Supplementary Information

Enhancing chemistry-intuitive feature learning to improve prediction performance of optical properties

Ming Sun, †[a] Caixia Fu, †[a] Haoming Su,[a] Ruyue Xiao,[a] Chaojie Shi,[a] Zhiyun Lu *[a] and Xuemei Pu*[a]

a. College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China. E-mail: xmpuscu@scu.edu.cn; luzhiyun@scu.edu.cn.

Contents

1. The model evaluation metrics

To evaluate the model's performance, we use four main metrics widely used to assess.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_{pred,i} - y_{exp,i}|$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{pred,\,i} - y_{exp,\,i})^2}{N}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{exp,i} - y_{pred,i})^2}{\sum_{i=1}^{N}(y_{exp,i} - \overline{y}_{exp})^2}$$

$$r = \frac{\sum_{i=1}^{N}(y_{exp,i} - \overline{y}_{exp})(y_{pred,i} - \overline{y}_{pred})}{\sqrt{\sum_{i=1}^{N}(y_{exp,i} - \overline{y}_{exp})^2}\sqrt{\sum_{i=1}^{N}(y_{pred,i} - \overline{y}_{pred})^2}}$$

Among these equations, $y_{pred,i}$ and $y_{exp,i}$ denote the predicted value and experimental value of sample *i*, respectively. $\overline{y}_{pred}$ and $\overline{y}_{exp}$ represent the average value of predicted values and experimental values, respectively. MAE is the mean absolute error. RMSE is the root mean square error. R^2 is the coefficient of determination, with a maximum value of 1. And the closer R^2 is to 1, the better the fit. r is the Pearson correlation efficient, with a maximum value of 1. And the closer r is to 1, the more correlated the predicted values are with the experimental values. In general, if MAE and RMSE are smaller, meanwhile, R^2 and r are closer to 1, then the model performs better.

2. Details of experiment characterizations

^{1}H NMR spectra were recorded on a Bruker Avance II-400 MHz spectrometer at 400 MHZ with DMSO-d_6 as the solvent and tetramethylsilane (TMS) as the internal reference. ^{13}C NMR spectrum was recorded on a Bruker Avance III-800 MHz spectrometer at 200 MHZ with DMSO-d_6 as the solvent and tetramethylsilane (TMS) as the internal reference. High resolution mass spectrometry (HRMS) spectrum was measured on a Q-TOF Premier ESI mass spectrometer (Micromass, Manchester, UK). Fluorescence spectra were collected on a Horiba Jobin Yvon-Edison Fluoromax-4 fluorescence spectrometer. UV-Vis absorption spectra were obtained on a UV 2600 spectrophotometer. The doped films of compounds were spin-coated from their corresponding 1,2-dichloroethane solutions with a concentration of 10 mg mL^{-1} at a speed of 1500 rpm on quartz substrates for 30 s. The absolute PLQY of the film sample were determined on a HORIBA Jobin Yvon Fluorolog-3 fluorescence spectrometer equipped with an integrating sphere (IS80 from Labsphere) and a digital photometer (S370 from UDT) under ambient conditions.

3. Quantum mechanics (QM) calculation

For the QM calculation, we performed the conformer search using Open Babel software with MMFF94 force field. The vertical transitions were considered for the excited state properties. Considering the solvent environment, we adopted the polarizable continuum model (PCM) with Gaussian 09 software for all the quantum chemistry calculations in this work, including optimizations of the ground state and S_1 state. Concretely, we optimized the ground state (S_0) with B3LYP hybrid functional and 6-31G(d) basis set first. Then S_1 state was calculated using time-dependent density functional theory (TD-DFT) with CAM-B3LYP/6-31G(d) method. In addition, the S_1 state was also optimized at CAM-B3LYP/6-31G(d) level. All the QM calculations were performed by Gaussian 09 package.[1]

4. Supporting tables and figures

Table S1. The ablation experiment for impact of each state feature on the model performance.[a]

	MPNN	MPNN-RB	MPNN-NO	MPNN-AA	MPNN-Al	MPNN-Ar	MPNN-State
λ_{Abs}	12.23	11.81	10.90	10.73	11.96	11.06	9.97
λ_{Emi}	16.83	14.79	15.04	15.45	14.52	15.34	14.55
FWHM	11.43	11.10	11.19	11.24	11.08	10.90	10.24
PLQY	0.131	0.131	0.126	0.123	0.123	0.128	0.120

[a]The evaluation metric is mean absolute error (MAE). For the absorption, emission and FWHM, the unit is nm. MPNN denotes the conventional MPNN architecture. MPNN-RB denotes MPNN coupled with the RotatableBond feature. MPNN-NO denotes MPNN coupled with the Fr_NO feature. MPNN-AA denotes MPNN couple with the Fr_AromAtoms feature. MPNN-Al denotes MPNN coupled with the AliphaticRings feature. MPNN-Al denotes MPNN coupled with the AromaticRings feature. MPNN-State represents the MPNN model coupled with all the five state features.

In the work, we chose the five state features based on the experimental findings. Specifically, the experiments reported that the spiral structures, aromaticity, and molecular rings play important roles in the four optical properties.[2–4] Thus, we selected the number of rotatable bonds (RotatableBond), the fraction of nitrogen and oxygen (Fr_NO), the fraction of aromatic atoms (Fr_AromAtoms), the number of aliphatic rings (AliphaticRings), and the number of aromatic rings (AromaticRings) as state features to characterize the three important structure factors. In order to evaluate impacts of the five state features, we compared the prediction performances of MPNN separately coupled with the five different state features on the four optical properties. It can be seen from Table S1 that each state feature improves the prediction performance with respect to the pure MPNN architecture, as evidenced by

lower MAE. These ablation experiment confirms the effectiveness of the five state features in improving the model prediction performance. After fusing all the five state features into MPNN (i.e., MPNN-State), the model achieves the best performance with respect to any single state function, showcasing the necessity of considering five state features.

Table S2. The similarity comparison among the five datasets based on duplicates.[a]

Datasets	Deep4Chem	ChemFluor	SMFluo1	BODIPYs	JCIM_Abs
Deep4Chem	/	0	0	0	16585
ChemFluor	0	/	0	4166	4170
SMFluo1	0	11	/	0	0
BODIPYs	0	4166	0	/	4166
JCIM_Abs	16585	4170	0	4166	/

[a] The optical properties are not only related to the emitter, but also influenced by the solvent. And all the experimental data of these five datasets were measured in different solvents. Thus, we should consider both the emitter SMILES and the solvent SMILES to assess the difference among the five datasets. Herein, following the method of removing any dye–solvent pairs with duplicate measurements proposed by Greenman *et al.*, [5] we find out any emitter–solvent pairs with duplicate measurements among these five datasets. Deep4Chem, ChemFluor, and SMFluo1 do not have any duplicates, indicating they are completely different datasets. BODIPYs and JCIM_Abs datasets are combined datasets, in which some data are aggregated from Deep4Chem and ChemFluor. BODIPYs and JCIM_Abs individually contain nearly 13300 and 27000 samples, which exhibit approximate 1/3 and 1/2 duplicates with respect to ChemFluor and Deep4Chem, respectively, thus still maintaining acceptable differences in data.

Table S3. The Pearson correlation coefficients between the computational values and the experimental ones for the new molecule PPI-2TPA.[a]

Pearson correlation coefficients	Abs_exp	Emi_exp
Predicted	0.89	0.99
Calculated	0.00	0.81

[a]Predicted and Calculated denote the results from our SubOptGraph and QM, respectively. Abs_exp and Emi_exp denote experimental absorption and emission wavelengths for PPI-2TPA, respectively. It can be seen that the predicted values have significantly higher correlation

coefficients with the experimental values than the calculated ones by QM, further confirming that our model can better predict the new molecule PPI-2TPA than the QM calculation.

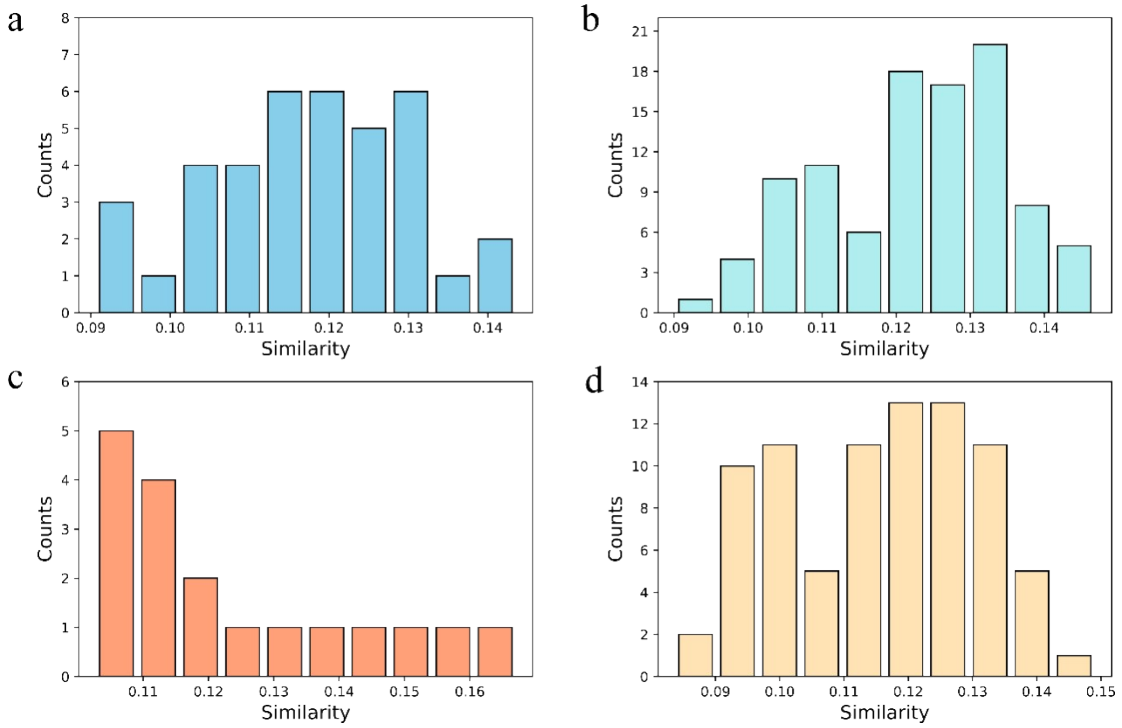

Fig. S1 The similarity histogram between 179 blue OLED emitters of the external test set and those in the training set of Deep4Chem for (a) the absorption wavelength. (b) Emission wavelength. (c) FWHM. (d) PLQY. We utilized the RDKit package to calculate the average values of Taminato similarity between every OLED emitter and those in the training set of Deep4Chem, based on the Morgan fingerprints. It can be seen that all the similarities are below 0.17, indicating low similarity between the OLED emitters and those emitters in the training set of Deep4Chem.

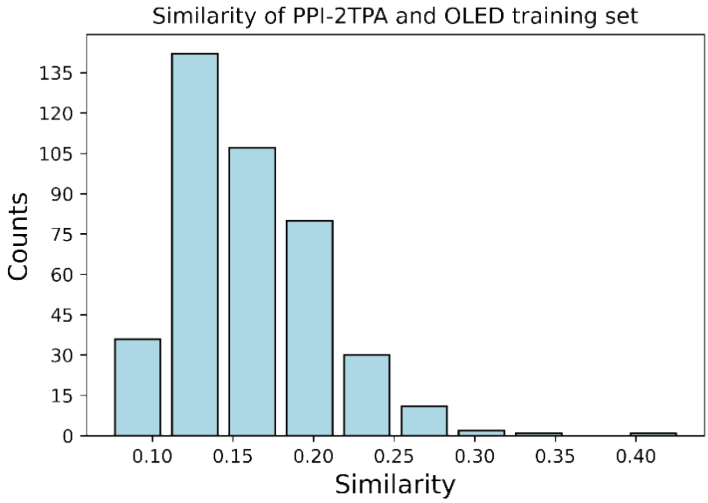

Fig. S2 The similarity histogram between PPI-2TPA and the 410 unique blue OLED emitters from the external test set (238 blue emitter/solvent combinations) and the training set of blue OLED emitters (1114 blue emitter/solvent combinations). We calculated the Taminato similarity between PPI-2TPA and these blue OLED emitters by using RDKit and the Morgan

fingerprints. The similarities are below 0.45 with majority ranging from 0.1 to 0.25, indicating that PPI-2TPA has a different structure from these OLED molecules.

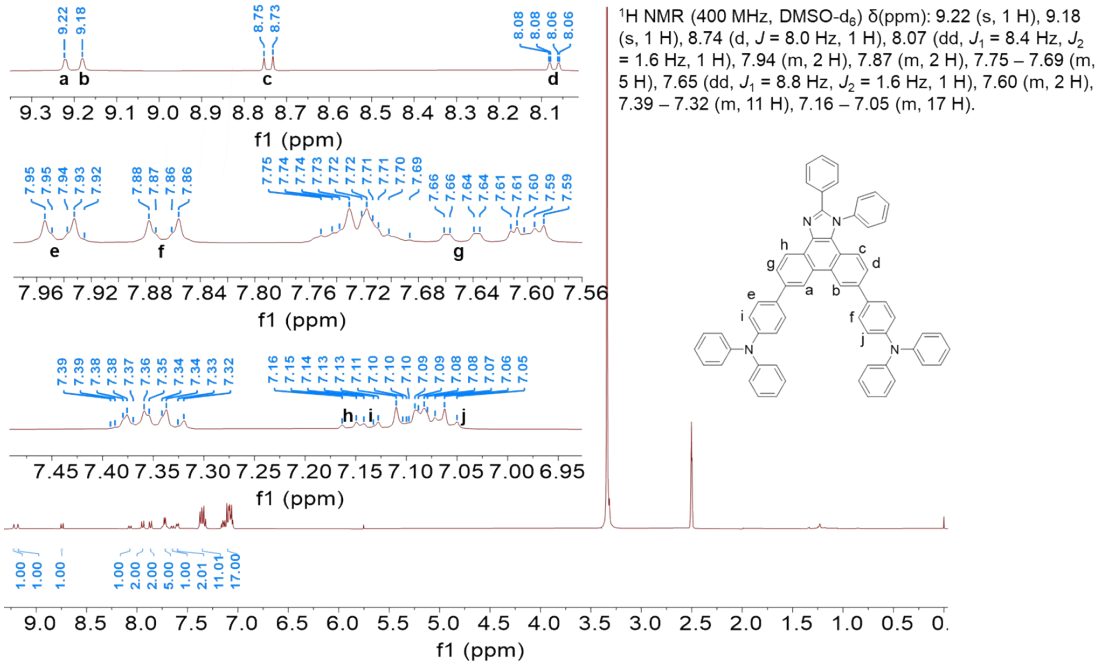

Fig. S3 ^{1}H NMR spectrum of PPI-2TPA

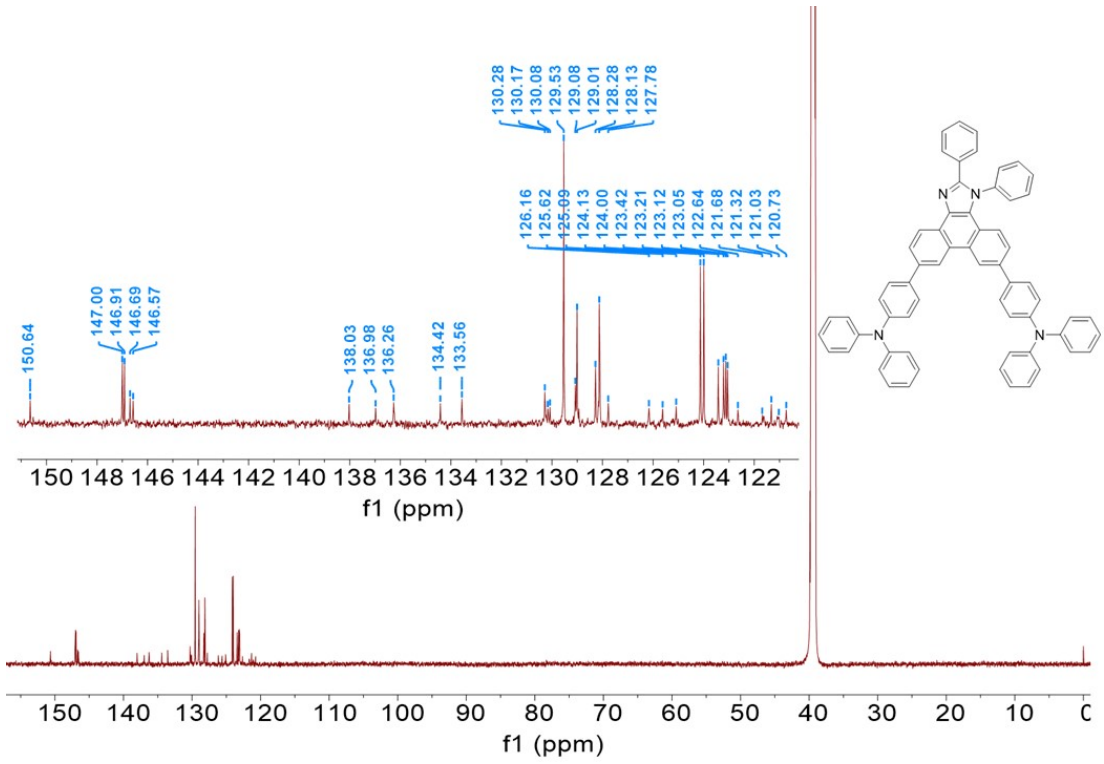

Fig. S4 ^{13}C NMR spectrum of PPI-2TPA

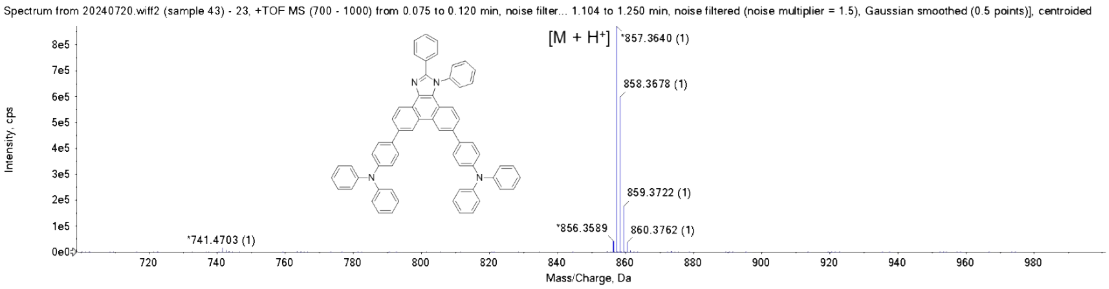

Fig. S5 HRMS spectrum of PPI-2TPA

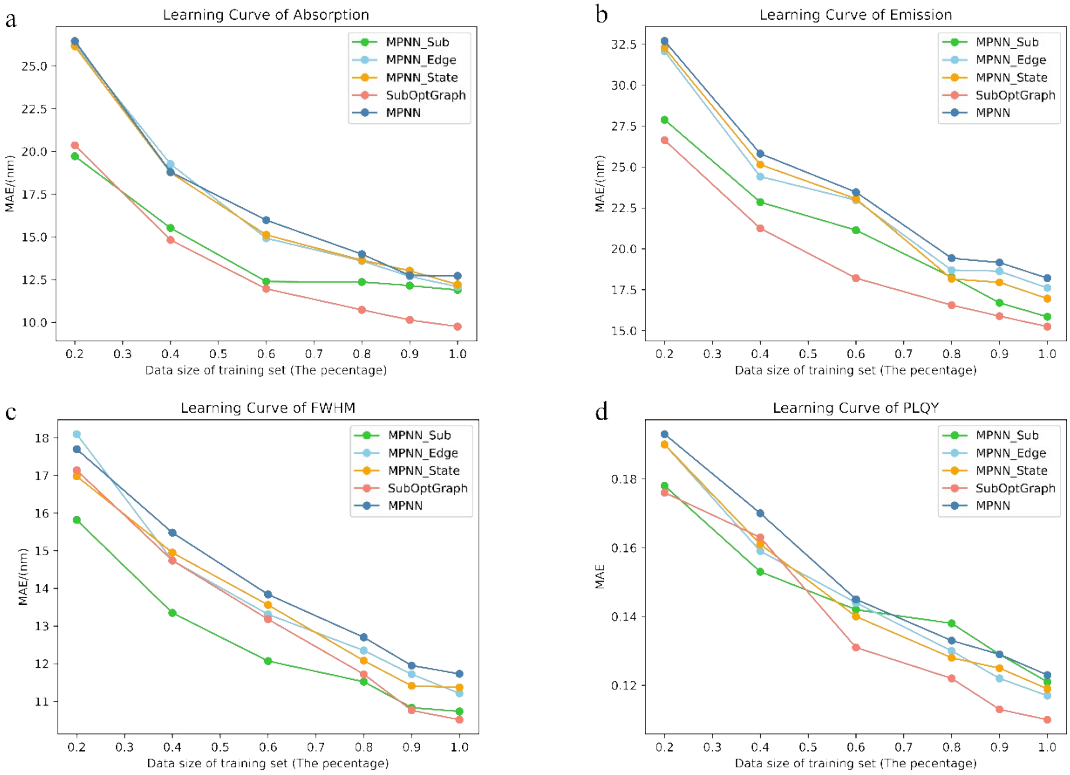

Fig. S6 Learning curves of different models for (a) the absorption wavelength, (b) the emission wavelength, (c) FWHM, and (d) PLQY. The mean absolute error (MAE) is served as the evaluation metric. MPNN-State, MPNN-Sub and MPNN-Edge denote MPNN separately coupled with the state function, the subgraph feature learning and edge feature updating.

The data was divided into the training and testing set in a ratio of 8 : 2. For the training set, we increased its size to assess the performance of the five models on the testing set. With increasing the percentage of training data from 0.2 to 1.0, all the five models perform better, in which SubOptGraph achieves the best performance. In addition, the performance of five models approach convergence when the training data is increased to be 100% percentage, suggesting that the data splitting of the 8:2 ratio is

reasonable for the model training.

References

(1) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.; Izmaylov, A.; Bloino, J.; Zheng, G.; Sonnenberg, J.; Hada, M.; Fox, D. Gaussian 09 (Revision A02). *Gaussian Inc. Wallingford CT* **2009**.

(2) Jousselin-Oba, T.; Mamada, M.; Wright, K.; Marrot, J.; Adachi, C.; Yassar, A.; Frigoli, M. Synthesis, Aromaticity, and Application of *Peri* -Pentacenopentacene: Localized Representation of Benzenoid Aromatic Compounds. *Angew Chem Int Ed* **2022**, *61* (1), e202112794. https://doi.org/10.1002/anie.202112794.

(3) Zhang, Y.-P.; Liang, X.; Luo, X.-F.; Song, S.-Q.; Li, S.; Wang, Y.; Mao, Z.-P.; Xu, W.-Y.; Zheng, Y.-X.; Zuo, J.-L.; Pan, Y. Chiral Spiro-Axis Induced Blue Thermally Activated Delayed Fluorescence Material for Efficient Circularly Polarized OLEDs with Low Efficiency Roll-Off. *Angewandte Chemie International Edition* **2021**, *60* (15), 8435–8440. https://doi.org/10.1002/anie.202015411.

(4) Guo, S.; Wang, L.; Deng, Q.; Wang, G.; Tian, X.; Wang, X.; Liu, Z.; Zhang, M.; Wang, S.; Miao, Y.; Zhu, J.; Wang, H. Exploiting Heterocycle Aromaticity to Fabricate New Hot Exciton Materials. *J. Mater. Chem. C*

2023, *11* (21), 6847–6855. https://doi.org/10.1039/D3TC01192E.

(5) Greenman, K. P.; Green, W. H.; Gómez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13* (4), 1152–1162. https://doi.org/10.1039/D1SC05677H.