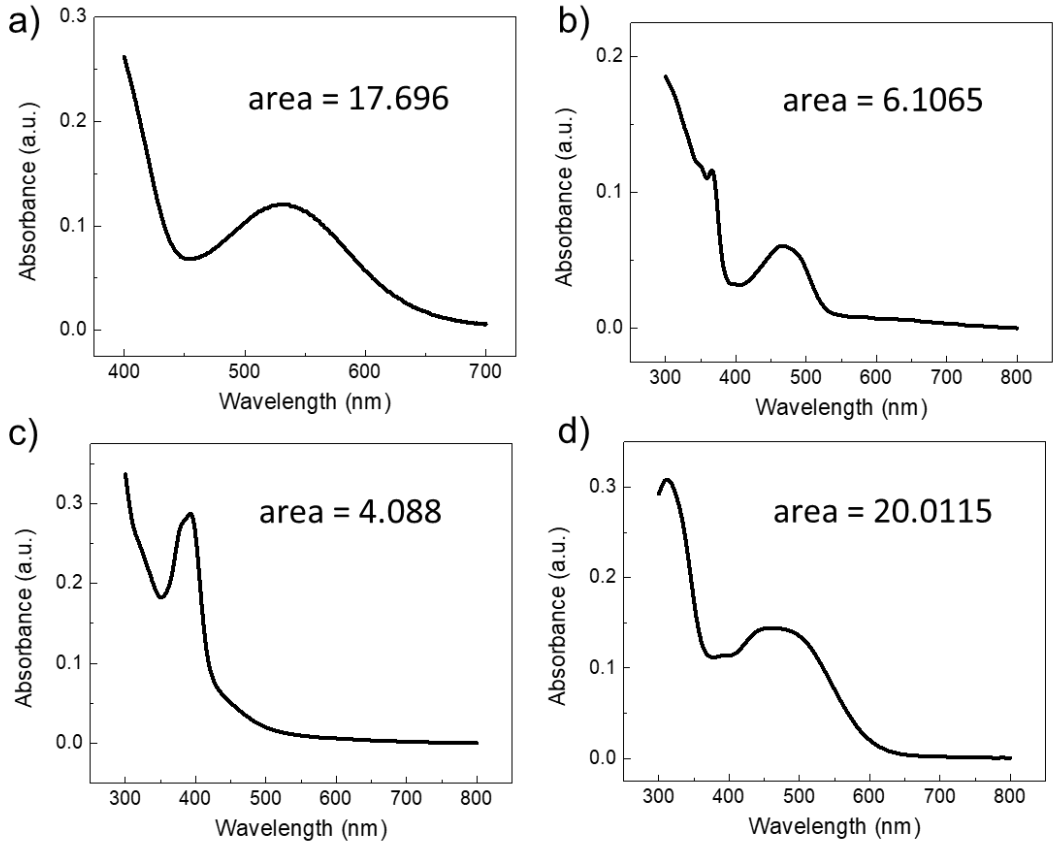



**Figure S4.** The absorption peak areas of **1-4** (a-d).

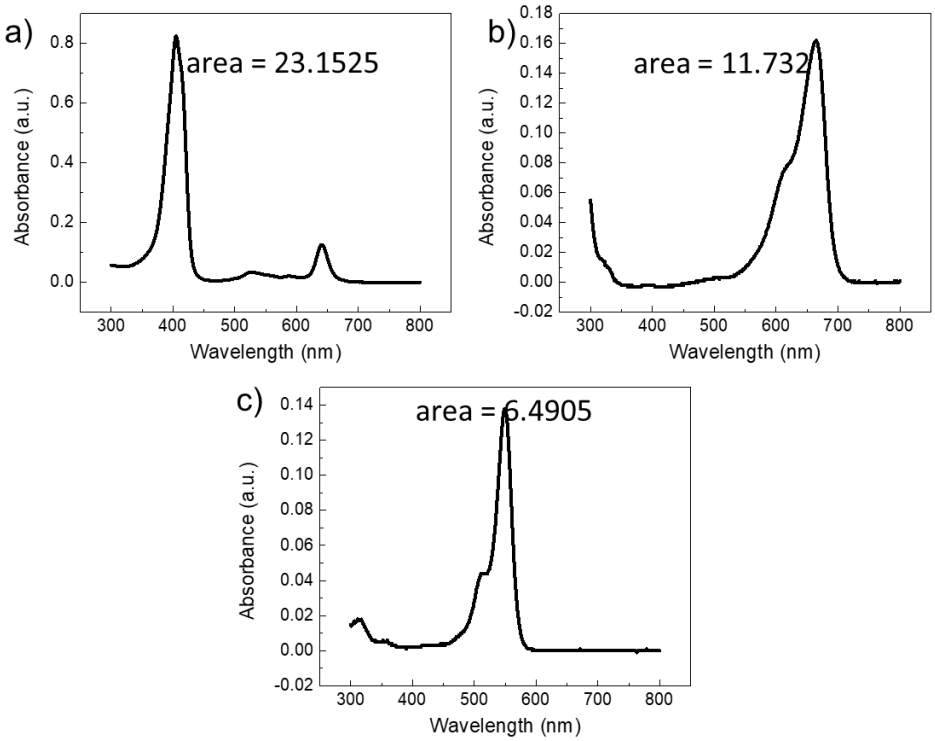



**Figure S5.** The absorption peak areas of Ce6, MB, and RB (a-c).

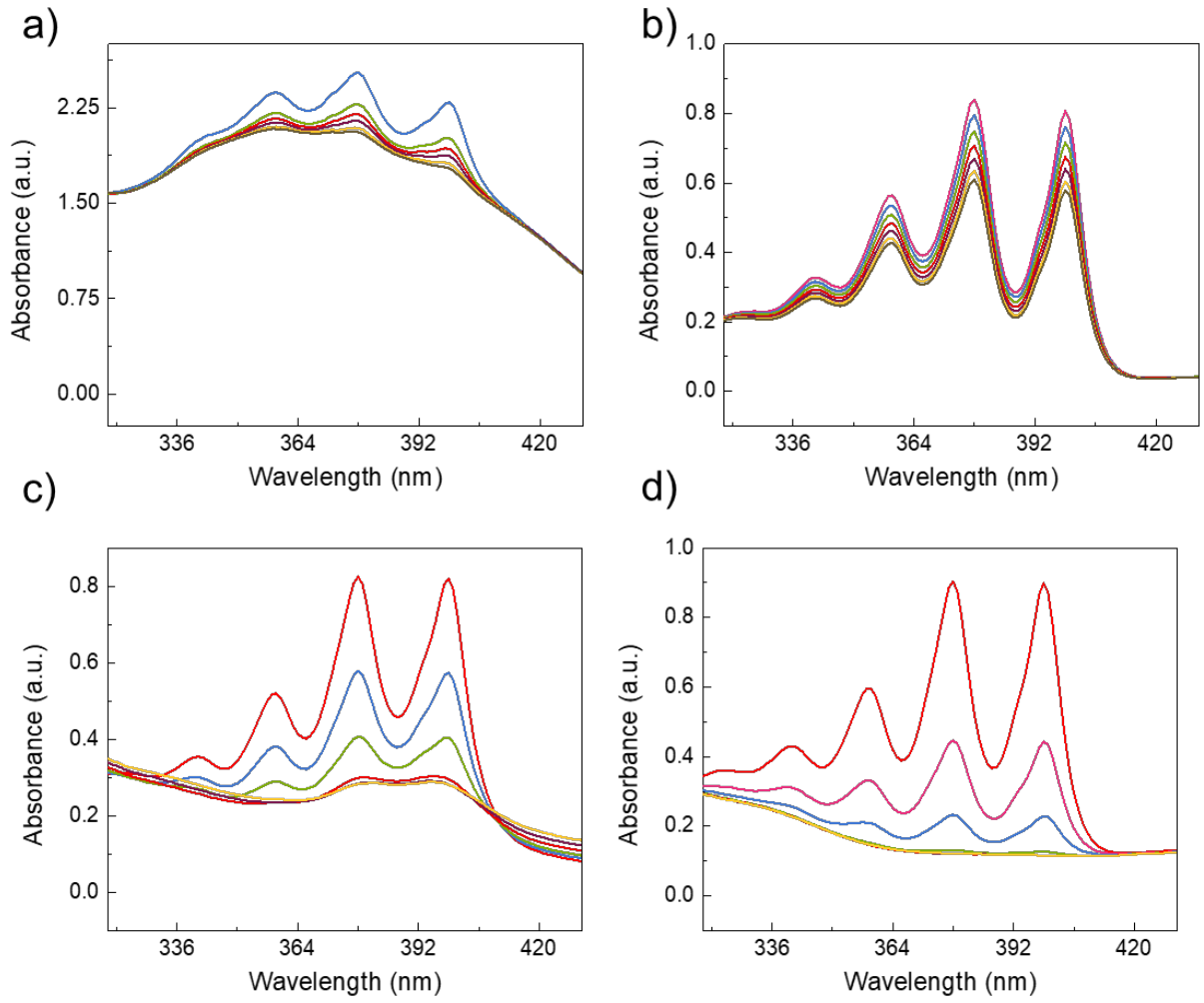



**Figure S6.** Photo-degradation of ABDA with **1-4** (a-d) in DMSO/water (v/v = 1/99) in five minutes, concentration of PSs: $5\times10^{-6}$ M.

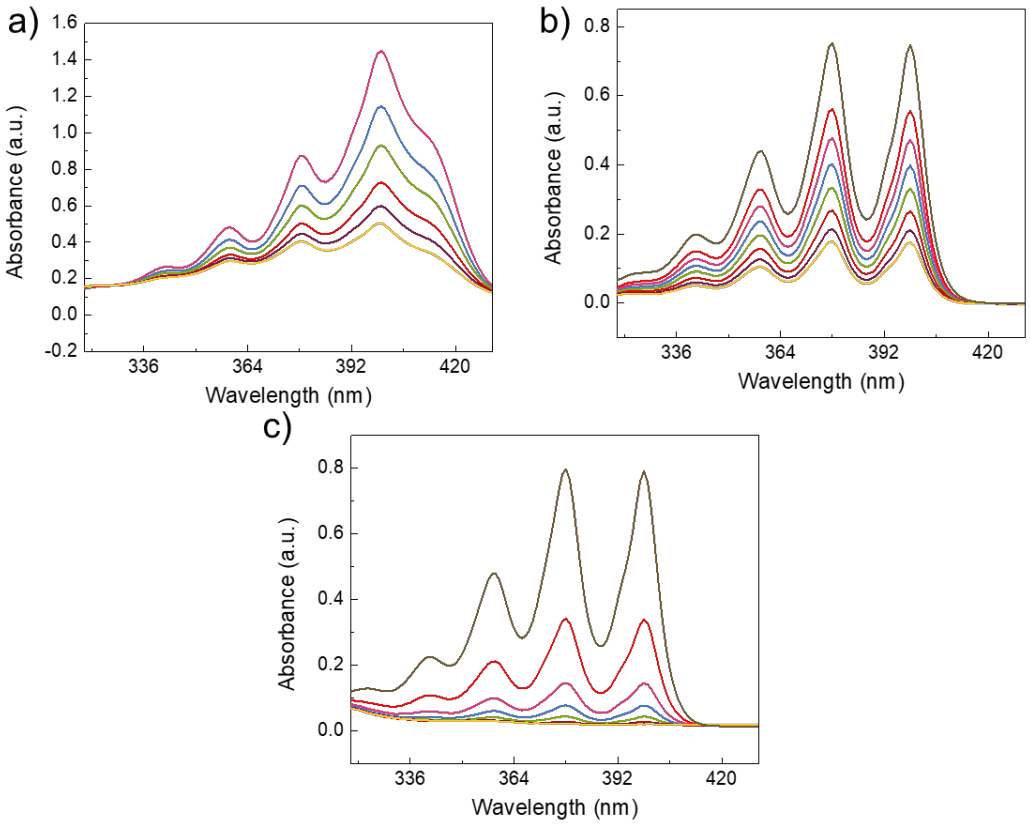
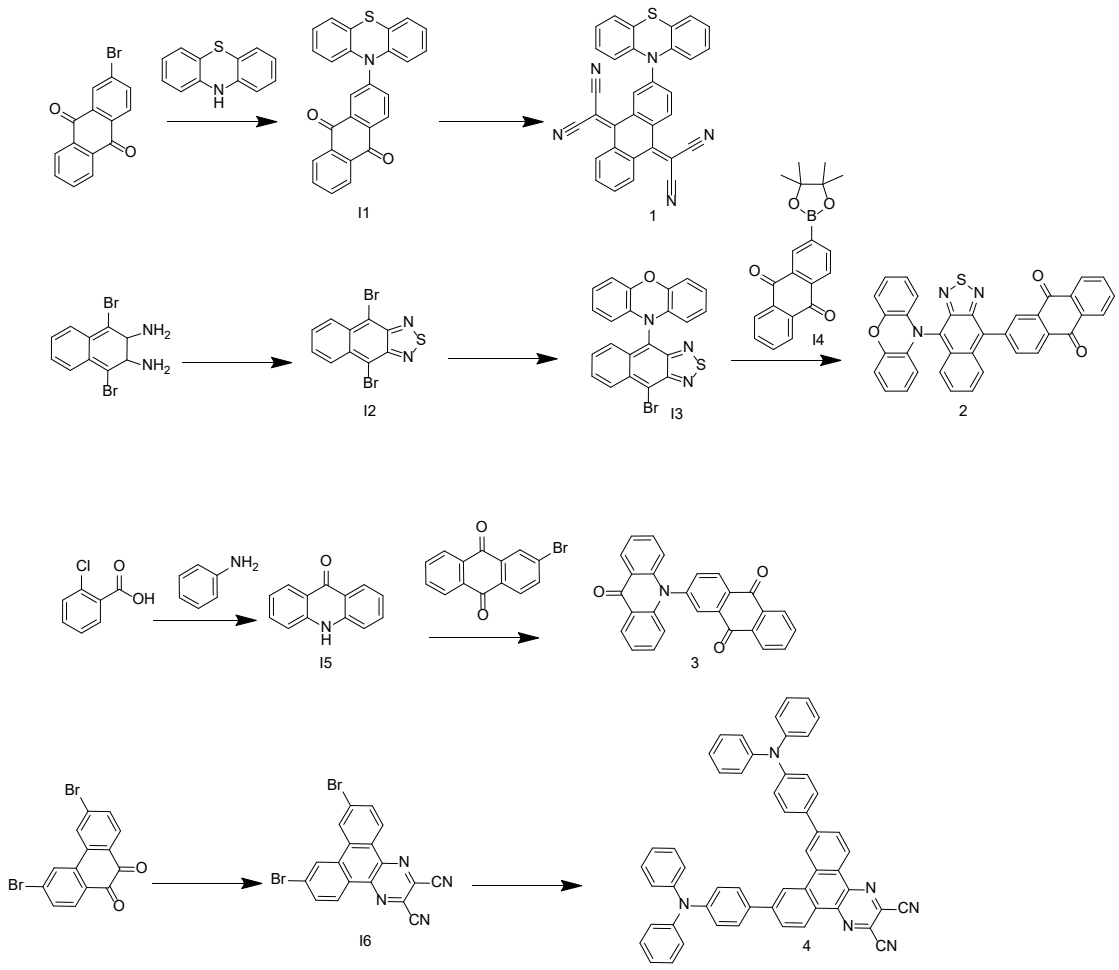
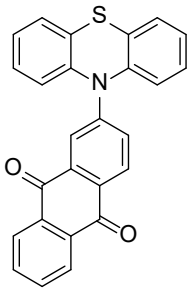



**Figure S7**. Photo-degradation of ABDA with Ce6, MB, and RB (a-c) in DMSO/water (v/v = 1/99) in five minutes, concentration of PSs: $5\times10^{-6}$ M.

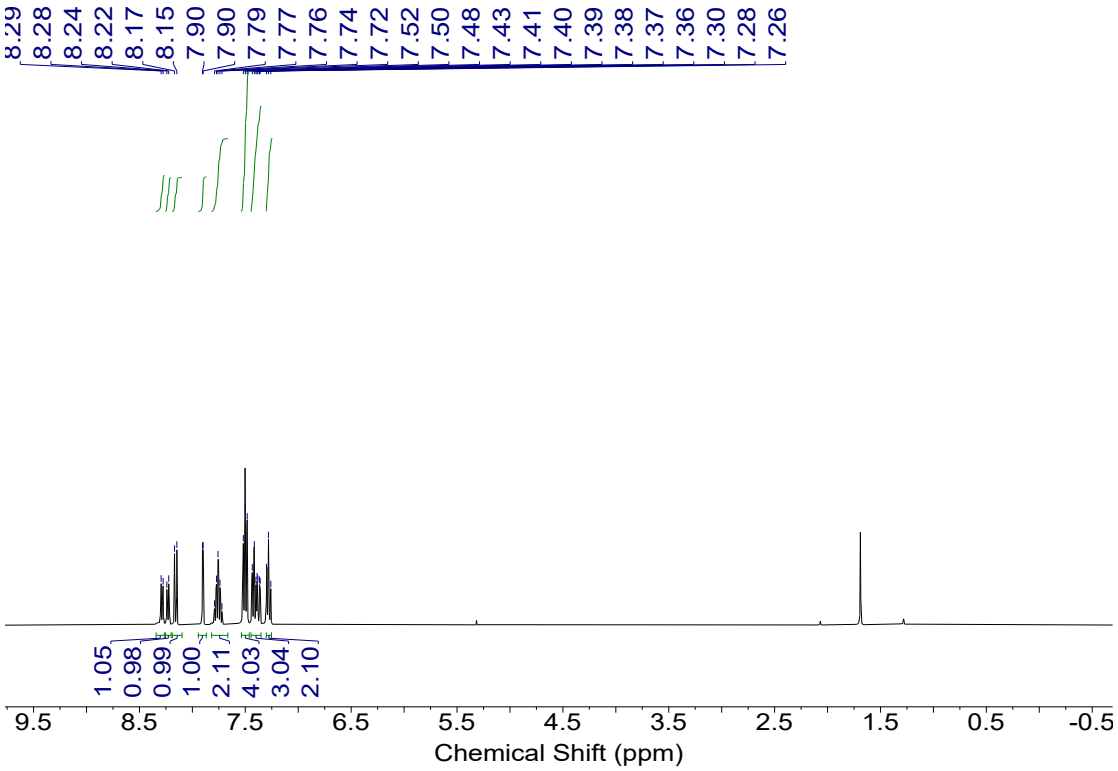
**Synthesis of 1-4.**

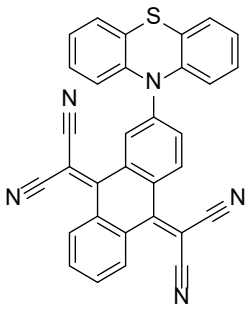



**Scheme S1.** The synthetic route towards compound 1-4.



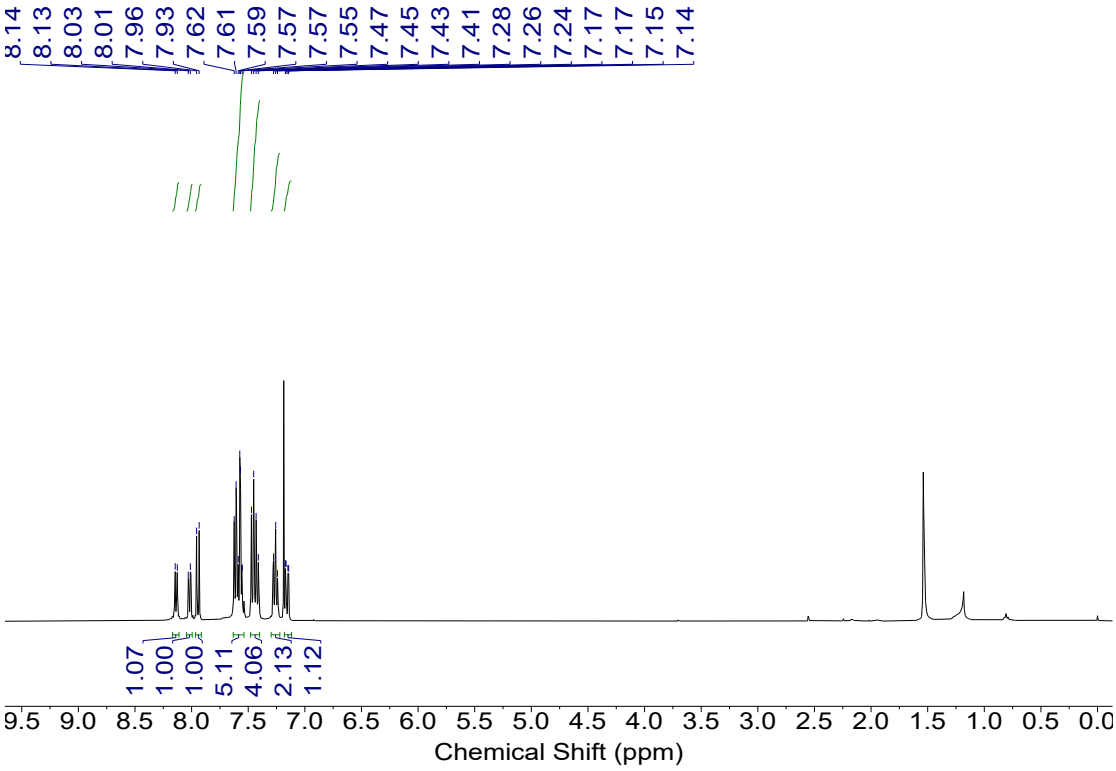
**Synthesis of I1.** A 100 mL round-bottom flask equipped with a magnetic stir bar was charged with 2-bromoanthracene-9,10-dione (375 mg, 1.29 mmol), phenothiazine (283 mg, 1.42 mmol), cesium carbonate (535 mg, 3.87 mmol) and toluene (15 mL). The solution was stirred at room temperature. After 10 min, 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, CDCl3) 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).

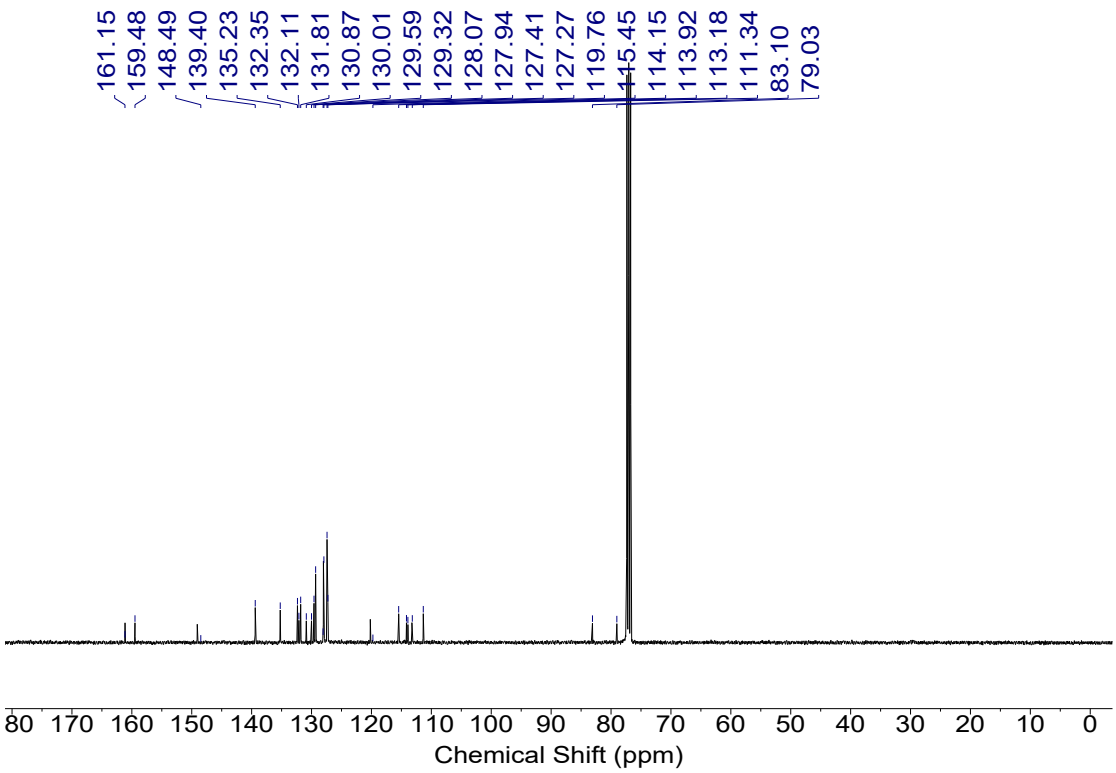


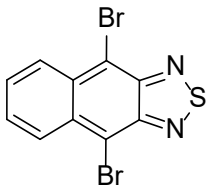




**Synthesis of 1.** To the solution of compound **I1** (40.5 mg, 0.10 mmol) and malononitrile (39.6 mg, 0.60 mmol) in dichloromethane (10 mL) was added titanium tetrachloride (0.08 mL, 0.7 mmol) slowly at 0 ºC. After the reaction mixture was stirred for 30 min, pyridine (0.06 mL, 0.7 mmol) was injected and stirred for another 30 min. Then the mixture was heated at 45 ºC 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, 63.5 % 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.





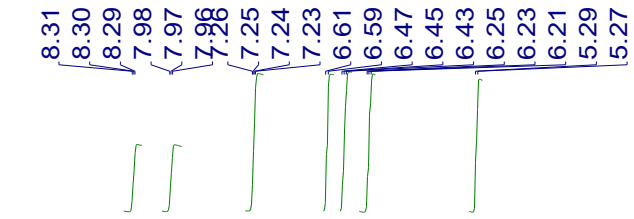



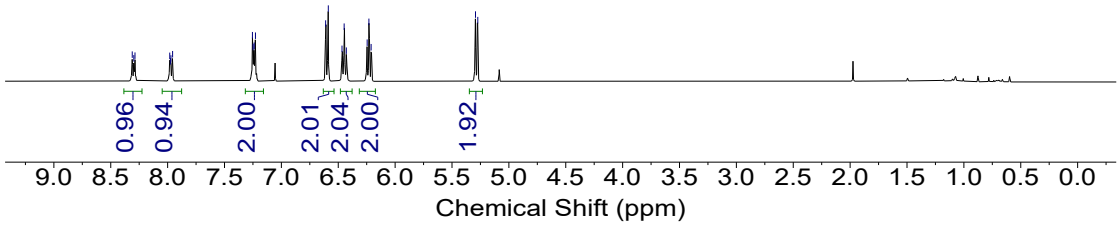

**Synthesis of I2**. In a 100 mL flask,1,4-dibromo-2,3-diaminonaphthalene (1.8 g, 5.5 mmol) was dissolved in 23 mL of anhydrous pyridine and then 1.31 mL (16.1 mmol) of thionylaniline and 7.0 mL (55 mmol) of chlorotrimethylsilane are added. The reaction was heated at 80 °C overnight with stirring. After the reaction 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).

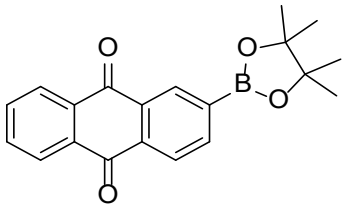




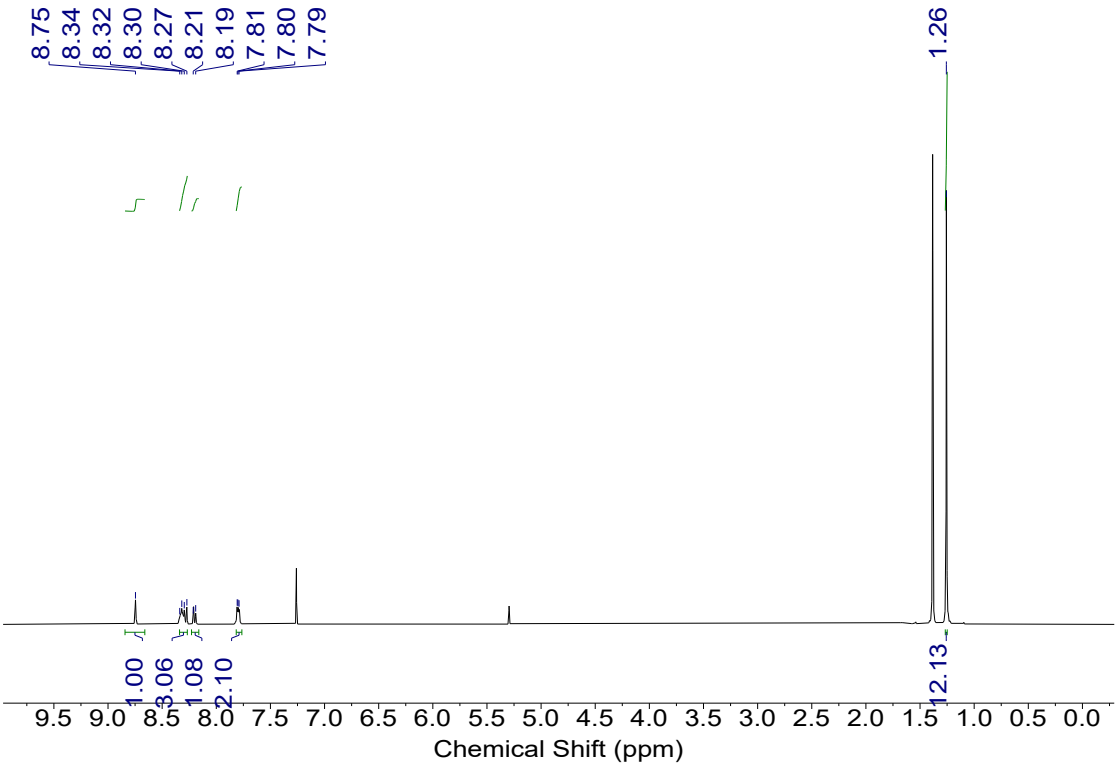
**Synthesis of I3**. A mixture of phenoxazine (0.5 g, 2.6 mmol), **I2** (400 mg, 2.4 mmol), palladium acetate (90 mg, 0.4 mmol), [(*t*-Bu)$_3$P]HBF$_4$ (348 mg, 1.2 mmol), and sodium *tert*-butoxide (0.92 g, 9.6 mmol) in 25 mL anhydrous toluene was stirred and reflux at 110 °C under argon atmosphere 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 removal of the solvent, 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).

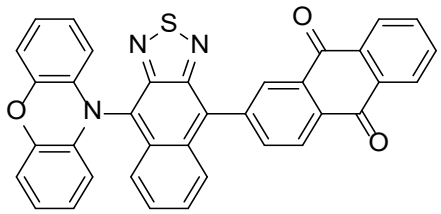


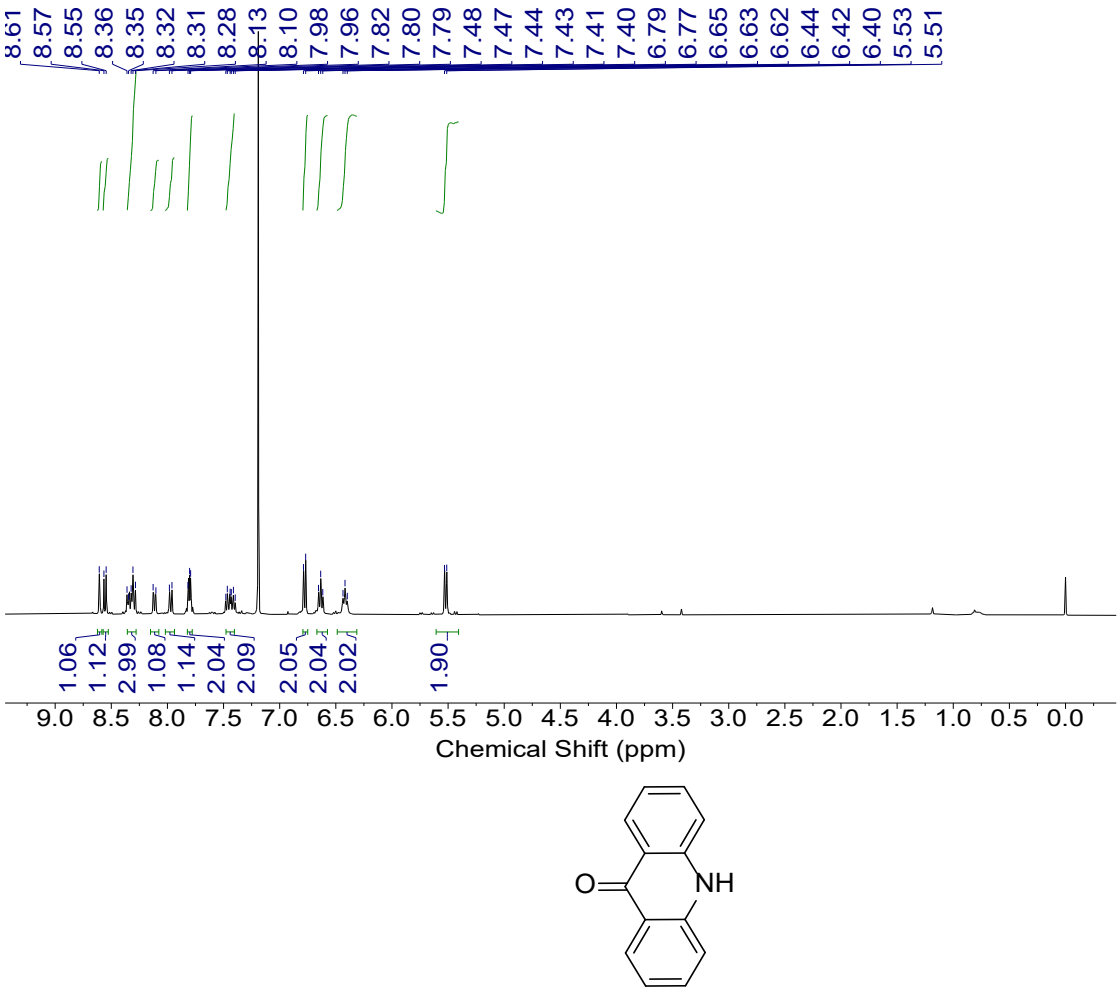

**Synthesis of I4**. 2-Bromoanthracene-9,10-dione (574 mg, 2.0 mmol), bis(pinacolato)diborane (1.02 g, 4.0 mmol), potassium acetate (687 mg, 7.0 mmol), Pd(dppf)Cl2 (73 mg, 0.23 mmol, dppf = 1,1'-bis (diphenylphosphanyl)ferrocene) and dioxane (20 mL) were mixed together in a 250 mL flask. After degassing, 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, 75.2% 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).




**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, 51.8% yield). $^1H$ 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).

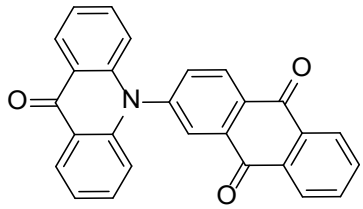




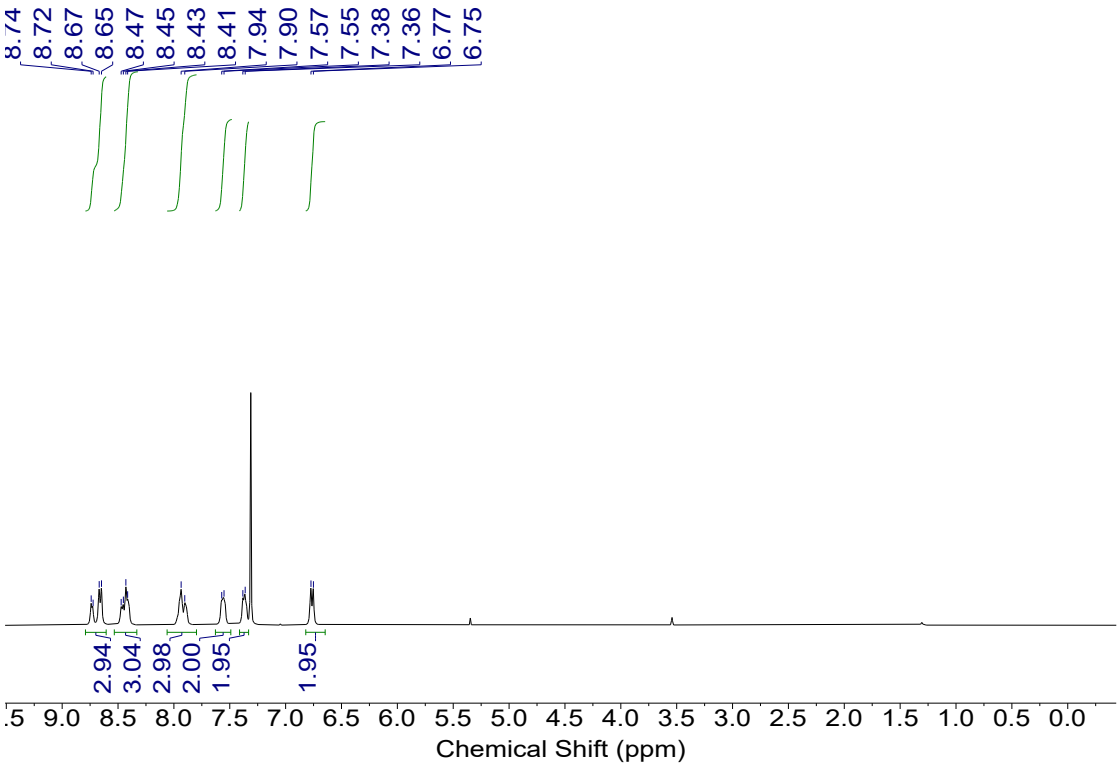


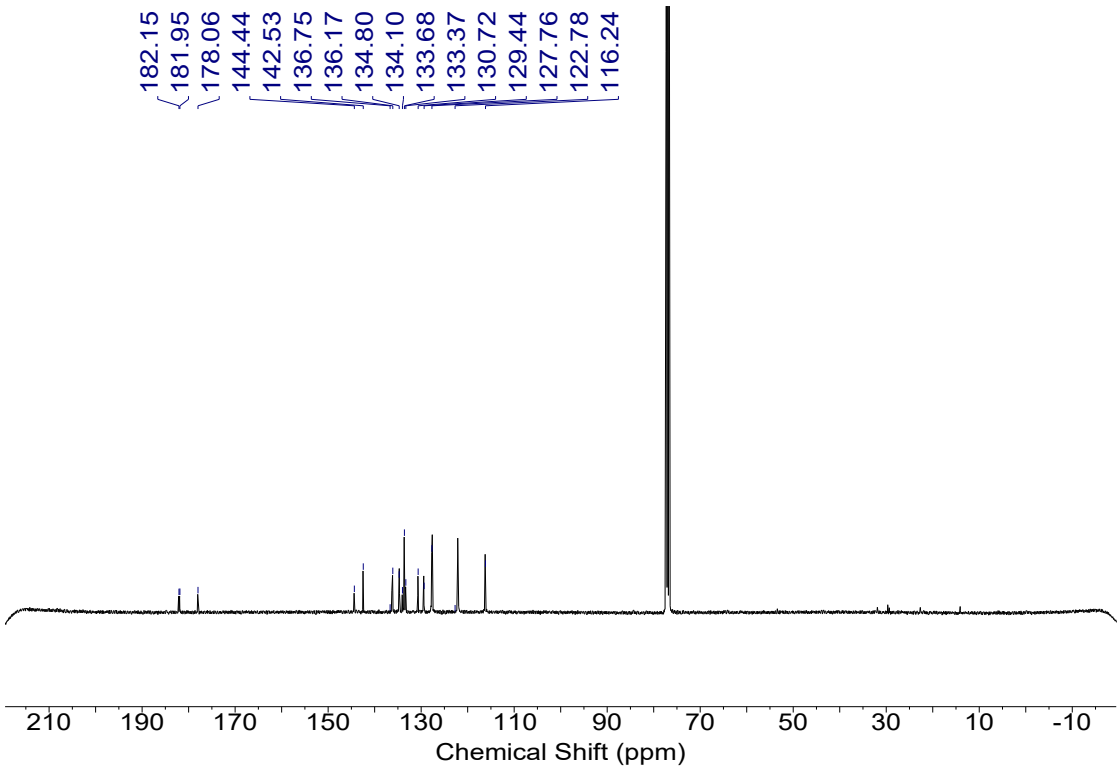
**Synthesis of I5.** *N*-phenylanthranilic acid (6.00 g, 27.8 mmol) was suspended in polyphosphoric acid (60 g) and heated to 120 ˚C in a round-bottom flask. which was equipped with a strong magnetic stirring bar. The dark green viscous mixture was also occasionally mixed thoroughly with a glass rod. 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

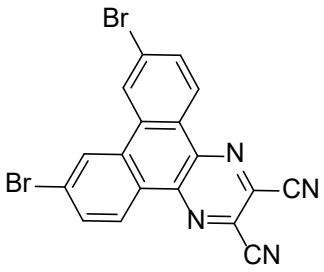
greenish yellow solid was dried in air overnight at 120 ˚C to give crude **I5** (5.4 g 27.8 mmol, 95 % yield), which was further used without purification.



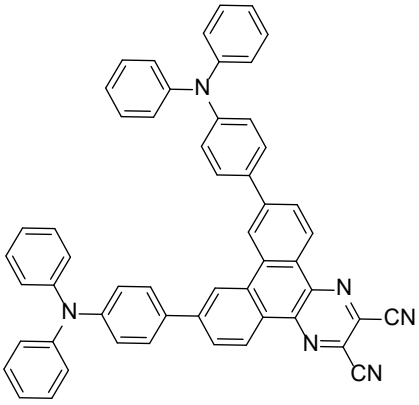

**Synthesis of 3**. A mixture of **I5** (256 mg, 2.6 mmol), 2-bromoanthracene-9,10-dione (700 mg, 2.4 mmol), palladium acetate (90 mg, 0.4 mmol), [(*t*-Bu)$_3$P]HBF$_4$ (348 mg, 1.2 mmol), and sodium *tert*-butoxide (0.92 g, 9.6 mmol) in 25 mL anhydrous toluene was stirred and reflux at 110 °C under 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.


8.74
8.72
8.67
8.65
8.47
8.45
8.43
8.41
7.94
7.90
7.57
7.55
7.38
7.36
6.77
6.75
2.94
3.04
2.98
2.00
1.95
1.95
.5 9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0
Chemical Shift (ppm)


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
210 190 170 150 130 110 90 70 50 30 10 -10
Chemical Shift (ppm)

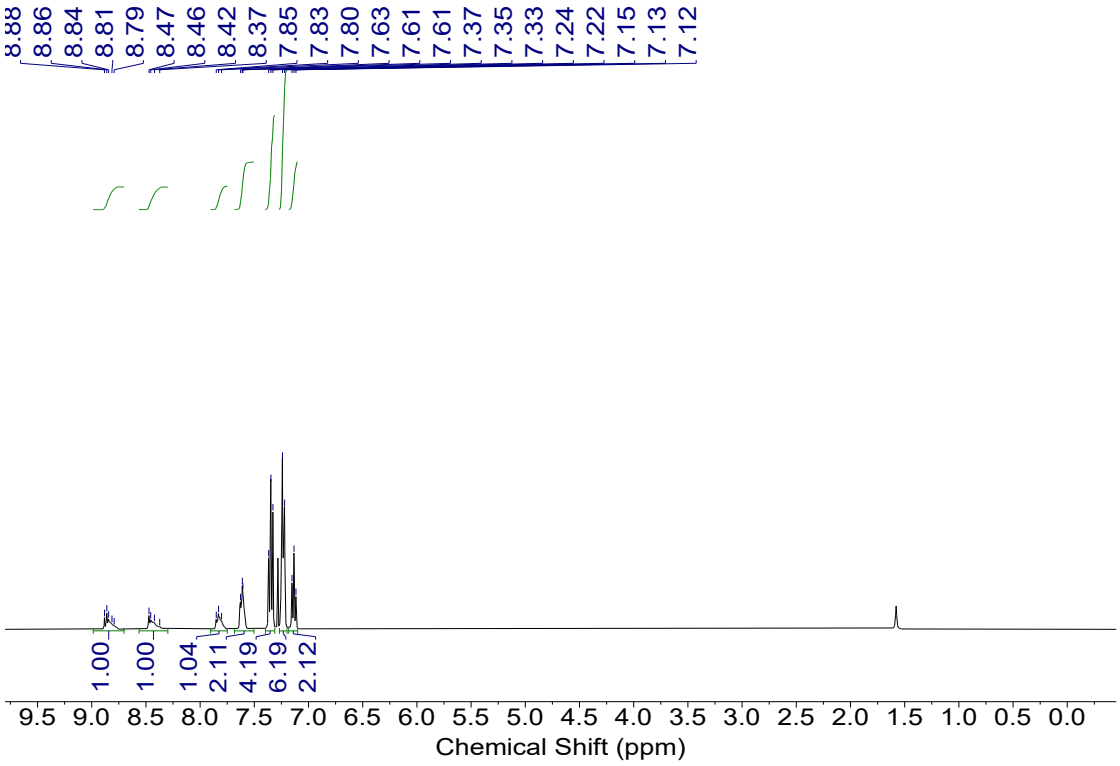


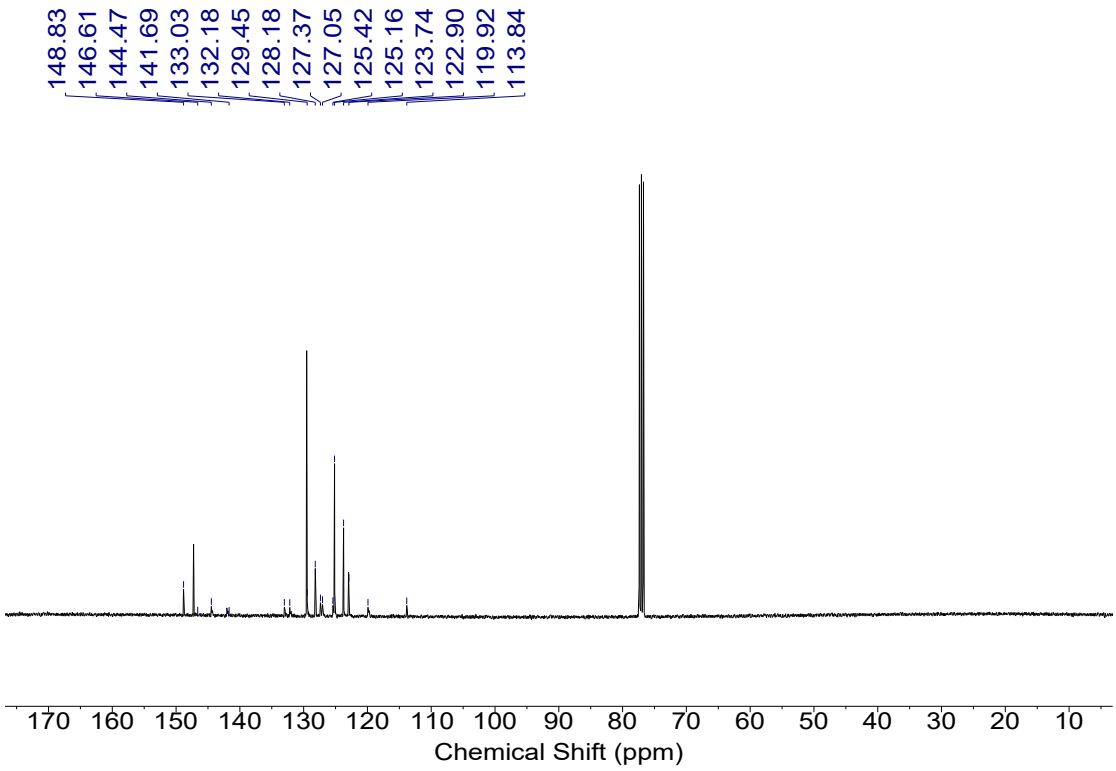

**Synthesis of I6.** A suspension of 3,6-dibromophenanthrene-9,10-dione (1.10 g, 3 mmol) and diaminomaleonitrile (0.32 g, 3 mmol) in acetic acid (10 mL) was heated to reflux for 8 hours. After cooling to room temperature, the resulting mixture was poured into ice water (100 mL) and then filtered. The 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 $^oC$ 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, 51.8% 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.






## 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).