Supplementary Materials

Data-driven OLED candidate design: a generative model from independent-property domains to the comprehensive performance enhancement

Xinxin Niu[1], Zhiyao Su[1], Luyu Wang[1], Wenbin Shi[1], Hengyue Zhang[1], Yanfeng Dang[1,*], Yuan Yuan[1,*], Yajing Sun[1,*], Wenping Hu[1,2]

[1]Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China.
[2]Joint School of National University of Singapore and Tianjin University, Fuzhou 350207, Fujian, China.

[*]**Correspondence to:** Dr. Yanfeng Dang, Dr. Yuan Yuan, Dr. Yajing Sun, Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, No. 92 Weijin Road, Nankai District, Tianjin 300072, China. E-mail: yanfeng.dang@tju.edu.cn; yyuan@tju.edu.cn; syj19@ tju.edu.cn

Content:

Supplementary Figure 1. The operating mechanism of the molecular generator and the SMILES processing method. 3

Supplementary Figure 2. Architecture of the GCN model with luminescent molecules and solvents as joint molecular inputs. 4

SupplementaryFigure 3. Visualization of three sampling strategies in the Elite chemical space. 6

SupplementaryFigure 4. Bayesian optimization iteration. 7

SupplementaryFigure 5. Fluorescence emission spectra of example molecules in MolElite calculated by TD-DFT. .. 8

SupplementaryFigure 6. Fluorescence emission spectra of example molecules in MolMediocritys calculated by TD-DFT. 9

SupplementaryFigure 7. Fluorescence emission spectra of MolElite based on ASBase calculated by TD-DFT. 10

SupplementaryFigure 8. Fluorescence emission spectra of MolMediocritys based on ASBase calculated by TD-DFT. 10

SupplementaryFigure 9. 1H-NMR 11

SupplementaryFigure 10. UV absorption spectra 12

SupplementaryFigure 11. Absolute quantum yield 12

SupplementaryFigure 12. ASBase dataset. 13

SupplementaryFigure 13. UMAP 13

SupplementaryFigure 14. The box plots of *Fsp*^{3} variations 14

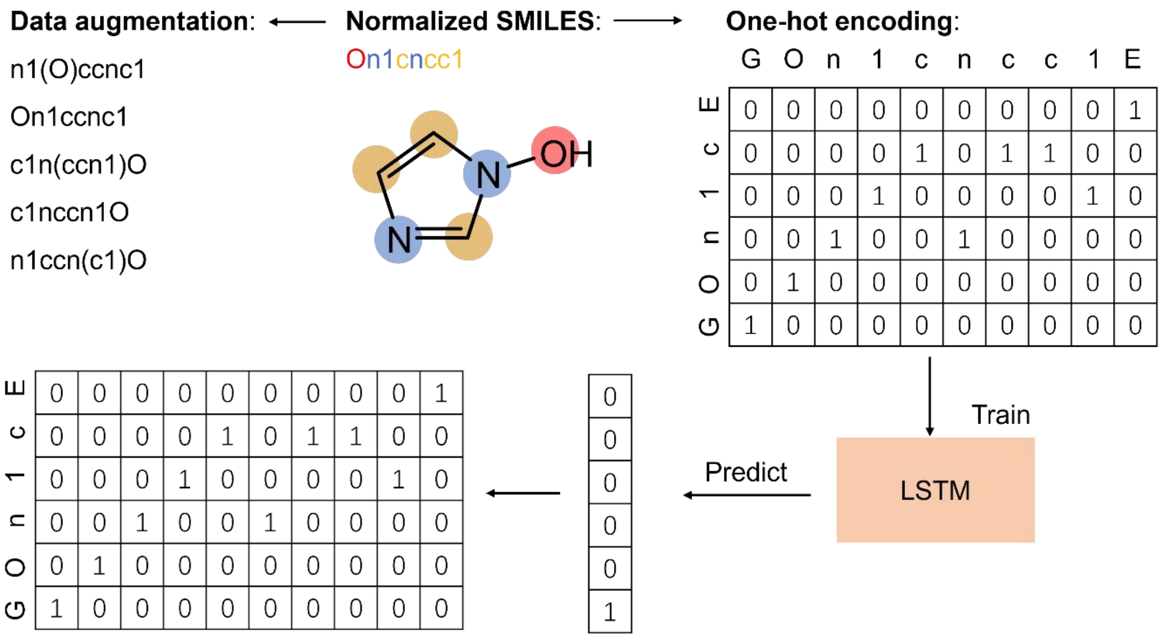
SupplementaryTable 1. Atomic and bond features used in the molecular graph. 15

SupplementaryTable 2. Performance of Union-GCN for predicting four optical properties.. 15

SupplementaryTable 3. Comparison of model performance for predicting the maximum emission wavelength in the DBexp using traditional GCN and Union-GCN. 16

SupplementaryTable 4. The radiative and non-radiative transition rates calculated by momap. 176

SupplementaryTable 5. Performance of Union-GCN for predicting four optical properties in ASBase. 17



	G	O	n	1	c	n	c	c	1	E
E	0	0	0	0	0	0	0	0	0	1
c	0	0	0	0	1	0	1	1	0	0
1	0	0	0	1	0	0	0	0	1	0
n	0	0	1	0	0	1	0	0	0	0
O	0	1	0	0	0	0	0	0	0	0
G	1	0	0	0	0	0	0	0	0	0

E	0	0	0	0	0	0	0	0	0	1
c	0	0	0	0	1	0	1	1	0	0
1	0	0	0	1	0	0	0	0	1	0
n	0	0	1	0	0	1	0	0	0	0
O	0	1	0	0	0	0	0	0	0	0
G	1	0	0	0	0	0	0	0	0	0

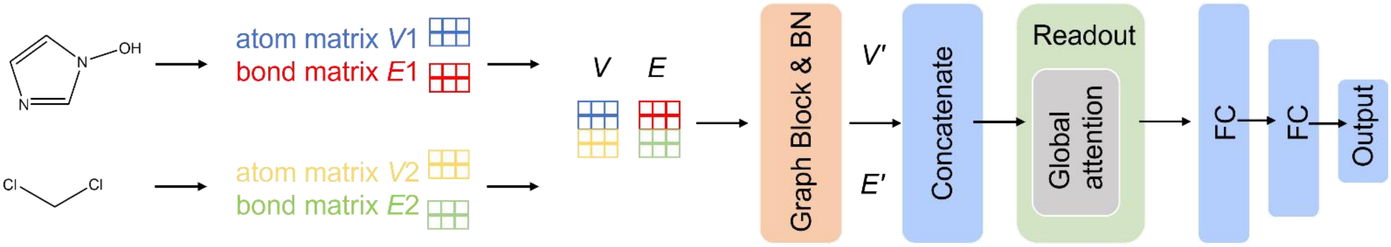
Supplementary Figure 1. The operating mechanism of the molecular generator and the SMILES processing method.

In the one-hot encoding format, each token in the SMILES string vocabulary has a unique numerical vector representation of a predefined length.[1] In this study, we only included the elements H, B, C, N, O, Si, P, S, F, Cl, Br, and I. During model training, the LSTM learns the conditional probability distribution of a token relative to all preceding tokens in the SMILES string. This encoding method assigns a unique mathematical vector to each token within the SMILES string, facilitating the model's learning process. Using multiple representations of the same entity (data augmentation) has been proposed as a strategy to work with small datasets and achieve generalized models.[2] We leverage the non-uniqueness of SMILES strings. Multiple valid SMILES strings representing the same molecular graph are constructed by starting from different non-hydrogen atoms in the molecule. This model was specifically trained on a dataset comprising 365,000 molecules from ChEMBL24.

The molecular generator was implemented as a LSTM network. The neural network consisted of four layers: layer 1, BatchNormalization; layer 2, LSTM with 1,024 units; layer 3, LSTM with 256 units; layer 4, BatchNormalization) and was trained with SMILES strings encoded as one-hot vectors. The categorical cross-entropy loss and the Adam optimizer with a learning rate of 0.001 are used for training the molecular generator. Transfer learning was performed by keeping the parameters of the first model layer constant and training the second layer with a smaller learning rate (equal to 10^{-4}).

To explore the chemical space surrounding high-quality luminescent molecules, we employed an iterative sampling approach. In each iteration of the generator's training, 5,000 molecules were sampled, ensuring comprehensive coverage of the chemical space around the parent molecules. Across multiple iterations, this process resulted in a total of 72,653 unique molecular structures. Notably, in the final sampling iteration, the overlap with the cumulative samples from previous iterations reached 63%, demonstrating the exhaustive nature of the sampling strategy. These results indicate the structural similarity of the sampled molecules to their parent molecules, while maintaining sufficient diversity to explore potential variations.

The Molecular Generator is implemented using a LSTM network, which is particularly well-suited for learning from sequential data such as SMILES strings. LSTM networks can effectively capture both short- and long-range dependencies within sequences, allowing them to learn the syntactic and chemical rules governing valid molecular structures. This capability is crucial for generating chemically valid and syntactically correct molecules, as SMILES encodes molecular graphs as linear character strings where atom and bond information are context-dependent. Unlike traditional RNNs, LSTMs incorporate gating mechanisms that mitigate the vanishing gradient problem, enabling them to preserve information across long sequences. Previous studies have successfully demonstrated the utility of LSTM-based architectures in *de novo* molecular design tasks, particularly for exploring chemical space and generating drug-like or optoelectronic molecules.[2–4] These advantages motivated our choice of LSTM as the backbone for the molecular generator in LumiGen.



Supplementary Figure 2. Architecture of the GCN model with luminescent molecules and solvents as joint molecular inputs.[5]

First-principles calculations.

To further verify the validity of the elite selection strategy, first-principles calculations are adopted. We commenced our study by randomly optimizing the geometric structures of molecules each from the MolElite and MolMediocritys using the Merck molecular force field (MMFF).[6] MMFF, known for its efficacy with organic molecules, provides reliable predictions of three-dimensional structures and molecular energies. These optimized structures form the basis for subsequent quantum chemical calculations. All quantum chemical calculations were conducted using DFT and time-dependent density functional theory (TD-DFT) at the B3LYP/6-31G(d) level with Gaussian 16 software.[7,8] We modeled solvation effects using the Integral Equation Formalism of the Polarizable Continuum Model (IEF-PCM).[9] This step involves separately optimizing the electronic ground states and excited states. TD-DFT is utilized to calculate transition energies to derive the molecular excited-state fluorescence emission spectra and ultraviolet (UV) absorption spectra. Spectra were visualized using Multiwfn.[10] In addition, we use the software package MOMAP, which is designed for predicting the luminescent properties and mobility of organic functional materials, to calculate the rates of radiative and non-radiative transitions of molecules.[11] This allows us to qualitatively describe the photoluminescence quantum yield of the molecules.

Validity = $\frac{N_{val}}{N_{all}}$

Uniqueness = $\frac{N_{uni}}{N_{val}}$

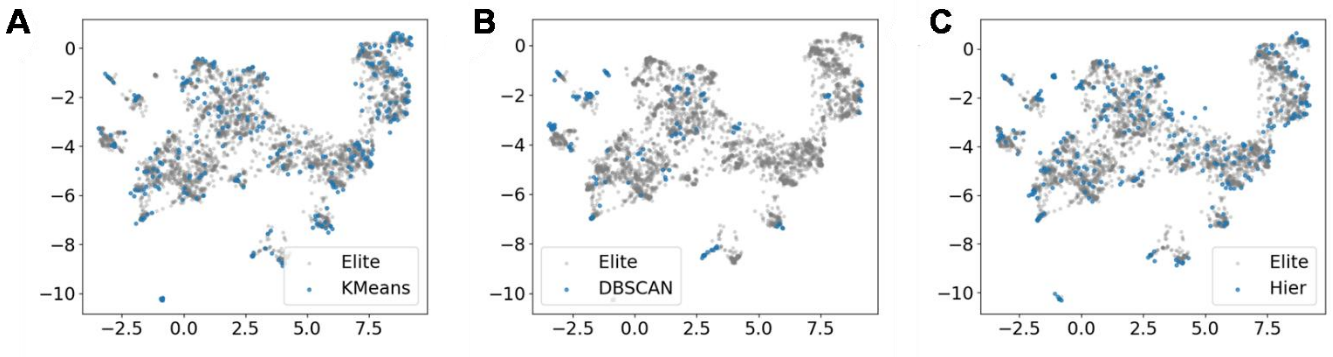
Novelty = $\frac{N_{nov}}{N_{val}}$

N_{all} is total number of generated molecules. N_{val} is the number of valid molecules. Check whether each generated molecule can be parsed as a valid chemical structure. N_{uni} is the number of unique molecules among the valid molecules. N_{nov} is the number of molecules among the valid ones that are not in the training set.

Additionally, we assessed the diversity of the generated molecular scaffolds using Shannon entropy scaled by the number of molecular scaffolds (SSE). Shannon entropy, a fundamental concept from information theory, measures the uncertainty or randomness of information, which, in this context, relates to the variety of molecular scaffolds generated. By normalizing Shannon entropy, we standardize this uncertainty measure to a scale of 0 to 1, making it easier to compare across different datasets.

$$SSE = \frac{\sum_{i=1}^{n} p_i \ln\left(\frac{c_i}{P}\right)}{log_2(n)}$$

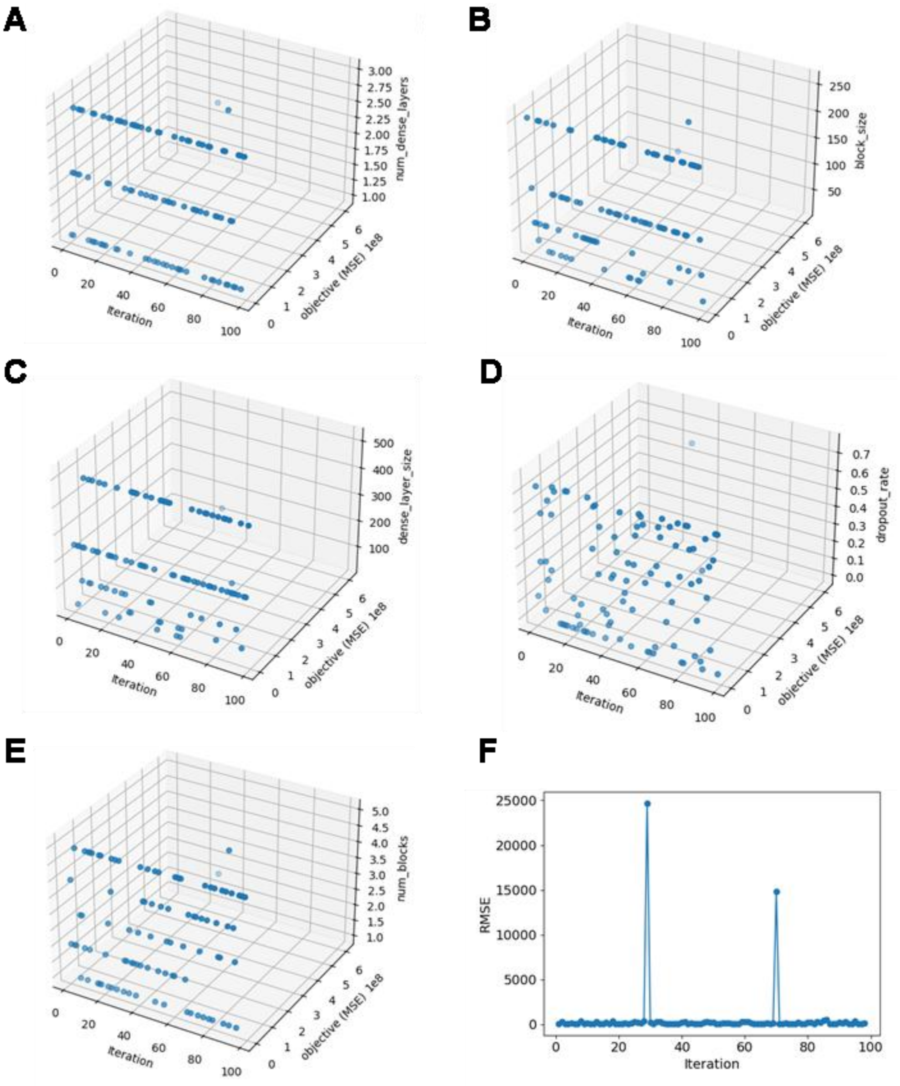
where the numerator is the Shannon entropy, *n* is the number of unique scaffolds considered, c_i is the number of compounds containing the *i*th scaffold and *P* is the total number of compounds among the considered *n* scaffolds.[13]



Supplementary Figure 3. Visualization of three sampling strategies in the Elite chemical space. (A) K-means clustered subset. (B) DBSCAN clustered subset. (C) Hierarchical clustering subset.

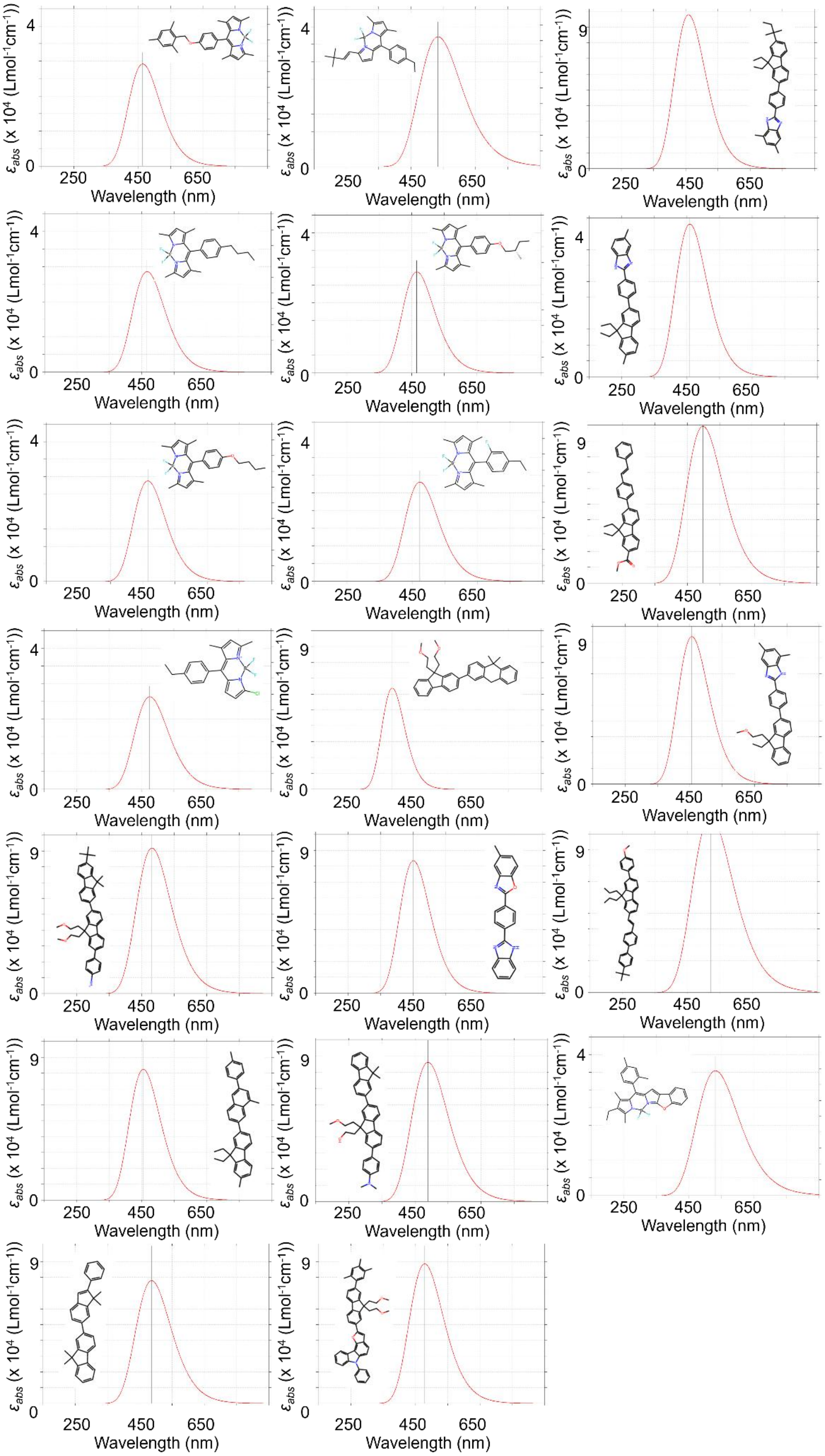
All points were first used to train a UMAP dimensionality-reduction model on the Elite set, after which the iterative data generated by the three sampling strategies were projected into the same two-dimensional coordinate system. Grey dots denote the entire Elite set, while blue dots represent the samples selected by each strategy. K-means and hierarchical clustering behave similarly: the colored samples from both methods almost completely trace the grey outline, uniformly capturing representative molecules from both high- and low-density regions. This stems from their objective of minimizing within-cluster variance, which naturally preserves the “prototype” structure of the chemical space and facilitates diverse exploration in subsequent iterations.

By contrast, DBSCAN’s colored points appear only on local islands of the UMAP map, leaving large grey regions unsampled. Because DBSCAN relies on fixed density thresholds, a small radius of neighborhood identifies only the densest micro-clusters and labels many points as noise, whereas a larger radius of neighborhood quickly merges clusters and sharply reduces category count. This makes it difficult to balance coverage and cluster granularity in high-dimensional, sparse fingerprint space. Overall, if broad and balanced exploration of chemical space is desired, K-means and hierarchical clustering are better suited as core clustering engines for the Sampling Augmentor.

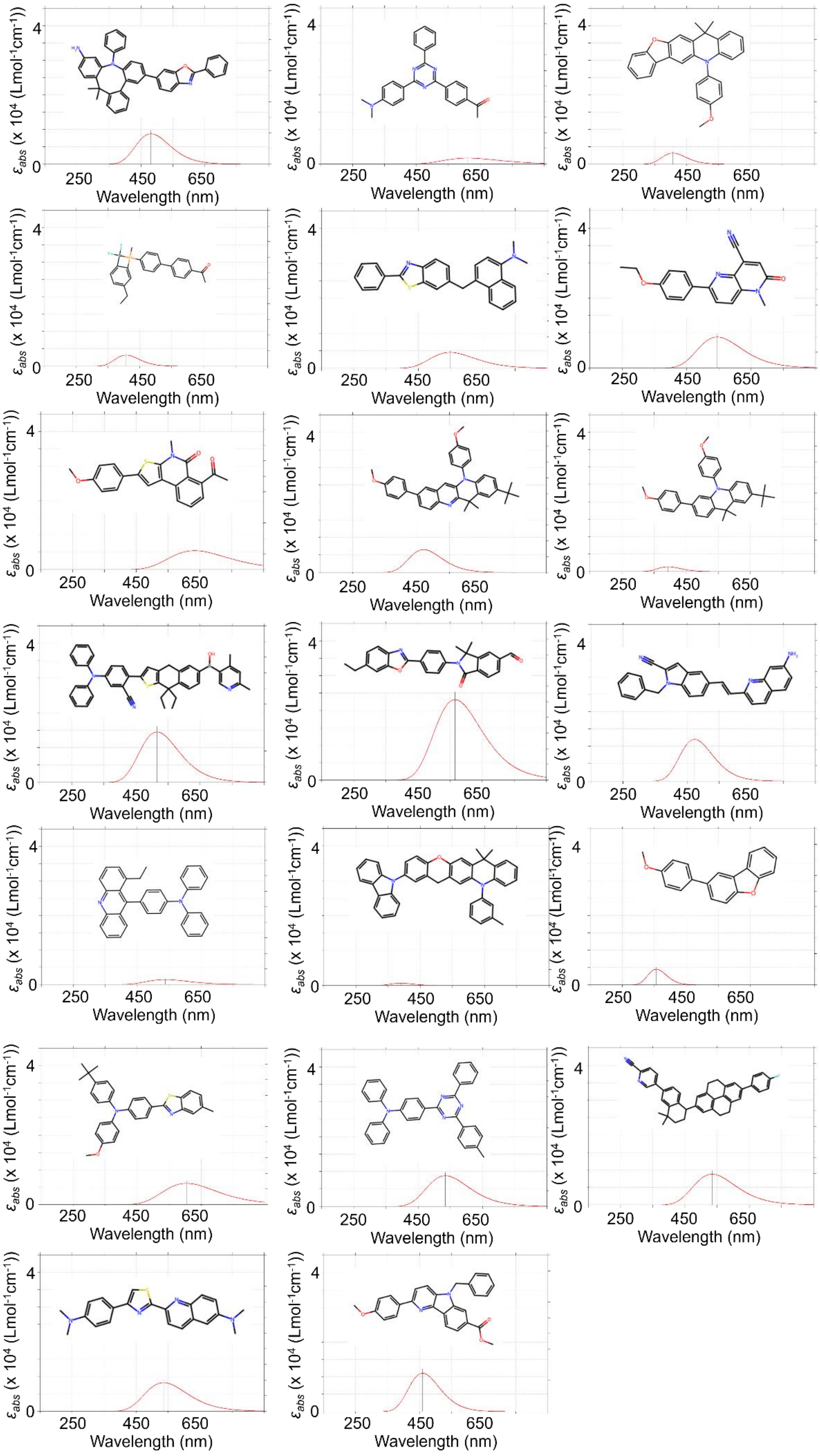


Supplementary Figure 4. 3D scatter‑trend plots showing the relationship among Bayesian optimization iteration number (x‑axis), validation MSE (y‑axis), and each key hyperparameter (z‑axis). (A) Number of dense layers. (B) Block size. (C) Dropout rate. (D) Dense layer size. (E) Number of blocks.

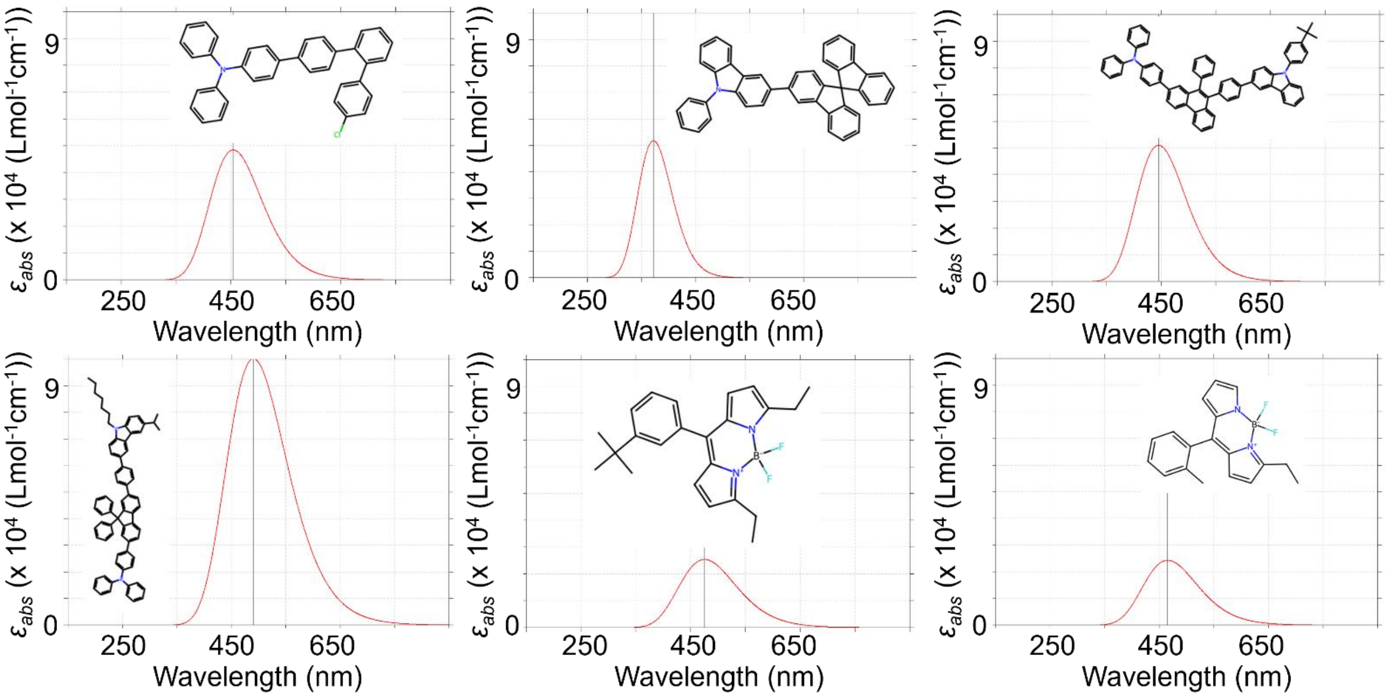
The search space for the number of dense layers was set to 1, 2, and 3. As shown in (A), this parameter has relatively little impact on validation loss, as low loss values were obtained across all three settings. We recommend using a medium to large block size, since Figure (B) shows a clear trend of increasing MSE as the block size decreases from 256 to 16. From Figure (F), we observe that while a few hyperparameter combinations result in higher loss, the majority fall within an acceptable range. This suggests that, overall, the model is relatively robust to variations in most hyperparameters.



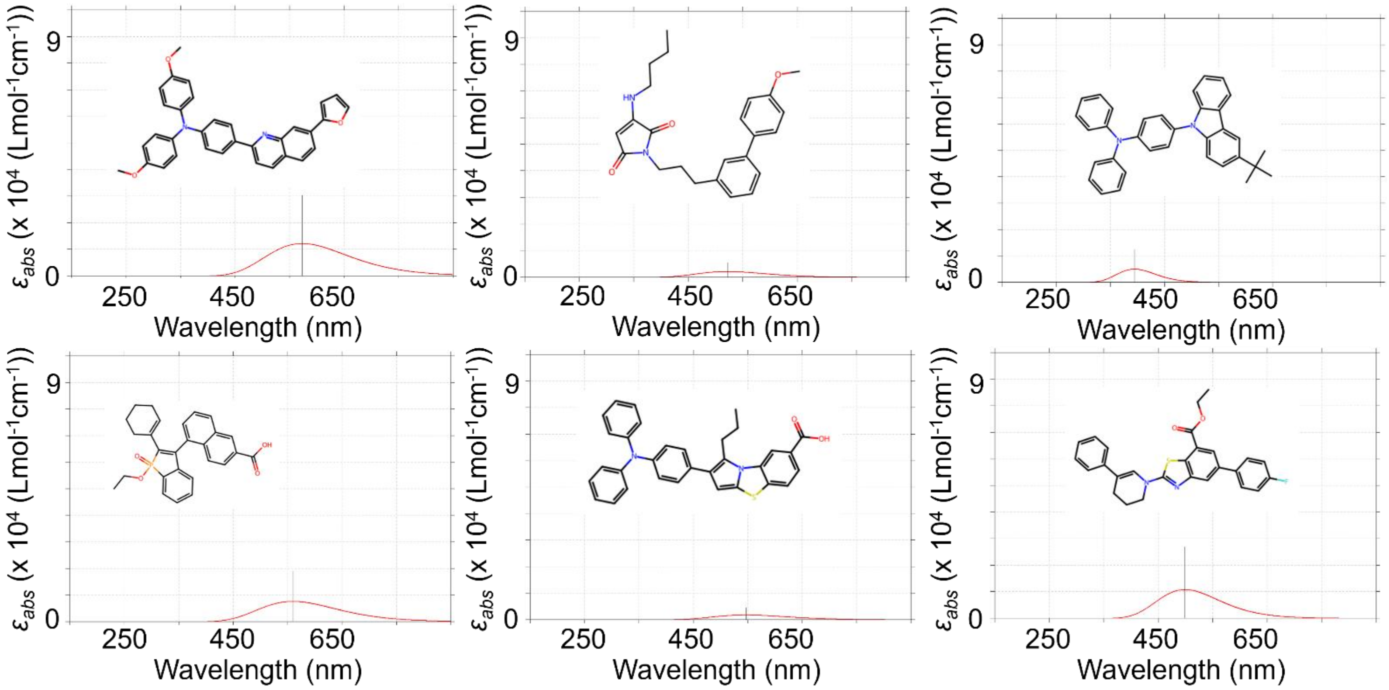
Supplementary Figure 5. Fluorescence emission spectra of example molecules in MolElite calculated by TD-DFT.



Supplementary Figure 6. Fluorescence emission spectra of example molecules in MolMediocritys calculated by TD-DFT.

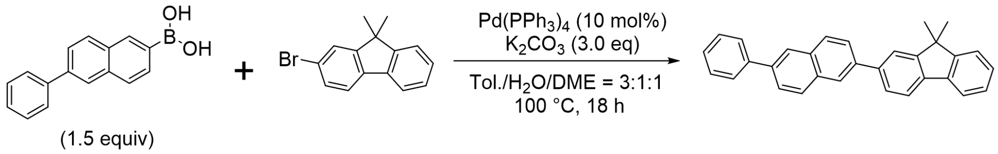


Supplementary Figure 7. Fluorescence emission spectra of example molecules in MolElite based on ASBase calculated by TD-DFT.

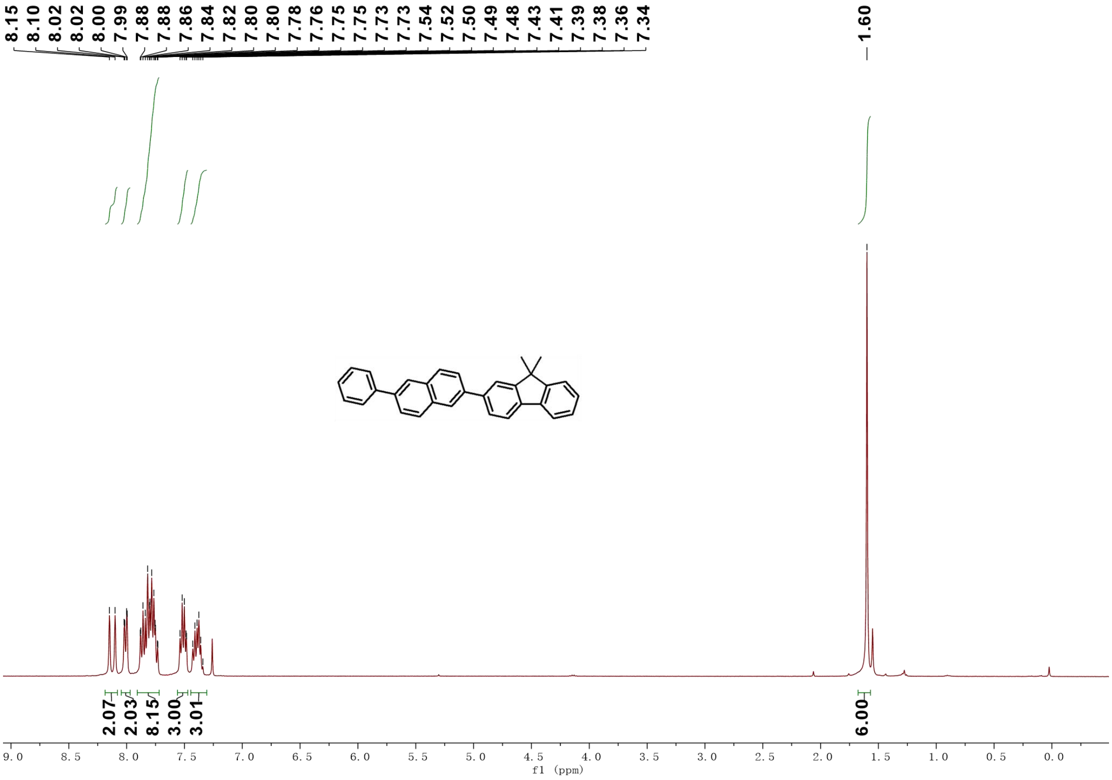


Supplementary Figure 8. Fluorescence emission spectra of example molecules in MolMediocritys based on ASBase calculated by TD-DFT.

The calculated fluorescence emission spectra of example molecules in the MolMediocritys exhibit poor luminescent performance, such as broad bandwidth and low extinction coefficients.



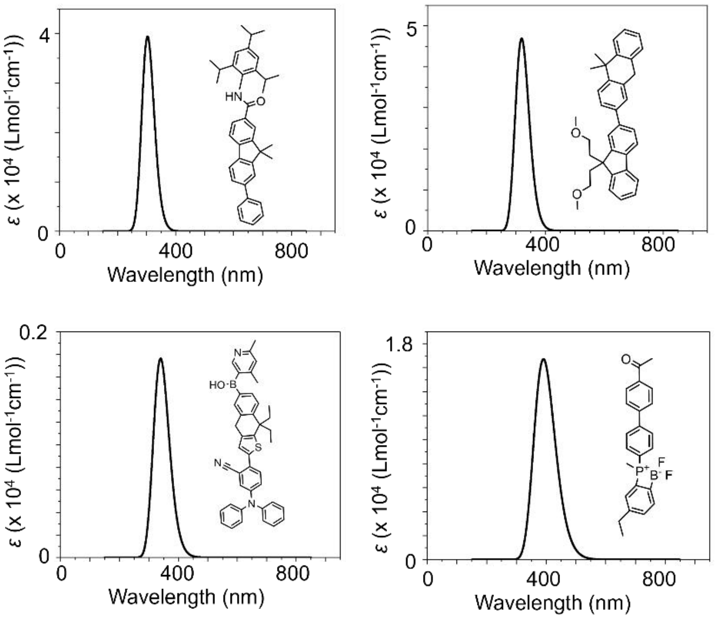
A 250 mL Schlenk flask was charged with 2-bromo-9,9-dimethyl-9H-fluorene (commercially available, 520 mg, 2.0 mmol, 1.0 equiv), (6-phenylnaphthalen-2-yl)boronic acid (744 mg, 3.0 mmol), $Pd(PPh_3)_4$ (236 mg, 0.02 mmol, 10 mol%), K_2CO_3 (830 mg, 6.0 mmol, 3.0 equiv) and degassed toluene/H_2O/DME = 3:1:1 (150 mL) under Ar. The mixture was heated to 100 °C and monitored by TLC. 18 h later, stop heating and the mixture was cooled to room temperature. The mixture was diluted with water and extracted with DCM. The aqueous phase was discarded and the organic phase was washed with brine and dried over anhydrous $MgSO_4$. The mixture was filtered and the filtrate was concentrated by rotary evaporation. The residue was purified by stirring the solid in DCM (50 mL) until a fine suspension was formed, filtered, washed with *n*-hexane (100 mL), and then dried under vacuum to afford the product as a white solid.



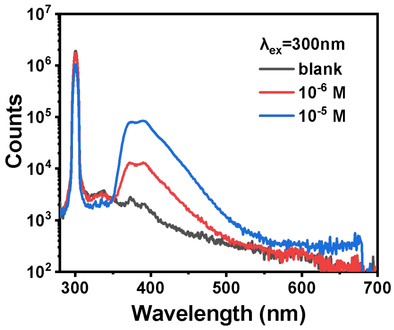
Supplementary Figure 9. 1H-NMR of 9,9-dimethyl-2-(6-phenylnaphthalen-2-yl)-9H-fluorene.

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.12 (d, *J* = 19.4 Hz, 2H), 8.01 (dd, *J* = 8.5, 2.3 Hz, 2H), 7.91 – 7.72 (m, 8H), 7.51 (q, *J* = 7.6 Hz, 3H), 7,38 (dq, *J* = 13.9, 7.5 Hz, 3H), 1.6 (s, 6H).

^{13}C NMR (101 MHz, $CDCl_3$) 13C NMR (101 MHz, CDCl3) δ 154.43, 153.97, 141.10, 140.21, 139.05, 138.89, 138.66, 138.58, 132.99, 132.88, 128.93, 128.76, 127.42, 127.36, 127.11, 126.49, 126.21, 126.08, 125.59, 125.47, 122.69, 121.62, 120.44, 120.17, 47.05, 27.32.



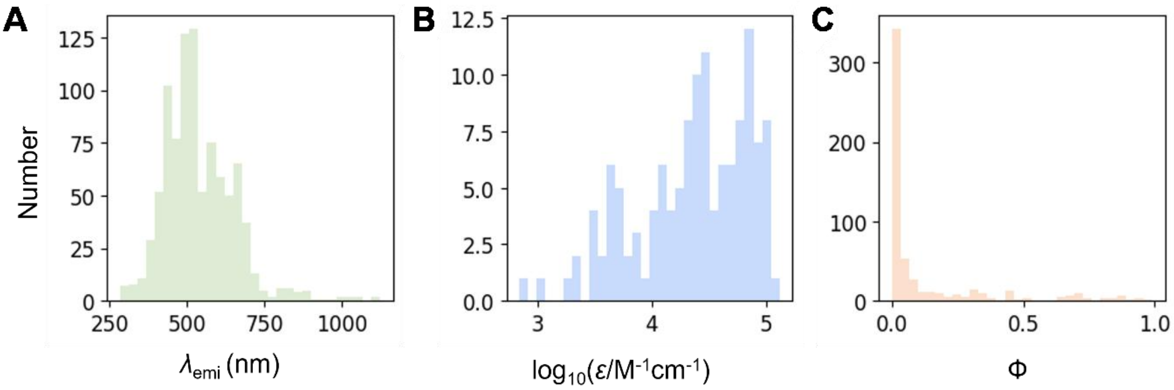
Supplementary Figure 10. UV absorption spectra of molecules in Figure 5.



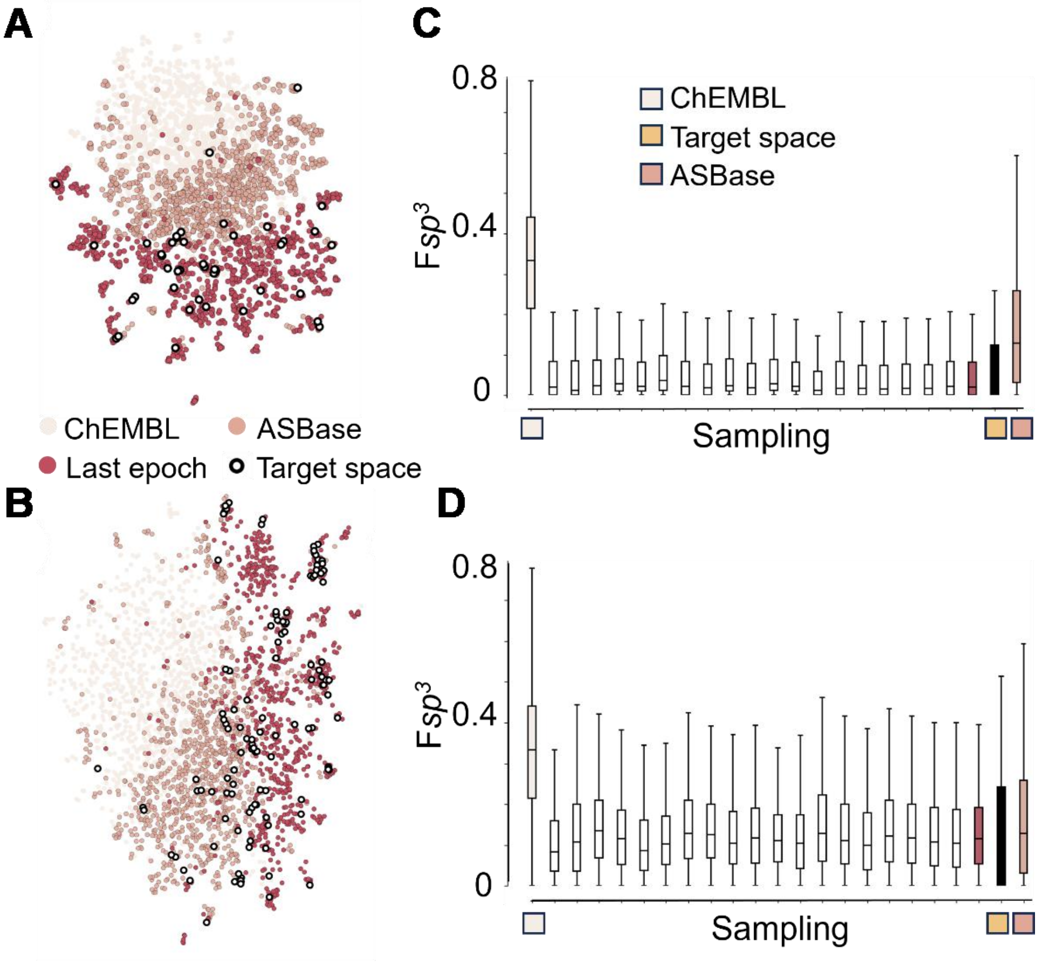
Supplementary Figure 11. Absolute quantum yield of 9,9-dimethyl-2-(6-phenylnaphthalen-2-yl)-9H-fluorene in 10^{-6} M dichloromethane and 10^{-5} M dichloromethane.

Fluorescence spectra were recorded on photoluminescence spectrometer (Tianjin Gangdong Sci & Tech Development Co., Ltd). Fluorescence quantum yields were measured on Photoluminescence Spectrometer (FLS 980, Edinburgh Instruments), and the UV-Vis spectra were performed using the UV-Vis Spectrophotometer (T6, Purkinje General Instrument).

To assess the stability of the fluorene-based molecule Mol1, both thermogravimetric analysis (TGA) and preliminary environmental stability tests were conducted. TGA measurements indicate that Mol1 exhibits a decomposition temperature (Td, defined at 5% weight loss) of approximately 341 °C under a nitrogen atmosphere, demonstrating excellent thermal stability suitable for vacuum deposition and device fabrication processes. The high Td can be attributed to the rigid and conjugated backbone, as well as the steric protection offered by the 9,9-dimethyl substitution on the fluorene core. Fluorene-based luminescent molecules generally exhibit moderate stability under ambient conditions due to the rigid and conjugated structure of the fluorene core. However, their environmental stability can be influenced by substitution patterns and functional groups. In particular, the C9-position of fluorene is known to be susceptible to oxidation, leading to the formation of fluorenone derivatives upon prolonged exposure to oxygen or UV light. To mitigate this, bulky substituents such as aryl or alkyl groups are often introduced at the 9-position, which sterically hinder oxidative degradation and improve stability in air. In our case, Mol1 features 9,9-dimethyl substitution, which is widely reported to enhance resistance to both moisture and oxygen. Preliminary observations indicate that Mol1 maintains its luminescence intensity with negligible change after one week of storage under ambient air without encapsulation, suggesting acceptable environmental robustness for potential optoelectronic applications.

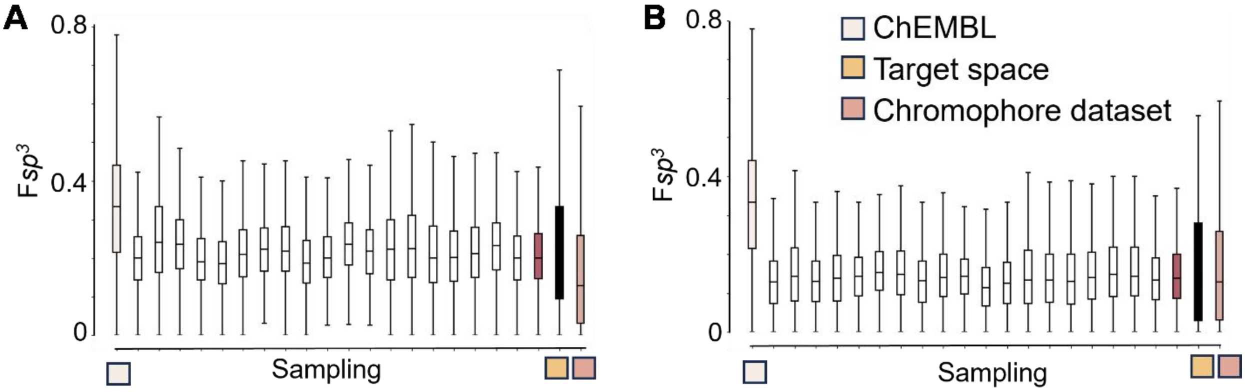


Supplementary Figure 12. Histograms of the distribution of three luminescent properties in the ASBase dataset. (A) Maximum emission wavelength. (B) Extinction coefficient. (C) Quantum yield.[12]



Supplementary Figure 13. UMAP for target space with (A) high Φ_{QY}. (B) high ε. The box plots of Fsp^3 variations corresponding to different target spaces are shown in (C) for high Φ_{QY} and (D) for high ε based on ASBase.

In our research, we implemented the UMAP technique to effectively visualize the dynamic shifts within the chemical space. UMAP, a nonlinear dimensionality reduction tool, maintains both local and global structures within high-dimensional data, thus offering a comprehensible visual representation in reduced dimensions.



Supplementary Figure 14. The box plots of Fsp3 variations corresponding to different target spaces are shown in (A) for high Φ_{QY} and (B).

Supplementary Table 1. Atomic and bond features used in the molecular graph

Type	Name	Description
Atom features	atomic identity	H, B, C, N, O, F, Si, P, S Cl, Br, I (one-hot)
	number of heavy neighbors	0, 1, 2, 3, 4, 5, 6 (one-hot)
	number of hydrogens	0, 1, 2, 3, 4 (one-hot)
	hybridization type	S, SP, SP2, SP3, SP3D, SP3D2, unspecified (one-hot)
	chirality	unspecified, CW, CCW, other (one-hot)
	is in a ring	true or false (binary)
	is aromatic	true or false (binary)
Bond features	bond type	single, double, triple, aromatic (one-hot)
	is in a ring	true or false (binary)
	is conjugated	true or false (binary)

Supplementary Table 1 summarizes the atomic and bond-level features used to construct the molecular graphs input to the Spectral Discriminator. Atom features include basic structural descriptors such as atomic identity (e.g., H, C, N, O, etc.), number of heavy neighbors, number of attached hydrogens, and hybridization type, all encoded in a one-hot format. Stereochemical information is also captured through chirality labels (e.g., CW, CCW), which are essential for modeling chiral-sensitive optical properties. Additional binary indicators, such as whether the atom is in a ring or is aromatic, further enhance the structural representation. Bond features include bond type (single, double, triple, aromatic), ring membership, and conjugation status, all encoded in a binary or one-hot fashion. This combination of topological and electronic features enables the model to capture essential chemical and structural cues influencing the spectral behavior of both solute and solvent molecules. Notably, solvent molecules are treated equivalently and represented as molecular graphs using the same feature set, ensuring a unified and chemically meaningful input representation across the dataset.

Supplementary Table 2. Performance of Union-GCN for predicting four optical properties. All results are evaluated on the test set

properties	MAE	RMSE	NMAE[a]
λ_{emi}	21.97±0.32 nm	30.85±0.65	2.7%
$\lg(\varepsilon_{max})$	0.17±0.01	0.26±0.02	3.0%
Φ_{QY}	0.15±0.00	0.20±0.00	15.0%
σ_{emi}	12.34±0.12 nm	17.86±0.09	5.0%

[a] NMAE represents the percentage of MAE in the range of optical properties within the DB_{exp}

Supplementary Table 3. Comparison of model performance for predicting the maximum emission wavelength in the DB_{exp} using traditional GCN and Union-GCN

Property	Model	MAE	RMSE
DB_{exp} (λ_{emi})	GCN	24.67 nm	34.78 nm
	Union-GCN	21.52 nm	29.76 nm

This table presents a quantitative evaluation of the predictive performance of two graph-based models—standard GCN and the proposed Union-GCN—for estimating λ_{emi} of molecules in the experimental dataset DB_{exp}. Performance is assessed using two commonly used regression metrics: MAE and RMSE, both expressed in nanometers (nm). The results show that the Union-GCN significantly outperforms the traditional GCN. Specifically, the MAE decreases from 24.67 nm (GCN) to 21.52 nm (Union-GCN), indicating an improvement in the average prediction accuracy. The RMSE drops from 34.78 nm to 29.76 nm, reflecting better handling of large prediction deviations. These improvements validate the benefit of ensemble learning in Union-GCN, which aggregates predictions from multiple independently trained GCN models to enhance robustness and generalization. The consistent reduction in both error metrics highlights Union-GCN's suitability for high-precision property prediction tasks in molecular design workflows.

Supplementary Table 4. The radiative and non-radiative transition rates calculated by momap. k_{ic} is non-radiative rate and k_r is radiative rate

Series	E_{ad} (a.u.)	EDMA (Debye)	EDME (Debye)	k_{ic} (s^{-1})	k_r (s^{-1})
Mediocrity1	0.0983	6.687859	6.409319	4.06E+11	2.11E+08
Mediocrity2	0.107424	6.746095	2.331586	2.68E+11	3.25E+07
Elite1	0.129086	9.51859	11.4214	2.67E+11	1.25E+09
Elite2	0.133049	1.697049	10.03051	3.21E+11	1.06E+09

The adiabatic excitation energy (E_{ad}) is the energy difference between the relaxed excited state energy and the ground state energy. The Gaussian log file for the optimized S1 excited state has already included the Dipole Square of Electronic Transition Dipole Absorption (EDMA) and the Dipole Square of Electronic Transition Dipole Emission (EDME) information.[11] Elite1 is the molecule shown in Figure 5a. Elite2 is the molecule shown in Figure 6c. Mediocrity1 and Mediocrity2 correspond to Figure 5e and f, respectively. Since the order of magnitude of the molecules in Figure 5c varies greatly, it is not listed here.

Supplementary Table 5. Performance of Union-GCN for predicting four optical properties

Dataset	properties	MAE	RMSE
DB_{exp}	λ_{emi}	21.52 nm	29.76 nm
	$\log_{10}(\varepsilon_{max})$	0.169	0.250
	Φ_{QY}	0.147	0.202
	σ_{emi}	13.18 nm	19.47 nm
ASBase	λ_{emi}	22.80 nm	37.79 nm
	$\log_{10}(\varepsilon_{max})$	0.293	0.381
	Φ_{QY}	0.4178	0.448

REFERENCES

[1] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, J. Cheminformatics 9 (2017) 48.

[2] M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, Nat. Mach. Intell. 2 (2020) 171–180.

[3] S. Mazraedoost, H. Sedigh Malekroodi, P. Žuvela, M. Yi, J.J. Liu, J. Chem. Inf. Model. 65 (2025) 3343–3356.

[4] A.T. Müller, J.A. Hiss, G. Schneider, J. Chem. Inf. Model. 58 (2018) 472–479.

[5] J. Guo, M. Sun, X. Zhao, C. Shi, H. Su, Y. Guo, X. Pu, J. Chem. Inf. Model. 63 (2023) 1143–1156.

[6] P. Tosco, N. Stiefl, G. Landrum, J. Cheminformatics 6 (2014) 37.

[7] Y. Zhao, D.G. Truhlar, Acc. Chem. Res. 41 (2008) 157–167.

[8] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Wallingford, CT 2016.

[9] J. Tomasi, B. Mennucci, R. Cammi, Chem. Rev. 105 (2005) 2999–3094.

[10] T. Lu, F. Chen, J. Comput. Chem. 33 (2012) 580–592.

[11] Y. Niu, W. Li, Q. Peng, H. Geng, Y. Yi, L. Wang, G. Nan, D. Wang, Z. Shuai, Mol. Phys. 116 (2018) 1078–1090.

[12] J. Gong, W. Gong, B. Wu, H. Wang, W. He, Z. Dai, Y. Li, Y. Liu, Z. Wang, X. Tuo, J.W.Y. Lam, Z. Qiu, Z. Zhao, B.Z. Tang, Aggregate 4 (2023) e263.

[13] J.L. Medina-Franco, K. Martínez-Mayorga, A. Bender, T. Scior, QSAR Comb. Sci. 28 (2009) 1551–1560.