Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 3728

Fluor-tools: an integrated platform for dye property prediction and structure optimization

Wenxiang Song,[a] Yuyang Zhang,[bc] Le Xiong,[a] Xinmin Li,[a] Jingwei Zhang,[a] Guixia Liu,[a] Weihua Li,[a] Youjun Yang *[bc] and Yun Tang *[a]

With the rapid advancement of fluorescent dye research, there is an urgent need for tools capable of accurately predicting dye optical properties while facilitating structural modification. However, the field currently lacks reliable and user-friendly tools for this purpose. To address this gap, we have developed Fluor-tools—an integrated platform for dye property prediction and structural optimization. The platform comprises two core modules: (1) Fluor-pred, a dye property prediction model that integrates domain-specific knowledge of fluorophores with a label distribution smoothing (LDS) reweighting strategy and an advanced residual lightweight attention (RLAT) architecture. This model achieves state-of-the-art performance in predicting four key photophysical properties of dyes. (2) Fluor-opt, a structural optimization module that employs a matched molecular pair analysis (MMPA) method enhanced with symmetry-aware and environment-adaptive modifications. This module derives 1579 structural transformation rules, enabling the directional optimization of non-NIR (non-near-infrared) dyes to NIR properties. In summary, Fluor-tools provides robust computational support for research in biomedical imaging and optical materials. The platform is freely accessible at https://lmmd.ecust.edu.cn/Fluor-tools/.

Received 9th September 2025
Accepted 28th October 2025

DOI: 10.1039/d5dd00402k

rsc.li/digitaldiscovery

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

1 Introduction

Fluorescent materials are compounds that absorb light and re-emit it as fluorescence. A defining feature of these materials is their conjugated π-electron systems, which enable fine-tuning of optical properties. They play pivotal roles across diverse applications, including pharmaceuticals, dyes and pigments, optoelectronics, organic light-emitting diodes, environmental monitoring, light-harvesting antennas, organic photovoltaics, information encryption, bioimaging, and high-throughput material discovery.[1–7] Notably, the near-infrared (NIR) region serves as an optimal optical imaging window, characterized by minimal tissue absorption and reduced light scattering.[8,9] This spectral range offers distinct advantages: low phototoxicity, deep tissue penetration, minimal autofluorescence interference, high signal-to-noise ratios, and low false-positive rates—all of which enable high-quality *in vivo* deep-tissue imaging.[10–16]

However, the development of fluorophores currently relies heavily on empirical knowledge and extensive trial-and-error experimentation, creating substantial barriers to designing high-performance dyes.[12] While existing empirical rules (*e.g.*, the Woodward–Fieser rules) can estimate the maximum absorption wavelength of dyes,[17] and time-dependent density functional theory (TD-DFT) allows calculation of absorption and emission spectral parameters,[18] *ab initio* calculations remain often time-consuming and cost-prohibitive, restricting their widespread application.

Recently, with the rapid advancement of artificial intelligence (AI), AI-based approaches have significantly transformed dye design by offering greater computational efficiency and predictive accuracy than traditional methods. These approaches are primarily applied in two domains: (1) dye property prediction, and (2) dye structure optimization. Property prediction represents the most widely studied application, where regression models are built to forecast key optical characteristics—including absorption wavelength (λ_{abs}), emission wavelength (λ_{em}), fluorescence quantum yield (Φ_{PL}), and molar absorption coefficient (ε_{max}).[19–23] For example, Joung *et al.* developed a multi-property prediction model using graph convolutional networks (GCNs), trained on an experimental dataset of 30 094 dye–solvent pairs; this model enabled accurate predictions of multiple optical properties.[20] Similarly, Wang *et al.* proposed NIRFluor, a multimodal prediction model trained on 5179 NIR fluorescent molecules, which achieved high accuracy in predicting four critical properties of NIR dyes.[23] In the field of dye structure optimization, conventional

[a]*Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*

[b]*State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China*

[c]*Shanghai Key Laboratory of Chemical Biology, Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China*

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

approaches typically involve generating a large number of derivatives based on target molecular scaffold, followed by screening using property prediction models to identify candidate molecules with desired characteristics.[24] Recently, Zhu *et al.* integrated the Reinvent4 to successfully synthesize a novel fluorescent compound exhibiting exceptional brightness.[25,26] Additionally, Han *et al.* developed DeepMoleculeGen, a generative deep learning (DL) model trained on a comprehensive experimental database containing 71 424 molecule–solvent pairs.[27]

However, current methodologies still suffer from notable limitations, which are elaborated as follows: (1) in terms of dye property prediction: existing models rely exclusively on molecular structures to forecast properties. They merely adopt generic molecular representations—such as molecular graphs or molecular fingerprints—while lacking the integration of dye-specific knowledge. Moreover, the imbalanced distribution of dye-related data leads to particularly poor prediction accuracy in spectral regions where data is sparse. (2) In the field of structural optimization: the currently prevalent "undirected random generation and subsequent screening" strategy suffers from certain accuracy limitations, primarily due to its reliance on the aforementioned property prediction models. Additionally, Reinvent4 is not specifically developed for dyes, making it difficult to generate structurally reasonable dye molecules. (3) Finally, previous models have not yet been truly delivered to end-users. Most previous studies failed to deploy their models, leaving these models unable to provide tangible support for the dye industry.

To address the aforementioned challenges, we have developed Fluor-tools—an innovative computational platform for the rational design of dyes. The platform architecture integrates two synergistic modules:

(1) Fluor-pred: a multimodal property prediction model that incorporates domain-specific knowledge of fluorophores. In terms of architectural design, we innovatively designed a residual lightweight attention (RLAT) architecture, coupled with a label distribution smoothing (LDS) reweighting strategy. These design elements collectively enhance model performance—particularly improving prediction accuracy in data-sparse regions. For dye representation, we integrated domain-specific knowledge of dyes, such as HOMO–LUMO gaps and custom-developed MMP fingerprints. Ultimately, Fluor-pred achieved state-of-the-art performance in predicting four core photophysical properties of dyes.

(2) Fluor-opt: this module enables the directed modification of non-NIR dyes to achieve NIR properties. It adopts an optimized quantitative structure–activity relationship-molecular matching pair (QSAR-MMP) algorithm, into which we have integrated symmetry modification methods and the applicability context of MMP transformation rules. This approach comprehensively captures the differences in structural transformation and structure–activity relationships (SARs) between non-NIR and NIR dyes. It supports the automated structural optimization of non-NIR dyes into NIR dyes and has been successfully applied in multiple cases.

Finally, we have integrated all the aforementioned research into the website (**https://lmmd.ecust.edu.cn/Fluor-tools/**), ensuring convenient access and utilization for researchers.

2 Results

2.1 Overall of Fluor-tools

Fluor-tools incorporates an accurate dye property prediction model, Fluor-pred. The model architecture of Fluor-pred is shown in Fig. 1a–c. This model was developed based on FluoDB,[26] the largest publicly available dye database to date, and employs a comprehensive set of effective molecular representations, as detailed in Fig. 1b. (1) First, we utilize the Attentive FP architecture to encode dye molecular graphs, effectively capturing structural information.[42] (2) To account for the distinct property distributions exhibited by different scaffold types, we incorporate scaffold-specific annotations as additional input features. (3) Building on the experimentally validated transformation rules extracted from Fluor-opt (which significantly alter dye properties), we identify 136 high-frequency substructures to construct dye-specific MMP fingerprints. (4) Recognizing the well-established correlation between dye properties and HOMO–LUMO gaps (where smaller gaps correspond to longer absorption wavelengths),[45–48] we integrate HOMO–LUMO gap predictions from the advanced quantum property prediction model Uni-mol+.[43] (5) Incorporates several RDKit-calculated molecular properties along with Morgan fingerprints—features that have been extensively adopted in previous studies.[19–23] For solvent representation, we employ Morgan fingerprints combined with solvent-type annotations.

Regarding the model architecture of Fluor-pred (Fig. 1c), we designed a RLAT framework for small molecule information extraction—a framework originally developed for protein sequence information extraction.[49,50] In Fluor-pred, the AttentionCNN module processes concatenated sequential information from dye molecules, which is then integrated with graph features and solvent representations. Compared to conventional deep attention models, RLAT maintains robust representational capacity while significantly reducing both parameter size and computational overhead, making it particularly suitable for dye molecular feature extraction. Notably, the four properties show severe distribution skewness (Fig. S1), causing significant accuracy drops in sparse-data regions. To address this, we introduced LDS to reweight the loss function based on estimated data density.[44] This method assigns higher weights to samples in low-density regions and lower weights to those in high-density regions, and fine-tunes the reweighting intensity *via* the hyperparameter α to achieve an optimal balance.

In addition, Fluor-tools also incorporates a module called Fluor-opt, which enables the directed transformation of non-NIR dyes into NIR dyes. Fluor-opt employs an optimized QSAR-MMP algorithm, where we incorporate symmetry modifications and the applicability context of transformation rules. This approach comprehensively captures the differences in structural transformations and SARs between non-NIR and NIR dyes, ultimately enabling the automated structural

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

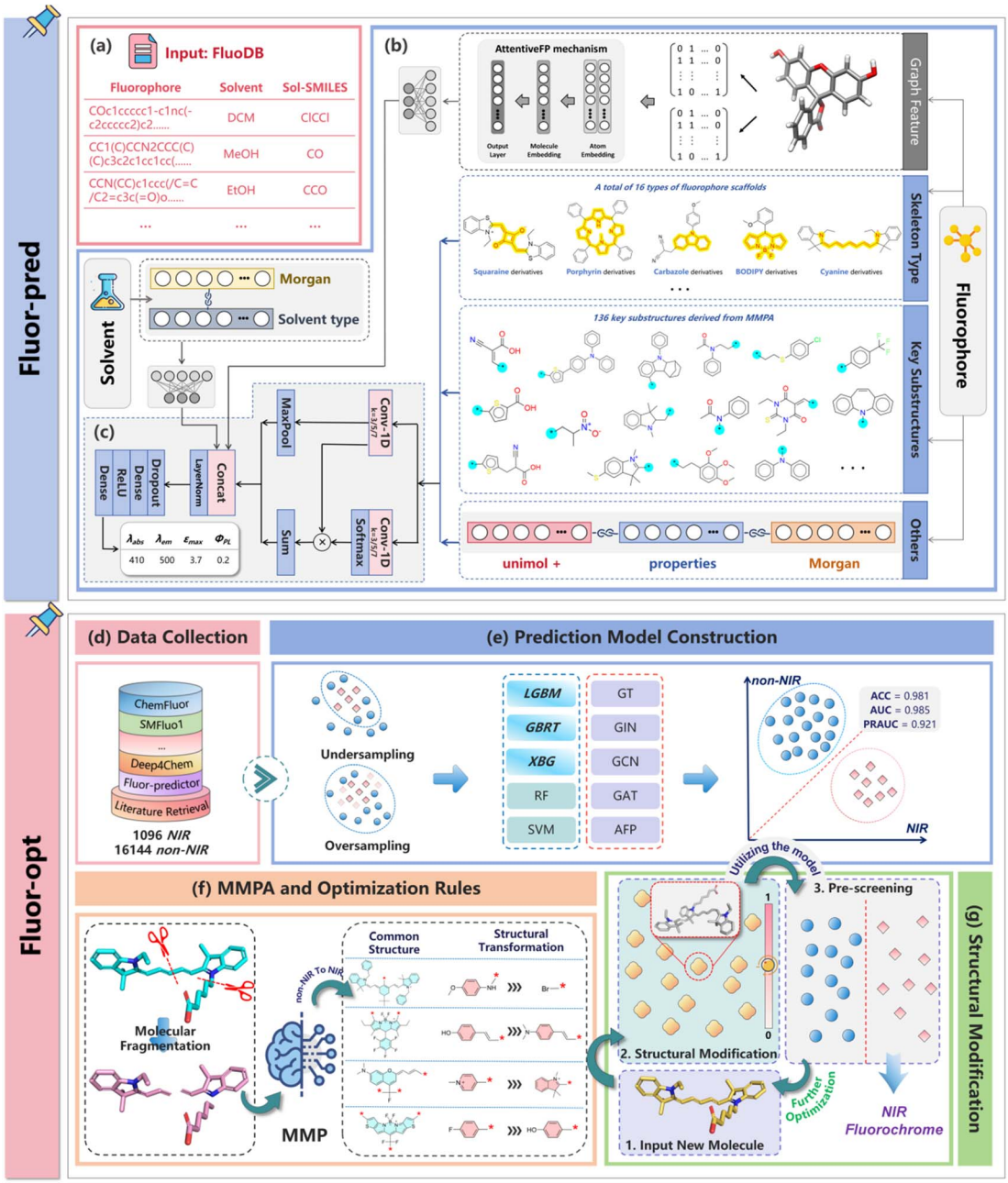

Fig. 1 Overview of Fluor-tools platform, integrating two core models: Fluor-pred (a–c) and Fluor-opt (d–g). The development process of Fluor-pred includes: (a) data collection and preprocessing: Fluor-pred is built upon the FluoDB database,[26] which, after preprocessing, contains 49 831 dye–solvent pairs; (b) molecular and solvent representations: Fluor-pred incorporates multiple representations, including molecular graphs, dye category labels, custom-designed MMP fingerprints, and HOMO–LUMO gap computed by Uni-mol+, and so on; (c) model architecture: Fluor-pred employs an optimized RLAT architecture for efficient feature extraction. The workflow of Fluor-opt includes: (d) data collection and preprocessing: data were obtained from multiple databases and literature mining. After processing, a total of 1096 NIR dye molecules and 16 144 non-NIR dye molecules were collected; (e) construction of NIR classification models: building a set of binary classifiers to distinguish between NIR and non-NIR dyes; (f) MMPA-based transformation rule extraction: deriving structural transformation rules for converting non-NIR dyes into NIR dyes; (g) structure optimization and candidate screening: applying appropriate transformation rules to target molecules to generate candidate compounds, which are then filtered by the binary classifiers. The molecules predicted as NIR dyes are retained as optimized outputs.

optimization of non-NIR dyes into NIR dyes. Compared with molecular generation methods, the MMPA approach implements minimal yet critical molecular modifications, which significantly enhances the synthetic feasibility of optimized dye molecules. The general development process of Fluor-opt includes: (1) building the largest open-source NIR dye database through literature mining and database integration (Fig. 1d); (2) creating a binary classification model with defined applicability domains to evaluate transformation rules and filter modified molecules (Fig. 1e); (3) developing an enhanced

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

MMP method incorporating dye symmetry features, yielding 1579 transformation rules (Fig. 1f); (4) validating rules with transformation context to establish an iterative optimization cycle for NIR dye data enhancement (Fig. 1g).

2.2 Data analysis

In the Fluor-pred study, we utilized the original dataset from FluoDB database,[26] comprising 49 831 validated dye–solvent pairs. Fig. S1 illustrates the dataset characteristics and distributions for the four prediction tasks.

For the Fluor-opt model, as illustrated in Fig. 2, we constructed a comprehensive dye database by integrating multiple data sources, which was used for extracting transformation rules *via* the MMPA method. Furthermore, due to the limited number of NIR dyes in existing databases, we conducted systematic literature mining using the keywords "near-infrared" and "dye"—screening for valid information from approximately 500 retrieved articles. After data processing, the final dataset comprises 17 240 experimentally validated dye structures, including 1096 NIR dyes and 16 144 non-NIR dyes.

Fig. 2 Workflow of data collection, processing, and analysis in Fluor-opt. (a) Overview of the data collection and analysis workflow for Fluor-opt; (b) correlation between dye properties and λ_{abs} in the Fluor-opt dataset; all six properties were calculated using RDKit; (c) comparison of NIR and non-NIR dye counts across different molecular scaffolds in the Fluor-opt dataset; (d) distribution of λ_{abs} and *t*-SNE-based dimensionality reduction in the Fluor-opt dataset. The *t*-SNE analysis was performed based on the Morgan fingerprints of the dyes.

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

In this study, we conducted systematic statistical analyses of the structural and property differences between NIR and non-NIR dyes within the Fluor-opt dataset. Fig. 2b demonstrates the relationship between dye properties and their classification as NIR or non-NIR dyes. NIR dyes tend to exhibit slightly higher molecular weights, log P values, ring counts, and degrees of unsaturation compared to non-NIR counterparts. This trend can be attributed to characteristic structural motifs of NIR dyes, such as: extended conjugated systems, as seen in porphyrins and cyanine dyes; and fused ring frameworks (*e.g.*, pentacene, perylene derivatives) or rigid bridging cores (*e.g.*, BODIPY scaffold). Collectively, these features account for the increased molecular weight and lipophilicity observed in NIR dyes. Based on the structural classification framework established by Zhu *et al.*, all dyes were categorized into 17 distinct scaffold types (Fig. 2c).[26] Among these, cyanine scaffolds contain the highest proportion of NIR dyes, followed by BODIPY derivatives. Notably, certain scaffold classes such as carbazole and triphenylamine show a significant skew toward non-NIR dyes. This scaffold-dependent variation underscores the pivotal influence of molecular scaffolds on the photophysical behavior of dyes. We provide detailed statistical data in Fig. S2.

To further elucidate the relationship between dye scaffolds and the characteristics of NIR dyes, we performed a statistical analysis of the λ_{abs} distributions across different scaffold types in the Fluor-opt dataset, as shown in Fig. 3. The results show that cyanine derivatives most frequently exhibit high λ_{abs}, followed by squaric acid, porphyrin, and BODIPY scaffolds. Notably, some scaffold types, such as triphenylamine, naphthalimide, and azo derivatives, exhibit a near-complete absence of NIR dyes, indicating that their structures impose substantial constraints on photophysical properties. Nonetheless, most scaffold types exhibit broad absorption wavelength ranges, demonstrating that rational structural optimization can effectively modulate their photophysical properties for the majority of scaffolds.

2.3 Results of fluor-pred

2.3.1 Performance of fluor-pred. Fluor-pred synergistically integrates multimodal molecular representations, innovative architectural design, and an advanced loss function. As evidenced in Fig. 4, the model achieves leading performance across all four critical photophysical property predictions—λ_{abs}, λ_{em}, Φ_{PL}, and ε_{max}—with R^2 values of 0.943, 0.943, 0.668, and 0.789, respectively. Moreover, LDS-weighted loss function addresses data sparsity by assigning higher weights to sparse regions, significantly reducing prediction errors in these intervals. As is evident from the rightmost histogram bars in Fig. 4a, the predicted values show strong agreement with the experimental values—even under low-data conditions. Furthermore, scaffold-performance analysis (Table S1) reveals a strong correlation between prediction accuracy and the volume of training data: scaffolds with sufficient training data (*e.g.*, BODIPY) maintain consistently high accuracy across all tasks, whereas data-deficient scaffolds (*e.g.*, porphyrins and azo compounds) exhibit significantly greater prediction variance.

Furthermore, to demonstrate the extrapolation capability of our model, we re-partitioned the entire dataset using scaffold-based split. Specifically, dye molecules were divided according to their Murcko scaffolds into training, validation, and test sets at a 7 : 1 : 2 ratio, ensuring that molecules sharing the same scaffold appeared in only one subset. We compared the results of random partitioning and scaffold partitioning, as shown in Fig. S3. It can be observed that the stricter scaffold split led to a moderate decline in model performance, yet the predictions remained robust, particularly for λ_{abs} and λ_{em}. However, the prediction of photoluminescence Φ_{PL} still showed similar limitations as in the random split, with performance noticeably lower than for the other three properties.

2.3.2 Study on quantum yield error. In response to the concern regarding the relatively low accuracy of the photoluminescence Φ_{PL} prediction task, we performed a quantitative

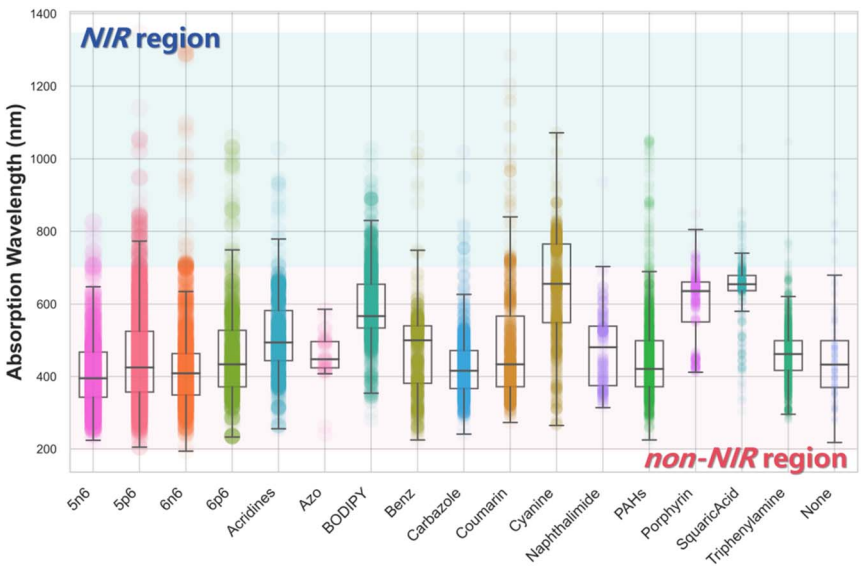

Fig. 3 Distribution of λ_{abs} across different dye scaffolds in the Fluor-opt dataset. Statistical distribution of λ_{abs} across 17 dye scaffolds, where different colors represent distinct scaffolds. The width of each bubble corresponds to the relative quantity of dyes within each scaffold category.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

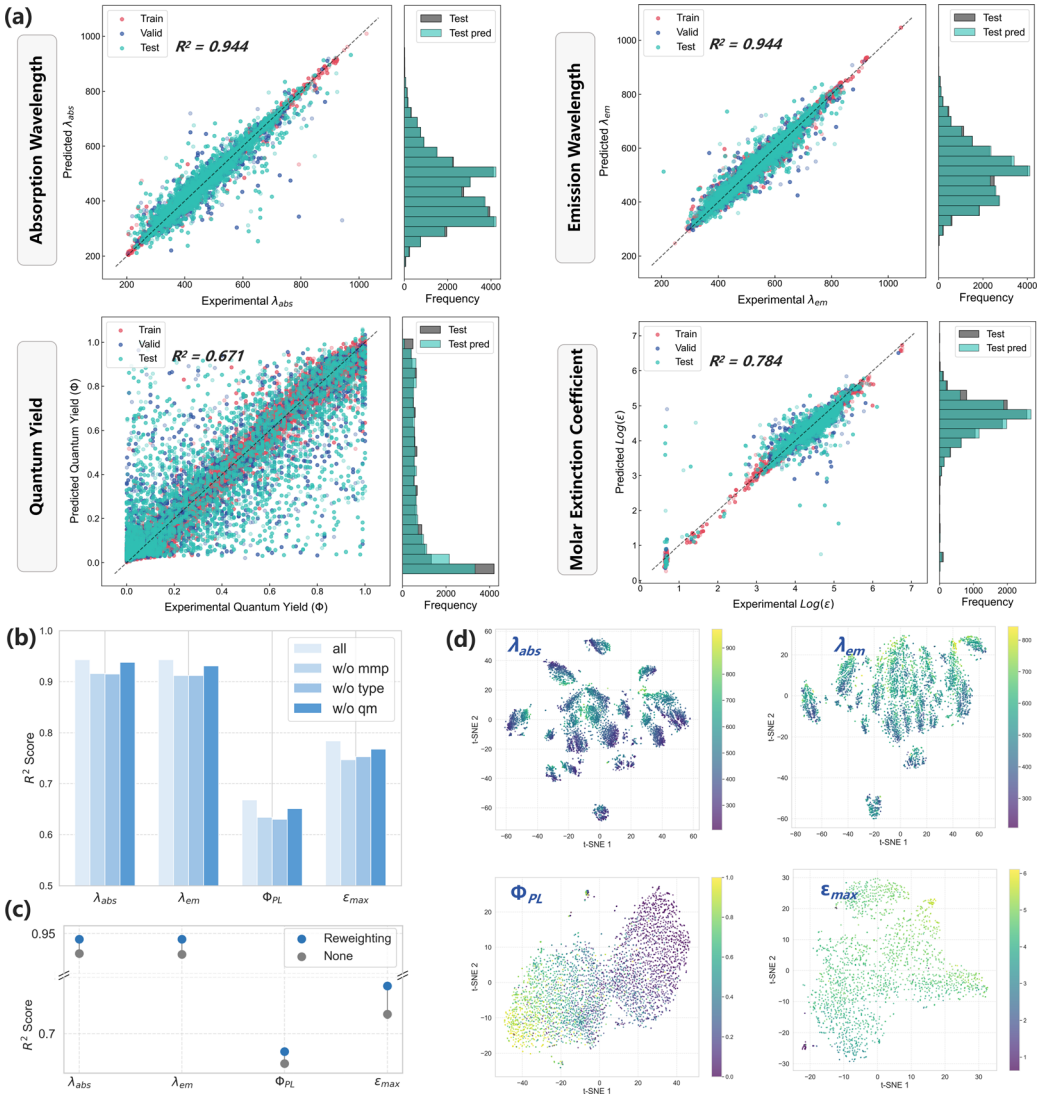

Fig. 4 (a) Fluor-pred performance across four prediction tasks: λ_{abs}, λ_{em}, Φ_{PL}, and ε_{max}. (b) Ablation experiments on the Fluor-pred model and corresponding performance. "all" indicates all features retained; "w/o mmp" represents the removal of MMP fingerprints; "w/o type" represents the removal of scaffold-type labels; "w/o qm" represents the removal of HOMO–LUMO gap values. (c) The impact of removing LDS reweighting on the prediction accuracy of the four tasks. (d) Feature extraction from the classification head of Fluor-pred on the test set and *t*-SNE-based dimensionality reduction analysis.

error analysis on the test set, as shown in Fig. S4. The analysis included residual distribution, binned error evaluation, and noise characteristic exploration.

First, the binned MAE analysis reveals the variation of errors with respect to the true Φ_{PL} values (Fig. S4a): in the low-Φ_{PL} range, the prediction errors are relatively small, whereas in the high-Φ_{PL} range, the errors increase markedly, with the MAE in the highest bin approaching 0.18. This indicates that predicting samples with high quantum yields is considerably more challenging. Second, the residual histogram shows that the overall residuals approximately follow a normal distribution, with the peak centered around zero, suggesting no severe systematic bias in the model (Fig. S4b). However, the distribution exhibits a slight negative skew and extended tails, indicating that a small number of samples still suffer from large prediction deviations. Such long-tail errors may arise from experimental measurement uncertainties or outliers present in the dataset. Finally, the residual *versus* predicted values scatter plot demonstrates that the residuals are not entirely random (Fig. S4c): in the low-prediction region, the model tends to slightly underestimate, while in the high-prediction region, it tends to overestimate. Moreover, the magnitude of residual fluctuations increases with larger predicted values, reflecting evident heteroscedasticity.

Therefore, the relatively low R^2 in Φ_{PL} prediction does not simply result from insufficient model performance. The primary contributing factors include imbalanced data distribution, the limited number of high-Φ_{PL} samples, and greater experimental noise in the high-Φ_{PL} region.

2.3.3 Ablation study. To assess the importance and contributions of individual modules in Fluor-pred, we

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

conducted ablation studies on several key components. As shown in Fig. 4b, the full-featured model demonstrated optimal performance across all four prediction tasks. The ablation of any feature component resulted in measurable performance degradation, with the most pronounced decline observed upon removal of MMP fingerprints. This compelling evidence validates the efficacy of our custom MMP fingerprints in capturing the critical structural determinants governing dye property variations. Interestingly, while the HOMO–LUMO gap exhibits clear linear correlations with multiple target properties (Fig. S5), the "w/o qm" condition showed the least impact on model performance. This limited impact may stem from the fact that when quantum chemical information is combined with higher-dimensional feature representations, its relatively low dimensionality causes its contribution to the overall model performance to be diluted by higher-dimensional features. Nevertheless, it is undeniable that Fluor-pred does exhibit improved predictive performance after integrating this feature.

To evaluate the importance of the LDS reweighting mechanism, as shown in Fig. 4c, removing the LDS reweighting led to a decrease in performance across all prediction tasks, with the most pronounced drop observed in the ε_{max} task. To further assess whether LDS reweighting genuinely improves prediction accuracy in low-density regions, we analyzed λ_{abs} as a representative property with a skewed distribution. The dataset was divided into three intervals: ≤300 nm, 300–700 nm, and ≥700 nm. The middle interval contains the majority of the data, while the lower and upper ends account for only 2.22% and 3.27%, respectively. As shown in Fig. S6, LDS improved accuracy across all intervals, with the most significant reduction in error observed in the NIR region (≥700 nm). Interestingly, in the data-rich 300–700 nm high-density region, the MAE also decreased. This may be due to the varying internal density within the 300–700 nm range, where certain subregions benefit more than others. Overall, LDS effectively improved the overall prediction accuracy of Fluor-pred, particularly enhancing the model's performance in sparse regions. Finally, we visualized the latent feature space from the final layer of the model using dimensionality reduction techniques. In Fig. 4d, smooth color gradients show each target property's variation trends and exhibit clear distinguishability, indicating the model has effectively distinguished and learned different properties.

2.3.4 Model comparison. To rigorously evaluate the performance advantages of Fluor-pred, we conducted a systematic comparative analysis against multiple recently published open-source dye property prediction models using the standardized FluoDB dataset. The models used for comparison include GBRT,[19] SMFluo,[37] SchNet,[38] ABT-MPNN,[39] Fluor-predictor,[40] and FLSF.[26] Among them, GBRT represents dye molecules using Morgan fingerprints, while the others employ various graph neural network (GNN) architectures for molecular representation. As shown in Table 1, Fluor-pred achieved the best overall performance across all four prediction tasks. Specifically, Fluor-pred achieves slightly better performance than other models on the λ_{abs}, λ_{em}, and Φ_{PL} tasks, and demonstrates a significant advantage on the ε_{max} task.

Table 1 Comparison between Fluor-pred and other models

	Algorithms	MAE[a]	RMSE[b]	R^{2c}
λ_{abs}	GBRT	13.67	28.71	0.93
	SMFluo	21.19	35.44	0.89
	SchNet	22.17	41.05	0.63
	ABT-MPNN	12.66	26.23	**0.94**
	Fluor-predictor	14.70	28.07	0.92
	FLSF	12.56	25.99	**0.94**
	Fluor-RLAT	**12.44**	**25.68**	**0.94**
λ_{em}	GBRT	14.56	25.91	0.92
	SMFluo	27.82	38.31	0.83
	SchNet	38.26	51.91	0.43
	ABT-MPNN	13.30	22.84	0.94
	Fluor-predictor	15.65	25.92	0.92
	FLSF	13.27	23.35	**0.94**
	Fluor-RLAT	**13.04**	**22.34**	**0.94**
Φ_{PL}	GBRT	0.12	0.18	**0.68**
	SMFluo	0.13	0.21	0.57
	SchNet	0.15	0.20	0.39
	ABT-MPNN	0.12	0.19	0.65
	Fluor-predictor	0.13	0.19	0.62
	FLSF	0.12	0.19	0.66
	Fluor-RLAT	**0.11**	**0.18**	0.66
ε_{max}	GBRT	0.20	0.31	0.66
	SMFluo	0.22	0.37	0.53
	SchNet	0.51	0.71	−2.01
	ABT-MPNN	0.32	0.45	0.31
	Fluor-predictor	0.17	0.34	0.70
	FLSF	0.23	0.34	0.59
	Fluor-RLAT	**0.15**	**0.29**	**0.78**

[a] MAE: mean absolute error. [b] RMSE: root mean square error. [c] R^2: Coefficient of determination. All models were trained and tested using the same data split from the FluoDB dataset, with some results directly taken from FLSF. The bolded numbers indicate the optimal results for the corresponding tasks.

However, in the fluorescence Φ_{PL} prediction task, none of the models achieved an R^2 value above 0.7, which is significantly lower than the performance observed in the other three tasks. The limitations in Φ_{PL} prediction primarily stem from two key factors: (1) the Φ_{PL} dataset is the most limited among the four tasks and exhibits a highly skewed distribution, with experimental values heavily clustered at one extreme of the range; (2) Φ_{PL} is influenced by molecular structure and solvent effects, and it is also highly sensitive to experimental conditions—including temperature, light source characteristics, filter selection, detector sensitivity, and sample-to-solvent ratios. These inherent complexities make Φ_{PL} prediction fundamentally more challenging than other optical properties. Despite these challenges in Φ_{PL} prediction, Fluor-pred maintains superior performance compared to existing models.

2.4 Results of Fluor-opt

2.4.1 Development of NIR/non-NIR dye classification models. Prior to conducting MMP analysis, we constructed a binary classification model to distinguish NIR from non-NIR dyes. This model served two primary purposes: evaluating the effectiveness of structural transformation rules and screening NIR candidates generated by Fluor-opt. Given the substantial

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

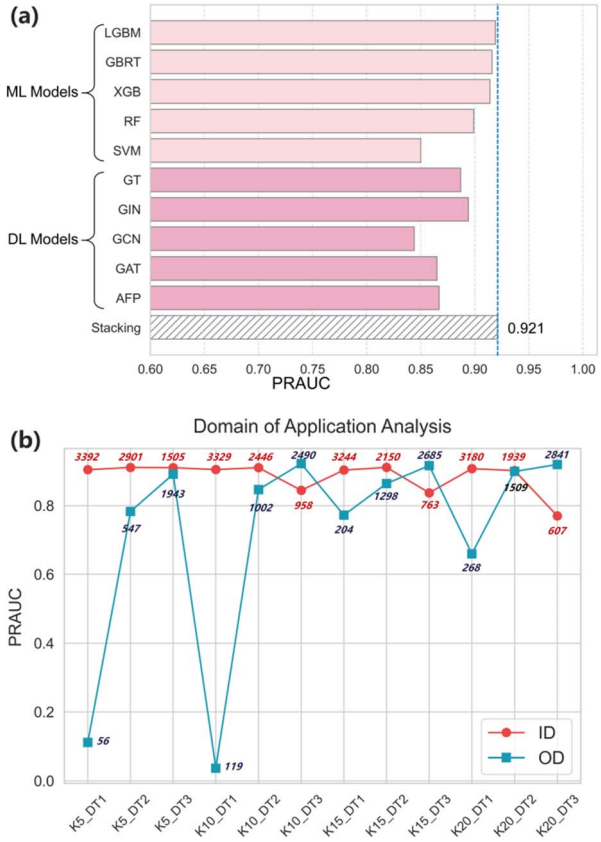

Fig. 5 (a) Predictive performance comparison between the ensemble model and baseline models, where the optimal model combines LGBM, GBRT and XGB. (b) Applicability domain analysis and selection for the ensemble model. Red and blue points represent the PRAUC values for compounds within and outside the application domain in the test set, respectively.

class imbalance (with non-NIR dyes far outnumbering NIR dyes), we selected PRAUC as the primary evaluation metric and employed both undersampling and oversampling strategies during training. Five machine learning algorithms and five deep learning models were evaluated. As shown in Fig. 5a, the final ensemble model—comprising LGBM, GBRT, and XGBoost—achieved high predictive performance on the test set (ACC = 0.981, AUC = 0.985, PRAUC = 0.921). This model enables accurate classification of NIR and non-NIR dyes, thereby providing a robust foundation for both MMP rule validation and downstream molecule screening. Detailed parameter settings and model comparisons are summarized in Tables S2–S5.

To further ensure prediction reliability, we conducted an applicability domain (AD) analysis. Based on the test set, the average Tanimoto similarity (γ) between test and training molecules was 0.3772 (σ = 0.1274). After jointly optimizing chemical space coverage and model performance, the optimal parameters (k = 20, Z = 2) yielded a similarity threshold of 0.632, which successfully captured 90% of the test set (Fig. 5b). For compounds within the applicability domain, the model maintained strong predictive power (PRAUC = 0.9077, MCC = 0.9789, F1-score = 0.8378).

Ultimately, we applied the model to the previously collected set of 17 194 dye molecules lacking experimental absorption wavelength data, successfully identifying 78 high-confidence NIR candidates (*i.e.*, AD-compliant compounds with prediction probabilities > 0.8) for MMP dataset augmentation. Due to the limited number of NIR dyes in the dataset, these molecules help further alleviate the shortage of NIR dyes and facilitate the extraction of more comprehensive structural transformation rules in MMPA analysis.

2.4.2 Deriving transformation rules from non-NIR to NIR. The MMPA analysis was conducted based on the augmented dataset derived from the aforementioned QSAR model, comprising a total of 17 287 dye molecules, including 1149 NIR dyes and 16 138 non-NIR dyes. Traditional MMP approaches overlook the contextual environment of structural transformations.[33] They often apply extracted transformation rules indiscriminately to all molecular structures. This practice is clearly inappropriate in the field of dyes. As shown in Fig. 3, dye molecules with different scaffolds exhibit significant differences in their properties; therefore, transformation rules derived from one scaffold, such as squaraine, may not be applicable to other types like cyanine dyes. We addressed these limitations by defining the shared scaffold fragments as the applicability context for each rule. During the optimization process, we introduced a hyperparameter, similarity_value, to identify the most compatible transformation rule for a given target molecule. This parameter quantifies how well a rule matches the target molecule; higher similarity implies stronger compatibility and a higher chance of valid NIR transformation. Furthermore, traditional MMP analysis is typically limited to single-site modifications. However, dye molecules are often subjected to symmetric modifications. To account for this, we extended the traditional MMP framework by incorporating symmetry-aware transformation rules, treating molecule pairs with consistent dual-site substitutions as valid matched molecular pairs. Detailed steps of the MMPA are provided in Section 4.2.4.

The first step in the MMP process is molecular fragmentation. To achieve comprehensive fragmentation coverage, we applied bond disconnection strategies involving one to three disconnection points. As shown in Table 2, three-point cuts resulted in significantly more fragments, and consequently, more MMPs. This outcome is primarily due to the inherent structural complexity of dye molecules, which often feature extended conjugated systems and polycyclic frameworks. These characteristics make simple single- or double-bond cuts insufficient for effective decomposition. Using this strategy, we

Table 2 Number of fragments and MMPs corresponding to molecular cutting strategy

Cutting strategy	Number of fragments	MMPs
Single cut	150 318 (2.76%)	168 030 (0.21%)
Double cut	1 149 542 (21.14%)	818 033 (1.05%)
Triple cut	4 136 517 (76.1%)	7 476 515 (98.74%)
Total	5 436 377	77 263 789

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

generated a total of 5 436 377 unique fragments and identified 8 462 578 matched molecular pairs.

To identify rules that favor the conversion of non-NIR dyes into NIR dyes, we analyzed the transformation frequency of each rule between non-NIR and NIR molecules and introduced an evaluation metric, label_value. As defined in eqn (1), $N(0 \rightarrow 1)$ indicates the number of times the rule converts non-NIR dyes into NIR dyes; $N(1 \rightarrow 0)$ indicates the number of times it converts NIR dyes into non-NIR dyes; indicates the total number of occurrences of the rule. Therefore, a label_value greater than 0 indicates that the transformation rule tends to promote the conversion from non-NIR to NIR dyes. After removing duplicate entries, we identified 1579 transformation rules that favor the non-NIR to NIR conversion (with label_value greater than 0) and extracted a total of 10 354 distinct transformation contexts.

$$\text{Label\_value} = \frac{N(0\rightarrow 1) \times 1 - N(1\rightarrow 0)}{N_{\text{all}}} \quad (1)$$

Table S6 presents several transformation rule examples, where frequency denotes the occurrence frequency of a given rule in converting non-NIR dyes into NIR dyes within the dataset, and label_value quantifies its effectiveness in promoting such conversions. For example, transformation rule 6 has a label_value of 1, indicating that all instances of its application resulted in forward transformations. Statistically, the majority of these forward rules (61.1%) appeared only once, while a smaller subset (10.7%) appeared more than 20 times. This imbalance is primarily attributed to the limited number of available NIR dye samples.

2.4.3 Validation and case study of Fluor-opt. To evaluate the reliability of transformation rules and the effectiveness of the context–aware parameter similarity_value, we selected 1923 non-NIR dye molecules with λ_{abs} values between 600–700 nm from our dataset as optimization targets. For each molecule, transformation rules were extracted and applied under varying similarity_value thresholds. As demonstrated in Fig. 6a, Fluor-opt consistently generated substantial numbers of NIR candidates across all threshold conditions. Notably, higher Similarity_Value thresholds reduced both the number of applicable rules per molecule and the total output structures. Nevertheless, these more stringent rules showed better alignment with the structural context of target molecules, resulting in significantly higher conversion success rates. The finding validate the effectiveness of the 1579 forward transformation rules, as well as the utility of the hyperparameter similarity_value. Considering that all generated molecules are ultimately screened by the NIR binary classification model, we recommend setting the

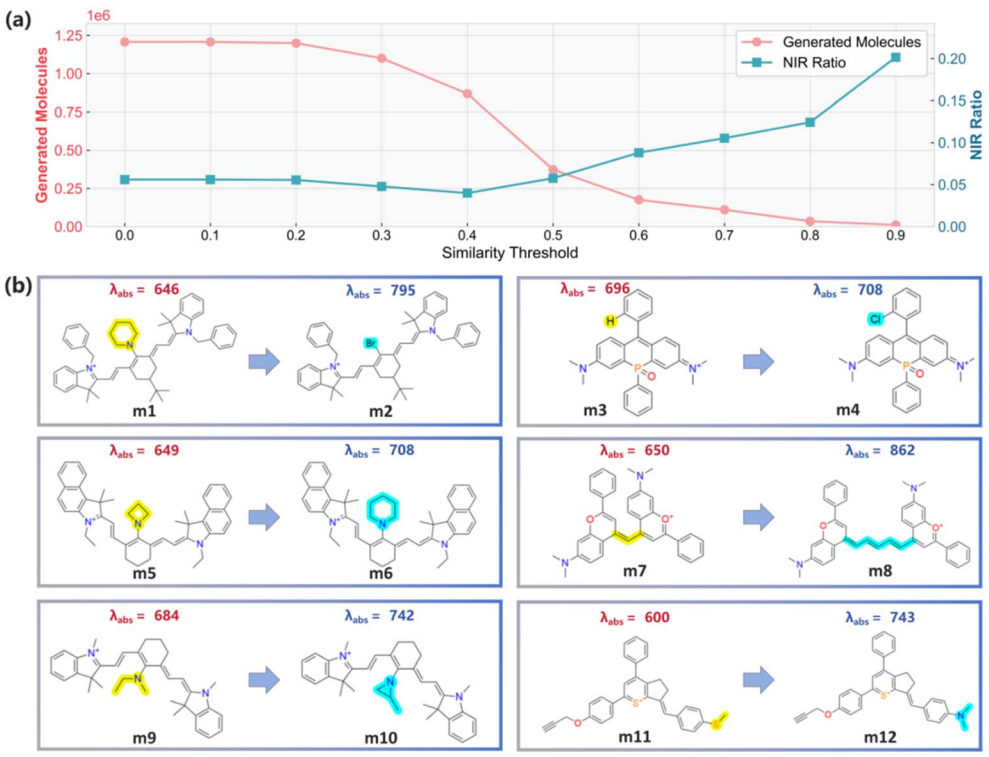

Fig. 6 (a) Line plot showing the total number of generated molecules and the corresponding proportion of NIR-classified compounds at different similarity thresholds. Higher similarity thresholds reduce the number of generated molecules but increase the success rate of conversion to NIR dyes. (b) Case studies: the compounds on the left side of the arrow represent the original input molecules, while the compounds on the right side correspond to the optimized molecules generated by Fluor-opt. Fluorescent markers highlight the transformed substructures, with all λ_{abs} values obtained from experimental measurements. **m1** and **m2** represent the original and modified dye molecules by Pascal *et al.*;[51] **m3** and **m4** denote the dye variants before and after modification by Li *et al.*;[52] **m5**, **m6**, **m9**, and **m10** correspond to Zhang *et al.*'s original and engineered dye molecules;[53] **m7** and **m8** indicate the pre- and post-modified dyes developed by Cosco *et al.*;[54,55] **m11** and **m12** refer to Zhou *et al.*'s molecular designs before and after optimization.[56]

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

similarity_value threshold between 0.2 and 0.4. This range achieves a balanced trade-off—ensuring reliable output quality while avoiding overly stringent filtering that may exclude promising NIR candidates.

Fluor-opt performs structural modifications including functional group substitution, functional group addition, and linker extension while preserving the core scaffold of the molecule. To validate its effectiveness, we conducted multiple case studies. As shown in Fig. 6b, the molecules on the left represent the original non-NIR dyes, while the molecules on the right correspond to their optimized structures. After optimization, the target dyes generally exhibited significant absorption redshift, typically ranging from 50 to 200 nm. For instance, when **m1** was used as the target molecule, optimization *via* Fluor-opt yielded the optimized molecule **m2**—a structure previously reported by Pascal *et al.*—with an absorption wavelength increase of nearly 150 nm.[51] Additionally, we optimized **m7**, a molecule developed by Cosco *et al.* in 2017, and the resulting optimized structure was identified as **m8**—a high-performance NIR dye first reported in 2021, which exhibited an absorption wavelength enhancement of over 200 nm.[52,53]

Remarkably, Fluor-opt achieves highly effective optimization through minor modifications, primarily by substituting a single key functional group. Compared to molecular generation methods, this strategy offers significant advantages in terms of synthetic accessibility. This tool provides valuable guidance for dye chemists, significantly expanding both the application scope and translational potential of fluorescent dyes in NIR-related technologies.

2.5 Website setup and services

As shown in Fig. 7, to establish a user-friendly and integrated platform for dye design and analysis, we integrated Fluor-opt and Fluor-pred into a unified tool—Fluor-tools, which has been deployed as a publicly accessible web application. Fluor-tools is freely accessible at **https://lmmd.ecust.edu.cn/Fluor-tools/**.

These modules form a synergistic closed-loop system through bidirectional knowledge transfer: the MMP fingerprints constructed from transformation rules extracted by Fluor-opt provide Fluor-pred with critical structural features, while Fluor-pred's high-accuracy predictions enable reliable evaluation of Fluor-opt's optimization results (Fig. 7d). Through this synergistic mechanism, researchers can first generate NIR candidate molecules *via* Fluor-opt, then efficiently screen them using Fluor-pred's predictive capabilities, enabling multi-parameter optimization that dramatically accelerates the development of high-performance fluorescent dyes.

3 Discussion and conclusions

This paper presents Fluor-tools, an integrated computational platform designed to address key challenges in the design and property prediction of fluorescent dyes. The platform comprises two core, synergistically working components: Fluor-pred and Fluor-opt. (1) Fluor-pred incorporates multimodal domain-specific knowledge of dyes, and by combining a specially designed RLAT architecture with a LDS reweighting strategy, it achieves state-of-the-art performance across all four critical photophysical property prediction tasks. Moreover, although

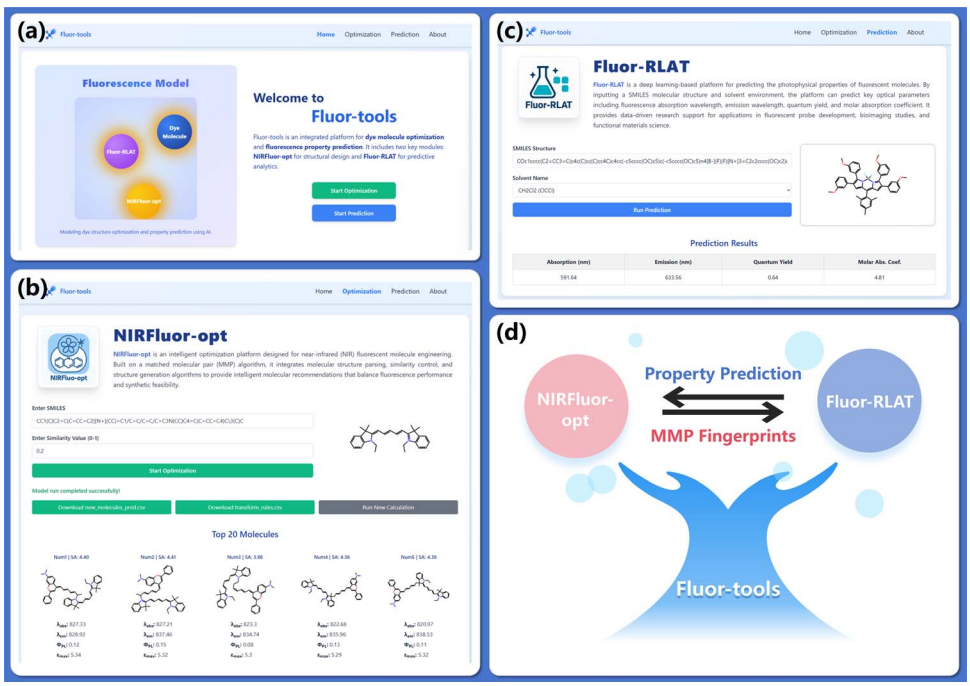

Fig. 7 Introduction to the Fluor-tools website: (a) Fluor-tools welcome page; (b) Fluor-pred, used for dye property prediction; (c) Fluor-opt, used for dye structure optimization; (d) the connection and functional transfer between Fluor-opt and Fluor-pred.

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

the multimodal architecture of Fluor-pred is tailored for dye, its use of the RLAT architecture and LDS reweighting strategy can also be explored in other molecular property prediction tasks. (2) Fluor-opt, on the other hand, leverages QSAR-enhanced MMPA and integrates dye-specific characteristics. Ultimately, it captures 1579 key structural transformations, enabling automated and targeted modification of dyes from non-NIR to NIR. Multiple case studies have validated its reliability.

However, Fluor-tools still has certain limitations. For example, due to the limited amount of training data for fluorescence quantum yield and inherent experimental measurement errors, the prediction accuracy of Fluor-pred for quantum yield remains relatively low. In addition, Fluor-opt currently focuses only on the directed optimization of absorption wavelength and has not yet undergone experimental validation. In future research, we will conduct in-depth investigations into the key factors influencing quantum yield and develop more rational methods to improve the prediction accuracy of Φ_{PL} For the structure optimization model Fluor-opt, we plan to explore incorporating target properties as constraints to establish a multi-objective generation framework and carry out corresponding experimental validations. In summary, Fluor-tools provides a reliable computational framework for the property prediction and rational optimization of dyes, and is well-positioned to play a significant role in biomedical imaging, materials science, and related fields.

4 Methods

4.1 Methods of Fluor-pred

4.1.1 Data collection and processing of Fluor-pred. In the Fluor-pred study, we utilized the original dataset from FluoDB,[26] which was compiled by integrating multiple public databases and collecting dye-related literature from PubMed. After removing invalid entries and duplicates, FluoDB provided 49 861 fluorophore-solvent pairs with experimental data on four optical properties: λ_{abs}, λ_{em}, Φ_{PL}, and ε_{max}.

On this basis, after further excluding molecules incompatible with Uni-mol+ calculations, the final dataset used in Fluor-pred contained 49 831 valid dye–solvent pairs. To ensure fairness in model comparison, we retained FluoDB's original 7 : 1 : 2 random partitioning scheme, resulting in 34 881 samples for training, 4982 for validation, and 9968 for testing, which were used for model training, hyperparameter optimization, and performance evaluation, respectively. In addition, to address the strong skew in the distribution of ε_{max} and improve model performance, a logarithmic transformation was applied to ε_{max} values prior to training.

4.1.2 Multimodal features of fluor-pred. Fluor-pred uses multimodal information to characterize dye and solvent molecules. As shown in Table S7, for dye molecules, we utilized molecular graph features, scaffold type labels, custom MMP fingerprints, Morgan fingerprints, quantum chemistry information (HOMO–LUMO gap), and five properties related to the dye molecules. For solvent molecules, solvent type labels and Morgan fingerprints were used for characterization.

The physicochemical properties of the dyes—namely molecular weight, log P, average Gasteiger charge, ring count, double bond count, and topological polar surface area (TPSA)—were all calculated using RDKit and custom scripts. For dye scaffold classification, we used the script provided by Zhu *et al.*, which defines 728 fluorescent dye scaffolds and categorizes the dyes into 17 scaffold types.[26] Detailed definitions of these scaffold categories can be found in Fig. S7.

The construction of MMP fingerprints in Fluor-pred was carried out as follows: among the 1579 key structural transformations extracted in Fluor-opt, we retained transformation rules that occurred more than 20 times and identified 136 associated substructures. These substructures were then used to build dye-specific, substructure-aware MMP fingerprints. The complete list of 136 substructures, along with the code for generating the MMP fingerprints, is freely available at **https://lmmd.ecust.edu.cn/Fluor-tools/**.

In this study, we employed Uni-mol+, a deep learning model that uses 3D molecular conformations for accurate HOMO–LUMO gap prediction. The Uni-mol + framework starts by generating an initial 3D conformation using RDKit, then iteratively refines this conformation to the DFT equilibrium state through neural network-based optimization. The final optimized conformation is then used to predict the HOMO–LUMO gap. Since our ablation studies showed that the HOMO–LUMO gap contributes minimally to the improvement in model performance, we chose to use the average HOMO–LUMO gap values in the web service to save computational time.

4.1.3 Model architecture of Fluor-pred. In terms of model architecture, our design was inspired by the RLAT framework used in protein property prediction, which is commonly employed to extract protein sequence information.[49,50] In the Fluor-pred model, as shown in Fig. 1F, we also used the AttentionCNN module to extract sequential information from dye molecules and concatenated it with other feature representations. Compared to traditional deep attention models, RLAT significantly reduces computational cost and model complexity while maintaining high performance, making it well-suited for large-scale data processing, especially when computational resources are limited. Furthermore, RLAT can automatically learn effective representations from data and dynamically adjust its attention to important features through the attention mechanism, enhancing the model's adaptability and optimization capability. Additionally, RLAT can be effectively integrated with other deep learning techniques, such as convolutional neural networks (CNN) and graph neural networks (GNN), further boosting the model's performance.

Hyperparameter optimization was performed using a grid search strategy, encompassing both general parameters (learning rate, weight decay, batch size) and architecture-specific parameters (number of network layers, dropout rate, graph feature dimension). To prevent overfitting and optimize computational efficiency, we implemented an early stopping strategy that terminated training if no improvement in validation metrics was observed for 20 consecutive epochs. For the four regression tasks predicted by Fluor-pred, we primarily used

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

MAE, RMSE, and R^2 as evaluation metrics. The definitions and formulas for all metrics are summarized in Table S8.

4.1.4 Reweighting method. In the fluorescence dye property prediction tasks, as illustrated in Fig. S1, the target variables are distributed unevenly across wide numerical ranges. For example, both λ_{abs} and λ_{em} are heavily concentrated in the mid-spectrum region, while Φ_{PL} are often clustered near zero. This skewed distribution may lead the model to focus primarily on the dominant regions, while underperforming on rare but critical boundary regions, ultimately limiting its generalization capability.

To mitigate this issue, we introduce the label distribution smoothing with inverse density strategy. The core idea of this approach is to estimate the smoothed label density distribution *via* kernel density estimation (KDE) and assign higher importance to samples located in low-density regions, thereby encouraging the model to learn more effectively from underrepresented targets. The detailed procedure is as follows:

(1) Label discretization and density estimation: for each task, the labels in the training set are uniformly discretized into a fixed number of bins (*e.g.*, 100), and Gaussian kernels are applied to smooth the label distribution. The smoothed density $\hat{p}(y_i)$ for a given label y_i is computed as follows:

$$\hat{p}(y_i) = \frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y_j - y_i)^2}{2\sigma^2}\right) \quad (2)$$

where h is the smoothing bandwidth, σ is the standard deviation of the Gaussian kernel, and n is the number of bins.

(2) Weight assignment: based on the smoothed density $\hat{p}(y_i)$, each sample is assigned a weight inversely proportional to its estimated density, encouraging higher importance for rare labels:

$$s_i = \frac{1}{\hat{p}(y_i) + \varepsilon} \quad (3)$$

where ε is a small constant (*e.g.*, 10^{-6}) to prevent division by zero.

(3) Weight normalization: to avoid gradient instability due to large variance in sample weights, all weights are normalized by their mean:

$$\overline{s}_i = \frac{s_i}{\frac{1}{N}\sum_{i=1}^{N} s_i} \quad (4)$$

where N is the total number of samples in the training set. This ensures that the reweighted loss remains on a comparable scale.

(4) Reweighted loss function: the final loss used during training is a reweighted Mean Squared Error (MSE), incorporating the normalized weights. To further stabilize training and make the loss more interpretable, we take the square root of the weighted average:

$$L = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\overline{s}_i \cdot (y_i - \hat{y}_i)^2} \quad (5)$$

where B is the batch size, y_i is the true label, and $\hat{y}_i$ is the predicted value.

By incorporating LDS-Inverse, the model can better learn features from low-frequency sample regions, thereby improving its performance on edge cases, especially for dye molecules with extreme spectral properties or uncommon photophysical behaviors.

4.2 Methods of Fluor-opt

4.2.1 Data collection and processing of Fluor-opt. The dataset used in Fluor-opt was constructed by integrating publicly available dye databases with targeted literature retrieval. Specifically, as summarized in Table 3, the collected sources include ChemFluor,[19] Deep4Chem,[20] SMFluo,[28] ChemDataExtractor,[31] Dye Aggregation,[29] DSSCDB,[30] experimental data from Ksenofontov *et al.*,[32] and Fluor-predictor.[40] In addition, we systematically retrieved and compiled relevant studies from Scopus, ScienceDirect, and Web of Science using the keywords "near-infrared" and "dye" to collect comprehensive experimental data. The obtained data underwent the following preprocessing steps to ensure consistency and applicability:

(1) Standardization: all datasets were unified in terms of format and units. Furthermore, dye and solvent molecules were converted into standardized SMILES representations using RDKit.

(2) Data Cleaning: since Fluor-opt focuses exclusively on optimizing λ_{abs}, we retained only λ_{abs} data from the above sources. In addition, identical dye–solvent pairs may exhibit conflicts across different datasets; therefore, we identified and removed entries with abnormal discrepancies (λ_{abs} > 5 nm). For entries within the acceptable range of variation, the average value was taken as a substitute.

(3) Removal of solvent conditions: as MMPA is a structure-based analytical method that does not account for solvent effects on dye properties, we averaged the λ_{abs} values of each dye molecule across different solvent conditions and adopted $\lambda_{abs} \geq$ 700 nm as the classification threshold for near-infrared dyes.

The final curated dataset consists of 17 240 experimentally validated dye structures, including 1096 NIR dyes and 16 144 non-NIR dyes. This dataset was initially used to build a binary classification model for NIR dye identification. Additionally, we retained 17 194 dye molecules without experimentally

Table 3 Data sources of Fluor-opt and the number of fluorescent dye–solvent Pairs

Data source	Number of entries
ChemFluor	4386
Deep4Chem	30094
SMFluo	1181
ChemDataExtractor	1915
Dye aggregation	3626
DSSCDB	2438
Fluor-predictor	36756
Ksenofontov's data	20608
Literature retrieval	1583

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

measured λ_{abs}. These molecules were predicted using the trained binary classification model, and those predicted as NIR dyes with high confidence and falling within the model's applicability domain were selected as supplementary data for subsequent MMPA analysis. For data analysis, we used RDKit and custom scripts to compute six physicochemical properties for both types of dyes, including molecular weight, log P, number of rings, number of double bonds, average Gasteiger charge, and TPSA. Dye scaffold classification was based on the structural classification framework established by Zhu *et al.*, which categorizes dyes into 17 distinct scaffold types.[26]

4.2.2 Binary classification model for NIR dyes. Before performing MMPA, we constructed a classification model to distinguish NIR dyes from non-NIR dyes using the aforementioned curated dataset, which includes 17 240 experimentally validated dye compounds. The dataset was randomly split into training (13 792 compounds), validation (1724 compounds), and test sets (1724 compounds) with an 8 : 1 : 1 ratio. We employed five ML models (LGBM, GBRT, XGBoost, RF, and SVM) and five DL models (GraphTransformer, GIN, GCN, GAT, and Attentive FP), applying both undersampling and oversampling strategies. ML models used Morgan fingerprints to represent molecules, while DL models used various GNN algorithms for molecular representation. The ensemble model was constructed by integrating individual models with PRAUC scores greater than 0.9, where the predicted probabilities from each individual model were used as inputs for logistic regression integration.

In the training of our binary classification model, we considered the severe imbalance between the two classes (with non-NIR dyes outnumbering NIR dyes). To address this issue, we employed both over-sampling and under-sampling strategies to balance the training set. (1) Over-sampling: we applied random over-sampling to the minority class (NIR dyes) by sampling with replacement, expanding its size to match that of the majority class. This ensured that the model was exposed to sufficient minority-class examples during training, thereby mitigating the risk of bias toward the majority class. (2) Undersampling: conversely, we also performed random undersampling of the majority class (non-NIR dyes), by randomly selecting a subset equal in size to the minority class. This yielded a smaller but balanced training set, which reduces the risk of overfitting introduced by over-sampling, though at the expense of potentially discarding part of the majority-class information.

Hyperparameter optimization was performed using a grid search strategy, encompassing both general parameters (learning rate, weight decay, batch size, dropout rate) and GNN-specific parameters (number of network layers, graph feature dimension). To prevent overfitting and optimize computational resource utilization, we implemented an early stopping strategy that terminated training if no improvement in validation set performance metrics was observed for 20 consecutive epochs. Detailed model parameters are provided in Table S5. For model evaluation, we used several metrics, including ACC (accuracy), MCC (Matthews correlation coefficient), F1 Score, recall, precision, SP (specificity), BA (balanced accuracy), AUC, and PRAUC. The definitions and formulas for all metrics are summarized in Table S8.

4.2.3 Definition of applicability domain. Defining the applicability domain is a crucial component of the five OECD principles for the development and validation of QSAR models, as it delineates the chemical space within which the model predictions are considered reliable.[34] In this study, we employed a Euclidean distance-based method (DM) grounded in structural similarity to estimate the AD of the model. Chemical structures were encoded using Morgan fingerprints, and the structural similarity between compounds was quantified using the Euclidean distance in the fingerprint space. The AD threshold was defined as follows:

$$\delta = \bar{d} + Z\sigma \tag{6}$$

d is the mean Euclidean distance between each compound in the training set and its nearest neighbor (also within the training set); σ is the standard deviation of these distances; Z is a user-defined parameter (typically between 0.5 and 1.0) that adjusts the significance level or strictness of the boundary. To evaluate whether a compound in the test set falls within the model's AD, we proceed as follows: for each compound in the test set, calculate the Euclidean distances to all compounds in the training set. Retain the k nearest neighbors and compute their average distance. If any of the k distances exceed the threshold δ, the compound is classified as outside domain; otherwise, it is inside domain.[41] This approach allows us to systematically assess the reliability of predictions based on structural proximity to known training data and prevents overinterpretation of extrapolated results.

4.2.4 Python-based MMPA and molecular optimization. MMPs refer to a pair of compounds with a single local structural change, which can reveal the dynamic relationship between molecules and their properties. In this work, MMPA and molecular optimization are divided into the following four parts: (1) molecular fragmentation, (2) obtaining matched molecular pairs, (3) extracting transformation rules, (4) and applying these transformation rules for molecular optimization.

(1) Molecular fragmentation: for molecular fragmentation, we implemented the algorithm proposed by Hussain and Rea automatically.[35,36] To comprehensively capture the property differences caused by structural changes, we performed extensive fragmentation of the molecules, including single-cut, double-cut, and triple-cut strategies, while preserving chirality during bond disconnection. As shown in Fig. 8a, the same molecule can be fragmented into different combinations of fragments, which are classified as either side chains or scaffolds.

(2) Identification of matched molecular pairs: two molecules are considered a matched molecular pair if they differ by only a single structural change. We identified the relevant MMPs by comparing the substructure fragments obtained in the previous step. If there is only one fragment difference between two fragment sequences, the corresponding molecules are considered a MMP. Additionally, since symmetric modifications are

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

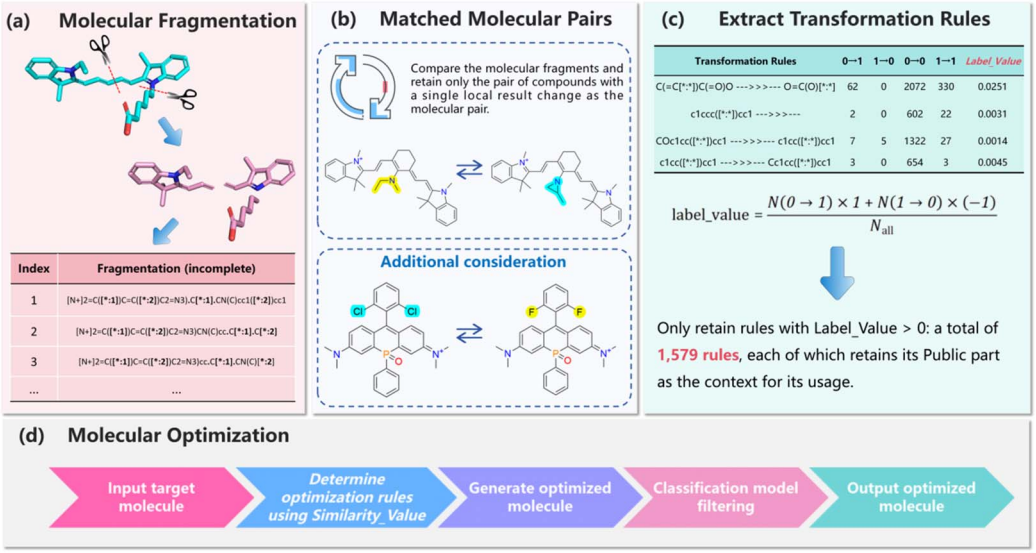

Fig. 8 MMPA and molecular optimization detailed methods. (a) molecular fragmentation; (b) obtaining matched molecular pairs; (c) extracting transformation rules; (d) and applying these transformation rules for molecular optimization.

commonly used in dye modifications, we introduced molecular symmetry changes as an additional criterion for MMPs. As shown in Fig. 8b, if two molecules undergo symmetric modifications at the same position, they are also considered a MMP. The MMP extraction process was carried out using custom scripts, which was time-consuming and took approximately 10 days to complete.

(3) Extracting transformation rules: as shown in Fig. 8c, after obtaining the matched molecular pairs, the next step is to extract the transformation rules from these pairs. We first quantified the label changes corresponding to each transformation rule. As described in Section 2.4.2, we introduced label_value to evaluate the transformation rules. If label_value is greater than 0, it indicates a positive rule that favors the conversion of non-NIR dyes to NIR dyes. The calculation method for label_value is shown in eqn (1). Ultimately, we obtained 1579 positive transformation rules, which were used for structural optimization in Fluor-opt. Traditional MMPA do not account for the transformation context, which could result in low conversion success rates. To address this issue, we retained the usage context for all positive transformation rules, meaning we preserved all molecules that successfully underwent transformation using the rule. These molecules then served as the context for applying the rule.

(4) Molecular optimization: as shown in Fig. 8d, during the optimization of the target molecule using transformation rules, we first introduced the hyperparameter similarity_value to select the most suitable transformation rules for the target molecule. similarity_value calculates the Tanimoto similarity between the target molecule and the environment molecules associated with each positive transformation rule, with the similarity computed based on Morgan fingerprints. The utility of similarity_value is described in Section 2.4.2, where it is shown that as similarity increases, the success rate of optimizing the molecule into an NIR dye also increases. Upon determining the transformation rules, Fluor-opt automatically substitutes the substructures of target molecules to achieve optimization, and evaluates the optimized molecules using the NIR dye prediction model, retaining only those predicted as NIR dyes. Finally, Fluor-pred performs detailed predictions of four key photophysical properties for the identified NIR dyes.

Author contributions

Y. T. and Y. Y. conceived and directed the work. W. S., Y. Z., L. X., X. L., J. Z., G. L., W. L., Y. Y., and Y. T. designed the study. W. S. collected and prepared the data and constructed the model. W. S., Y. Y., and Y. T. interpreted the results and wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

Fluor-tools is freely available for access and use at **https://lmmd.ecust.edu.cn/Fluor-tools/**. All data and related codes, as well as user instructions, are freely available at: **https://github.com/wenxiang-Song/fluor_tools** or **https://zenodo.org/records/17489518** (DOI: **https://doi.org/10.5281/zenodo.17489518**). The FluoDB and dye scaffold classification scripts are available for free at **https://github.com/ChemloverYuchen/FLAME** (DOI: **https://doi.org/10.1038/s41467-025-58881-5**). Uni-mol + can be freely installed and used *via* **https://github.com/deepmodeling/Uni-Mol** (DOI: **https://doi.org/10.48550/arXiv.2303.16982**). ChemFluor's DOI: **https://doi.org/10.1021/acs.jcim.0c01203**. Deep4Chem's DOI: **https://doi.org/10.1021/jacsau.1c00035**. SMFluo's DOI: **https://**

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

doi.org/10.1021/acs.jcim.1c01449. ChemDataExtractor's DOI: https://doi.org/10.1038/s41597-019-0306-0. Dye Aggregation's DOI: https://doi.org/10.3390/data5020045. DSSCDB's DOI: https://doi.org/10.1186/s13321-018-0272-0. Fluor-predictor's DOI: https://doi.org/10.1021/acs.jcim.5c00127. Experimental data from Ksenofontov *et al.* is a publicly available fluorescent dye database, sourced from DOI: https://doi.org/10.1016/j.saa.2022.121442.

Supplementary information is available. See DOI: https://doi.org/10.1039/d5dd00402k.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant U23A20530) and Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism (Shanghai Municipal Education Commission).

Notes and references

1 Y. Cai, W. Si, W. Huang, P. Chen, J. Shao and X. Dong, *Small*, 2018, **14**, 1704247.
2 W. Cheng, H. Chen, C. Liu, C. Ji, G. Ma and M. Yin, *VIEW*, 2020, **1**, 20200055.
3 C. Ji, L. Lai, P. Li, Z. Wu, W. Cheng and M. Yin, *Aggregate*, 2021, **2**, e39.
4 M. Kapil Dev and K. Uday, *Mini-Rev. Org. Chem.*, 2023, **20**, 655–674.
5 K. D. Mahato and U. Kumar, *Methods Appl. Fluoresc.*, 2023, **11**, 035003.
6 A. Tkaczyk, K. Mitrowska and A. Posyniak, *Sci. Total Environ.*, 2020, **717**, 137222.
7 N. Tomar, G. Rani, V. S. Dhaka and P. K. Surolia, *Int. J. Energy Res.*, 2022, **46**, 11556–11573.
8 M. Yang, J. Huang, J. Fan, J. Du, K. Pu and X. Peng, *Chem. Soc. Rev.*, 2020, **49**, 6800–6815.
9 R. Wang, J. Chen, J. Gao, J. A. Chen, G. Xu, T. Zhu, X. Gu, Z. Guo, W. H. Zhu and C. Zhao, *Chem. Sci.*, 2019, **10**, 7222–7227.
10 F. Salehpour, P. Cassano, N. Rouhi, M. R. Hamblin, L. De Taboada, F. Farajdokht and J. Mahmoudi, *Photobiomodulation, Photomed., Laser Surg.*, 2019, **37**, 581–595.
11 M. Wolf, M. Ferrari and V. Quaresima, *J. Biomed. Opt.*, 2007, **12**, 062104.
12 H. Li, H. Kim, F. Xu, J. Han, Q. Yao, J. Wang, K. Pu, X. Peng and J. Yoon, *Chem. Soc. Rev.*, 2022, **51**, 1795–1835.
13 H. Chen, B. Dong, Y. Tang and W. Lin, *Acc. Chem. Res.*, 2017, **50**, 1410–1422.
14 Y. Wang, H. Yu, Y. Zhang, C. Jia and M. Ji, *Dyes Pigm.*, 2021, **190**, 109284.
15 J. I. Scott, Q. Deng and M. Vendrell, *ACS Chem. Biol.*, 2021, **16**, 1304–1317.
16 Z. Lei and F. Zhang, *Angew Chem. Int. Ed. Engl.*, 2021, **60**, 16294–16308.
17 J. F. Joung, M. Han, M. Jeong and S. Park, *J. Chem. Inf. Model.*, 2022, **62**, 2933–2942.
18 C. Adamo and D. Jacquemin, *Chem. Soc. Rev.*, 2013, **42**, 845–856.
19 C.-W. Ju, H. Bai, B. Li and R. Liu, *J. Chem. Inf. Model.*, 2021, **61**, 1053–1065.
20 J. F. Joung, M. Han, J. Hwang, M. Jeong, D. H. Choi and S. Park, *JACS Au*, 2021, **1**, 427–438.
21 K. P. Greenman, W. H. Green and R. Gómez-Bombarelli, *Chem. Sci.*, 2022, **13**, 1152–1162.
22 S. G. Jung, G. Jung and J. M. Cole, *J. Chem. Inf. Model.*, 2024, **64**, 1486–1501.
23 X. Wang, H. Wu, T. Wang, Y. Chen, B. Jia, H. Fang, X. Yin, Y. Zhao and R. Yu, *Anal. Chem.*, 2025, **97**, 1992–2002.
24 A. D. Gorse, *Curr. Top. Med. Chem.*, 2006, **6**, 3–18.
25 H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, *J. Cheminf.*, 2024, **16**, 20.
26 Y. Zhu, J. Fang, S. A. H. Ahmed, T. Zhang, S. Zeng, J.-Y. Liao, Z. Ma and L. Qian, *Nat. Commun.*, 2025, **16**, 3598.
27 M. Han, J. F. Joung, M. Jeong, D. H. Choi and S. Park, *ACS Cent. Sci.*, 2025, **11**, 219–227.
28 J. Shao, Y. Liu, J. Yan, Z.-Y. Yan, Y. Wu, Z. Ru, J.-Y. Liao, X. Miao and L. Qian, *J. Chem. Inf. Model.*, 2022, **62**, 1368–1375.
29 V. Venkatraman and L. Kallidanthiyil Chellappan, *Data*, 2020, **5**, 45.
30 V. Venkatraman, R. Raju, S. P. Oikonomopoulos and B. K. Alsberg, *J. Cheminf.*, 2018, **10**, 18.
31 E. J. Beard, G. Sivaraman, Á. Vázquez-Mayagoitia, V. Vishwanath and J. M. Cole, *Sci. Data*, 2019, **6**, 307.
32 A. A. Ksenofontov, M. M. Lukanov and P. S. Bocharov, *Spectrochim. Acta Mol. Biomol. Spectrosc.*, 2022, **279**, 121442.
33 D. Gurvic, A. G. Leach and U. Zachariae, *J. Med. Chem.*, 2022, **65**, 6088–6099.
34 S. Bhatia, T. Schultz, D. Roberts, J. Shen, L. Kromidas and A. Marie Api, *Regul. Toxicol. Pharmacol.*, 2015, **71**, 52–62.
35 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.
36 A. Dalke, J. Hert and C. Kramer, *J. Chem. Inf. Model.*, 2018, **58**, 902–910.
37 J. Shao, Y. Liu, J. Yan, Z.-Y. Yan, Y. Wu, Z. Ru, J.-Y. Liao, X. Miao and L. Qian, *J. Chem. Inf. Model.*, 2022, **62**, 1368–1375.
38 S.-H. Hung, Z.-R. Ye, C.-F. Cheng, B. Chen and M.-K. Tsai, *J. Chem. Theory Comput.*, 2023, **19**, 4559–4567.
39 C. Liu, Y. Sun, R. Davis, S. T. Cardona and P. Hu, *J. Cheminf.*, 2023, **15**, 29.
40 W. Song, L. Xiong, X. Li, Y. Zhang, B. Wang, G. Liu, W. Li, Y. Yang and Y. Tang, *J. Chem. Inf. Model.*, 2025, **65**, 2854–2867.
41 T. Alexander and G. Alexander, *Curr. Pharm. Des.*, 2007, **13**, 3494–3504.
42 Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang and M. Zheng, *J. Med. Chem.*, 2020, **63**, 8749–8760.
43 S. Lu, Z. Gao, D. He, L. Zhang and G. Ke, *Nat. Commun.*, 2024, **15**, 7104.
44 M. Steininger, K. Kobs, P. Davidson, A. Krause and A. Hotho, *Mach. Learn.*, 2021, **110**, 2187–2211.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 10 November 2025. Downloaded on 6/24/2026 1:48:02 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

45 D. Jacquemin, V. Wathelet, E. A. Perpète and C. Adamo, *J. Chem. Theory Comput.*, 2009, **5**, 2420–2435.
46 A. Dreuw and M. Head-Gordon, *Chem. Rev.*, 2005, **105**, 4009–4037.
47 J. R. De Lile, S. G. Kang, Y.-A. Son and S. G. Lee, *ACS Omega*, 2020, **5**, 15052–15062.
48 H. Luo and M. Mohammadnia, *Inorg. Chem. Commun.*, 2022, **141**, 109522.
49 J. E. Gado, M. Knotts, A. Y. Shaw, D. Marks, N. P. Gauthier, C. Sander and G. T. Beckham, *Nat. Mach. Intell.*, 2025, **7**, 716–729.
50 S. Qiu, B. Hu, J. Zhao, W. Xu and A. Yang, *Briefings Bioinf.*, 2025, **26**(2), DOI: **10.1093/bib/bbaf114**.
51 S. Pascal, A. Haefele, C. Monnereau, A. Charaf-Eddin, D. Jacquemin, B. Le Guennic, C. Andraud and O. Maury, *J. Phys. Chem.*, 2014, **118**, 4038–4047.
52 N. Li, T. Wang, N. Wang, M. Fan and X. Cui, *Angew. Chem., Int. Ed.*, 2023, **62**, e202217326.
53 J. Zhang, M. Moemeni, C. Yang, F. Liang, W.-T. Peng, B. G. Levine, R. R. Lunt and B. Borhan, *J. Mater. Chem. C*, 2020, **8**, 16769–16773.
54 E. D. Cosco, J. R. Caram, O. T. Bruns, D. Franke, R. A. Day, E. P. Farr, M. G. Bawendi and E. M. Sletten, *Angew. Chem., Int. Ed.*, 2017, **56**, 13126–13129.
55 E. D. Cosco, B. A. Arús, A. L. Spearman, T. L. Atallah, I. Lim, O. S. Leland, J. R. Caram, T. S. Bischof, O. T. Bruns and E. M. Sletten, *J. Am. Chem. Soc.*, 2021, **143**, 6836–6846.
56 H. Zhou, X. Zeng, A. Li, W. Zhou, L. Tang, W. Hu, Q. Fan, X. Meng, H. Deng, L. Duan, Y. Li, Z. Deng, X. Hong and Y. Xiao, *Nat. Commun.*, 2020, **11**, 6183.

© 2025 The Author(s). Published by the Royal Society of Chemistry