Electronic Supplementary Material (ESI) for ChemComm.
This journal is © The Royal Society of Chemistry 2020

Electronic Supplementary Information

Exploiting Deep Learning for Predictable Carbon Dots Design

Xiao-Yuan Wang[#], Bin-Bin Chen[#], Jie Zhang, Ze-Rui Zhou, Jian Lv, Xiao-Peng Geng, and Ruo-Can Qian*

East China University of Science and Technology

E-mail: ruocanqian@ecust.edu.cn

Supplementary Methods

Materials and reagents, Apparatus, Synthesis of CDs, Cell culture, Imaging of CDs in HeLa cells, Construction of DCNN, Information extraction from the images

Supplementary Table S1

Feature extraction from reactants.

Supplementary Table S2

The parameters and performance of different models.

Supplementary Table S3

Reaction conditions for CDs synthesis.

Supplementary Figure S1

The structure of the DCNN.

Supplementary Figure S2

The performance of different models.

Supplementary Figure S3

3D fluorescence spectra of CDs.

Supplementary Figure S4

CDs solutions under 365 nm UV lamps.

Supplementary Figure S5

A screenshot of the webpage sharing the codes and database in our work.

Supplementary Figure S6

The treatment of 2D pseudo-color images by Matlab software.

References

Supplementary Methods

Materials and reagents. Ethylenediamine (EDA), 1,4-benzoquinone, phenyl-p-benzoquinone, 2,5-diphenyl-p-benzoquinone, chloranil, alanine, valine, leucine, citric acid, sodium citrate, and L-cystine were purchased from Aladdin Reagent Co., Ltd. (Shanghai, China). Thiourea and N,N-dimethylformamide (DMF) were purchased from Shanghai Lingfeng Chemical Reagent Co., Ltd. L-serine and L-tryptophan were purchased from Damas-Beta Co., Ltd. (Shanghai, China). Ethanol was purchased from Sinopharm Chemical Reagent Co., Ltd. Triethylenetetramine (TETA) was purchased from Energy-Chemical Co., Ltd (Shanghai, China). All solutions were prepared using 18.2 MΩ cm of ultrapure water (EMD Millipore, TONDINO, Shanghai). HeLa cervical cancer cells were obtained from Shanghai Moxi Boil Co., Ltd.

Apparatus. FL images of the CDs were obtained using a laser confocal microscope (Leica, TCS SP8). The FL spectra of the CDs were recorded using an FL spectrophotometer (LS-55 Lumine).

Synthesis of CDs. A series of CDs (twelve CDs, labeled in numerical order) were synthesized using a one-pot hydrothermal method. The reactants and synthesis conditions are listed in Table S2. The CDs were formed through hydrothermal carbonization. Residual amounts of reactants were removed using a cellulose ester dialysis membrane (500-1000 MWCO) over 48 h. The resulting CDs were dried via lyophilization and dispersed in water for further use.

Cell culture. HeLa cells were cultured in DMEM (Gibco) with 10 % fetal bovine serum (FBS, Sigma), streptomycin (100 μg/mL), and penicillin (100 μg/mL). The culture dishes were placed in a humid atmosphere at 37 ^{o}C with 5 % CO_2.

Imaging of CDs in living cells. HeLa cells in DEME supplemented with 10 % FBS were added to culture dishes. Then cells were cultured for 24 h in an incubator (37 ^{o}C, 5 % CO_2). After 24 h incubated, the culture medium was replaced with 5 mL DEME containing 2 mM CDs. The cells were cultured for 2 h in an incubator (37 ^{o}C, 5 % CO_2), and then rinsed with PBS buffer three times, and transferred for FL imaging.

Construction of DCNN. The structure of the DCNN was depicted in Figure S1**.** Cross Entropy Error Function was introduced as loss function. The equation was defined as:

$$\mathrm{L} = \frac{1}{N}\sum_{i} -[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

y_i refers to the label of sample i, while p_i is the related prediction. N is the number of sample.

the bounds of the grid search are listed below:

(1) Batch size and epochs: batch size [8, 16, 24, 32], epochs [100, 200, 300, 400, 500]

(2) Training Optimization Algorithm: ['SGD', 'RMSprop', 'Adagrad', 'Adadelta', 'Adam', 'Adamax', 'Nadam']

(3) Learning rate: [0.0001, 0.001, 0.01, 0.1, 0.2]

(4) Network Weight Initialization: ['uniform', 'lecun_uniform', 'normal', 'zero', 'glorot_normal', 'glorot_uniform', 'he_normal', 'he_uniform']

(5) Neuron Activation Function: ['softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid', 'linear']

Supplementary Note 1. Feature extraction from reactants.

Reactants are very important in terms of the structure and optical properties of CDs, thus the numbers of atomic components in the reactants were selected as the first and foremost feature to be considered. To maintain the convenience of calculation and generalizability for other reaction predictions, the number of reactants was limited to two. Regarding the volume of solvent, reaction temperature, and reaction time, relative numeric values were directly used as input features for uniform units. Solvents are difficult to be transformed into data that can be understood by a machine. One convenient and feasible solution is to number all the solvents because of the limited categories of different solvents. The features were arranged and shown in Table S1 after Python processing.

Table S1. Feature extraction from reactants.

Feature	Description	Type
C	The number of C atoms	Reagent properties
H	The number of H atoms	Reagent properties
O	The number of O atoms	Reagent properties
N	The number of Na atoms	Reagent properties
B	The number of B atoms	Reagent properties
Cl	The number of Cl atoms	Reagent properties
Na	The number of Na atoms	Reagent properties
S	The number of S atoms	Reagent properties
Zn	The number of Zn atoms	Reagent properties
F	The number of F atoms	Reagent properties
P	The number of P atoms	Reagent properties
Fe	The number of Fe atoms	Reagent properties
Cu	The number of Cu atoms	Reagent properties
I	The number of I atoms	Reagent properties
K	The number of K atoms	Reagent properties
Si	The number of Si atoms	Reagent properties
Mass	The mass of the reagent	Reaction condition
Solvent	The type of solvent	Reaction condition
Volume	The volume of solvent	Reaction condition
Temperature	The reaction temperature	Reaction condition
Time	The reaction time	Reaction condition

Supplementary Note 2. Construction of DCNN.

A typical 1D CNN classification model is formed by two convolution layers, two full connected layers, a dropout layer, and a output layer. The input dimension of the two convolution layers is 40 (corresponding to extract 40 features) with 32 and 64 kernels of size 3, respectively. The convolution layers are followed by a flattened and two hidden layers with 64 and 32 units respectively. Finally, a dropout layer with a 10% dropout rate was added. Deep learning combines low-level features to generate more abstract high-level representation attribute categories or features, allowing effective discovery of the distributed feature representations from data. Therefore, a multi-layer neural network structure is extremely suitable for revealing the hidden rules behind the synthesis of CDs. We adopted a solver using Adaptive moment estimation (Adam) and data shuffling.

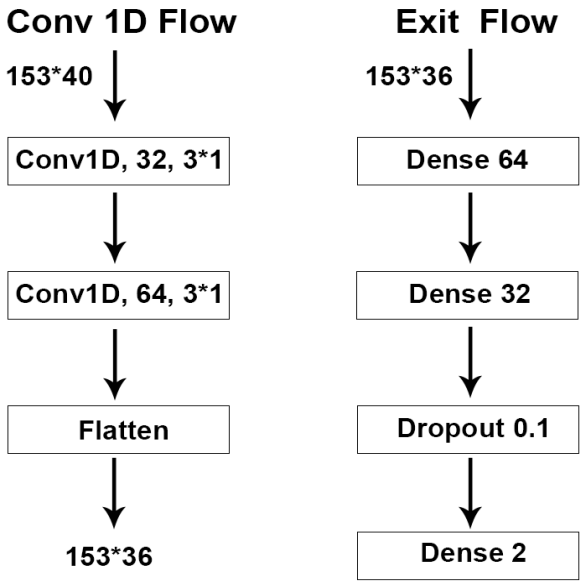

Figure S1. The structure of the DCNN.

Supplementary Note 3. Other machine learning models

Other machine learning models were explored. Among these models, DCNN possess higher accuracy whether in training or validation dataset. All models construction were performed by sklearn.[1]

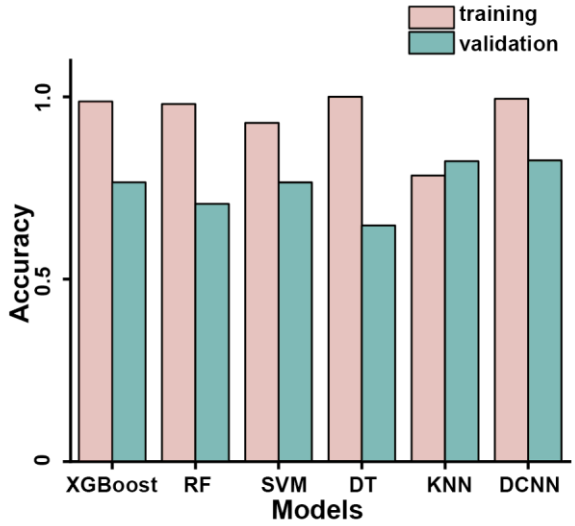

Figure S2. The performance of different models.

Table S2. The parameters and performance of different models.

Model	Parameters	Training accuracy	Validation accuracy
XGBoost	XGBClassifier(n_estimators=100,max_depth = 2)	0.987	0.765
RF	RandomForestClassifier()	0.980	0.706
SVM	SVC(C = 10, kernel='rbf', probability=True,decision_function_shape='ovo')	0.928	0.765
DT	DecisionTreeClassifier()	1.000	0.647
KNN	KNeighborsClassifier()	0.784	0.823
DCNN	Described in Figure S1	0.994	0.824

Supplementary Note 4. Reaction conditions for CDs synthesis.

Table S3. Reaction conditions for CDs synthesis.

CDs	Reactant 1	Reactant 2	Solvent	Temperature	Time
CDs-1	1,4-Benzoquinone (0.1 g)	EDA (200 μL)	Water (5 mL)	160 °C	5 h
CDs-2	Phenyl-p-benzoquinone (0.1 g)	EDA (200 μL)	Water (5 mL)	160 °C	5 h
CDs-3	1,4-Benzoquinone (0.1 g)	TETA(200 μL)	Water (5 mL)	160 °C	5 h
CDs-4	Phenyl-p-benzoquinone (0.1 g)	TETA(200 μL)	Water (5 mL)	160 °C	5 h
CDs-5	Tetrachlorobenzoquinone (0.1 g)	EDA (200 μL)	Water (5 mL)	160 °C	5 h
CDs-6	Tetrachlorobenzoquinone (0.1 g)	TETA (200 μL)	Ethanol (5 mL)	160 °C	5 h
CDs-7	2,5-Diphenyl-p-benzoquinone (0.1 g)	TETA (200 μL)	DMF (5 mL)	160 °C	5 h
CDs-8	6,13-Pentacenequinone(0.1g)	TETA (200 μL)	Water (5 mL)	200 °C	5 h
CDs-9	L- Cystine (0.1 g)	L-Cysteine (0.1g)	Water (5 mL)	190 °C	4 h
CDs-10	L- Cystine (0.1 g)	L-Cysteine (0.1g)	Ethanol (5 mL)	190 °C	4 h
CDs-11	L- Cystine (0.1 g)	L-Cysteine (0.1g)	DMF (5 mL)	190 °C	4 h
CDs-12	Sodium citrate (0.3 g)	Thiourea (0.1 g)	Water (10 mL)	180 °C	5 h
CDs-13	Citric acid (0.1 g)	Tryptophan (0.1g)	Water (6 mL)	120 °C	5 h
CDs-14	Citric acid (0.1 g)	Alanine(0.1g)	Water (6 mL)	120 °C	5 h
CDs-15	Citric acid (0.1 g)	Valine(0.1g)	Water (6 mL)	120 °C	5 h
CDs-16	Citric acid (0.1 g)	Leucine(0.1g)	Water (6 mL)	120 °C	5 h

Supplementary Note 5. 3D fluorescence spectra of CDs

The 3D FL spectra of CDs were obtained by origin 2017. The X-axis is the wavelength of emission, while Y-axis is the wavelength of excitation. The depth of the color represents the intensity.

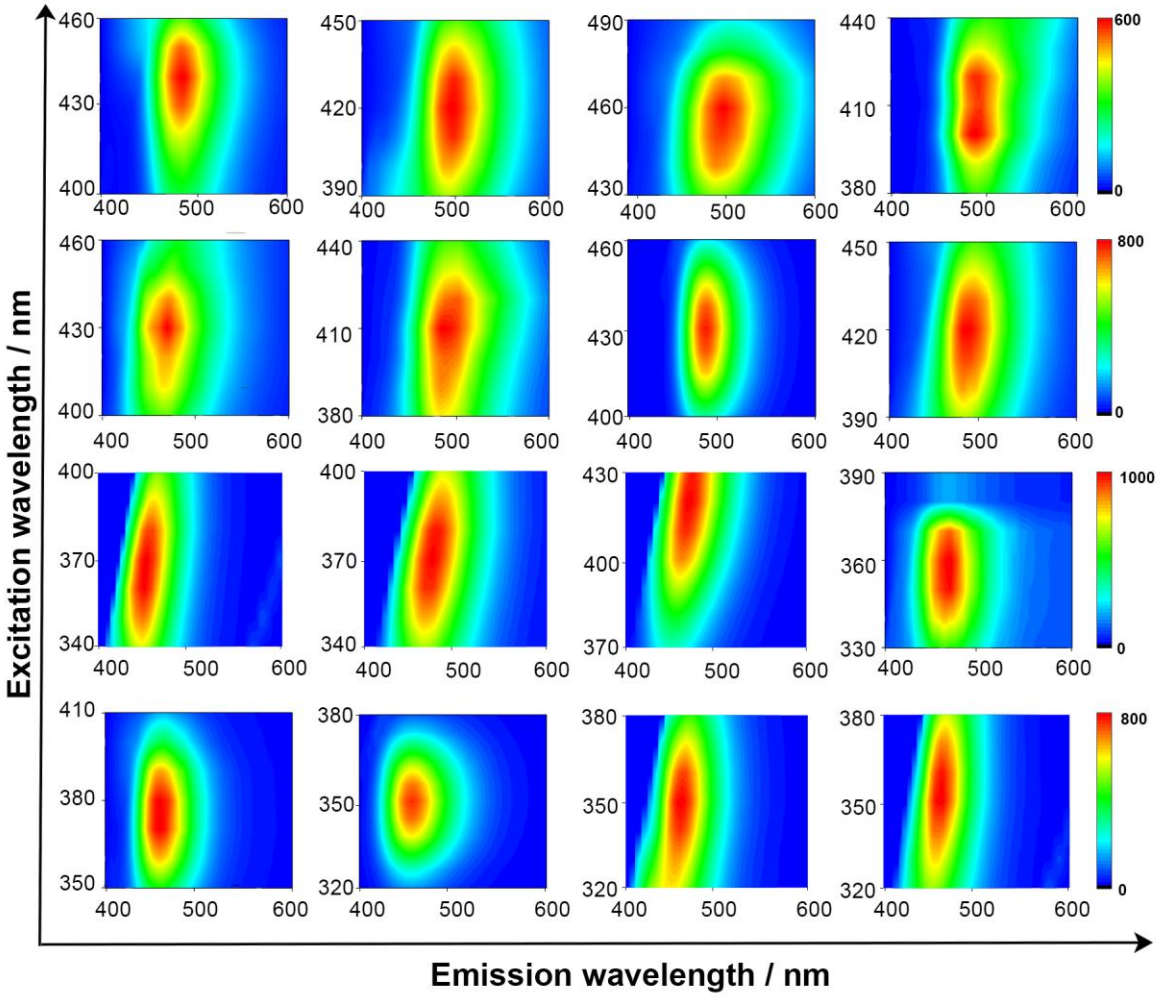

Figure S3. 3D fluorescence spectra of CDs.

Supplementary Note 6. CDs solutions under 365 nm UV lamps

The CDs solution was added to the colorimetric dish, then photos were acquired by smart phone under 365 nm UV lamps.

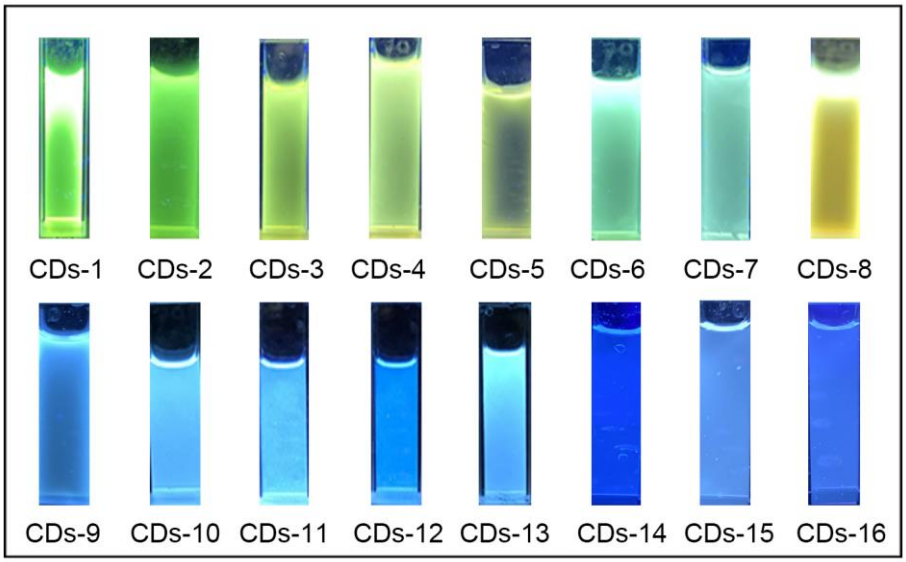

Figure S4. CDs solutions under 365 nm UV lamps.

Supplementary Note 7. Codes and database published through Github

All codes and data for training were uploaded at https://github.com/xiaoyuan-sudo/CDs.

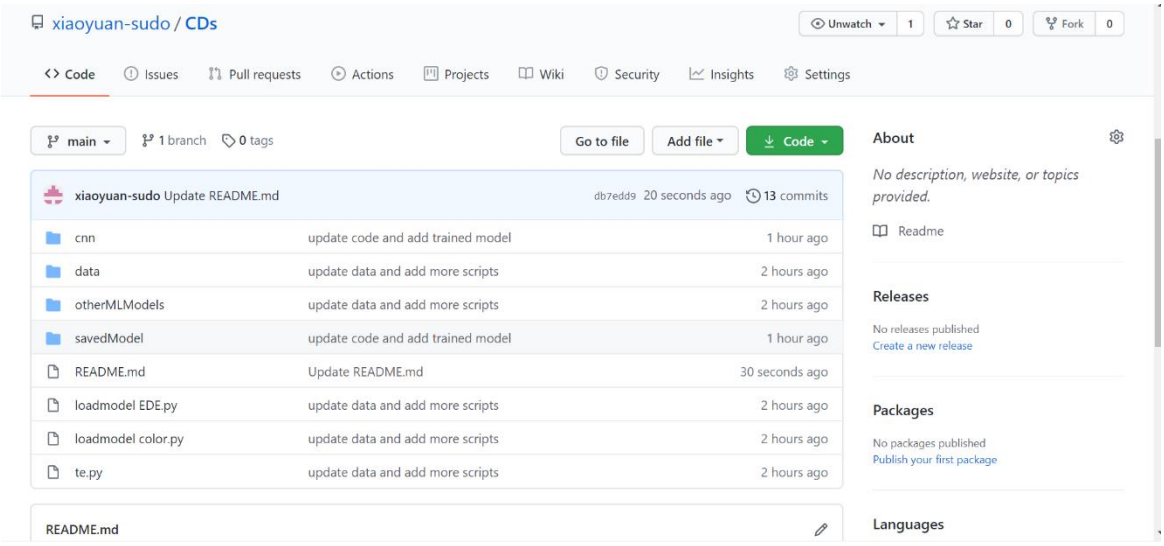

Figure S5. A screenshot of the webpage sharing the codes and database in our work.

Supplementary Note 8. Treatment of 2D pseudo-color images using Matlab software

The 2D pseudo-color blue and green channel intensity corresponding to HeLa cells could be obtained by Matlab software. The codes (left column) and corresponding comments (right column) were as follows:

```
I = imread('D:\files name');        %read image
J = imresize(I,[100 100]);          %change the image size to 100*100 pixels
g = J(:,:,2);                       %extract green channel
contour(g)                          %contour map
colormap(jet)                       %fill the contour map with color
colorbar                            %add color bar
```

Figure S6. The treatment of 2D pseudo-color images Matlab software.

Reference

[1] Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.