Supporting Information

Data-driven discovery of near-infrared type I photosensitizers for RNA-targeted tumor photodynamic therapy

Wen Chen[1]*, Xiao-Qiong Mao[1], Xiao-Zhi Wang[2], Ya-Cong Liao[3], Xiao-Yue Yin[2], Hai-Long Wu[2], Tai-Yi Chen[1], Meng-Qing Liu[1], Tong Wang[2]*, Ru-Qin Yu[2]

[1]Key Laboratory of Functional Organometallic Materials of College of Hunan Province, College of Chemistry and Materials Science, Hengyang Normal University, Hengyang 421008, P.R. China

[2]State Key Laboratory of Chemo and Biosensing, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China

[3]Department of Ultrasound Diagnosis, The Second Xiangya Hospital, Central South University, Changsha 410011, P. R. China

Email: wenchen@hnu.edu.cn (Wen Chen); wangtong@hnu.edu.cn (Tong Wang)

Table of contents

1. Computational Section

1.1. Data featurization

The molecules in Data set 1 and Data set 2 are stored in the form of SMILES strings. RDKit was utilized to compute the Morgan fingerprints (radius = 3, nBits = 2048) and 17 physicochemical descriptors for each compound. Additionally, molecular graph representations were generated using RDKit as another input for the 1-PS-GCN model. The feature vectors for each atom in the molecular graph were also computed. The detailed physicochemical properties and node (atom) features are summarized in Table S1. Finally, all atomic features were concatenated into a feature vector with a dimensionality of 131.

1.2. 1-PS-GCN

In this study, we developed the 1-PS-GCN model by integrating graph convolutional network (GCN) with molecular Morgan fingerprints to accurately distinguish type I PSs from non-type I PSs. GCN is deep learning model commonly applied to graph-structured data, with their core mechanism relying on message-passing. This mechanism facilitates the exchange of information and the updating of features between connected nodes, enabling the model to learn high-dimensional node representations as well as the overall structural properties of the graph.

The molecular graph serves as one of the inputs to the 1-PS-GCN model. The molecular atomic features are refined using GCNConv, a graph convolutional layer. The operation of each graph convolutional layer is mathematically represented by the following equation:

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (1)$$

where σ indicates an activation function, such as ReLU; $H^{(l)}$represents the activation matrix of the l^{th} layer, $H^{(0)} = X$(feature matrix, $N \times F$); $W^{(l)}$denotes the learning weight matrix of the corresponding layer. $\hat{A} = A + I_N$, the design of **A** (adjacency matrix) plus $\mathbf{I}_N$ (identity matrix) is to retain its information when message passing. $\hat{D}$ is the diagonal degree matrix of $\hat{A}$, the role of $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is to standardize $\hat{A}$ and maintain the scale of the $H^{(l)}$.

After performing graph convolution, we applied global average pooling (GAP) to compute the average of all node features within each graph, thereby generating a global representation with a feature dimension of 64. The 64-dimensional molecular graph features were then concatenated with the 2048-dimensional Morgan fingerprints and 17 physicochemical property features, resulting in a 2129-dimensional feature vector. This concatenated vector was passed through two fully connected layers. In the first layer, the 2129-dimensional features were mapped to 128 dimensions, followed by a ReLU

activation function to introduce nonlinearity and a dropout operation (with a probability of 0.5) to improve the model's generalization. The second layer further reduced the 128-dimensional features to a 2-dimensional output, completing the classification task. Detailed hyperparameters and the optimization process are provided in Table S2.

1.3. Baseline models

Deep neural networks (DNN). DNN is a simple deep learning model, which is widely used to process complex data and solve nonlinear problems. DNN has multiple hidden layers between the input layer and the output layer. After the data is fed into the network, each neuron aggregates the information of the connected neurons, and then applies nonlinear activation function to process and transform the input. Through layer-by-layer processing, the DNN is able to capture increasingly abstract features, thereby effectively modeling complex input-output relationships. During the training process, the parameters (weights and biases) are adjusted to minimize the error between the output and the true label, and the weights and biases are updated using backpropagation and gradient descent algorithms to minimize the prediction error.

Convolutional neural networks (CNN). CNN is one of the most popular and widely used deep learning networks. Its main advantage is that it can automatically detect important features without any human supervision. CNN is classified as a multi-layer feedforward neural network. Its core design includes convolutional layers, activation functions, pooling layers, fully connected layers, etc. During training, CNN extracts local features of the image through convolutional layers, reduces the dimension using pooling layers, and performs classification or regression through fully connected layers. Its local connection and weight sharing characteristics reduce the number of parameters and improve training efficiency. This makes CNNs particularly well-suited for processing grid-structured data, such as images and videos, where they have demonstrated outstanding performance in computer vision tasks.

***k*-nearest neighbor (KNN).** KNN is an instance-based learning method for classification task. The core idea follows the principle that "similar instances are grouped together". By calculating the distance (e.g., Euclidean distance) between the unknown sample and each training sample, the closer the distance to the unknown sample is, the higher the weight of the sample. Then, the class label assigned to the unknown sample is determined based on the majority class among these k nearest neighbors. Although KNN is a simple and intuitive algorithm that does not require an explicit training phase, it may be computationally intensive during prediction and is highly sensitive to outliers. Therefore, selecting an appropriate k value and distance metric is critical to optimizing the performance of a KNN model.

Random forest (RF). RF is an ensemble learning algorithm based on the idea of bagging[S1]. It trains and predicts samples by constructing multiple decision trees (DTs). Each tree randomly selects samples and features during the training process, and finally establishes a strong regression or classification model suitable for different data sets by voting or averaging the prediction results of each decision tree. In the modeling process of RF, randomness is reflected in two key aspects: random sampling and random feature selection. By performing bootstrap sampling on the training data and randomly selecting feature subsets at each split, the diversity between RF-based learners can be improved, the risk of overfitting of a single decision tree can be reduced, and the generalization performance of the model can be enhanced.

Support vector machine (SVM). SVM is an efficient and powerful supervised learning algorithm used for classification, regression, and outlier detection[S2]. It maximizes the margin between two classes of samples by finding an optimal hyperplane, which is a boundary that is equidistant from the nearest positive and negative sample points (support vectors). Its versatility covers both linearly separable and inseparable scenarios. For linearly inseparable problems, the original data needs to be mapped to a high-dimensional space through kernel function techniques, making the originally linearly inseparable data separable in the high-dimensional space. SVM is called SVR when used for regression and SVC when used for classification. It is well suited to handle high-dimensional problems with less training data. Its ability to perform well relies on the principle of structural risk minimization, which makes it widely used in different fields such as structural reliability analysis and drug discovery.

1.4. Model evaluation

To evaluate the predictive performance of the model, we selected several effective metrics: accuracy (ACC), precision (Pre) and the area under the receiver operating characteristic curve (ROC-AUC). The equations for the first two metrics are expressed as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Pre = \frac{TP}{TP + FP} \tag{3}$$

where TP (true positive), TN (true negative), FP (false positive), and FN (false negative) represent the corresponding counts for each class. The values of these metrics range from 0 to 1, and higher values indicate better model performance.

In addition, we used the area under the receiver operating characteristic curve (ROC-AUC) to evaluate model performance. The ROC-AUC score is a commonly used metric for binary classifiers and represents the area under the ROC curve. The ROC curve plots the true positive rate (TPR) against

the false positive rate (FPR), illustrating the classifier's performance across different decision thresholds. The ROC-AUC score ranges from 0 to 1, where an ROC-AUC of 0.5 indicates poor performance (equivalent to random guessing), and an ROC-AUC of 1 signifies excellent performance with perfect classification. The formulas for calculating TPR and FPR are as follows:

$$TPR = \frac{TP}{TP + FN} \quad (4)$$

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

1.5. Multi-stage screening

In this study, we designed a multistage screening process to identify PSs with multiple desired properties. In the first stage, high-throughput screening of type I PSs was performed using the 1-PS-GCN model. In the second stage, to further refine potential type I photosensitizer candidates with good synthetic accessibility and low molecular weight, we used SAscore to evaluate the feasibility of molecule synthesis. The SA score ranges from 1 to 10, with 1 representing high synthetic accessibility (easy to synthesize) and 10 indicating low synthetic accessibility (difficult to synthesize). Additionally, molecular weight was considered a critical criterion, as smaller molecular weights often enhance biological activity and synthetic feasibility. In the context of tumor therapy, small molecules targeting RNA have demonstrated superior photodynamic efficacy in prior studies. Therefore, we employed RSAPred, an open-source machine learning tool, for the third stage of screening. This tool calculates various features based on RNA sequences and small molecule structures and uses a multiple linear regression (MLR) algorithm to construct a regression model that predicts the binding affinity between small molecules and six RNA subtypes. The binding affinity is reflected by the log-scale dissociation constant (pKd) for each RNA-small molecule pair. Using this tool, each candidate molecule and a widely prevalent RNA G4 sequence in cancer cells (Terra 5'-UUAGGGUUAGGGUUAGGGUUAGGG-3') were input into the model. The resulting pKd values were ranked to identify type I PS candidates with the highest potential for RNA targeting.

1.6. t-SNE and visualization of chemical space

t-SNE (t-distributed stochastic neighbor embedding) is a widely used dimensionality reduction technique for visualizing high-dimensional data in low-dimensional spaces. It is particularly effective for uncovering and analyzing complex patterns and structures within datasets. The method works by modeling the similarity between data points in high-dimensional space and their counterparts in low-dimensional space through the construction of similar probability distributions in both spaces and

minimizing their divergence. This process enables the effective mapping of data points into a lower-dimensional space for visualization. In this study, t-SNE was applied to the Morgan fingerprints of molecules to visualize their chemical space.

1.7. Model interpretability method

To further understand the decision-making process of the deep learning model and open the "black-box" of deep learning, we employed a series of interpretability methods to analyze the internal working mechanisms of the 1-PS-GCN model. Since 1-PS-GCN is a hybrid model combining molecular fingerprint and GCN, we separately investigated the feature extractors of these two parts.

To investigate the contribution of molecular fingerprint features to the model's predictions, we employed the Integrated Gradients method for interpretability analysis on the fully connected layers. This technique quantifies the influence of input features on prediction outcomes by calculating the gradient-weighted average, resulting in an importance score for each feature dimension. The contributions of individual fingerprint features were visualized using bar charts (Fig. 2d). Additionally, to explore the connection between fingerprint features and molecular structural information, we utilized the Draw.DrawMorganBit function in the RDKit package to depict the molecular structures corresponding to the top three most important fingerprint features (Fig. 2e).

For interpreting molecular graph features, we adopted GNNExplainer[S3], a graph neural network visualization method. GNNExplainer is specifically designed for graph neural network (GNN) and generates explanations for any GNN and graph mining task by analyzing both network structures and node attributes. In this study, we set the training iterations for GNNExplainer to 100 and the learning rate to 0.001 to effectively capture the importance of graph structures and node features. Finally, the top 10 most important features identified by the model were visualized using bar charts (Fig. 2f).

1.8. Expert-AI consensus to select best candidate

Top 10 molecules were shown in Fig. S2. PYD was identified through an expert-AI consensus. Literature analysis revealed that the absorption wavelengths of compounds No. 2, 5, 7, 8, and 9 are all below 600 nm. Structurally, compound No. 6 is not resistant to photobleaching. Although compounds No. 1, 3, and 4 meet the criteria for NIR absorption and photobleaching resistance, their synthesis routes are relatively complex. Ultimately, compound No. 10 (PYD) was chosen as the optimal candidate for synthesis and performance validation due to its rigid structure, which confers resistance to photobleaching. Additionally, its structural similarity to the commercial RNA detection reagent Pyronine highlights its strong potential as a high-performance photosensitizer.

1.9. Molecular docking and density functional theory (DFT) calculations

Molecular docking. Nucleic acid crystal structures for docking were obtained by RNAComposer server prediction (https://rnacomposer.cs.put.poznan.pl/). The nucleic acid structures were processed using PyMol 2.5.5 before docking began[S4, 5]. All processed small molecules as well as nucleic acids were converted into the PDBQT format required for AutoDock Vina 1.1.2 docking, which was done using ADFRsuite 1.0[S6]. For docking, the global search thoroughness was set to 32 and the remaining parameters were left at their default settings. We considered the highest scoring output docking conformation to be the binding conformation and finally visualised and analysed the docking results using PyMol 2.5.5.

DFT calculations. a Gaussian package was used for all density functional theory calculations. The ground state geometries for PYD was optimized at the B3LYP/6-311++G(d,p) level. The vertical excitation (UV-vis absorption) of PYD was obtained based on the optimized ground state geometries. Geometry optimizations were performed at TD-B3LYP/B3LYP/6-311++G(d,p) for the singlet (S1) and triplet (T1) excitated states of the compounds. The solvent (water) effect was included in all calculations using the solvation model based on the density (SMD). The spin-orbit coupling (SOC), <SnIHSOITm>, between singlet and triplet excited states for PYD was calculated by PYSOC program[S7].

2. Experimental Section

2.1. Reagents and instruments

6-amino-1,2,3,4-tetrahydronaphthalen-1-one and methyl iodide were purchased from Titan Scientific. (Shanghai, China). Phosphorus oxychloride, sulfuric acid and perchloric acid were obtained from Sinopharm Chemical Reagents Co., Ltd. Organic solvents of analytical grade were obtained from Energy Chemical. MCF-7 cells (Human breast cancer cell line) and 4T1 cells (Mouse breast cancer cell line) were obtained from the cell bank of Central Laboratory at Xiangya Hospital (Changsha, China). RPMI 1640 medium, DMEM high glucose medium, penicillin, streptomycin and 10% heat-inactivated fetal bovine serum were purchased from Thermo Fisher (MA, USA). Thin-layer chromatography (TLC) was performed on silica gel aluminum sheets with an F-254 indicator. The column chromatography was conducted using 200-300 mesh SiO_2 (Shanghai Haohong Biomedical Technology Co., Ltd.). Dihydrorhodamine 123, 9,10-anthracenediyl-bis(methylene)-dimalonic acid (ABDA), Singlet Oxygen Sensor Green (SOSG) were obtained from Life Technologies. 2,7-Dichlorodihydrofluorescein diacetate (DCFH-DA), 2,7-Dichlorodihydrofluorescein (DCFH) and 3'-(4-hydroxyphenyl) fluorescein (HPF, Molecular Probes Invitrogen) were purchased from Sigma-Aldrich. Dihydroethidium (DHE), Hoechst, Annexin V-FITC apoptosis detection kit, RNase A and DNase I were purchased from Beyotime biotechnology. Calcein-AM/propidium iodide (PI) Double

Stain Kit was obtained from KeyGen Biotech. All nucleic acids were purified by high performance liquid chromatography and obtained from Sangon Biotech (Shanghai, China). The nucleic acids were dissolved in Tris-HCl buffer (10 mM, pH 7.4) and their concentrations were determined based on absorbance at 260 nm. For dissolving RNAs, the solutions were treated with diethyl pyrocarbonate. Stock solutions of **PYD** (10 mM) were dissolved in DMSO and stored at -20°C. Ultrapure water with an electric resistance >18.25 MΩ was obtained from a Millipore Milli-Q water purification system (Billerica, MA, USA) and used throughout the experiments.

^{1}H NMR and ^{13}C NMR spectra were recorded on a Bruker Avance-III 400 instrument (Bruker) using tetramethylsilane (TMS) as an internal standard. High-resolution mass spectrometry analysis was performed on LCQ advantage ion trap mass spectrometry (Thermo Fisher Scientific, Bremen, Germany). UV-Vis absorbance spectra were recorded on UV-2450 spectrophotometer (Shimadzu, Japan) with an interval of 2.0 nm. Fluorescence spectra were recorded on an FLS1000 spectrofluorometer (Edinburgh Instruments, United Kingdom) with excitation and emission slits of 5.0 nm. Fluorescence lifetime (τ) measurements were carried out with a time correlated single photon counting (TCSPC) nanosecond fluorescence spectrometer (Edinburgh FLS920, United Kingdom) at ambient temperature (298 K). Electron spin resonance (ESR) spectra were performed on a JES-FA200 spectrometer (JEOL, Japan). Confocal fluorescence imaging was performed on a Nikon A1+ confocal microscope (Nikon, Japan) with 20× or 100× objective lens. In vivo fluorescence imaging was performed on an IVIS Lumina XR small animal imaging system (Caliper, Switzerland).

2.2. In vitro Assays

Spectral experiment. For absorption measurements, PYD was mixed with or without RNA (dissolved in DEPC) in Tris-HCl buffer. Fluorescence spectra of PYD with or without RNA were recorded in Tris-HCl buffer using an excitation wavelength of 640 nm. In selectivity studies, PYD was co-incubated with nucleic acids of different structures (Table S5) including ssDNA (single stranded DNA)), dsDNA (double strand DNA), dsRNA (double strand RNA), DNA G4s, RNA G4s, ssRNA (single stranded RNA), ions, amino acid and reactive oxygen and fluorescence spectra were recorded. In fluorescence titrations, PYD was incubated with different concentrations of RNA and fluorescence spectra were obtained. For photostability studies, PYD incubated with or without RNA in buffer were irradiated on a Quanta Master 8000 for 2 h and fluorescence intensities at 690 nm recorded in real time. For reference, cyanine 5 (Cy5) was also exposed to the same irradiation conditions, and the fluorescence intensities at 660 nm were documented.

Determination of ROS in vitro. To determine the ability to produce ROS, PYD (10 μM) with or without RNA were incubated with DCFH (10 μM). For $^1O_2$ detection, PYD with or without RNA were incubated with ABDA (50 μM in DMSO) or SOSG (10 μM in MeOH). For O_2^{-}• and OH• detection, PYD with or without RNA G4s were incubated with DHR123 (10 μM) and HPF (10 μM), respectively. The samples were irradiated with an LED light (660 nm, 10 mW/cm^2, 10 min). For dynamic monitoring of ROS production, the samples were irradiated for different times and fluorescence or absorption spectra for the corresponding indicators were acquired. Excitation wavelength for DCFH: 504 nm, emission wavelength: 529 nm. Excitation wavelength for SOSG: 490 nm, emission wavelength: 525 nm; Excitation wavelength for DHR123: 488 nm, emission wavelength: 526 nm.

2.3. Cellular Studies

Cellular assays. MCF-7 cells and 4T1 cells were obtained from Xiangya Hospital (Changsha, China) and cultured in configured medium of DMEM or RPMI-1640. After MCF-7 cells were cultured in cell confocal dishes (1×10^5 cells) for 12 h, cells were washing three times with PBS and incubated with PYD (5.0 μM) in culture medium for 30 min at 37℃ before imaging. For enzyme digestion imaging experiments, MCF-7 cells were first fixed with 4% paraformaldehyde and then permeabilised with PBS containing Triton X-100, stained with PYD (5.0 μM) and hoechst for 30 min, and washed again twice with PBS. Finally**,** cells were treated with DNase I (100 U/mL), RNase A (20 μg/mL) or PBS for 4 h before fluorescence imaging. PYD: (Ex:640 nm, Em: 660 nm-750 nm); Hoechst: (Ex: 405 nm, Em: 430 nm-490 nm).

Intracellular detection of ROS. To determine the ability of PYD to generate ROS in cells, MCF-7 cells were pretreated with or without PYD (5.0 μM) for 30 min, followed by incubation with 10 μM of DCFH-DA, DHE or HPF for 30 min. The cells were washed with 1×PBS and irradiated with an LED light (660 nm, 100 mW/cm^2, 10 min). To test the ability of PYD to produce ROS in hypoxic conditions, the cells were cultured in 2% oxygen atmosphere, treated with PYD and incubated with DCFH-DA and irradiated. To quench the effect of ROS, the cells were pretreated with N-acetylcysteine (NAC, 5.0 mM) for 2 h before staining with PYD and DCFH-DA. To inhibit the activity of superoxide dismutase (SOD), the cells were treated with 2-methoxyestradiol (2ME2, 5.0 μM). To quench the effect of O_2^{-}•, pre-treated the cells with vitamin C (50 μM) for 2.0 h prior to staining with PYD and DHE. DCFH-DA and HPF: Ex:488 nm, Em: 530 ± 15 nm; DHE: Ex:488 nm, Em: 600 ± 30 nm.

Cytotoxicity studies. The cytotoxicity of PYD was assessed under both light and non-light conditions using the WST-1 cell proliferation assay. Various cell lines were seeded into 96-well cell culture plates and incubated at 37°C for 24 h. The cells were then treated with different concentrations of PYD (0,

0.25, 0.5, 1.0, 2.0, 5.0 μM) for 30 min, washed, and transferred to fresh medium. They were further incubated for 6 h under normoxic or hypoxic conditions and subsequently irradiated with a 660 nm LED at 100 mW/cm² for 10 min. After irradiation, the cells were allowed to incubate for an additional 24 h. Cell viability was evaluated by adding WST-1 reagent and incubating the cells for 4 h. Absorbance was measured at 450 nm using a microplate reader.

Live/dead cell assay. MCF-7 cells were seeded and grown in confocal cell culture dishes for 24 h. Cells were treated with PYD (5.0 μM) for 30 min at 37°C followed by illumination with and without illumination. Before confocal fluorescence imaging cells were co-stained with calcein AM and PI Calcein. AM Ex: 488 nm, Em: 515 ± 10 nm; PI Ex: 561 nm, Em: 625 ± 15 nm.

Flow cytometric assay. To investigate the cell killing mechanism of PYD, MCF-7 cells were cultured in cell plates for 24 h. The cells were incubated with PYD (5.0 μM) for 30 min, replaced with a fresh medium and irradiated for different time intervals (0, 2, 5, 8, 10 min). The treated cells were incubated for another 4 h, and costained with Annexin V-FITC and PI.

2.4. Vivo Studies

Tumor in vivo model. The animal studies were all authorised by the Hunan Provincial Department of Science and Technology and passed the animal ethical review (Ethics No. CSU-2024-0144). Female BALB/c mice were provided by the Animal Breeding Room of Xiangya Medical College. To establish a mouse subcutaneous transplantation tumor model, 4T1 cells (5.0×10^6 cells) were implanted subcutaneously into the abdomen of each mouse. Before imaging or treatment, the tumors grew to approximately 80 - 90 mm^3. Tumor size was calculated as follows: V = 1/2 × length × width × width.

Blood circulation analysis. Healthy female Balb/c mice (n = 3) were intravenously injected with PYD (50 μM, 50 μl), blood samples were collected at different time points via the orbital venous plexus and incubated at 4 °C. Fluorescence spectra for PYD were recorded. The blood concentrations of PYD at different times were determined based on their standard curves. Pharmacokinetic parameters were calculated as following [S8]:

$C = A'e^{-\alpha t} + B'e^{-\beta t} - C'e^{-k_a t}$

C is the concentration of the PYD at a specific time, t is the specific time, α is distribution rate constant, β is elimination rate constant. When t is long enough, $k_a \gg \alpha$, $\alpha \gg \beta$. So , the above equation is simplified as follows:

$e^{-\beta t} \rightarrow 0$, $e^{-ka} \rightarrow 0$

$C = B'e^{-\beta t}$, so $\lg C = \lg B' - 0.4343\ \beta t$

A semi-logarithmic graph was plotted for lg C *vs* t, and the elimination rate constant (β) was calculated based on the slope. For PYD, β= 0.1911, so $t_{1/2}$ = 0.693/β = 3.6 min

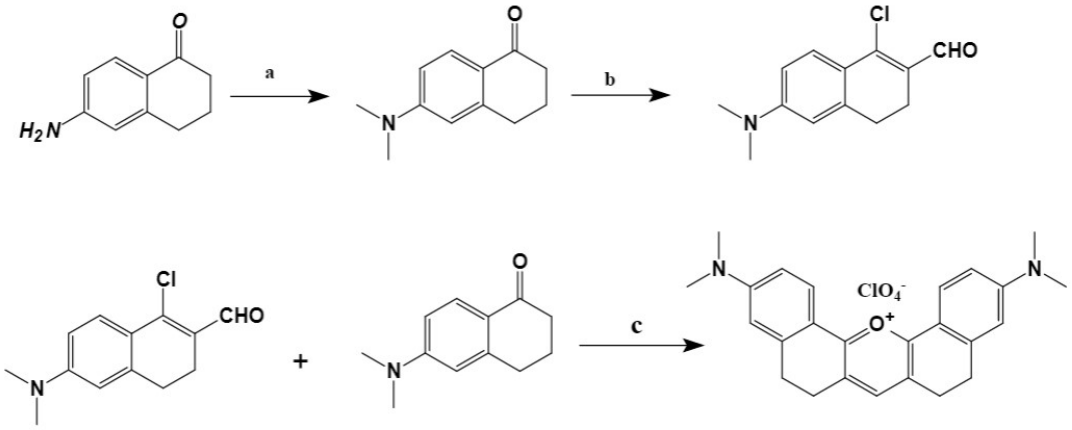
In vivo tumor treatment via PDT. To investigate the antitumor efficacy of PYD-mediated PDT, tumor-bearing mice were randomly divided into four groups (n = 4). Group I and group II received intratumoral injections of PBS (10 μL), while group III and group IV were injected with PYD (10 μL). Both groups were injected every two days for a total of three times. Mice in group II and group IV were irradiated with LED light (660 nm, 100 mW/cm², 10 min) 1.0 h post-injection. The tumor sizes were measured every other day for 14 days, and on day 15, the mice were euthanized. The tumors and major organs (hearts, livers, spleens, lungs and kidneys) were harvested. Fluorescence images were acquired with the IVIS imaging system. For histological analysis, blood biochemistry analysis, the tumor tissues and major organs were excised on day 14 after treatment. The excised organs and tumors were fixed with 4% formaldehyde, embedded in paraffin, sliced into 5.0 μm in thickness and stained with haematoxylin and eosin (H&E).

2.5. Synthesis of compounds

6-**(dimethylamino)-3,4-dihydronaphthalen-1(2H)-one**: Methyl iodide (3.0 eq) was added to a solution of 6-amino-3,4-dihydro-1(2H)-naphthalenone (1.0 eq) in anhydrous ethanol, potassium carbonate (1.0 eq) as a catalyst. The mixture was refluxed for 12 hours. Subsequently, the product was filtered and washed with diethyl ether (Et_2O), then evaporated to yield the crude product. The crude product was purified using silica gel chromatography. (EA: PE = 1: 10) to obtain 6-(dimethylamino)-3,4-dihydronaphthalen-1(2H)-one as a white solid (58 %). ^{1}H NMR (400 MHz, DMSO-d6) δ (ppm): 7.704 (d, J = 8.8, 1H), 6.640-6.636 (m, 1H); 6.475 (d, J = 2.4, 1H); 2.995 (s, 6H); 2.821 (t, J = 6.0, 2H); 2.434(t, J = 6.8, 2H); 1.982-1.920 (m, 2H). ^{13}C NMR (100 MHz, DMSO-d6): 195.70, 153.74, 146.71, 128.80, 121.39, 110.53, 109.84, 40.03, 30.22, 23.64.

1-**chloro-6-(dimethylamino)-3,4-dihydronaphthalene-2-carbaldehyde**: Phosphoryl trichloride (1.5 eq) was added to a solution of 6-(dimethylamino)-3,4-dihydronaphthalen-1(2H)-one (1.0 eq) in anhydrous N, N-dimethylformamide. The mixture was reaction under ice bath conditions for 4 h. After that the reaction solution was introduced into ice water, the organic phase was extracted with ethyl acetate and rotary evaporation. The crude product was purified by silica gel chromatography (ethyl acetate: petroleum ether = 1: 4) to obtain product as a yellow solid (56%). ^{1}H NMR (400 MHz, DMSO-d6) δ (ppm): 10.127 (s, 1H); 7.612 (d, J = 8.8, 1H), 6.664 (d, J = 8.8, 1H); 6.620 (s, 1H); 3.014 (s, 6H); 2.745 (t, J = 8.0, 2H); 2.481(t, J = 7.6, 2H). ^{13}C NMR (100 MHz, DMSO-d6): 189.34, 152.87, 146.51, 141.33, 128.24, 126.97, 126.25, 1225.19, 119.06, 117.01, 110.94, 110.38, 45.24, 27.74, 21.96.

Synthesis of PYD: 1-chloro-6-(dimethylamino)-3,4-dihydronaphthalene-2-carbaldehyde (1.0 eq) and 6-(dimethylamino)-3,4-dihydronaphthalen-1(2H)-one (1.0 eq) was added to a round-bottomed flask, 10 mL of sulphuric acid was added to the round-bottomed flask, and the mixtures were reacted at 90℃ for 4 h. After completed of the reaction, the mixture was introduced into ice water and 1mL perchloric acid was added into the mixture immediately. The mixed system was filtered and washed with water, and the crude product was purified by silica gel chromatography (MeOH: DCM = 1: 10) to give **PYD** as a black solid (62% yield). ^{1}H-NMR (400 MHz, DMSO-d6) δ (ppm): 8.025 (s, 1H), 7.970 (s, 1H), 7.948 (s, 1H), 6.842 (d, *J* = 2.4, 1H); 6.8195 (d, *J*=1.4, 1H); 6.722(d, *J*=1.2, 2H); 3.129 (s, 12H); 2.972-2.906 (m, 8H), ^{13}C NMR (100 MHz, DMSO-d6):162.83, 154.41, 146.90, 143.78, 127.54, 124.66, 113. 03, 111.93, 111.13, 27.24, 25.35. MS (ESI) m/z: calcd for $C_{25}H_{27}N_2O^+$ $[M]^+$ m/z 371.21, found 371.21.



Scheme S1 Synthetic routes for PYD

3. Additional Figures

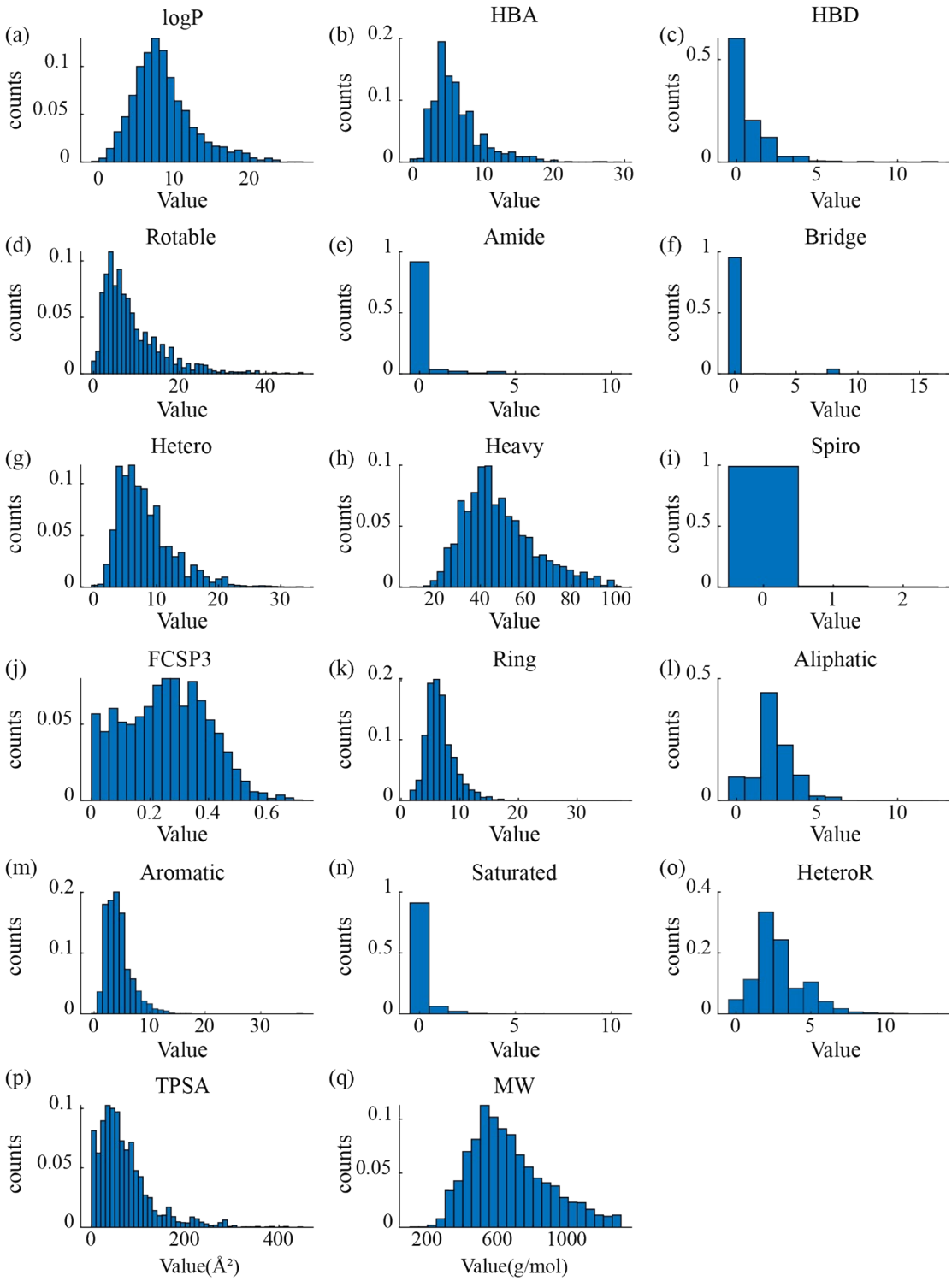


Fig. S1 Distribution plots of the 17 physicochemical properties of data set 2 used in this study. Detailed definitions of each descriptor can be found in **Table S1**.

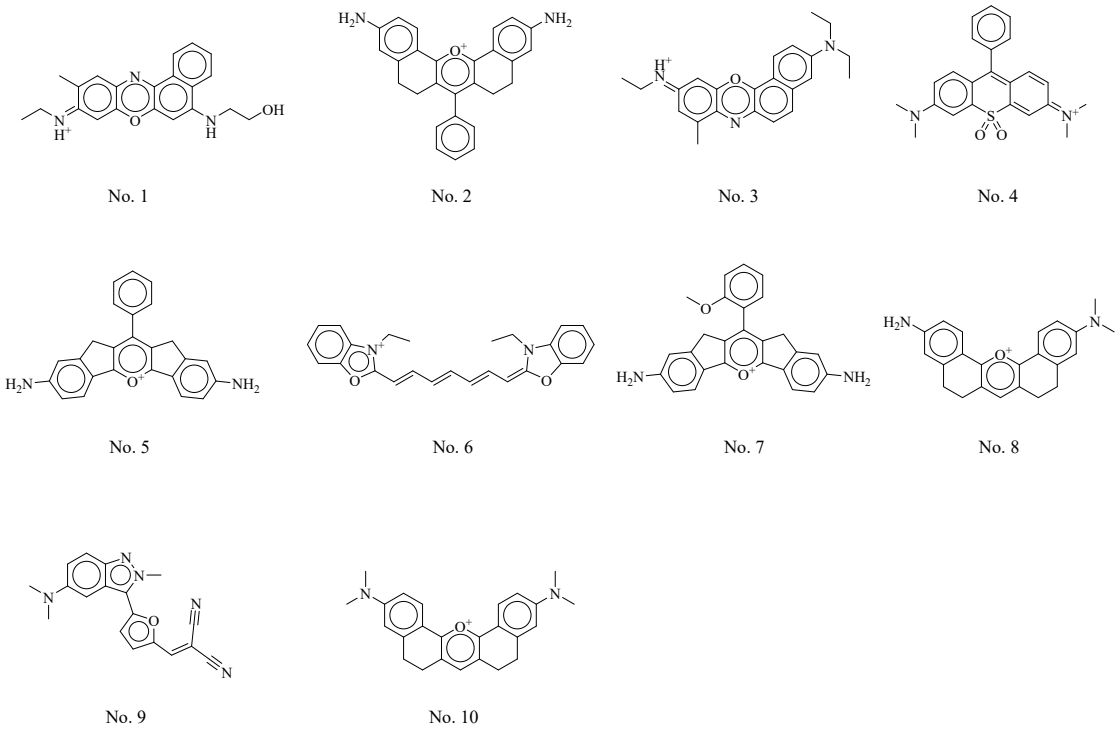


Fig. S2 The top 10 molecules identified through multi-stage screening.

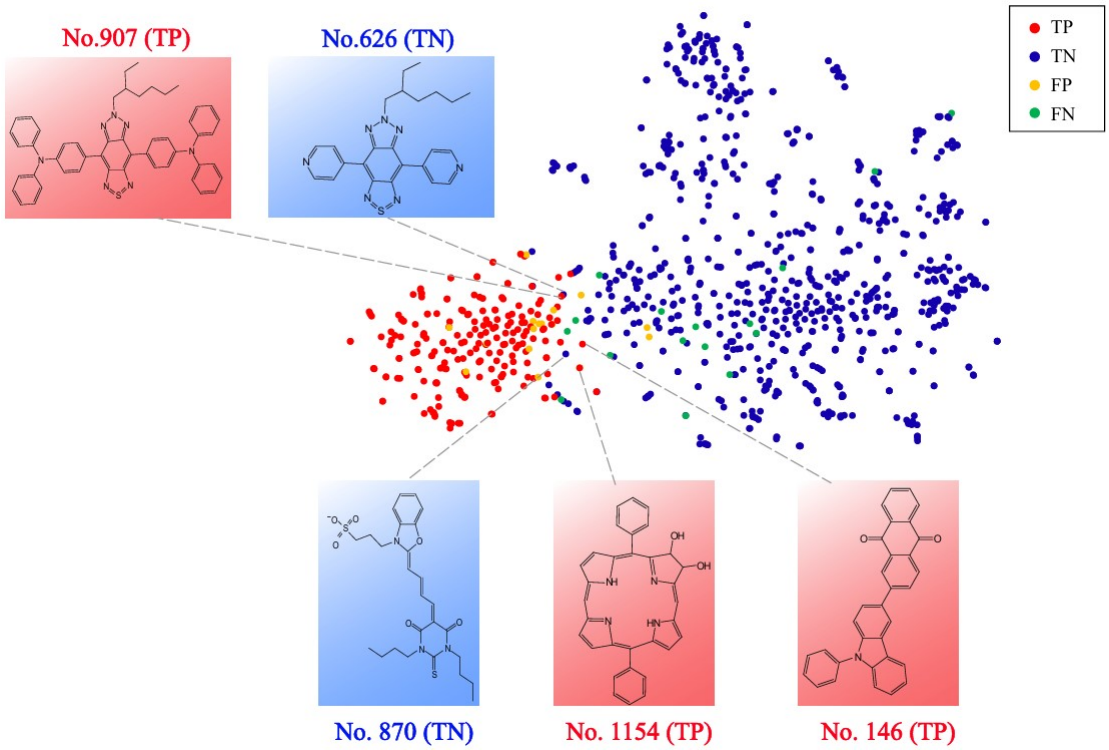


Fig. S3 t-SNE visualization of molecules in Data set 1 in the 1-PS-GCN feature space. TP: correctly predicted type I; TN: correctly predicted non-type I; FP: non-type I misclassified as type I; FN: type I misclassified as non-type I.

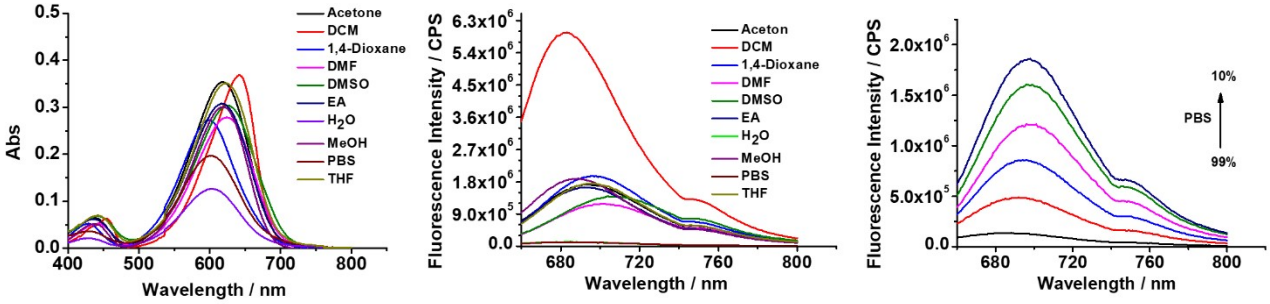


Fig. S4 (a)Absorption and (b) emission spectra of PYD in different solvents and (c) its water-dependent fluorescence characteristics in a 1,4-dioxane/water system.

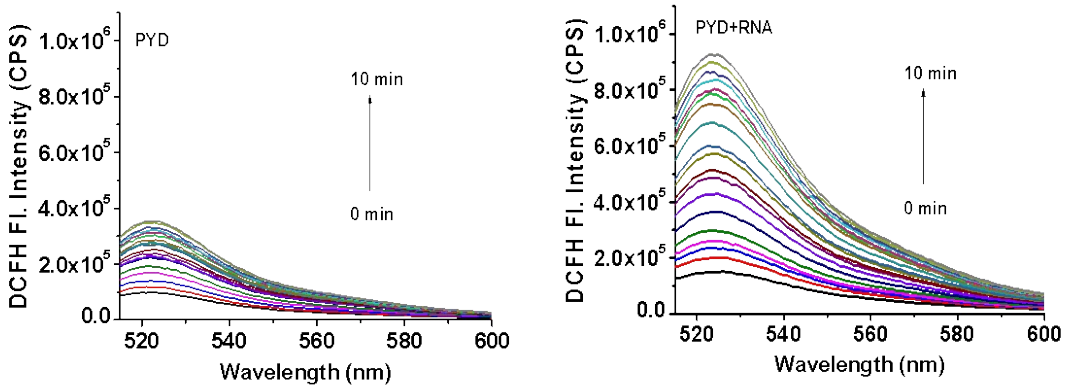


Fig. S5 Time-dependent fluorescence spectra of DCFH for **PYD** (10 μM) in the absence or presence of RNA in Tris-HCl buffer (10 mM, pH = 7.4) upon irradiation (660 nm, 10 mW/cm^2).

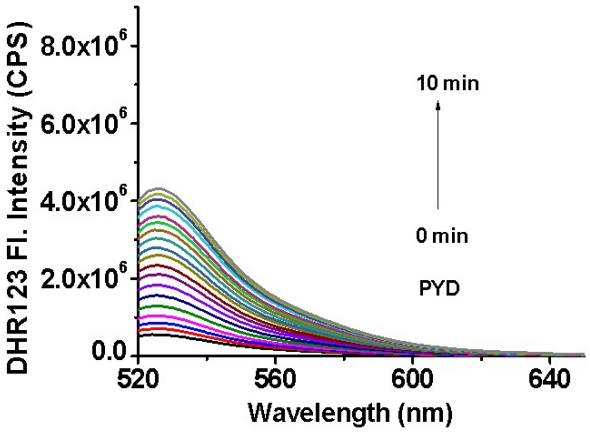


Fig. S6 Time-dependent fluorescence spectra of DHR123 for **PYD** (10 μM) in the absence of RNA in Tris-HCl buffer (10 mM, pH = 7.4) upon irradiation (660 nm, 10 mW/cm^2).

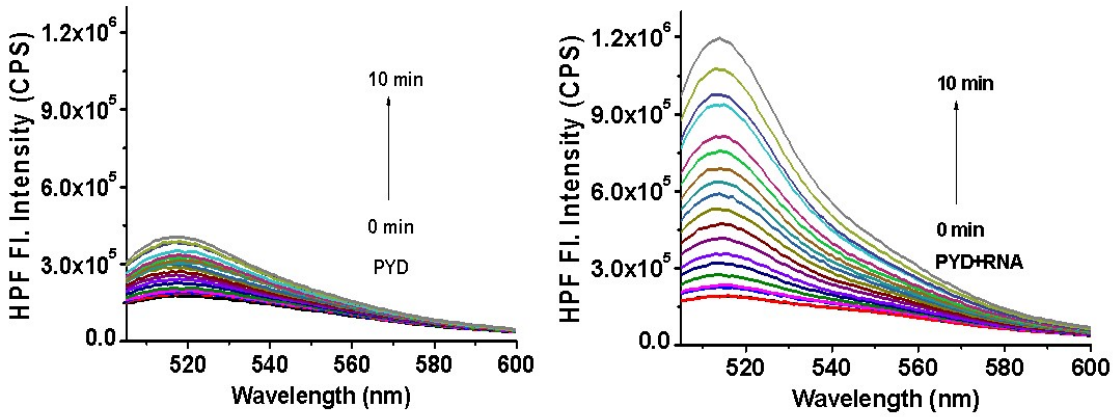


Fig. S7 Time-dependent fluorescence spectra of HPF for **PYD** (10 μM) in the absence or presence of RNA in Tris-HCl buffer (10 mM, pH = 7.4) upon irradiation (660 nm, 10 mW/cm^2).

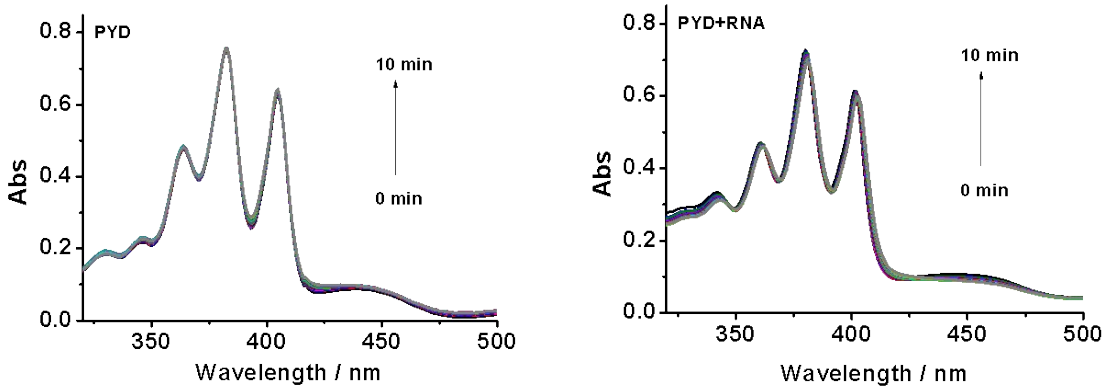


Fig. S8 Time-dependent absorption spectra of ABDA for **PYD** (10 μM) in the absence or presence of RNA in Tris-HCl buffer (10 mM, pH = 7.4) upon irradiation (660 nm, 10 mW/cm^2).

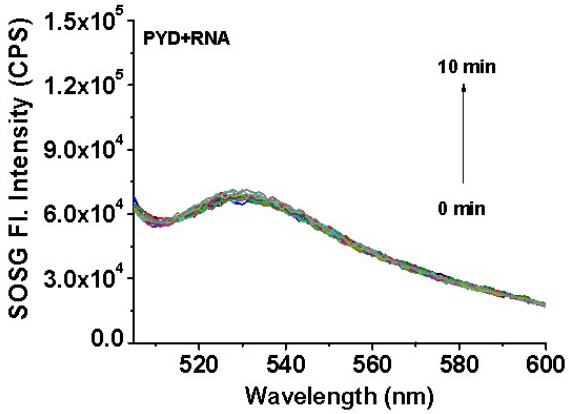


Fig. S9 Time-dependent fluorescence spectra of SOSG for **PYD** (10 μM) in the presence of RNA in Tris-HCl buffer (10 mM, pH = 7.4) upon irradiation (660 nm, 10 mW/cm^2).

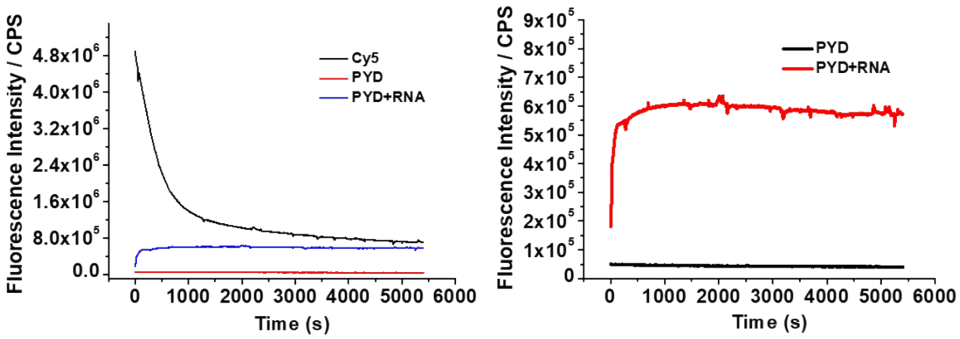


Fig. S10 Photobleaching experiment of PYD and Cy5 and magnified curves of PYD and PYD+RNA.

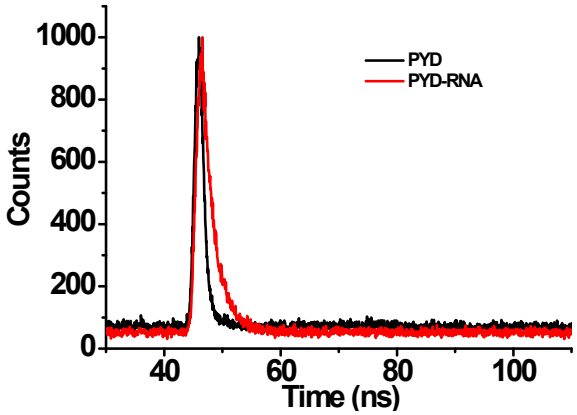


Fig. S11 Fluorescence lifetime of **PYD** with or without RNA.

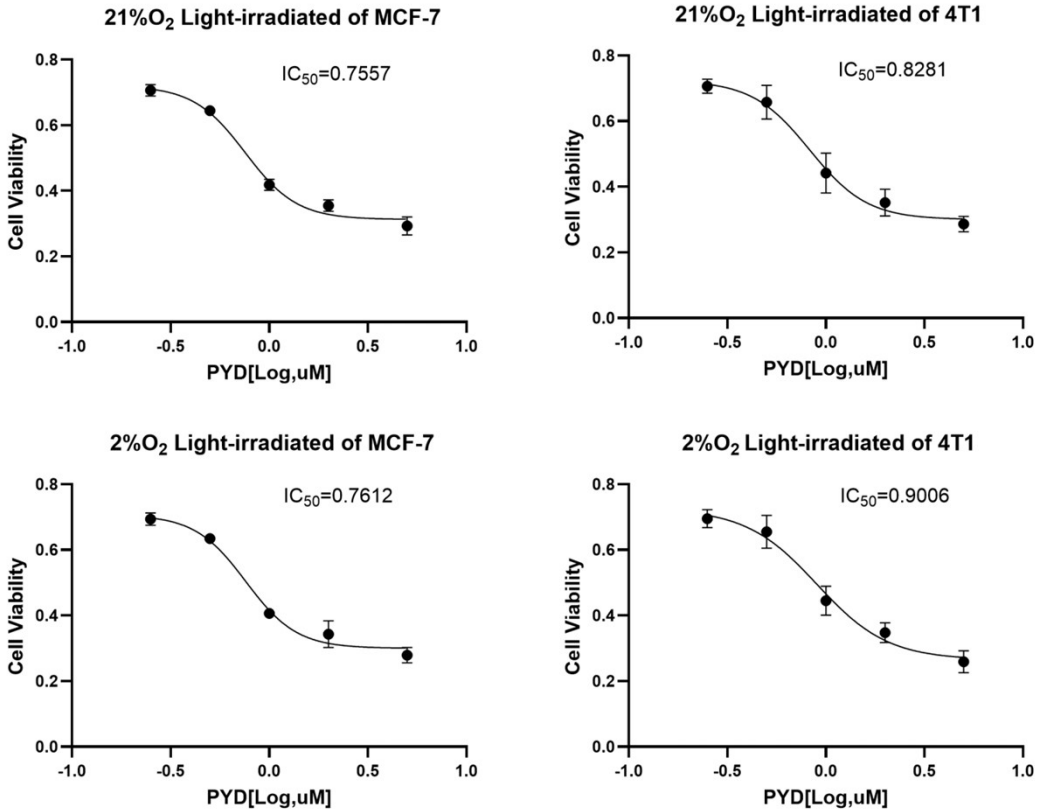


Fig. S12 Under normal oxygen and hypoxic conditions, the IC_{50} values of light **PYD** for MCF-7 and 4T1 cells.

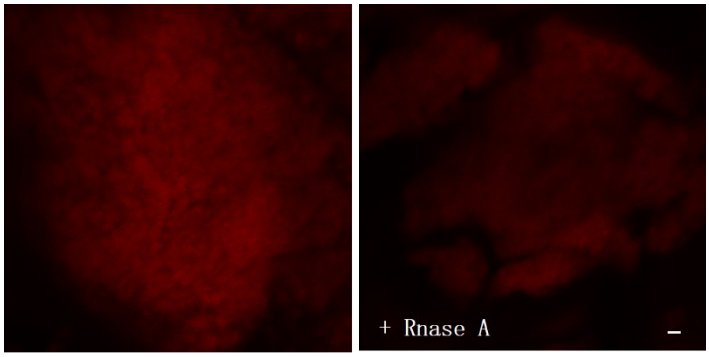


Fig. S13 Whether to perform **PYD**(50 μM, 50 μL) fluorescence imaging on the fixed tumor tissue sections after RNase A treatment.

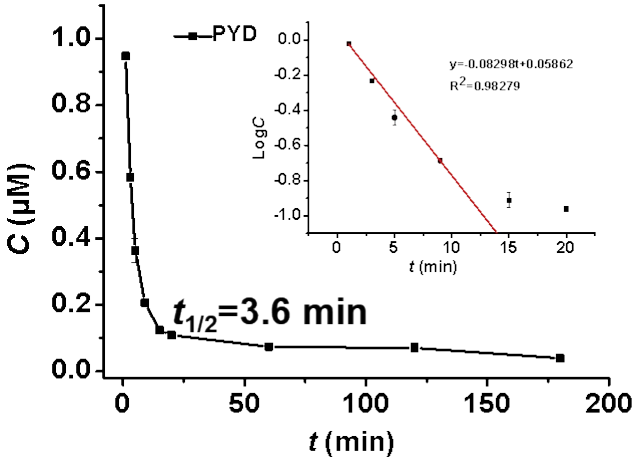


Fig. S14 Blood circulation for PYD (50 μM, 50 μL) in healthy BALB/c mice (n = 3). Inset: linear correlation between the logarithm of PYD concentrations and time.

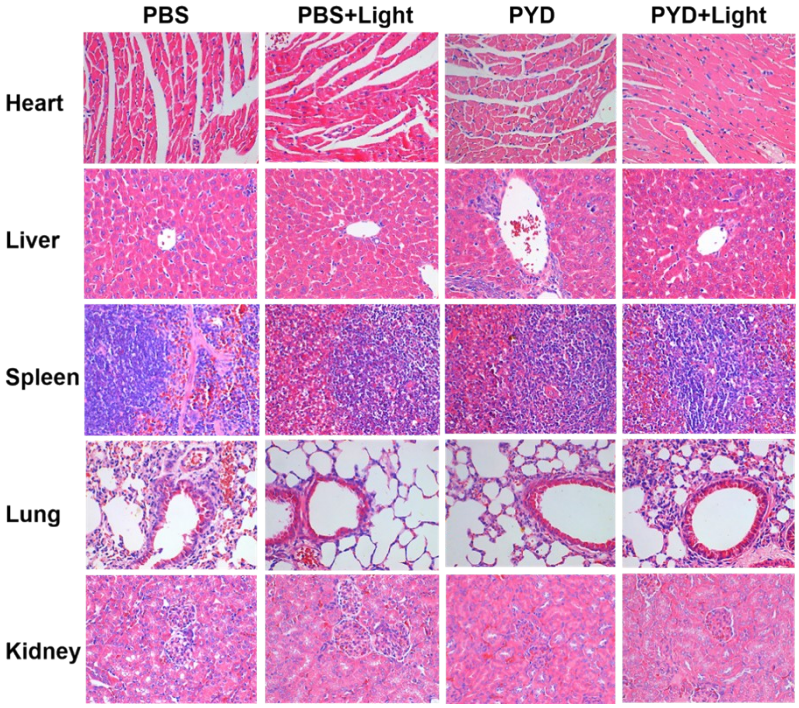


Fig. S15 Representative hematoxylin-eosin staining images for organs from mice bearing 4T1 tumors after injection with **PYD** (50 μM, 20 μL) or PBS with or without irradiation.

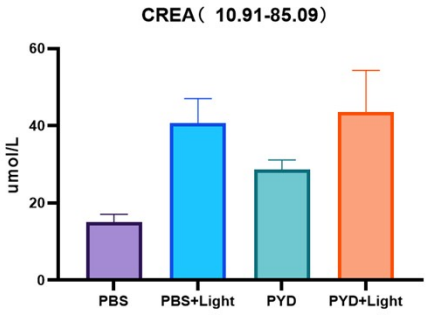


Fig. S16 After PYD injection, the cardiac blood tests of mice showed kidney function (kidney function: CREA).

4. NMR and MS spectra

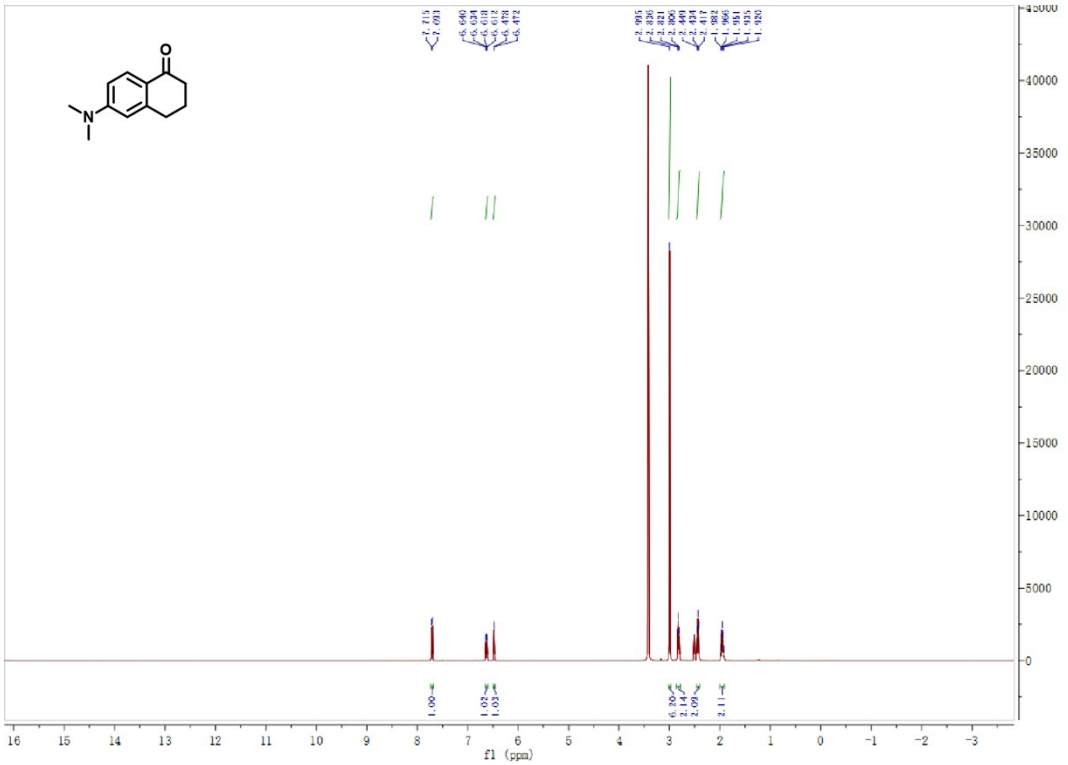


Fig. S17 ^{1}H NMR spectrum of 6-(dimethylamino)-3,4-dihydronaphthalen-1(2H)-one.

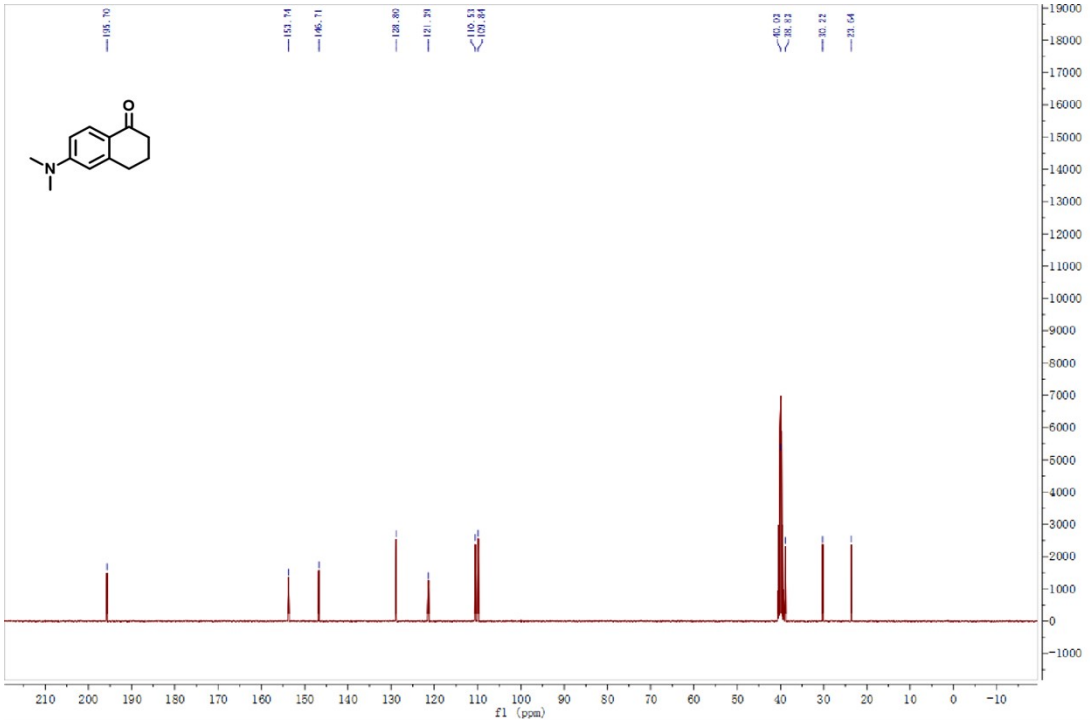


Fig. S18 ^{13}C NMR spectrum of 6-(dimethylamino)-3,4-dihydronaphthalen-1(2H)-one.

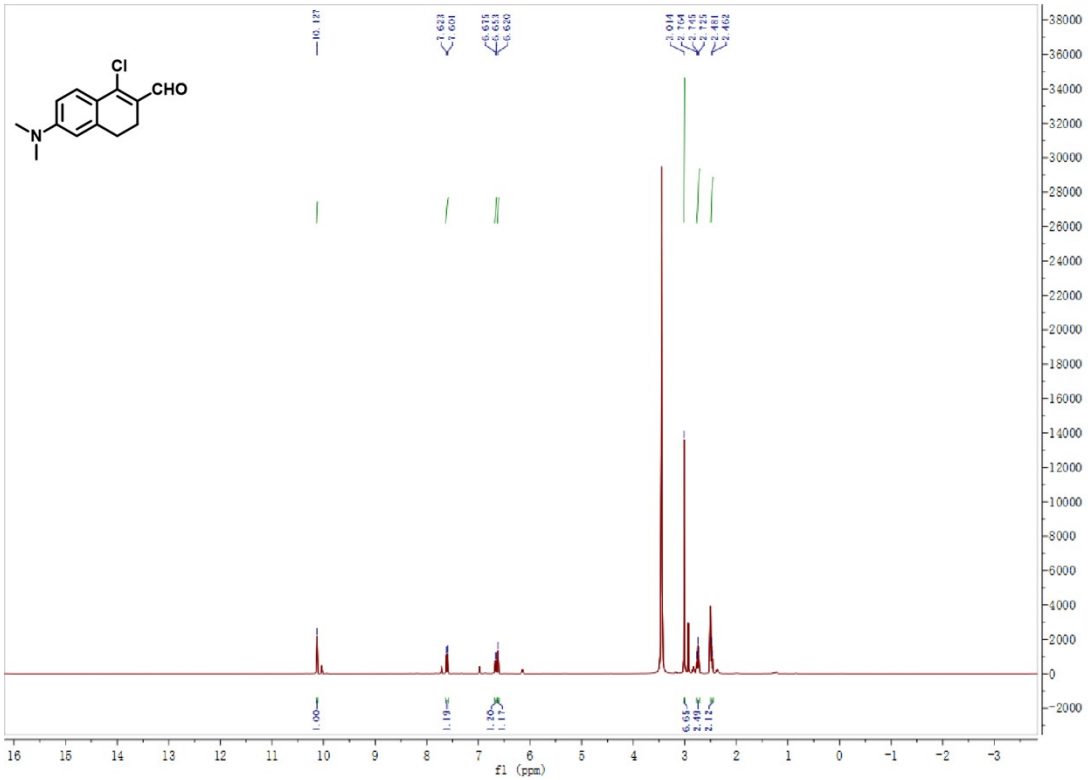


Fig. S19 ^{1}H NMR spectrum of 1-chloro-6-(dimethylamino)-3,4-dihydronaphthalene-2-carbaldehyde.

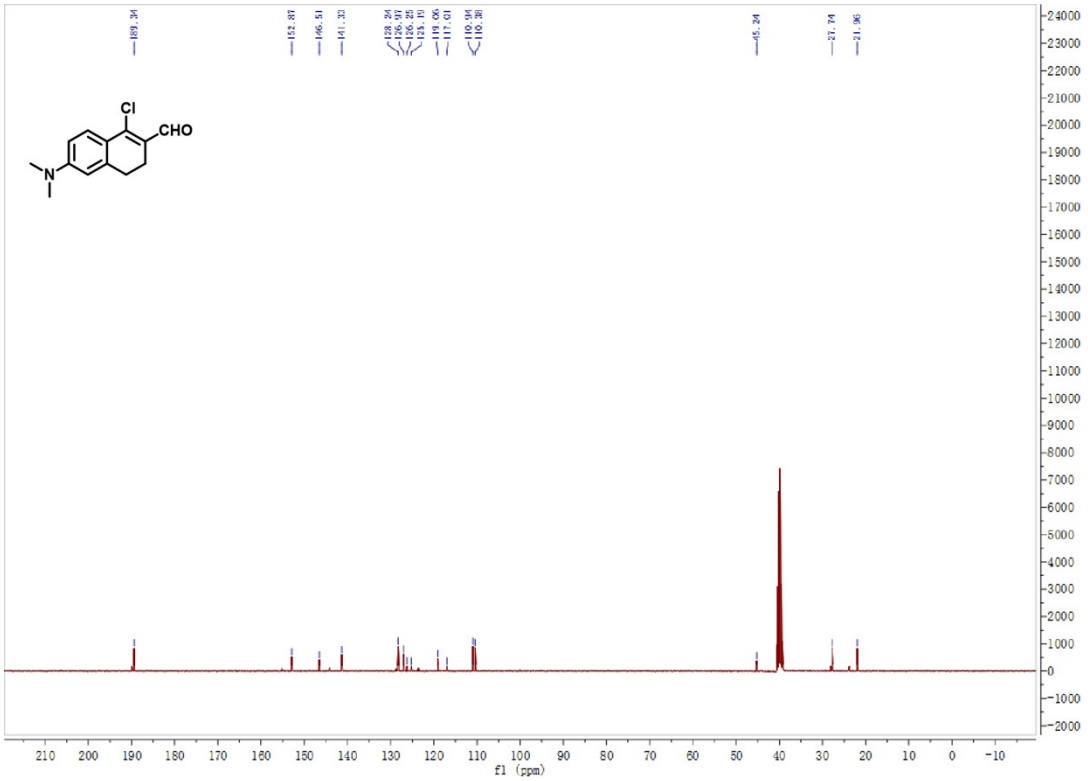


Fig. S20 ^{13}C NMR spectrum of 1-chloro-6-(dimethylamino)-3,4-dihydronaphthalene-2-carbaldehyde.

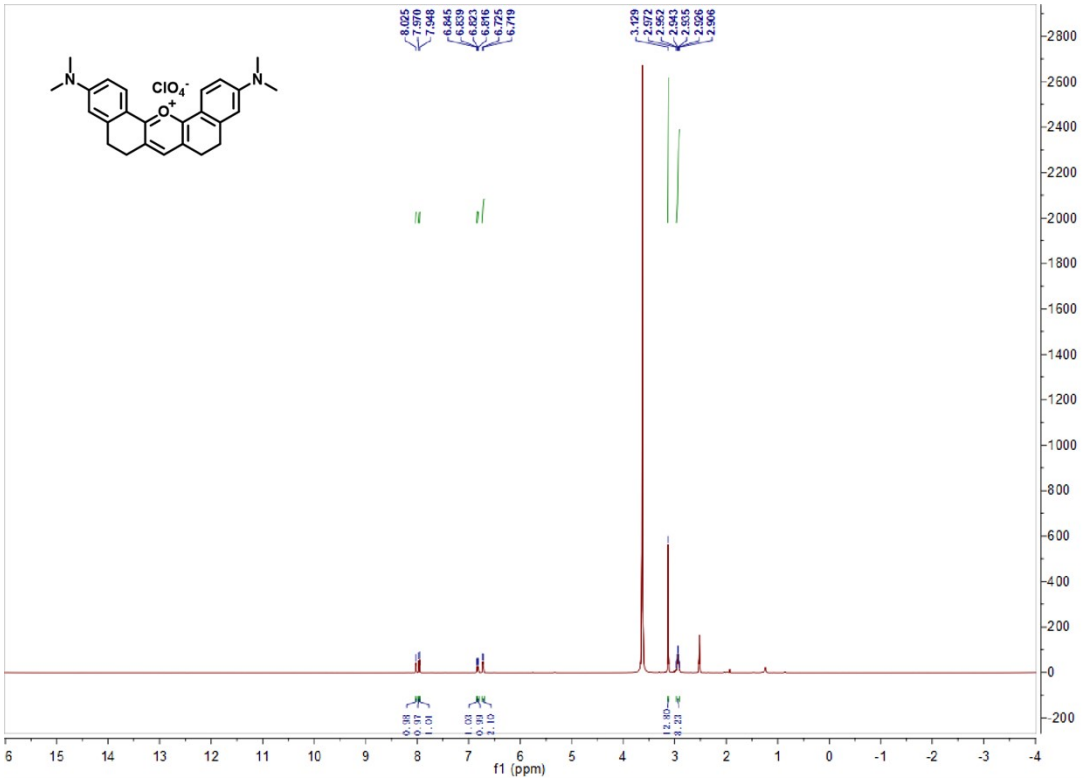


Fig. S21 ^{1}H NMR spectrum of **PYD**.

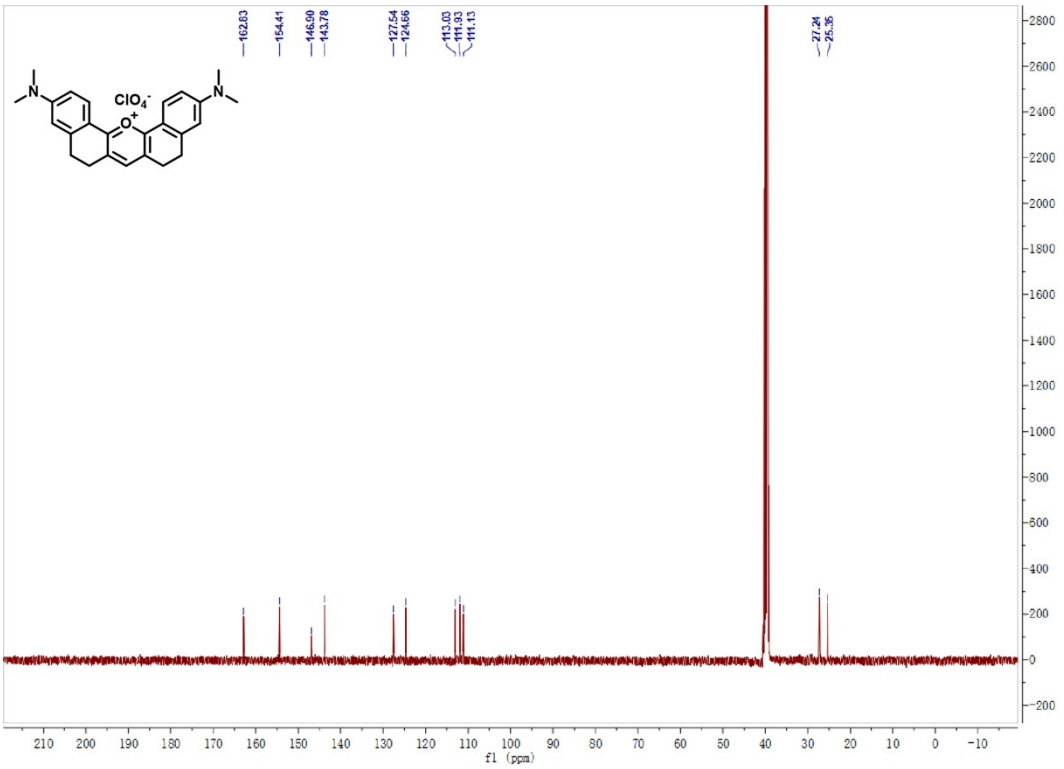


Fig. S22 ^{13}C NMR spectrum of **PYD**.

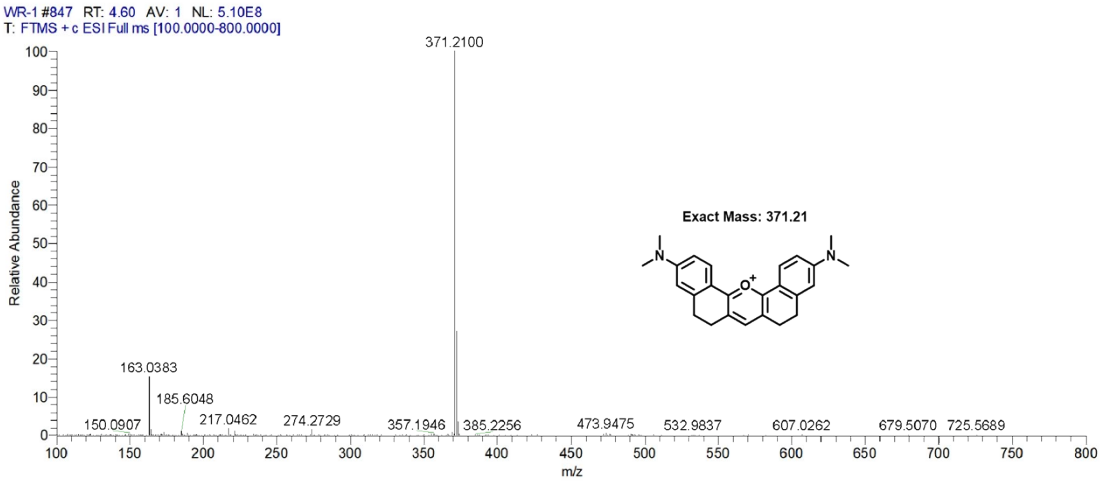


Fig. S23 HR-MS spectrum of **PYD**.

5. Additional Tables

Table S1 The node (atom) features and physicochemical properties of small molecules.

Node (atom) features	Size	Description
Atomic number	101	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,…,99, 100, other] (one-hot)
Degree	7	Number of bonds for each atom [0, 1, 2, 3, 4, 5, other] (one-hot)
Formal charge	8	Electrical charge [-3, -2, -1, 0, 1, 2, 3, other] (one-hot)
Hydrogens	6	Number of bonded hydrogens [0, 1, 2, 3, 4, other] (one-hot)
Hybridization	6	[sp, sp2, sp3, sp3d, sp3d2, other] (one-hot)
Ring	1	Whether the atom is in ring [0/1] (one-hot)
Aromaticity	1	Whether the atom is part of an aromatic system [0/1] (one-hot)
Atomic mass	1	Atomic mass
Physicochemical properties	**Size**	**Description**
logP	1	Partition coefficient
HBA	1	Number of hydrogen bond acceptors
HBD	1	Number of hydrogen bond donors
Rotable	1	Number of rotable bonds
Amide	1	Number of amide bonds
Bridge	1	Number of bridgehead atoms
Hetero	1	Number of hetero atoms
Heavy	1	Number of heavy atoms
Spiro	1	Number of spiro atoms
FCSP3	1	The fraction of sp3 carbon
Ring	1	Number of rings
Aliphatic	1	Number of aliphatic rings
Aromatic	1	Number of aromatic rings
Saturated	1	Number of saturated rings
HeteroR	1	Number of heterocycles
TPSA	1	Topological polar surface area
MW	1	Molecular weight

Table S2 Hyper-parameters for 1-PS-GCN model.

Model	Parameters to be optimized	Results	Package
1-PS-GCN	The number of GCN layer: [1, 2]	2	PyG (2.4.0)
	The number of nodes of each GCN hidden layer: [256, 128, 64, 32]	[64, 64]	
	The pooling layers: [global_add_pool, global_mean_pool, global_max_pool]	global_mean_pool	
	The number of FC hidden layer: [2, 3, 4]	2	
	The neurons of FC1 hidden layer: [2048, 1024, 512, 256, 128, 64, 32]	128	
	The dropout rate of each FC hidden layer: [0, 0.1, 0.2, 0.3, 0.4, 0.5]	0.5	
	The learning rate: [1e-3, 1e-4, 1e-5]	1e-4	
	The activation function: ReLU	ReLU	
	The optimizer: Adam	Adam	
	The number of epochs: 100	100	
	The patience of early stopping: 10	10	
	The batch size of DataLoader: [32, 16]	16	

Table S3 Hyper-parameters for DNN and CNN models.

Model	Parameters to be optimized	Results	Package
DNN	The number of FC hidden layer: [2, 3, 4]	2	torch (2.0.1)
	The neurons of FC1 hidden layer: [1024, 512, 256, 128, 64, 32]	128	
	The dropout rate of FC hidden layer: [0, 0.1, 0.2, 0.3, 0.4, 0.5]	0.5	
	The learning rate: [1e-3, 1e-4, 1e-5]	1e-4	
	The activation function: ReLU	ReLU	
	The optimizer: Adam	Adam	
	The number of epochs: 1000	1000	
	The patience of early stopping: 10	10	
	The batch size of DataLoader: 16	16	
CNN	The number of FC hidden layer: [2, 3, 4]	2	torch (2.0.1)
	The neurons of FC1 hidden layer: [1024, 512, 256, 128, 64, 32]	128	
	The dropout rate of FC hidden layer: [0, 0.1, 0.2, 0.3, 0.4, 0.5]	0.5	
	The learning rate: [1e-3, 1e-4, 1e-5]	1e-4	
	The activation function: ReLU	ReLU	
	The optimizer: Adam	Adam	
	The number of epochs: 1000	1000	
	The patience of early stopping: 10	10	
	The batch size of DataLoader: 16	16	

Table S4 Hyper-parameters for three classical machine learning classifiers.

Model	Parameters to be optimized	Results	Package
RF	n_estimators: range (50, 301, 50)	200	scikit-learn (1.3.0) RandomForestClassifier
	min_samples_split: range (2, 11, 1)	3	
	min_samples_leaf: range (1, 11, 1)	8	
	max_depth: range (2, 11, 1)	10	
	max_features: range (sqrt, log2, None)	log2	
SVM	C: [1, 10, 50, 100]	100	scikit-learn (1.3.0) SVC
	gamma: [1e-2, 1e-3, 1e-4]	1e-3	
	kernel: [linear, poly, rbf]	rbf	
KNN	n_neighbors: range (1, 31, 1)	23	scikit-learn (1.3.0) KNeighborsClassifier
	weights: [uniform, distance]	distance	
	algorithm: [auto, ball_tree, kd_tree, brute]	auto	

Table S5 Sequences for nucleic acids of different structures used in this study.

Structure	Names	Sequences(5'to3')
RNA	NRAS (RNA G4)	5'-GGGAGGGGCGGGUCUGGG-3'
	Terra (RNA G4)	5'-UUAGGGUUAGGGUUAGGGUUAGGG-3'
	Bcl2 (RNA G4)	5'-GGGGGCCGUGGGGUGGGAGCUGGGG-3'
	Hp26	5'-r(CAGUACAGAUCUGUACUG)-3'
	dsRNA-2	5'-r(UUUUUAAAAA)-3'
	ssRNA-1	5'-r(UUUUUGGGGGG)-3'
DNA	VEGF (DNA G4)	5'-GGGAGGGTTGGGGTGGG-3'
	C-Myc (DNA G4)	5'-TGAGGGTGGGGAGGGTGGGGAA-3'
	Bcl2 (DNA G4)	5'-GGGCGGGCGCGGGAGGAAGGGGGCGGG-3'
	Bom17 (DNA G4)	5'-GGTTAGGTTAGGTTAGG-3'
	Tel26 (DNA G4)	5'-AAAGGGTTAGGGTTAGGGTTAGGGAA-3'
	ds-15GC-2	5'-GAAAAAAGAGAGAGG-3'
	ds26	5'- CAATCGGATCGAATTCGATCCGATTG-3'
	ds-DNA-9-1	5'-CATGCGCGCATG-3'
	ss-21T	5'-TTTTTTTTTTTTTTTTTTTT-3'
	ss-DNA-3	5'-CCTCTCTCTTTTTTC-3'

6. References

S1. L. Breiman, Random forests, *Mach. Learn.*, 2001, **45**, 5-32.

S2. R. G. Brereton and G. R. Lloyd, Support vector machines for classification and regression, *Analyst*, 2010, **135**, 230-267.

S3. Z. Ying, D. Bourgeois, J. You, M. Zitnik and J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, *Advances in Neural Information Processing Systems*, 2019, **32**.

S4. O. Trott and A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.*, 2010, **31**, 455-461.

S5. W. L. DeLano, Pymol: An open-source molecular graphics tool, *CCP4 Newsl. Protein Crystallogr.*, 2002, **40**, 82-92.

S6. P. A. Ravindranath, S. Forli, D. S. Goodsell, A. J. Olson and M. F. Sanner, AutoDockFR: advances in protein-ligand docking with explicitly specified binding site flexibility, *PLOS Comput. Biol.*, 2015, **11**, e1004586.

S7. X. Gao, S. Bai, D. Fazzi, T. Niehaus, M. Barbatti and W. Thiel, Evaluation of spin-orbit couplings with linear-response time-dependent density functional methods, *J. Chem. Theory Comput.*, 2017, **13**, 515-524.

S8. C. Yao, Y. Chen, M. Zhao, S. Wang, B. Wu, Y. Yang, D. Yin, P. Yu, H. Zhang and F. Zhang, A Bright, Renal-Clearable NIR-II Brush Macromolecular Probe with Long Blood Circulation Time for Kidney Disease Bioimaging, Angew. Chem. Int. Ed., 2022, 61, e202114273.