Electronic Supplementary Information (ESI)

for

High-efficiency synthesis of red carbon dots using machine learning

*Jun Bo Luo, † Jiao Chen, ‡ Hui Liu, *‡Cheng Zhi Huang,*, ‡ Jun Zhou *, †‡*

† a Key Laboratory of Luminescence Analysis and Molecular Sensing (Southwest University), Ministry of Education, College of Computer and Information Science, Southwest University, Chongqing 400715, P. R. China.

‡ Key Laboratory of Luminescent and Real-Time Analytical System (Southwest University), Chongqing Science and Technology Bureau, College of Pharmaceutical Sciences, Southwest University, Chongqing 400715, P. R. China.

1. The specific processing process of the proposed machine learning model.

We collected the synthesis conditions and results of CDs from a large number of studies, and divided these synthesis results into two categories, such as red CDs and non-red CDs. This is used as the target variable when building the machine learning model, and the synthesis conditions such as precursor, solvent, temperature, and reaction time are used as input data. Since the precursors and solvents cannot be directly input into the model, we converted precursors and solvents into numerical features such as atomic content, and the scale of the entire dataset was 151 samples, each containing 22 features.

Then, we started building machine learning model. The model is divided into two parts, the data preprocessing pipeline and the classifier. Among them, the function of the data preprocessing pipeline is to perform feature extraction and transformation on the input raw data. It contains three different techniques, such as XGBoost feature extractor, one-hot encoding technology, and PCA dimensionality reduction technique. We input the data into the preprocessing pipeline for feature extraction. The data first was input the XGBoost feature extractor to fit the model, and obtain the converted data. These data are samples on the leaf nodes of the basic learner of the XGBoost model. These data are the positions of the samples on the leaf nodes of the XGBoost model, and have the form of discrete numerical features. The dimension of the sample is increased to 70 dimensions (the number of dimensions is the same as the number of base learners of XGBoost). Then, we perform one-hot encoding on these data to obtain higher-dimensional sparse data. At this time, the dimension of the sample is expanded to 537 dimensions, and the data becomes more sparse. Finally, we use PCA technique to reduce the dimensionality of these data to remove redundant information. After dimensionality reduction, the dimension of the sample becomes 52 dimensions. The above is the working process of the data preprocessing pipeline.

Finally, we trained six different machine learning models using the preprocessed data and compared their performance on AUC and F1-score. The specific training process is as follows. We use the ten-fold cross-validation method to randomly divide the entire dataset into 90% training set and 10% test set, and then train and test six different classifiers, respectively. Finally, the logistic regression classifier is used to

data classify. It shows the best performance on these two metrics. We chose the logistic regression classifier as the final classifier, and designed ten sets of CDs synthesis conditions by ourselves, and simultaneously synthesized ten different CDs (4 of which are red CDs, and the rest are non-red CDs). We use these ten combinations of synthesis conditions and the corresponding synthesis results as independent validation sets to test the generalization performance of the model. The results show that our model can accurately predict whether the color of synthetic CDs is red or not according to the input synthesis conditions. We selected a red CD from the synthesized CDs, and carried out a series of analytical experiments to detect the properties of the CDs. We applied this CD in cell imaging experiments and achieved good results.

2. **We use the 10-fold cross validation method to evaluate the performance of six different classifiers. The specific parameters and comparison results of six different classification models are as follows.**

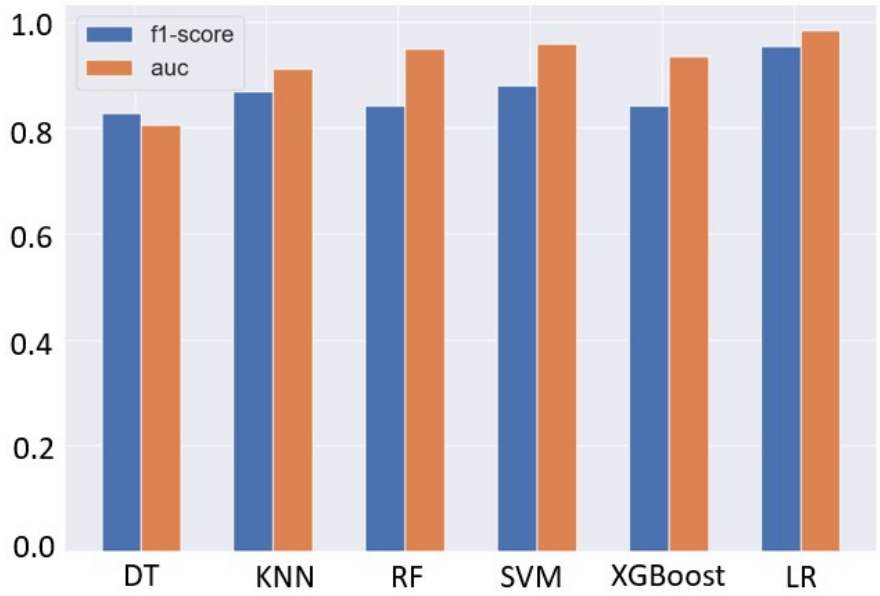

Figure S1. F1-score and AUC of different classifiers

Tabel S1. The parameter combination of different classifiers

Model	Parameters	F1-score	AUC
DT	Max_depth:3	0.735	0.790
	Max_depth:5	0.802	0.782
	Max_depth:8	0.806	0.828
	Max_depth:10	0.806	0.828
	.....	.....	.....
KNN	n_neighbors:1	0.858	0.875
	n_neighbors:3	0.814	0.901
	n_neighbors:5	0.834	0.893
	n_neighbors:7	0.801	0.900
	.....	.....	.....
RF	n_estimators:50, max_depth:5	0.847	0.937
	n_estimators:60, max_depth:5	0.868	0.947
	n_estimators:80, max_depth:5	0.841	0.949
	n_estimators:80, max_depth:6	0.849	0.948
	.....	.....	.....
SVC	C: 1000, kernel: rbf	0.880	0.959
	C: 1000, kernel: sigmoid	0.840	0.936
	C: 1000, kernel: linear	0.940	0.975
	C: 100, kernel: linear	0.940	0.975
	.....	.....	.....
Xgboost	n_estimators: 120, max_depth: 5,learning_rate:0.1	0.841	0.934
	n_estimators: 100, max_depth: 7,learning_rate:0.1	0.854	0.928
	n_estimators: 100, max_depth: 5,learning_rate:0.1	0.841	0.934
	n_estimators: 100, max_depth: 5,learning_rate:0.15	0.860	0.934
	.....	.....	.....
Lr	Penalty: l2,solver: liblinear,max_iter: 50,C: 0.5	0.934	0.976
	Penalty: l1,solver: liblinear,max_iter: 50,C: 5	0.928	0.982
	Penalty: l1,solver: liblinear,max_iter: 50,C: 0.5	0.948	0.981
	Penalty: l1,solver:saga,max_iter: 50,C: 0.5	**0.954**	**0.983**
	.....	.....	.....

3. Statistical analysis of the data used for machine learning modeling

First, we performed a correlation analysis on the original data and some features of the data after data preprocessing pipeline.

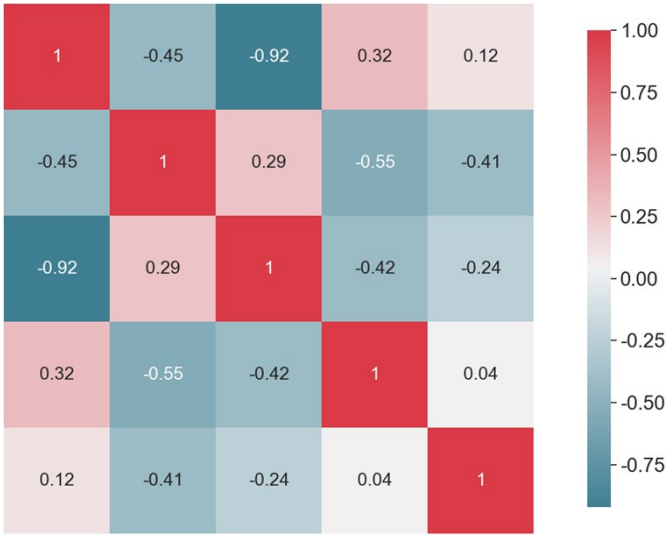

Figure S2. Correlation matrix between some features in the original data.

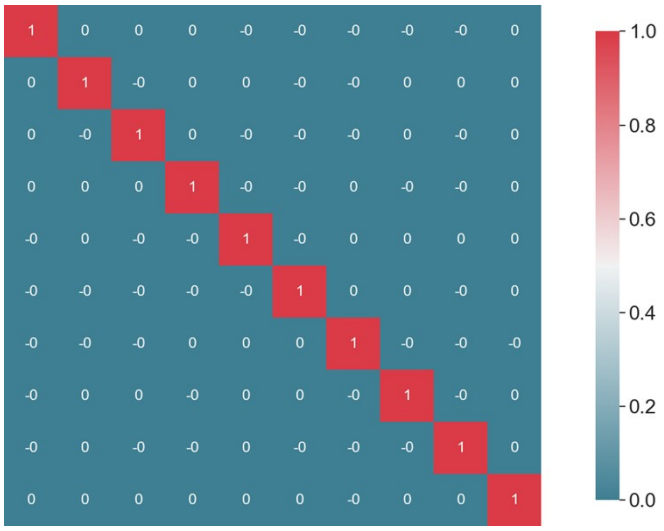

Figure S3. Correlation matrix between some features in the data after data preprocessing pipeline

As shown in Figure S2 and S3, the results of correlation analysis show that the features in the data transformed by the data processing pipeline are independent of each other, eliminating the collinearity within the data. It is more conducive to regression analysis.

Then, we performed normality analysis on the original data and some features of the data processing pipeline.

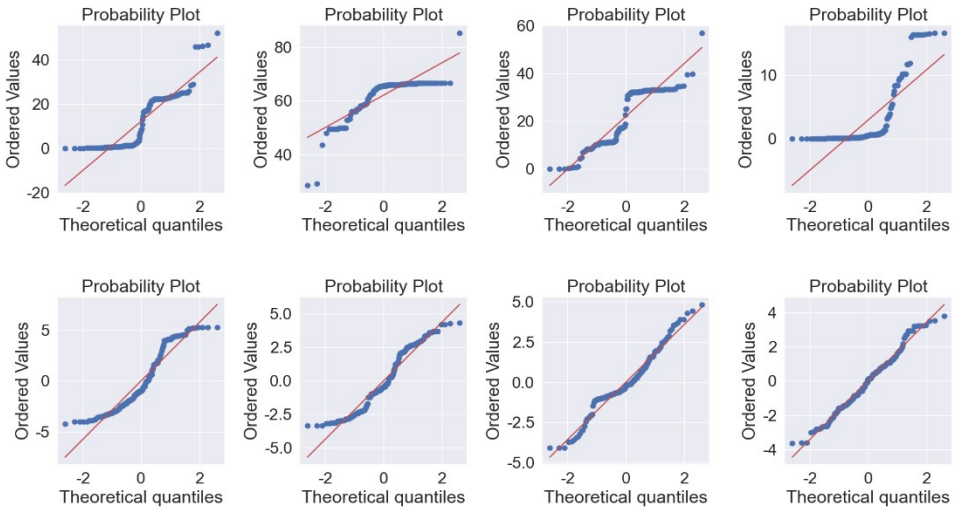

Figure S4. The normality comparison chart of some features before and after data preprocessing pipeline

In Figure S4, the first row is part of the features in the original data, and the second row is part of the features in the preprocessed data corresponding to the first row. The results of normality analysis show that each feature in the data transformed by the data processing pipeline has better normality, which is more conducive to linear classifiers such as logistic regression.

4. Experimental materials, reagents and instruments

Materials and reagents. M-phenylene diamine, bought o-phenylendiamine from Aladdin reagent co., LTD. (Shanghai, China), hydrochloric acid, urea, spermine, dopamine bought dicyandiamide from McLean reagent co., LTD. (Shanghai, China), citric acid for purchased from Sigma Aldrich trade co., LTD. (Shanghai, China), thiourea purchased from Beijing chemical reagents co., LTD. (Beijing, China), Neutral red was purchased from Chengdu Jinshan Chemical Reagent Co., LTD. (Chengdu, China), ethylenediamine, anhydrous ethanol, DMF, ethyl acetate and sulfuric acid from Chongqing Chuandong Chemical (Group) Co., LTD. (Chongqing, China), and p-aminobenzoic acid from Chengdu Kelon Chemical Reagent Factory (Chengdu, China). Ultra-pure water (18.2 M ω ·cm) from the

Millipore Water Purification System was used in all experiments.

The instruments. All fluorescence spectra were recorded using a Hitachi F-7100 fluorescence spectrophotometer (Tokyo, Japan) and confocal fluorescence images were obtained using an Olympus IX-81 inverted microscope and a Nikon A1R laser scanning confocal microscope, respectively.

5. Synthesis experiments of carbon dots

One-pot hydrothermal method is used for the synthesis of 10 carbon points. **Table S2** lists the reactants and synthesis conditions. Silica gel is used for column purification to remove the residual impurities in the reaction.

Tabel S2. Conditions for synthesis of carbon dots

CDs	Precursor 1/g	Precursor 2/g	Temperature	Time/h	Solvent/ml
CDs-1	Dopamine hydrochloride(0.184)	o-phenylendiamine(0.108)	200℃	6	H_20(10) H_2SO_4(1)
CDs-2	Citric acid(1.44)	The urea(0.6)	200℃	12	DMF(10)
CDs-3	Thiourea(0.2)	Citric acid(0.1)	180℃	10	DMF(5)
CDs-4	Neutral red(0.1)	Thiourea(0.1)	180℃	3	H_20(10)
CDs-5	Citric acid(1.05)	Ethylenediamine(0.335)	150℃	10	H_20(10)
CDs-6	Citric acid(0.07)	Dicyandiamide(0.25)	180℃	3	H_20(4)
CDs-7	Para aminobenzoic acid(0.1)	/	180℃	12	ethyl alcohol(5)
CDs-8	Citric acid(0.1)	Spermine(0.05)	160℃	3	H_20(5)
CDs-9	o-phenylendiamine(0.072)	/	160℃	8	H_20(10)
CDs-10	M-phenylene diamine(0.05)	Para aminobenzoic acid(0.05)	180℃	12	ethyl alcohol(10)

6. Color of carbon dots solution under 365nm UV lamp

Synthesis results showed that four carbon dots solution were red or near red, while the rest were bule, yellow, and green.

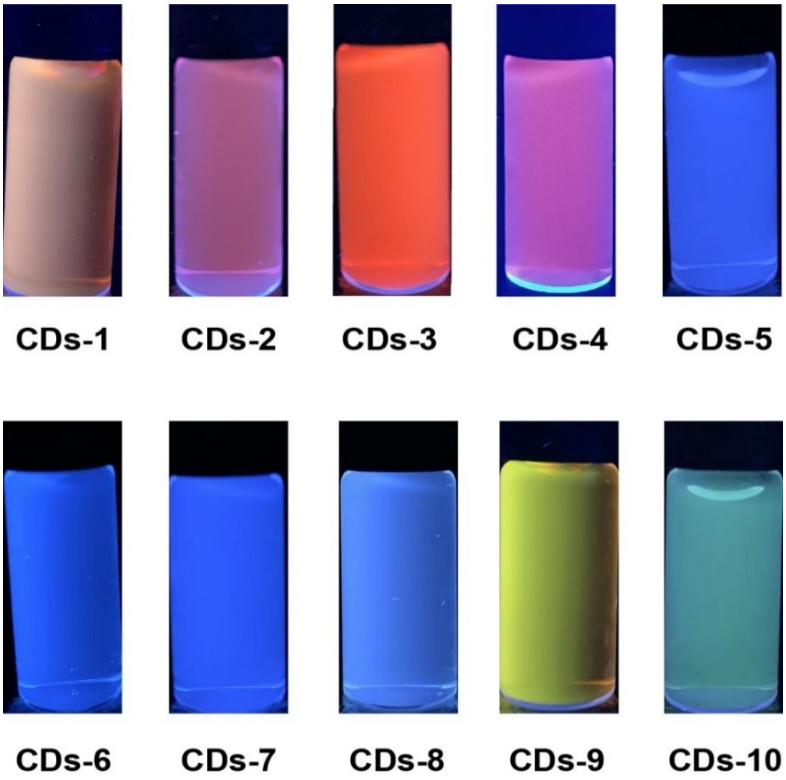

Figure S5. CDs solutions under 365 nm UV lamps

7. 3D fluorescence plots of CDs

Their 3D fluorescence spectra were measured by FL-7100 fluorescence spectrophotometer.

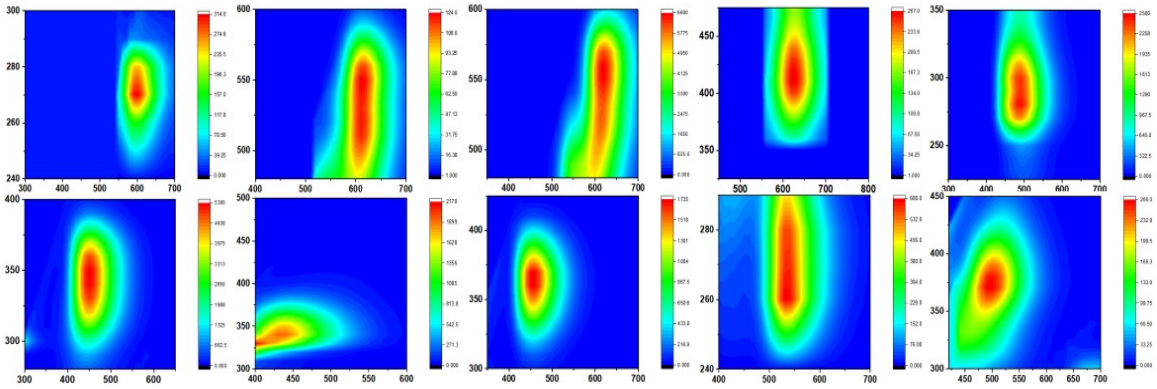

Figure S6. FL spectra of 10 CDs under same excitation wavelengths

8. Properties of red carbon dots for living cell imaging

Preparation of R-CDs. R-CDs were synthesized by hydrothermal method. In short, 100 mg of neural red and 100mg of thiourea were dissolved into 10 mL of ultrapure water and mix well by vortex blender. Then, the mixture was transferred to a Teflon-lined stainless steel autoclave (25mL) and heated at 180 °C for 3 hours. After being cooled down to room temperature naturally, the reaction solution was centrifuged at 8500 rpm for 10min to remove incomplete large particles. The supernatant was purified by silica gel column chromatography. Finally, a red solution was collected under the guidance of a 365nm UV hand lamp and freeze-dried to obtain R-CDs.

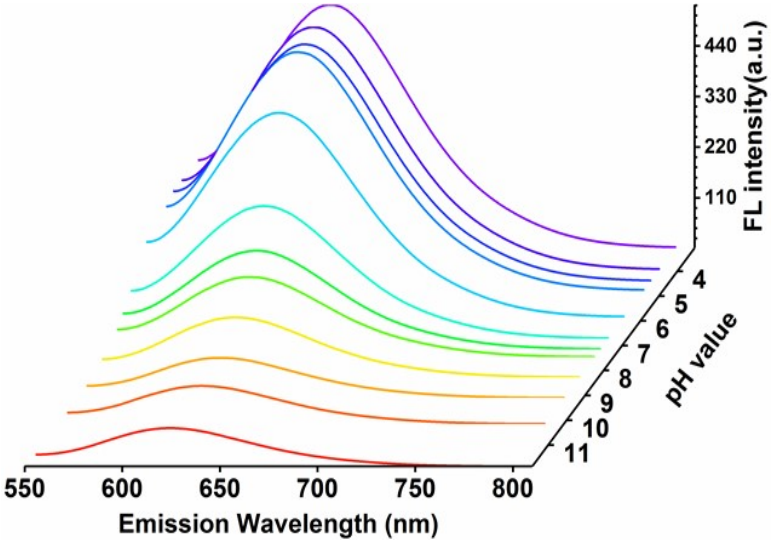

Figure S7. Fluorescence spectral changes of R-CDs with the pH value reducing from 11.86 to 3.07 (Ex = 536 nm).

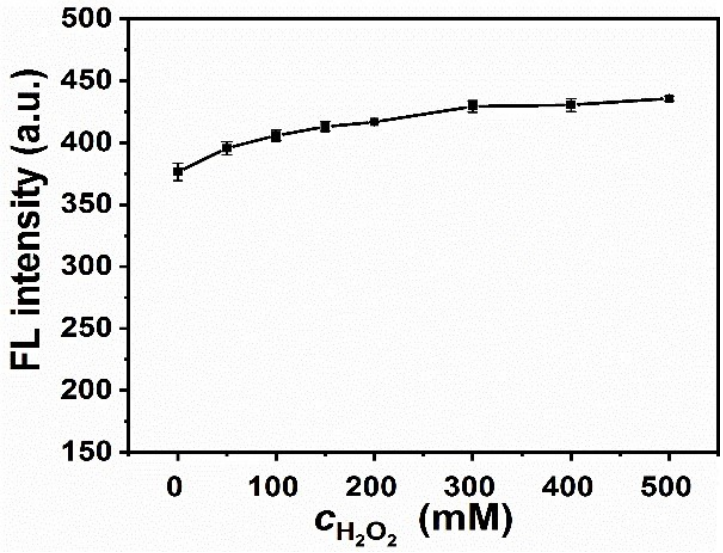

Figure S8. Fluorescence intensity of R-CDs at different concentrations of H_2O_2 (0−500 mM).

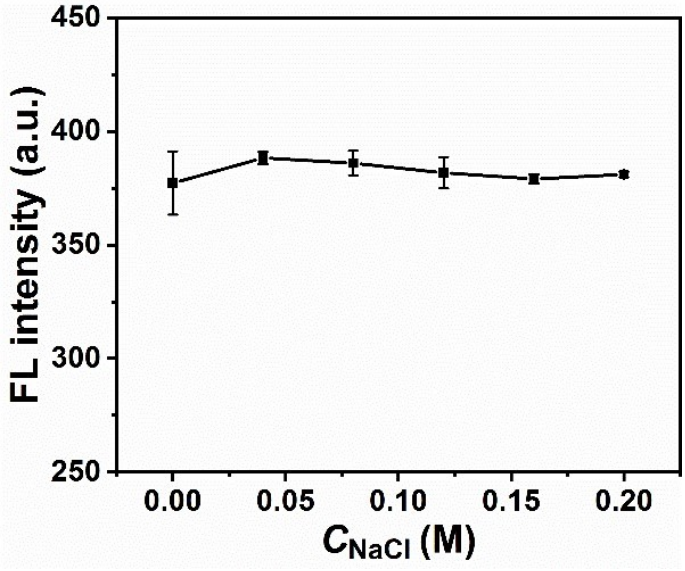

Figure S9. Fluorescence intensity of R-CDs at different concentrations of NaCl (0−0.20 M).

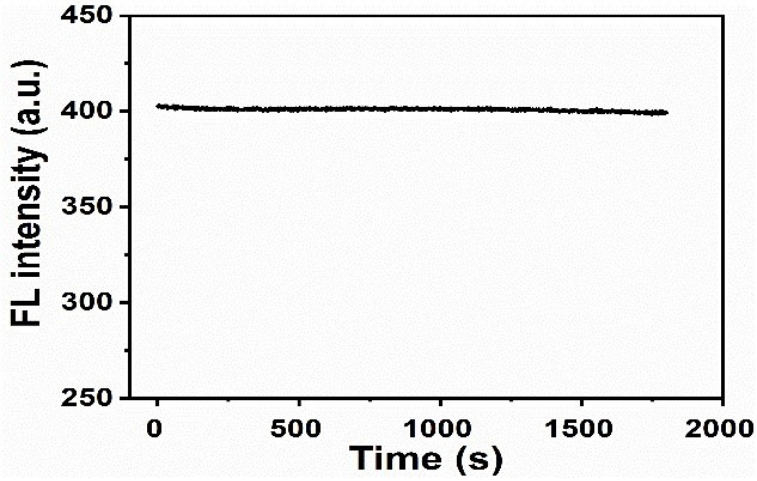

Figure S10. The fluorescence intensity of 1800 s was irradiated by CDs under 536 nm

excitation light.

9. Living cell imaging of carbon dots

Cell culture. HeLa cells were cultured in DMEM containing 10% fetal bovine serum (FBS, Sigma), streptomycin (100μg/mL) and penicillin (100μg/mL). The petri dish was placed in a humid environment containing 5% CO_2 at 37°C.

Cytotoxicity Assay. Cell counting kit-8 (CCK-8) assay was employed for evaluating the Cytotoxicity of R-CDs. Generally, Hela cells were seeded in 96-well plates at a density of 1 × 105 cells per well in 100μL DMEM culture medium supplemented with 2% FBS and cultured for 24 h. Then the cells were incubated with different concentrations of R-CDs for 24h. After wiping off culture medium and washing with phosphate-buffered saline (PBS) three times, 100 μL of CCK-8 solutions (10%) was added to each well and incubated for another 15 minutes. The absorbance of each well at 450 nm was measured with a microplate reader. Cell viability was calculated by the following formula:

$$Cell\ viability(\%) = \frac{OD_{Treated} - OD_{PBS}}{OD_{Control} - OD_{PBS}}$$

$OD_{Treated}$ represents the optical density adding R-CDs, $OD_{Control}$ represents the optical density without R-CDs, OD_{PBS} refer to the absorbance of PBS without cells in the 96 well-plates.

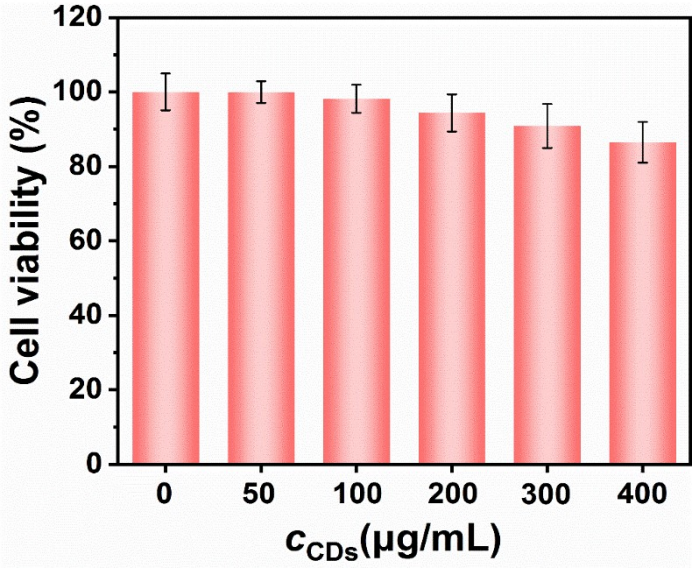

Figure S11. Cytotoxicity of R-CDs for HeLa cells at different concentrations using CCK-8 assay. (n=6)

Figure S11 shows the cytotoxicity of R-CDs for HeLa cells at different concentrations using CCK-8 assay. We can see that the cell viability remains above 85% at 400 mg/mL. It indicates that the red CDs have good biocompatibility and can be used for cell imaging analysis.

Confocal Fluorescence Imaging in Live Cells. Hela cells were used for confocal imaging, as a rule, 1 mL cells with 1×10^5 cells/mL density seeded on glass-bottom culture dishes (35 mm) with 2% FBS in a media at 37 °C for 24 h. After wiping off culture medium and washing with PBS three times, culture dishes were replaced by 10μL of R-CDs solution(2.93mg/mL) and 90μL of DMEM culture medium without FBS. After incubation for 2 hours, wash culture dishes with PBS three times and then add 1mL medium. Finally, Olympus IX-81 microscopy was employed for visualizing the cellular imaging of R-CDs.