nature communications

Article

https://doi.org/10.1038/s41467-025-58881-5

A modular artificial intelligence framework to facilitate fluorophore design

Received: 20 July 2024

Accepted: 2 April 2025

Published online: 16 April 2025

Check for updates

Yuchen Zhu[1,5], Jiebin Fang[2,3,5], Shadi Ali Hassen Ahmed[1], Tao Zhang[4], Su Zeng[1], Jia-Yu Liao[1] ✉, Zhongjun Ma[2,3] ✉ & Linghui Qian[1] ✉

Fluorescence imaging, indispensable for fundamental research and clinical practice, has been driven by advances in fluorophores. Despite fast growth over the years, many available fluorophores suffer from insufficient performances, and their development is highly dependent on trial-and-error experiments due to subtle structure-property effects and complicated solvent effects. Herein, FLAME (FLuorophore design Acceleration ModulE), an artificial intelligence framework with a modular architecture, is built by integrating open-source databases, multiple prediction models, and the latest molecule generators to facilitate fluorophore design. First, we constructed the largest open-source fluorophore database to date (FluoDB), containing 55,169 fluorophore-solvent pairs. Then FLSF (FLuorescence prediction with fluoroScaFfold-driven model) with a domain-knowledge-derived fingerprint for characterizing fluorescent scaffolds (called fluoroscaffold) was designed and demonstrated to predict optical properties quickly and accurately, whose reliability and potential have been verified via molecular and atomistic interpretability analysis. Further, a molecule generator was incorporated to provide new compounds with desired fluorescence. Representative 3,4-oxazole-fused coumarins were synthesized and evaluated, creating an unreported compound with bright fluorescence.

Luminescent molecules have found widespread applications in numerous fields[1–3], among which fluorophores have attracted increasing attention in bioimaging due to their small size, chemical tractability, and low cost[4,5]. To meet specific requirements such as the penetration depth and detection sensitivity in bioimaging, the underlying structure-property relationship (SPR) of fluorophores is important for designing compounds with proper excitation wavelength and desired brightness[6–8]. However, our knowledge of this relationship remains limited[9–13], which is largely due to two reasons: (1) Data sparsity. That is, all possible structural modifications should be tested to illustrate the SPR of a specific fluorophore, but met with synthetic challenges. Moreover, the problem also lies in the limited access to complete, comparable, and meaningful photophysical data of existing fluorophores[14]. (2) Multiple interrelated factors may affect the fluorescence. A subtle modification in the structure may lead to significant optical changes, and the fluorescence may further be influenced by the surrounding environment, leaving the rational design of fluorophores difficult[6,8,15].

Recently, machine learning-based data-driven science has shown tremendous potential to become a very useful tool across various disciplines[16–18], such as predicting molecular properties[19], virtual screening[20], and molecular generation[21]. In the case of fluorescence,

[1]Institute of Drug Metabolism and Pharmaceutical Analysis, Research Center for Clinical Pharmacy, College of Pharmaceutical Sciences, State Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou 310058, China. [2]Hainan Institute of Zhejiang University, Sanya 572025, China. [3]Institute of Marine Biology and Pharmacology, Ocean College, Zhejiang University, Zhoushan 316021, China. [4]Institute of Intelligent Innovation, Henan Academy of Sciences, Zhengzhou 451162, China. [5]These authors contributed equally: Yuchen Zhu, Jiebin Fang. ✉e-mail: jyliao@zju.edu.cn; mazj@zju.edu.cn; lhqian@zju.edu.cn

multiple intricately intertwined properties need to be considered for the molecular design, including maximum absorption wavelength—λ_{abs}, maximum emission wavelength—λ_{em}, photoluminescence quantum yield—Φ_{PL}, and molar absorption coefficient—ε_{max}[6,15]. Pioneered by Tsuda et al., a massively parallelized version of de novo molecule generator (ChemTS) was employed to design fluorophores with absorption/emission wavelengths and oscillator strengths calculated by quantum chemical computation, generating 3643 candidate fluorophores by using 1024 cores for 5 days[22]. Though powerful, the high computational cost must be considered.

Due to its end-to-end paradigm, machine learning can learn directly from the data to identify implicit patterns and make predictions without any prior knowledge, promising for fluorescence prediction. For instance, Ju et al. established a database (ChemFluor) recording the optical properties of over 4300 solvated fluorophores[23]. Both the fluorophore and solvent were characterized using molecular descriptors or fingerprints, which were combined as the input for predicting photophysical parameters using the Gradient Boosted Regression Trees (GBRT) model. Similarly, Park's group developed a graph convolutional network (GCN)-based model[24] and employed the integrated gradients method[25] sequentially to predict seven optical properties and obtain attributions of atoms/functional groups/solvents to the optical properties. Very recently, Tsai et al. modified the SchNet model to introduce the solvation embedding outside the interaction layers so as not to overly amplify the solute-solvent interaction and provided enhanced prediction for ΔE_{abs} and ΔE_{emi}[26].

As researchers in fluorescent probes[27–29], we're keen to make an easy-operation toolkit that allows the generation of structure-new fluorophores with desired optical performance efficiently, to explore the frontiers of fluorophores with a minimized burden on chemical synthesis and experimental tests[30].

Very recently, Park et al. developed a generative deep learning (Gen-DL) model to generate molecules with seven predefined optical properties[31]. Alternatively, to fully exploit the chemical space, optical property prediction models can be introduced to molecule generators for efficient sampling to select optimal structures with desired optical properties.

Herein, we systematically compiled experimental data to build a new fluorophore database named FluoDB (Fig. 1), consisting of 55,169 fluorophore-solvent pairs, as the machine learning algorithm asks for large volumes of data to acquire effective information. Compared with existing databases, FluoDB improves in both data volume and molecular diversity, categorized with 16 core fluorescent scaffolds and 728 subgroups. Then we proposed a new prediction model, FLSF (FLuorescence prediction with fluoroScaFfold-driven model), in which a domain-knowledge-derived fingerprint encoded by 728 fluorescent-scaffold subgroups (called fluoroscaffold) is fused to traditional message passing neural networks (MPNN; reported to outperform SchNet, DTNN, and the Transformer in predicting UV-Vis spectra[32]) using the gated recurrent unit (GRU). In benchmarking tests, FLSF is advantageous at quickly and accurately predicting optical properties over previous state-of-the-art (SOTA) models. Its reliability and potential were further validated through a series of interpretability analyses. To guide the fluorophore design directly, we set up an artificial intelligence (AI) framework, FLAME (FLuorophore design Acceleration ModulE), by integrating different open-source databases, prediction models, and molecule generators. Using Reinvent 4[33] as a representative molecule generator, a series of compounds with predicted properties were generated. Among them, 3,4-oxazole-fused coumarins were synthesized using a novel one-pot synthetic methodology, giving an unreported compound with bright fluorescence, and exhibiting the potential of FLAME in accelerating fluorophore design.

Results

Data collection and processing

In our previous study, a database (SMFluo1) focusing on near infrared fluorophores was constructed, containing five widely used fluorescent scaffolds[29]. The limitation in data volume of SMFluo1 makes it difficult to meet the requirements of deep learning algorithms, particularly those based on graph neural networks (GNN), including GCN and Attentive FP[34], leading to moderately good prediction accuracy. In addition, to evaluate a fluorophore for bioimaging, four key photophysical parameters (λ_{abs}, λ_{em}, Φ_{PL}, and ε_{max}) are needed, where λ_{abs} and λ_{em} are related to the penetration depth and $\Phi_{PL}\times\varepsilon_{max}$ indicates the brightness[15]. Of note, other factors including the blinking, thermal stability, photobleaching, and labeling specificity should be taken into consideration when designing probes for bioimaging, but parameters for these properties were not included in FluoDB due to limited access[6]. Taking these into consideration, data collection was carried out as follows (Fig. 2a): (i) Literature survey via searching the name of fluorescent scaffolds on PubMed; (ii) Retrieval of experimental data from various open-source databases[23,35–40] and supplement with four photophysical parameters & solvent information from the original literature. These data were processed after the combination (see "Data processing" in Methods).

Most fluorescent compounds are derived from some basic scaffolds and they may share common optical characteristics; thus,

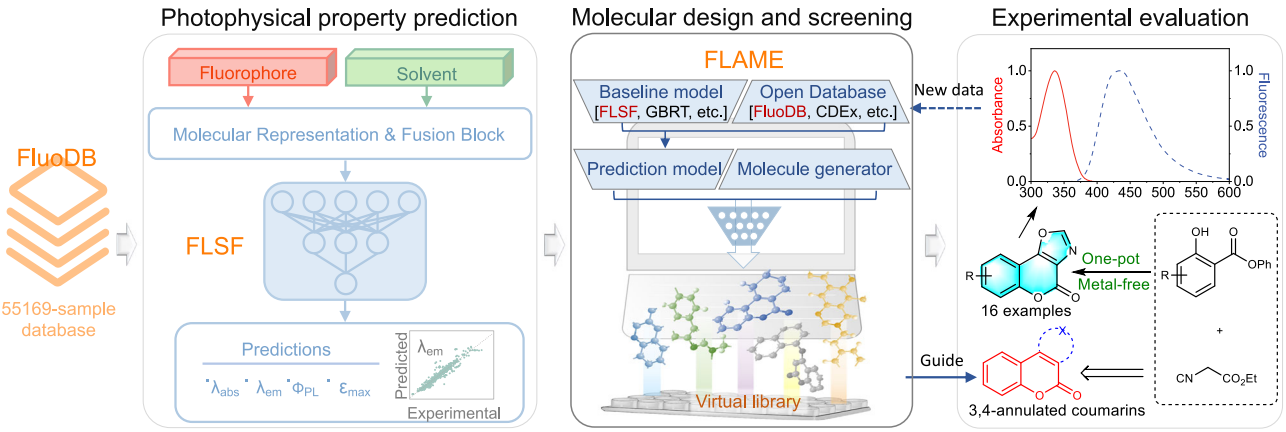

Fig. 1 | Facilitating fluorophore design with FLAME (FLuorophore design Acceleration ModulE). Overview of the user-friendly framework, FLAME, assembled from the latest databases (including FluoDB, the database constructed in the current study), prediction models (i.e., FLSF (FLuorescence prediction with fluoroScaFfold-driven model) constructed in the current study together with other state-of-the-art prediction models), and molecule generators, together with its application in the design of unreported fluorophores with desired fluorescence followed by experimental evaluation.

https://doi.org/10.1038/s41467-025-58881-5

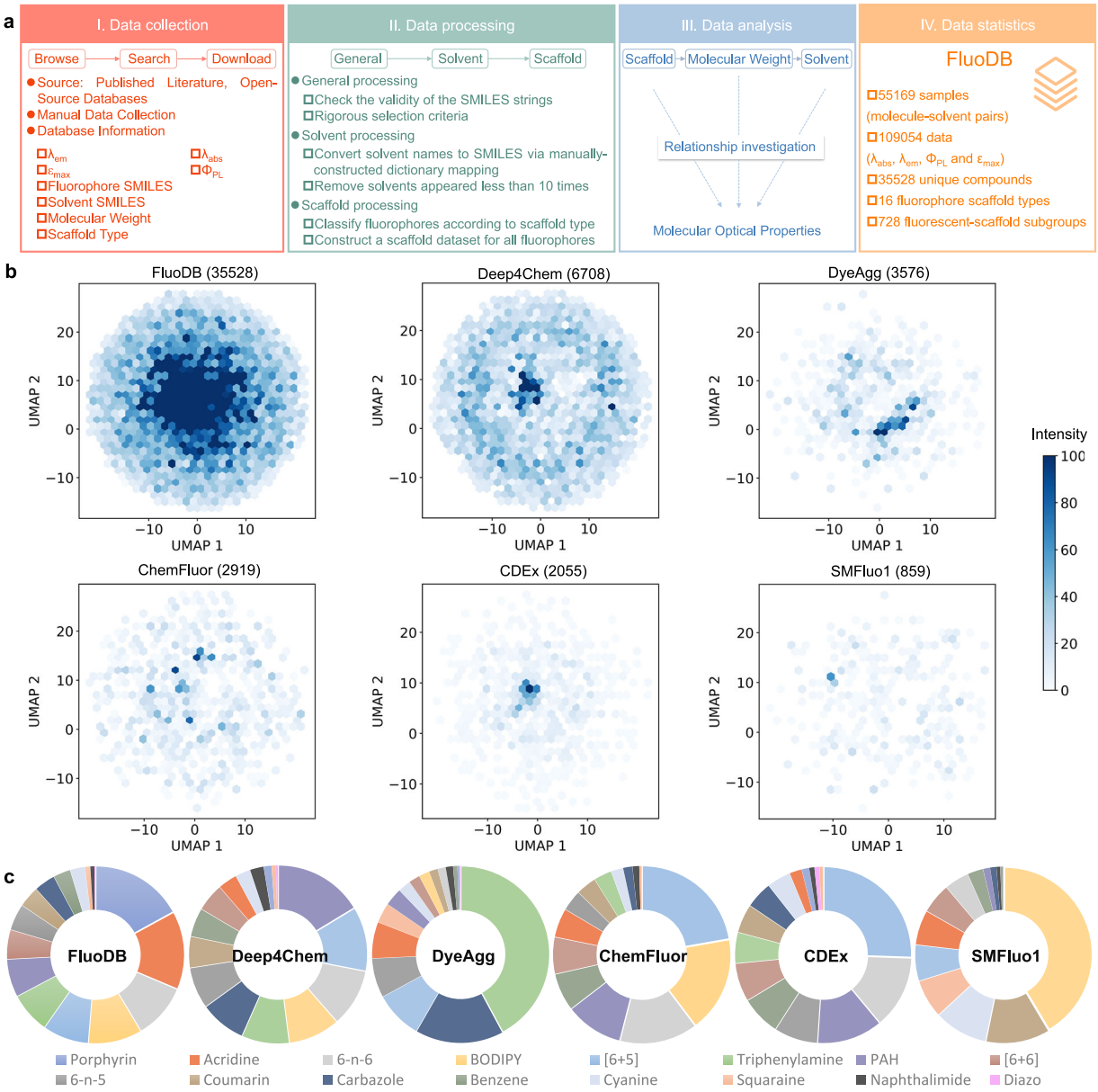

Fig. 2 | Construction and analysis of the fluorophore database, FluoDB. **a** General pipeline for data collection and processing to construct the new database, FluoDB, followed by systematic analysis and statistics of FluoDB to visualize the optical properties of various fluorophores in different solvents. **b** UMAP (Uniform Manifold Approximation and Projection) of different databases (Deep4Chem[36], DyeAgg[38], ChemFluor[23], ChemDataExtrator (CDEx)[35], and SMFluo1[29]) using Morgan fingerprints. The UMAP algorithm was applied with a neighborhood size of 10 and a minimum distance of 0.3. The number of unique compounds in each database is listed in the bracket. **c** Distribution of various fluorescent scaffolds in different databases. Source data are provided as a Source data file.

we categorized the fluorophores into twelve classic fluorescent scaffolds and four non-classical scaffolds (Fig. S1; detailed skeletal structures for 728 subgroups are shown in Table S1). FluoDB, a new database containing 35,528 unique fluorophores and 55,169 fluorophore-solvent pairs, was therefore constructed with SMILES of fluorophores/solvents, category of fluorescent scaffolds, experimental photophysical data, and original reference.

Compared to representative open-source databases (e.g., Deep4Chem[36], DyeAgg[38], ChemFluor[23], ChemDataExtrator (CDEx)[35], and SMFluo1[29]), FluoDB gets improved in the number of molecules and the richness in optical information (Fig. 2b and Fig. S2), exhibiting much higher molecular diversity according to the data distribution and structural analysis (Fig. S3 and Tables S2–3). In addition, different scaffolds distribute relatively even in both FluoDB and Deep4Chem, while the data of each category is largely enriched in FluoDB (Fig. 2c and Table S4).

Data analysis with FluoDB

With FluoDB, the correlations between different parameters were investigated (Fig. S4) and indicated an obvious positive correlation between λ_{abs} and λ_{em}. In addition, molecular weight (MW) also has a certain positive correlation with λ_{abs}, λ_{em}, and ε_{max}, which is consistent with the scatter plot analysis (Fig. S5) and experimental results (i.e., introducing large π bridging moieties or strong electron acceptors/donors is commonly used for longer absorption and emission wavelengths[41], and these modifications often increase the MW). External factors such as the surrounding solvent are reported to influence the optical properties of certain fluorophores[42].

To investigate this in a large scope, fluorophores with experimental data available for different solvents (≥5) were selected from FluoDB. The variance distribution of each photophysical parameter for selected molecules in different solvents is shown in Fig. S6, where these parameters do vary along the change of solvent type, underscoring the importance of solvents when predicting optical properties.

As mentioned earlier, we divided the fluorescent scaffolds into 16 types (Fig. S1 and Table S1), and any input fluorophore can be quickly classified accordingly to explore potential commonalities from the same group. As indicated in Figs. S7–22, a discrepancy was found in λ_{abs} and λ_{em} distribution with different scaffolds, where most groups centered in the UV-Vis range, while BODIPY, porphyrin, and squaraine lie in longer wavelength (above 550 nm)[43,44]. In addition, larger wavelength tunability was found from acridine, naphthalimide, coumarin, and cyanine. Besides, $\Delta\lambda$ (Stokes shift, $\Delta\lambda = \lambda_{em} - \lambda_{abs}$) of BODIPY, porphyrin, and squaraine is relatively small (~25 nm). Statistical analysis of such large-scale data is indicative for choosing the ideal fluorescent scaffold to start with, and we also prepared a toolkit where users can search for fluorophores with desired similarity as the molecule of interest from the database.

Workflow and prediction performance of FLSF

With FluoDB, open-source prediction models, including GBRT[23], SMFluo[29], UVVisML[45], SchNet[26], and ABT-MPNN[46], were tested. Data in FluoDB-Lite (SMILES in the mixture/complex form were removed from FluoDB) was divided randomly in a ratio of 7:1:2 for training, validation, and testing, respectively (Table S5). As shown in Table 1 and Table S6, ABT-MPNN, a general molecular property prediction model based on an atom-bond Transformer, performed best in predicting λ_{abs} and λ_{em}, highlighting the advantage of combining Transformers with GNN for molecular representation. However, the introduction of attention mechanisms in ABT-MPNN led to 10 times slower training than UVVisML (a Directed MPNN, D-MPNN), despite the improved MAE for λ_{abs} and λ_{em} by 9.18% and 4.86%, respectively. For practicability, it is desirable to replace the attention mechanism in ABT-MPNN with a new way to speed up the training while maintaining the prediction accuracy.

As all fluorophores in FluoDB were classified into 16 core scaffolds and distribution discrepancy in optical properties among them was observed (Figs. S7–21), a special molecular fingerprint—fluoroscaffold (a 728-dimensional digital fingerprint encoded by 728 fluorescent-scaffold subgroups listed in Table S1), fused with the current feature extraction method based on MPNN, was designed for better molecular representation of fluorophores. A new prediction model (FLSF) was constructed based on it (Fig. 3a). As shown in Table 2, FLSF predicted well for different fluorophores, especially for BODIPY-based compounds (the largest proportion in FluoDB) with MAE of 6.44 nm/7.37 nm for $\lambda_{abs}/\lambda_{em}$. For non-classical scaffolds (i.e., [6 + 5], [6 + 6], 6-n-5, 6-n-6), FLSF also has a good performance, promising for dealing with novel fluorophores. Overall, FLSF performs well at predicting λ_{abs} and λ_{em} ($R^2 = 0.94$) and needs improvement at Φ_{PL} and ε_{max} ($R^2 \approx 0.6$; Fig. 3b). Then we conducted benchmark tests of FLSF and summarized the results in Table 1 and Table S6 for direct comparison with reported SOTA models. Obvious improvements were seen in the prediction accuracy of λ_{abs}, λ_{em}, and ε_{max} by FLSF than ABT-MPNN (the same of Φ_{PL}), at a much faster speed, indicating its great potential for high-throughput screening of candidate fluorophores. To check whether FLSF can capture the solvent effect, a multi-solvent test set (fluorophores with experimental data available for ≥4 different solvents) together with the control test set (fluorophores in the same solvent) was selected and the prediction performance of FLSF was compared with other baseline models. As shown in Tables S7–8, FLSF has the best prediction performance on the multi-solvent test set. To our delight, FLSF can also predict λ_{abs} and λ_{em} of fluorophores showing solvatochromism with high accuracy (Tables S9–10), further supporting its potency to capture solvent effects.

Table 1 | Prediction performance with different models (including previously reported GBRT[23], SMFluo[29], UVVisML[45], SchNet[26], and ABT-MPNN[46]) towards FluoDB

Object	Algorithms	MAE	MSE	RMSE	R^2
λ_{abs}	GBRT	13.67	824.24	28.71	0.93
	SMFluo	21.19	1255.71	35.44	0.89
	UVVisML	13.94	716.91	26.78	0.94
	SchNet	22.17	1684.74	41.05	0.63
	ABT-MPNN	12.66	687.97	26.23	0.94
	FLSF_MACCS	12.96	713.33	26.71	0.94
	FLSF_Morgan	14.75	853.52	29.22	0.92
	FLSF	12.56	675.34	25.99	0.94
λ_{em}	GBRT	14.56	671.52	25.91	0.92
	SMFluo	27.82	1467.36	38.31	0.83
	UVVisML	13.98	518.02	22.76	0.94
	SchNet	38.26	2695.06	51.91	0.43
	ABT-MPNN	13.30	521.65	22.84	0.94
	FLSF_MACCS	13.88	560.29	23.67	0.94
	FLSF_Morgan	15.66	746.21	27.32	0.92
	FLSF	13.27	545.12	23.35	0.94
Φ_{PL}	GBRT	0.12	0.03	0.18	0.68
	SMFluo	0.13	0.04	0.21	0.57
	UVVisML	0.13	0.04	0.19	0.64
	SchNet	0.15	0.04	0.20	0.39
	ABT-MPNN	0.12	0.03	0.19	0.65
	FLSF_MACCS	0.13	0.04	0.20	0.61
	FLSF_Morgan	0.12	0.04	0.19	0.64
	FLSF	0.12	0.03	0.19	0.66
ε_{max}	GBRT	0.20	0.10	0.31	0.66
	SMFluo	0.22	0.14	0.37	0.53
	UVVisML	0.26	0.14	0.37	0.51
	SchNet	0.51	0.51	0.71	-2.01
	ABT-MPNN	0.32	0.20	0.45	0.31
	FLSF_MACCS	0.25	0.13	0.36	0.56
	FLSF_Morgan	0.23	0.11	0.33	0.61
	FLSF	0.23	0.12	0.34	0.59

The unit of MAE/RMSE is nm for λ_{abs} and λ_{em}, not applicable for Φ_{PL} and ε_{max} (used in $\log_{10}\varepsilon_{max}$). GBRT, SMFluo, UVVisML, SchNet, and ABT-MPNN are implemented from their open-source codes. The difference between FLSF (FLuorescence prediction with fluoroScaFfold-driven model) and FLSF_MACCS/FLSF_Morgan is the feature extraction method, as fluoroscaffold was introduced in FLSF, while the conventional MACCS/Morgan fingerprints were used for the latter. λ_{abs} maximum absorption wavelength, λ_{em} maximum emission wavelength, Φ_{PL} photoluminescence quantum yield, ε_{max} molar absorption coefficient, *MAE* mean absolute error, *MSE* mean-square error, *RMSE* root-mean-square error, R^2 the coefficient of determination.

Time-dependent density functional theory (TD-DFT) used to be the most widely used tool for predicting optical properties[47]. However, such traditional theoretical calculations require high computational and time costs[48], faced with insufficient accuracy in predicting λ_{abs} and λ_{em}, much less in parameters like Φ_{PL} involved in various radiation and non-radiation processes[49]. For direct comparison with FLSF, we collected 162 fluorophore-solvent pairs from FluoDB (Table S11) and used TD-DFT to calculate their λ_{abs} and λ_{em} (Fig. 3c and Table S12). The MAE of FLSF decreased by more than 0.2 eV for predicting λ_{abs} and λ_{em} than TD-DFT. Of note, FLSF can provide all prediction results in less than one second, while the average calculation time of TD-DFT exceeds 200 CPU hours in the current test set.

https://doi.org/10.1038/s41467-025-58881-5

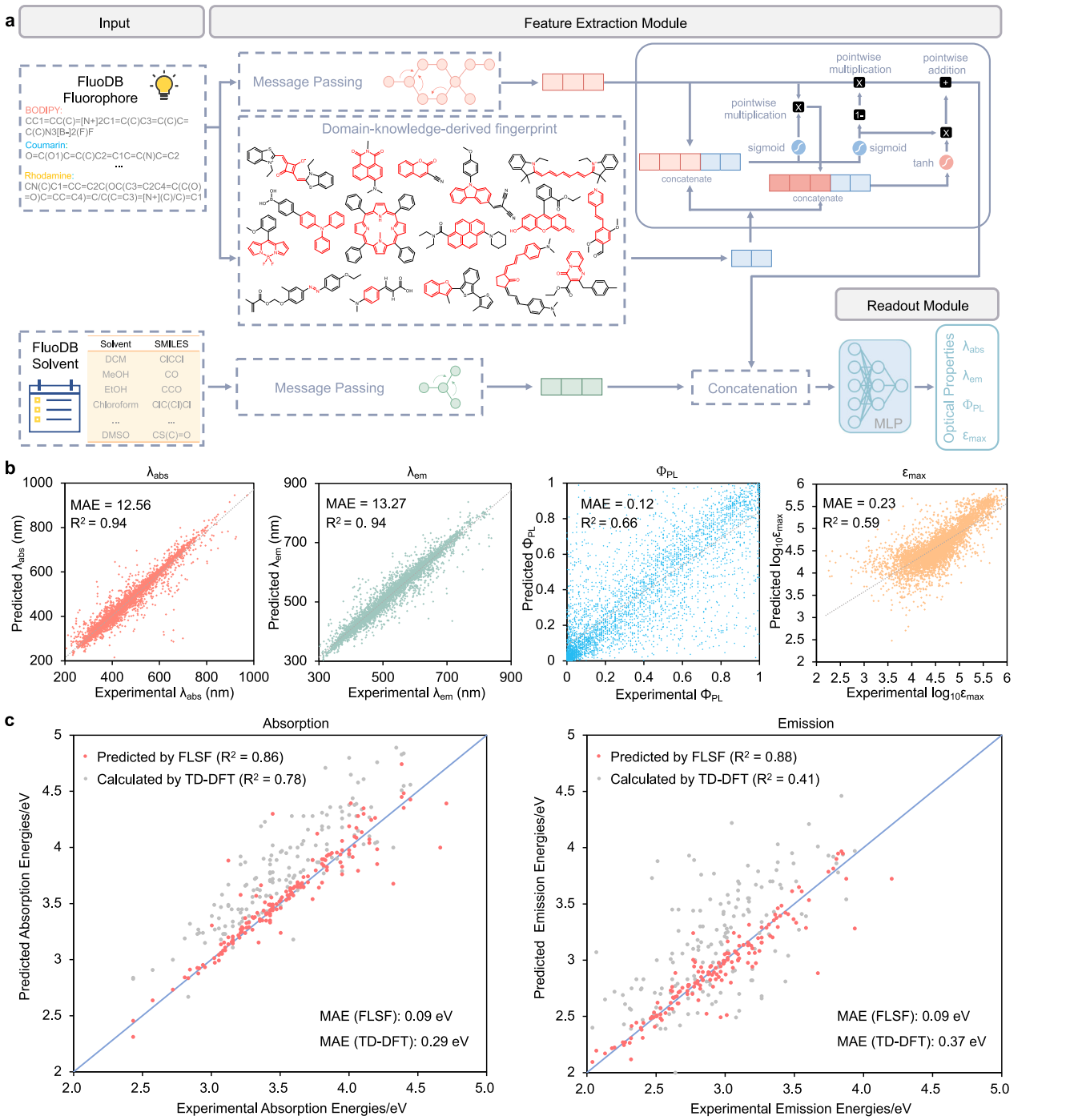

Fig. 3 | Schematic overview of our newly developed FLSF (FLuorescence prediction with fluoroScaFfold-driven model) and its prediction performance. **a** The model architecture of FLSF. A domain-knowledge-derived fingerprint based on the 728 fluorescent-scaffold subgroups (called fluoroscaffold) is fused with a message-passing neural network (MPNN) for the feature extraction of the input fluorophore. The feature extraction of the solvent molecule is based on MPNN. The feature vectors of both the fluorophore and the solvent are input together to output a prediction of the property of interest. MLP: multilayer perceptron. **b** The overall prediction performance of FLSF for different photophysical parameters. λ_{abs}: maximum absorption wavelength; λ_{em}: maximum emission wavelength; Φ_{PL}: photoluminescence quantum yield; ε_{max}: molar absorption coefficient. **c** Comparison between FLSF (red points) and TD-DFT (time-dependent density functional theory) calculations (gray points) for λ_{abs} (left) and λ_{em} (right) prediction. MAE mean absolute error, R^2 the coefficient of determination. Source data are provided as a Source data file.

Interpretability analysis of FLSF

The interpretability of a model, illustrating how it makes decisions and achieves related results, helps to verify the reliability of the model and excavate valuable information from the data. First, we analyzed the interpretability of FLSF from the molecule-level perspective[50]. The embedding vectors from three states of FLSF were studied, namely, the state only treated by D-MPNN without fluoroscaffold integration, the state with fluoroscaffold integration but before solvent incorporation, and the state after solvent incorporation. According to the 2D-PCA (two-dimensional principal component analysis) dimension

Table 2 | Performance of FLSF (FLuorescence prediction with fluoroScaFfold-driven model) in predicting four photophysical parameters towards different fluorescent scaffolds

Scaffold	MAE (Number of data)			
	λ_{abs} (nm)	λ_{em} (nm)	Φ_{PL}	ε_{max} (in $\log_{10}\varepsilon_{max}$)
Squaraine	15.43 (83)	15.52 (30)	0.17 (14)	0.24 (24)
Naphthalimide	12.36 (88)	17.46 (72)	0.21 (70)	0.21 (40)
Coumarin	10.77 (272)	13.58 (220)	0.10 (165)	0.22 (248)
Carbazole	12.59 (388)	14.77 (309)	0.11 (180)	0.21 (99)
Cyanine	11.79 (310)	16.20 (258)	0.12 (185)	0.16 (209)
BODIPY	6.44 (1412)	7.37 (1315)	0.11 (1215)	0.14 (515)
Triphenylamine	16.91 (641)	18.11 (353)	0.17 (211)	0.22 (162)
Porphyrin	29.43 (31)	11.15 (20)	0.06 (18)	0.20 (825)
PAH	11.70 (630)	16.02 (591)	0.13 (465)	0.30 (322)
Acridine	16.59 (242)	14.42 (182)	0.14 (99)	0.27 (720)
[6 + 5]	15.21 (689)	16.77 (540)	0.12 (390)	0.25 (349)
[6 + 6]	14.86 (259)	17.02 (210)	0.13 (159)	0.27 (305)
6-n-5	12.00 (361)	12.68 (282)	0.11 (214)	0.25 (233)
6-n-6	13.30 (591)	12.91 (521)	0.10 (405)	0.25 (599)
Benzene	20.42 (197)	13.97 (161)	0.12 (129)	0.34 (138)

MAE mean absolute error, λ_{abs} maximum absorption wavelength, λ_{em} maximum emission wavelength, Φ_{PL} photoluminescence quantum yield, ε_{max} molar absorption coefficient.

reduction distribution diagram (Fig. 4a), there is a clear difference in the distribution between short-wavelength and long-wavelength fluorophores in λ_{abs} and λ_{em} prediction tasks, indicating that FLSF can effectively identify their structural features with different wavelengths. Interestingly, the integration of fluoroscaffold makes this difference more significant, and the data distribution dispersion is further improved after the introduction of solvent, highlighting the importance of scaffold information for the prediction and implying that FLSF is sensitive in capturing subtle differences caused by solvents.

Subsequently, the explicability analysis of FLSF at the atom-level perspective was conducted[50]. To be specific, each atom in the fluorophore was masked, and the prediction values (e.g., λ_{em}) before and after masking were compared to reveal the attribution of each atom. Coumarin was taken as the example since it has been derived to cover a wide range of wavelengths, providing invaluable SPR information (Fig. 4b, left) for validating the reliability of FLSF[51,52]. With a classic D-π-A structure, the introduction of electron-donating groups (EDG) on the phenyl ring and electron-withdrawing groups (EWG) on the lactone ring can effectively achieve redshift of coumarin according to experimental experience. Representative examples in Fig. 4b (right) demonstrate that FLSF has grasped such rules. In addition, researchers found that the replacement of ketone with imine at position 2 can also produce redshift[53] (e.g., compound e-g), and FLSF has also mastered it. Of note, although coumarin derivatives with substitution other than oxygen at position 1 are not recorded in FluoDB, FLSF can indicate the contribution of oxygen at this position to the redshift, which is also supported by recent experimental results[54]. It implies that FLSF has good generalization ability/reliability and may provide new structural modification suggestions for fluorophore design.

Construction of FLAME for fluorophore design

While large databases and various property prediction models have significantly advanced our knowledge of the SPR of certain molecules, a gap exists in their direct applications for molecular design. Therefore, we aimed to build a multifunctional software package, FLAME, to meet the practical needs of researchers for novel fluorophore design by integrating the database, prediction models, and molecule generators into one framework (Fig. 5a). FLAME provides six open-source fluorophore databases, including FluoDB (Fig. 5b). Users can input the molecule of interest to search for related information of existing molecules in the database, as well as to train the model with different databases for illustrating the data impact. Meanwhile, FLAME offers six open-source prediction models (i.e., FLSF, UVVisML, ABT-MPNN, SchNet, SMFluo, and GBRT), which can be combined with the above datasets to meet various requirements from different users. In-parallel comparison between different combinations also helps to identify the best settings for specific parameter prediction (Table 3).

To provide structure-new compounds with predicted optical properties directly, a newly reported open-source generative AI framework, Reinvent 4[33], was introduced. As a scoring tool embedded in FLAME for molecular design, the speed of training and predicting is critical for the prediction model. FLSF proven good at these two aspects was coupled with Reinvent 4 herein (users can make their own choice). With the help of FLAME, both de novo molecular generation and structural modifications can be achieved. For example, if users are interested in the development of novel BODIPY derivatives, they can set the desired photophysical parameters (single or multiple parameters) with FLAME. Then, newly generated molecules (not recorded in FLAME's built-in database) belonging to BODIPY with predicted properties will be screened out. Alternatively, users can input a parent structure of interest with desired parameters into FLAME to obtain optimized structures. Of note, both processes used to be highly dependent on specialized knowledge and years of experience, while FLAME is promising to think out of the box and offer fluorophore candidates more efficiently.

Experimental evaluation

With increasing interest in coumarin-based fluorescent probes due to their excellent biocompatibility, good structural flexibility, and tunable fluorescence[52], FLAME is employed to guide the development of novel coumarin derivatives for concept proof (Fig. 6a and Fig. S23). Four optical parameters (λ_{abs}, λ_{em}, Φ_{PL}, and ε_{max}) were set as scoring targets. We trained the generative model and sampled one million molecules, with a focus on coumarin-type compounds during screening. From the virtual library generated by FLAME, 3,4-oxazole-fused coumarins attracted our attention due to their structural novelty and synthesizability. A variety of oxazole-containing dyes were reported to possess attractive photophysical properties, such as high fluorescence quantum yields[55–57], while the fluorescence properties of 3,4-oxazole-fused coumarins have not been reported yet.

Available strategies for the synthesis of this scaffold include (a) heating of 7-*N*,*N*-dimethylamino-4-hydroxycoumarin in the presence of nitromethane and DABCO[58], (b) synthesis from 4-hydroxy-3-nitrocoumarin and benzyl alcohol under gold nanoparticle or $FeCl_3$ catalysis[59], and (c) synthesis from 4-hydroxy-3-nitrocoumarin and acids in the presence of triphenylphosphine and phosphorus pentoxide under microwave irradiation[60] (Fig. S24). The lack of structural diversity on the phenyl ring using reported strategies, together with the demand for simple and efficient synthetic procedures to construct diverse 3,4-oxazole-fused coumarins from readily available starting materials, drives us to develop new synthetic methodology. Inspired by our previous work in isocyanide chemistry[61–63], we proposed a one-pot approach to synthesize 3,4-oxazole-fused coumarins from ethyl isocyanoacetates and phenyl salicylates promoted by base (Fig. 6b and Fig. S25). Under the optimized conditions (Table S13), 16 oxazole-fused coumarins were synthesized successfully (Figs. S26–27), carrying electron-donating or electron-withdrawing substituents on the phenyl ring. With these compounds in hand, their optical properties were evaluated (Figs. S28–29). Consistent with the prediction result from FLSF, the introduction of an amino group at the 6- or 7-position of the coumarin scaffold (i.e., **3h, 3o**) led to a redshift in λ_{abs} and an increase

https://doi.org/10.1038/s41467-025-58881-5

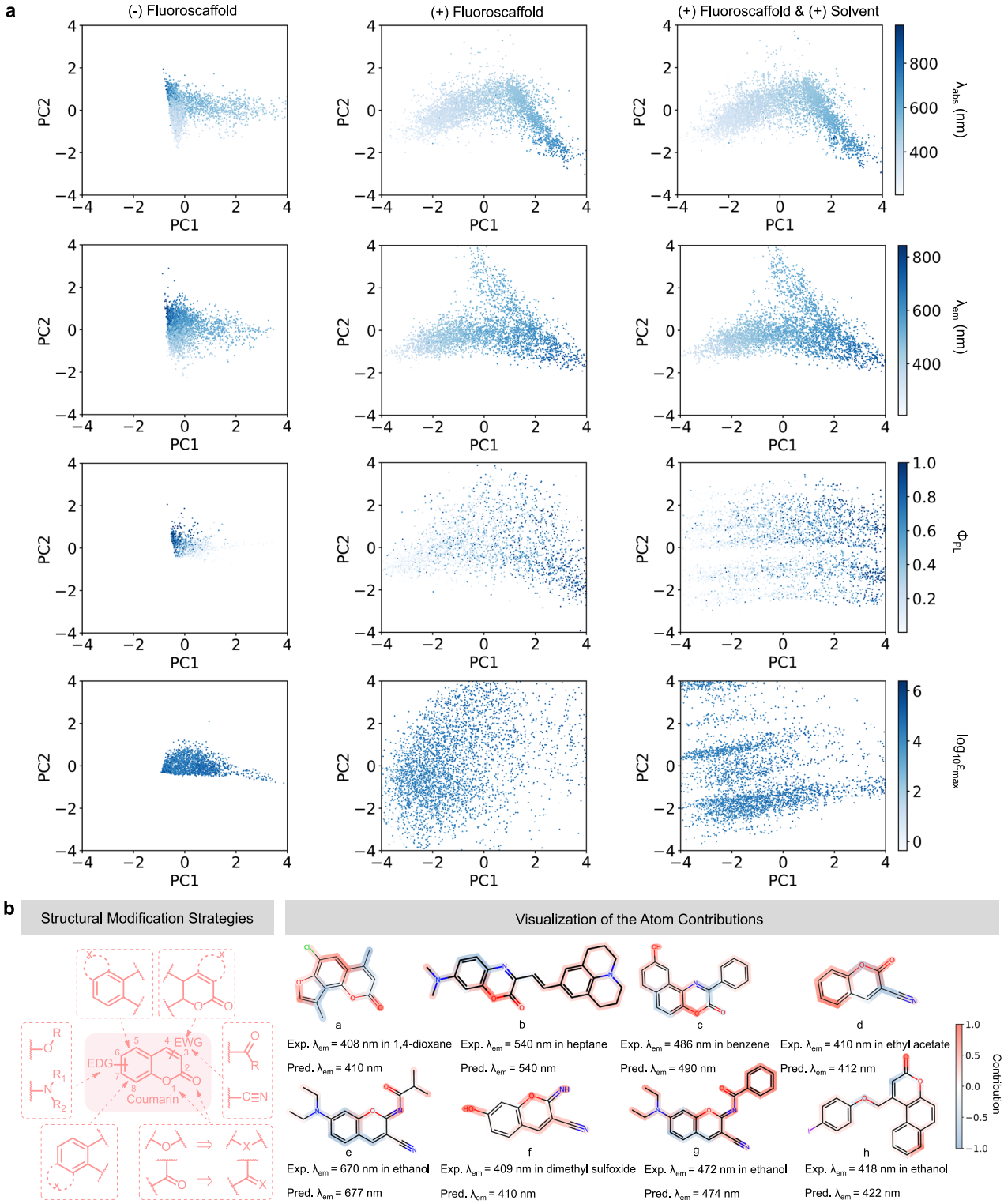

Fig. 4 | Interpretability analysis of FLSF (FLuorescence prediction with fluoroScaFfold-driven model). a FLSF embedding interpretability through PCA (Principal Component Analysis). 2D-PCA plots of the molecular embeddings at different stages: (Left) before integrating fluoroscaffold information, (Center) after integrating fluoroscaffold information, and (Right) after further integration of solvent information. Each dot is colored by experimental values. PC1: Principal Component 1; PC2: Principal Component 2. The intensity of the color scale represents the magnitude of the experimental values of the molecular parameters, with darker colors indicating higher experimental values. Source data are provided as a Source data file. λ_{abs}: maximum absorption wavelength; λ_{em}: maximum emission wavelength; Φ_{PL}: photoluminescence quantum yield; ε_{max}: molar absorption coefficient. **b** Summary of reported structural modification strategies for coumarin-based fluorophores (left) and atomic contributions learned by FLSF (right). The color bar values represent the normalized difference in the predicted values before and after masking specific atoms. EDG electron-donating group, EWG electron-withdrawing group, Exp. experimental, Pred. predicted.

https://doi.org/10.1038/s41467-025-58881-5

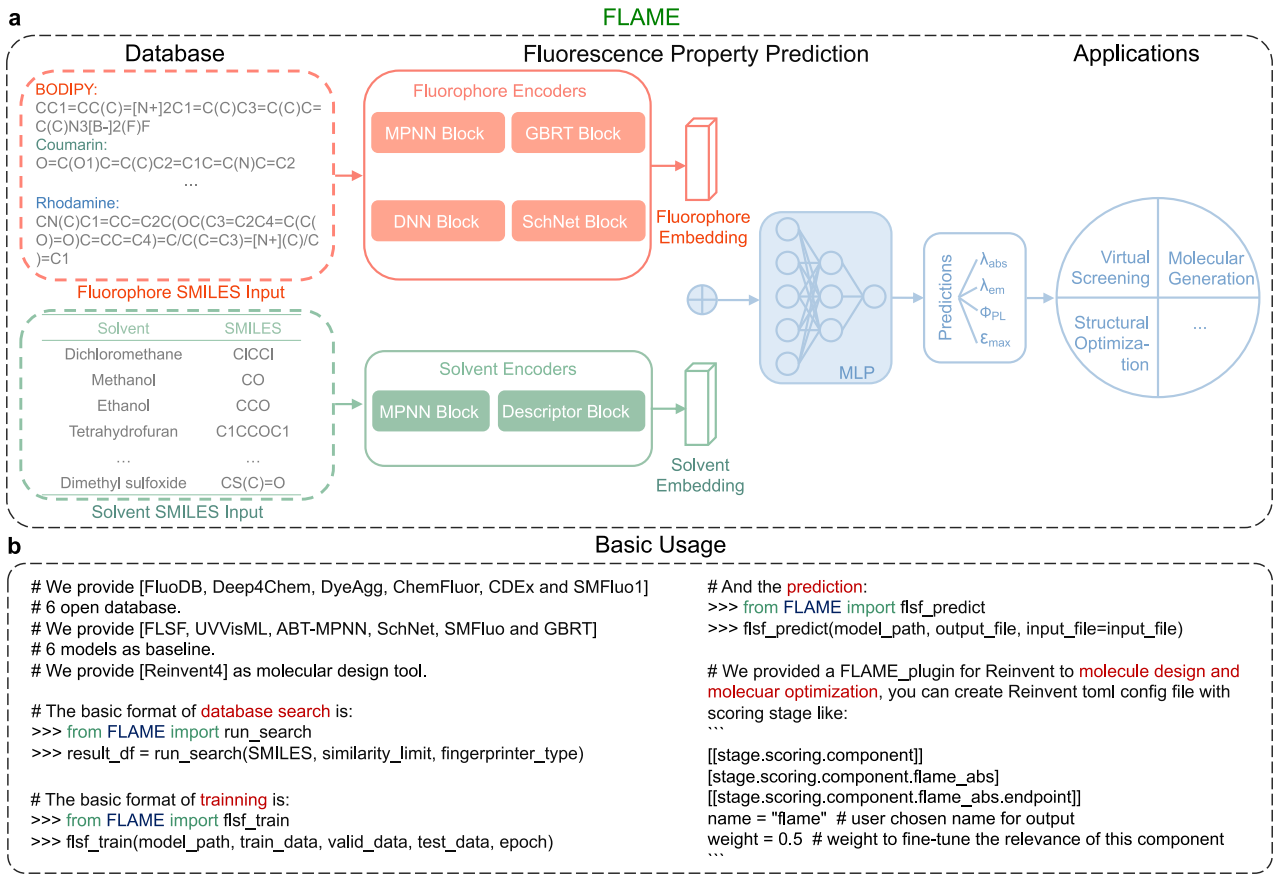

Fig. 5 | Schematic overview of FLAME (FLuorophore design Acceleration ModulE) and its usage. a The framework of FLAME to facilitate the fluorophore design. FLAME, assembled from the latest databases and prediction models, is feasible for various applications, including virtual screening, molecular generation, and structural optimization. SMILES: Simplified Molecular Input Line Entry System; MLP: multilayer perceptron; λ_{abs}: maximum absorption wavelength; λ_{em}: maximum emission wavelength; Φ_{PL}: photoluminescence quantum yield; ε_{max}: molar absorption coefficient. **b** The basic workflow of FLAME for various applications, including database search, photophysical property prediction, and creating unreported molecules with predicted optical properties by integrating different fluorophore databases, prediction models, and molecule generators. Databases including Deep4Chem[36], DyeAgg[38], ChemFluor[23], CDEx[35], SMFluo1[29], and our FluoDB; prediction models including previously reported UVVisML[45], ABT-MPNN[46], SchNet[26], SMFluo[29], GBRT[23], and our FLSF (FLuorescence prediction with fluoroScaFfold-driven model).

in Φ_{PL} (Table S14). Then, solvent effects on these two fluorophores were investigated (Fig. 6c and Table S15). As expected, the solvent polarity has a significant impact on their absorption/emission wavelengths, and **3o** liberated much stronger emission than **3h** in all solvents, which was selected for bioimaging. Brilliant fluorescence was observed in HeLa cells after 30-min incubation (Fig. 6d), indicating its potential for live-cell imaging.

Discussion

FLAME, a modular AI-assisted framework for fluorophore design, was developed to help researchers design de novo molecules with desired optical performance efficiently. To achieve this, we expanded the available fluorophore database by supplementing data from various aspects, such as fluorescent scaffold types and photophysical parameters, to give the biggest open-source fluorophore database to date, FluoDB, which contains 55,169 solvated fluorophores and 109,054 data entries including four key photophysical parameters, λ_{abs}, λ_{em}, Φ_{PL}, and ε_{max}. Compared with reported databases, FluoDB exhibits higher molecular diversity and data volume. By conducting a series of data analyses on FluoDB, insights were gained into the correlation between different photophysical parameters, as well as their relationships with molecular weight and solvent type.

To meet the requirement of scoring for molecular generation, the prediction model needs to be accurate and fast. FLSF with a domain-knowledge-derived fingerprint for characterizing fluorescent scaffolds (called fluoroscaffold: a 728-dimensional digital fingerprint) was designed and exhibited encouraging accuracy with a training speed 10 times faster than ABT-MPNN. In addition, FLSF's predictive power was tested through a series of molecule-level and atom-level interpretability analyses. The attribution of each atom learned by FLSF is highly consistent with expertise. Based on FLSF, Reinvent 4 as a molecule generator was employed for de novo creation of fluorophore candidates. A series of 3,4-oxazole-fused coumarins yet-to-be-developed for fluorophores were synthesized using our newly developed metal-free approach via base-promoted tandem reaction of phenyl salicylates with isocyanoacetates. The predicted optical performance of these compounds is highly consistent with the experimental results (MAE = 13.3 nm for λ_{abs}, 0.093 for Φ_{PL}, and 0.430 for $\log_{10}\varepsilon_{max}$), and an unreported coumarin derivative with brilliant fluorescence (Φ_{PL} = 0.541, $\log_{10}\varepsilon_{max}$ = 4.314 in water) promising for bioimaging was obtained.

The above results exemplify the advance of FLAME in facilitating the design of new fluorophores, which can reduce the burden on trial-and-error experiments by simply inputting the desired photophysical

Table 3 | Direct comparison of the prediction performance from different combinations of databases and prediction models via FLAME (FLuorophore design Acceleration ModulE)

Dataset 1: FluoDB				
Object	**Algorithms**	**MAE**	**MSE**	**RMSE**
λ_{abs}	GBRT	13.67	824.24	28.71
	SMFluo	21.19	1255.71	35.44
	UVVisML	13.94	716.91	26.78
	SchNet	22.17	1684.74	41.05
	ABT-MPNN	12.66	687.97	26.23
	FLSF	12.56	675.34	25.99
Dataset 2: Deep4Chem				
λ_{abs}	GBRT	24.97	1972.37	44.41
	SMFluo	34.2	2992.06	54.7
	UVVisML	24.59	1833.93	42.82
	SchNet	23.68	2071.75	45.52
	ABT-MPNN	22.07	1614.01	40.17
	FLSF	24.26	1930.89	43.94

MAE mean absolute error, *MSE* mean-square error, *RMSE* root-mean-square error, databsets including our FluoDB and Deep4Chem[36]; prediction models including previously reported GBRT[23], SMFluo[29], UVVisML[45], SchNet[26], ABT-MPNN[46], and our FLSF (FLuorescence prediction with fluoroScaFfold-driven model); λ_{abs}: maximum absorption wavelength. The unit of MAE/RMSE is nm for λ_{abs}.

parameters into the black box of FLAME. Multi-step computational processes can thus be executed automatically and can be handled by anyone without a prerequisite for expertise in either fluorescence or computation. With its modular architecture, FLAME can be further updated with new data/algorithms to advance with time in accelerating fluorophore development. Moreover, synthetic accessibility predicting models[64–68] can be further integrated into FLAME for synthesizability scoring during the sampling, and retrosynthetic analysis tools (e.g., AiZynthFinder[69], Retro*[70], and ASKCOS[71]) can also be incorporated into FLAME which can help with retrosynthesis planning towards the fluorophore candidate, making fluorophore design and synthesis more efficiently.

Methods

Data collection

In addition to searching literature via PubMed by using fluorescent scaffolds as keywords, we also compiled and supplemented multiple open-source databases, including Deep4Chem[36], ChemFluor[23], Dye Aggregation (DyeAgg)[38], ChemDataExtrator (CDEx)[35], DYES[39], PhotochemCAD[40], and Dye-Sensitized Solar Cell Database (DSSCDB)[37]. FluoDB is currently available for the experimental photophysical data, including maximum absorption wavelength (λ_{abs}), maximum emission wavelength (λ_{em}), photoluminescence quantum yield (Φ_{PL}), and molar absorption coefficient (ε_{max}), which are key factors in photochemical studies. During the data collection, if multiple peaks were found in the absorption/emission spectra for the same fluorophore-solvent pair, the peak with the longest wavelength/largest intensity was collected for λ_{abs} and λ_{em}. The majority of the experimental values were obtained at 298 K, so the effect of temperature was not considered in the model development.

Data processing

First, we removed the invalid data: (1) remove data without solvent information or with gas as the solvent; (2) remove data whose SMILES cannot be converted to valid chemical structures. Then we did some general processing to limit the data range of each photophysical parameter: (1) remove data with Φ_{PL} above 1; (2) remove data with λ_{abs} or λ_{em} below 200 nm or above 1500 nm; (3) remove data with ε_{max} above ten million. During the redundant data processing, we set difference thresholds for each parameter. The difference threshold is 5 nm for λ_{abs} and λ_{em}, 0.1 for Φ_{PL}, and 0.02 for $\log_{10}\varepsilon_{max}$. For each fluorophore-solvent pair, redundant data from different resources were removed if exceeding the difference threshold, and the average value of the remaining data was put into the database.

Finally, all fluorophores were standardized using SMILES notation, and a dictionary mapping to convert solvent names and acronyms into SMILES was constructed. Furthermore, the number of solvent types was streamlined from 393 to 72 by removing those that occurred less than 10 times. Since these less-used solvents account for a small portion of the original data (~2000 entries), their removal will not affect the data diversity.

Model training and evaluation

We trained and tested four optical property parameters separately. Data containing fluorophores in the mixture/complex form (containing water, metal ions, etc. in the SMILES) were removed from FluoDB (6309 entries were removed) to give FluoDB-Lite (the original FluoDB is also named FluoDB-Full to differentiate it from FluoDB-Lite when applicable), before a random split with a ratio of 7:1:2 (detailed in Table S5). Of note, some implausible data ($\varepsilon_{max} < 100$) were removed from the test set during ε_{max} prediction (Table S6). The hyperparameters for FLSF were tuned by Bayesian optimization on the validation set. All regression models are evaluated by MAE (mean absolute error), MSE (mean-square error), and RMSE (root-mean-square error). The training was conducted on two servers—OCHPC and SYHPC. The OCHPC server has 2 Intel Skylake Gold 6132 processors and 192GB RAM, along with an NVIDIA Tesla K80 24GB GPU. The SYHPC server has 4 Intel 8360H processors, 3TB RAM, and an NVIDIA A100-40GB GPU. The hyperparameters for FLSF are shown in Table S16.

TD-DFT calculations

The molecules used for the TD-DFT (time-dependent density functional theory) tests were sourced from the previously divided test set. We selected data that contained all four parameters, excluded molecules containing ions, and restricted the number of heavy atoms to less than 30. Initial geometries were refined using semi-empirical tight-binding density functional theory (GFN2-xTB) followed by geometry optimizations at the B3LYP/6-31 + G(d)/IEFPCM level of theory in the Gaussian 16 software package. TD-DFT calculations[72] were performed with CAM-B3LYP/6-31 + G(d)/IEFPCM level of theory.

General information for chemical synthesis

^{1}H, ^{13}C, ^{19}F NMR spectra were recorded using JNM-ECZ 400S (400 MHz) spectrometer. Chemical shifts were reported in parts per million (ppm), and the residual solvent peak was used as an internal reference: ^{1}H (chloroform δ 7.26; DMSO δ 2.50), ^{13}C (chloroform δ 77.16; DMSO δ 39.52). Data are reported as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, br = broad, dd = doublet of doublets, ddd = doublet of doublet of doublets), coupling constants (Hz) and integration. Melting point (MP) was obtained on Buchi M-560. For thin layer chromatography (TLC), Huanghai TLC plates (HSGF 254) were used, and compounds were visualized with a UV light at 254 nm. High-resolution mass spectra (HRMS) were obtained on an Agilent G6545 spectrometer using an electron spray ionization time-of-flight (ESI-TOF) source. Unless otherwise noted, all reactions were carried out under an ambient atmosphere; exclusion of air or moisture was not required. Anhydrous and deuterated solvents were purchased from commercial suppliers and used as received without further purification. Phenyl salicylates **1a-1p** (Fig. S26) were prepared according to literature[73]. Ethyl isocyanoacetate (2) was purchased from commercial suppliers and used without further purification.

https://doi.org/10.1038/s41467-025-58881-5

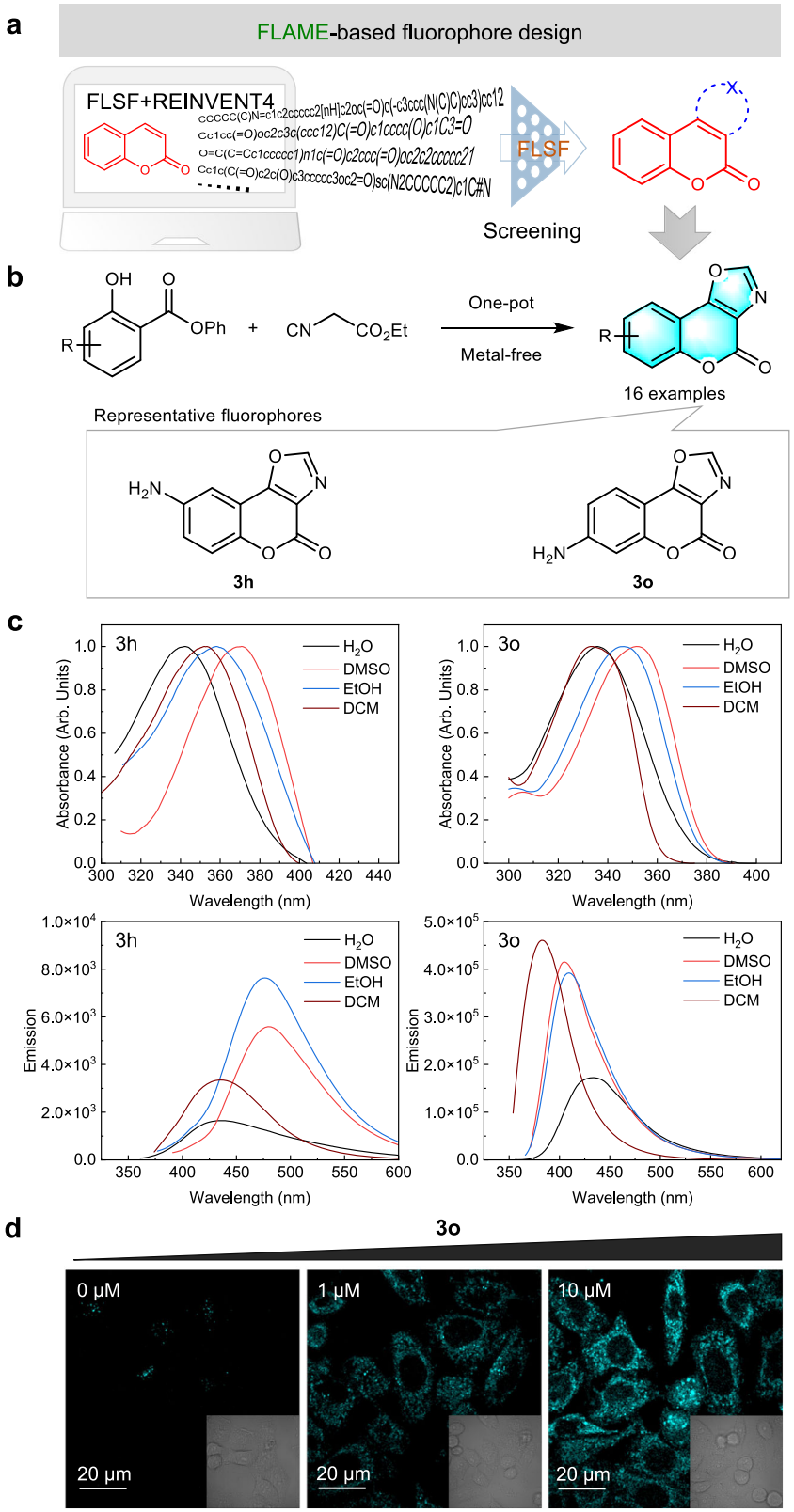

Fig. 6 | Development of new fluorophores assisted by FLAME (FLuorophore design Acceleration ModulE). a Generation of unreported heterocyclic-fused coumarins with predicted optical properties by FLAME. FLSF: FLuorescence prediction with fluoroScaFfold-driven model; Reinvent 4[33]: a newly reported open-source generative AI framework. **b** The new strategy for one-pot synthesis of 3,4-oxazole-fused coumarins. **c** Absorption (up) and emission (down) spectra of **3h** and **3o** recorded in different solvents. H_2O: water; DMSO: dimethyl sulfoxide; EtOH: ethanol; DCM: dichloromethane. Source data are provided as a Source data file. **d** Confocal fluorescence images of living HeLa cells treated with different concentrations of **3o**. The cell imaging was performed three times with similar results.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets and prediction results are available at Figshare (https://doi.org/10.6084/m9.figshare.26317933)[74]. All data generated in this study are provided in the Source data file. Source data are provided with this paper.

Code availability

All codes used in this study are available at Zenodo (https://zenodo.org/records/14842448)[75] and GitHub (https://github.com/ChemloverYuchen/FLAME).

References

1. Chen, Y., Wang, S. & Zhang, F. Near-infrared luminescence high-contrast in vivo biomedical imaging. *Nat. Rev. Bioeng.* **1**, 60–78 (2023).
2. Bryden, M. A. & Zysman-Colman, E. Organic thermally activated delayed fluorescence (TADF) compounds used in photocatalysis. *Chem. Soc. Rev.* **50**, 7587–7680 (2021).
3. Kumar, R. et al. Revisiting fluorescent calixarenes: from molecular sensors to smart materials. *Chem. Rev.* **119**, 9657–9721 (2019).
4. Wang, K. et al. Fluorescence image-guided tumour surgery. *Nat. Rev. Bioeng.* **1**, 161–179 (2023).
5. Wu, L., Huang, J., Pu, K. & James, T. D. Dual-locked spectroscopic probes for sensing and therapy. *Nat. Rev. Chem.* **5**, 406–421 (2021).
6. Grimm, J. & Lavis, L. D. Caveat fluorophore: an insiders' guide to small-molecule fluorescent labels. *Nat. Methods* **19**, 149–158 (2022).
7. Hong, G., Antaris, A. L. & Dai, H. Near-infrared fluorophores for biomedical imaging. *Nat. Biomed. Eng.* **1**, 0010 (2017).
8. Remmel, A. How to keep the lights on: the mission to make more photostable fluorophores. *Nature* **630**, 258–260 (2024).
9. Wang, S. et al. Anti-quenching NIR-II molecular fluorophores for in vivo high-contrast imaging and pH sensing. *Nat. Commun.* **10**, 1058 (2019).
10. Wang, C. et al. Twisted intramolecular charge transfer (TICT) and twists beyond TICT: from mechanisms to rational designs of bright and sensitive fluorophores. *Chem. Soc. Rev.* **50**, 12656 (2021).
11. Yan, K. et al. Ultra-photostable small-molecule dyes facilitate near-infrared biophotonics. *Nat. Commun.* **15**, 2593 (2024).
12. Zhou, J., Ren, T.-B. & Yuan, L. The strategy to improve the brightness of organic small-molecule fluorescent dyes for imaging. *Chin. Chem. Lett.* https://doi.org/10.1016/j.cclet.2024.110644 (2024).
13. Lovell, T. C., Branchaud, B. P. & Jasti, R. An organic chemist's guide to fluorophores–understanding common and newer non-planar fluorescent molecules for biological applications. *Eur. J. Org. Chem.* **27**, e202301196 (2024).
14. Cavazos-Elizondo, D. & Aguirre-Soto, A. Photophysical properties of fluorescent labels: a meta-analysis to guide probe selection amidst challenges with available data. *Anal. Sens.* **2**, e202200004 (2022).
15. Jiang, G. et al. Chemical approaches to optimize the properties of organic fluorophores for imaging and sensing. *Angew. Chem. Int. Ed.* **63**, e202315217 (2024).
16. Puszkarska, A. M. et al. Machine learning designs new GCGR/GLP-1R dual agonists with enhanced biological potency. *Nat. Chem.* **16**, 1436–1444 (2024).
17. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
18. Du, Y. et al. Machine learning-aided generative molecular design. *Nat. Mach. Intell.* **6**, 589–604 (2024).
19. Lewis, L. et al. Improved machine learning algorithm for predicting ground state properties. *Nat. Commun.* **15**, 895 (2024).
20. Gentile, F. et al. Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672–697 (2022).
21. Zhang, O. et al. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nat. Mach. Intell.* **5**, 1020–1030 (2023).
22. Sumita, M. et al. De novo creation of a naked eye-detectable fluorescent molecule based on quantum chemical computation and machine learning. *Sci. Adv.* **8**, eabj3906 (2022).
23. Ju, C. W., Bai, H., Li, B. & Liu, R. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: emission wavelengths and quantum yields. *J. Chem. Inf. Model.* **61**, 1053–1065 (2021).
24. Joung, J. F. et al. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. *JACS Au* **1**, 427–438 (2021).
25. Joung, J. F., Han, M., Jeong, M. & Park, S. Beyond Woodward–Fieser rules: design principles of property-oriented chromophores based on explainable deep learning optical spectroscopy. *J. Chem. Inf. Model.* **62**, 2933–2942 (2022).
26. Hung, S.-H., Ye, Z.-R., Cheng, C.-F., Chen, B. & Tsai, M.-K. Enhanced predictions for the experimental photophysical data using the featurized Schnet-bondstep approach. *J. Chem. Theory Comput.* **19**, 4559–4567 (2023).
27. Qian, L., Li, L. & Yao, S. Q. Two-photon small molecule enzymatic probes. *Acc. Chem. Res.* **49**, 626–634 (2016).
28. Wang, W. et al. Real-time imaging of cell-surface proteins with antibody-based fluorogenic probes. *Chem. Sci.* **12**, 13477–13482 (2021).
29. Shao, J. et al. Prediction of maximum absorption wavelength using deep neural networks. *J. Chem. Inf. Model.* **62**, 1368–1375 (2022).
30. Koscher, B. A. et al. Autonomous, multi-property-driven molecular discovery: From predictions to measurements and back. *Science* **382**, eadi1407 (2023).
31. Han, M., Joung, J. F., Jeong, M., Choi, D. H. & Park, S. Generative deep learning-based efficient design of organic molecules with tailored properties. *ACS Cent. Sci.* https://doi.org/10.1021/acscentsci.4c00656 (2024).
32. McNaughton, A. D. et al. Machine learning models for predicting molecular UV–Vis spectra with quantum mechanical properties. *J. Chem. Inf. Model.* **63**, 1462–1471 (2023).
33. Loeffler, H. H. et al. Reinvent 4: Modern AI-driven generative molecule design. *J. Cheminform.* **16**, 20 (2024).
34. Xiong, Z. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2020).
35. Beard, E. J., Sivaraman, G., Vazquez-Mayagoitia, A., Vishwanath, V. & Cole, J. M. Comparative dataset of experimental and computational attributes of UV/Vis absorption spectra. *Sci. Data* **6**, 307 (2019).
36. Joung, J. F., Han, M., Jeong, M. & Park, S. Experimental database of optical properties of organic compounds. *Sci. Data* **7**, 295 (2020).
37. Venkatraman, V., Raju, R., Oikonomopoulos, S. P. & Alsberg, B. K. The dye-sensitized solar cell database. *J. Cheminform.* **10**, 18 (2018).
38. Venkatraman, V. & Chellappan, L. K. An open access data set highlighting aggregation of dyes on metal oxides. *Data* **5**, 45 (2020).
39. Ksenofontov, A. A., Lukanov, M. M. & Bocharov, P. S. Can machine learning methods accurately predict the molar absorption coefficient of different classes of dyes? *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **279**, 121442 (2022).
40. Taniguchi, M. & Lindsey, J. S. Database of absorption and fluorescence spectra of >300 common compounds for use in PhotochemCAD. *Photochem. Photobiol.* **94**, 290–327 (2018).
41. Wang, S., Li, B. & Zhang, F. Molecular fluorophores for deep-tissue bioimaging. *ACS Cent. Sci.* **6**, 1302–1316 (2020).

42. Klymchenko, A. S. Solvatochromic and fluorogenic dyes as environment-sensitive probes: design and biological applications. *Acc. Chem. Res.* **50**, 366–375 (2017).
43. Resch-Genger, U., Grabolle, M., Cavaliere-Jaricot, S., Nitschke, R. & Nann, T. Quantum dots versus organic dyes as fluorescent labels. *Nat. Methods* **5**, 763–775 (2008).
44. Würth, C., Geißler, D., Behnke, T., Kaiser, M. & Resch-Genger, U. Critical review of the determination of photoluminescence quantum yields of luminescent reporters. *Anal. Bioanal. Chem.* **407**, 59–78 (2015).
45. Greenman, K. P., Green, W. H. & Gomez-Bombarelli, R. UVVisML (0.0.2). *Zenodo* https://doi.org/10.5281/zenodo.5986671 (2022).
46. Liu, C., Sun, Y., Davis, R., Cardona, S. T. & Hu, P. ABT-MPNN: an atom-bond transformer-based message-passing neural network for molecular property prediction. *J. Cheminform.* **15**, 29 (2023).
47. Adamo, C. & Jacquemin, D. The calculations of excited-state properties with time-dependent density functional theory. *Chem. Soc. Rev.* **42**, 845–856 (2013).
48. Rubesova, M., Muchova, E. & Slavicek, P. Optimal tuning of range-separated hybrids for solvated molecules with time-dependent density functional theory. *J. Chem. Theory Comput.* **13**, 4972–4983 (2017).
49. Charaf-Eddin, A., Le Guennic, B. & Jacquemin, D. Excited-states of BODIPY–cyanines: ultimate TD-DFT challenges? *RSC Adv.* **4**, 49449–49456 (2014).
50. Wu, J. et al. ALipSol: An attention-driven mixture-of-experts model for lipophilicity and solubility prediction. *J. Chem. Inf. Model.* **62**, 5975–5987 (2022).
51. Sharma, S. J. & Sekar, N. Deep-red/NIR emitting coumarin derivatives—synthesis, photophysical properties, and biological applications. *Dyes Pigm.* **202**, 110306 (2022).
52. Cao, D. et al. Coumarin-based small-molecule fluorescent chemosensors. *Chem. Rev.* **119**, 10403–10519 (2019).
53. Rabahi, A. et al. Synthesis and optical properties of coumarins and iminocoumarins: Estimation of ground- and excited-state dipole moments from a solvatochromic shift and theoretical methods. *J. Mol. Liq.* **195**, 240–247 (2014).
54. Matikonda, S. S., Ivanic, J., Gomez, M., Hammersley, G. & Schnermann, M. J. Core remodeling leads to long wavelength fluorocoumarins. *Chem. Sci.* **11**, 7302–7307 (2020).
55. Takechi, H., Oda, Y., Nishizono, N., Oda, K. & Machida, M. Screening search for organic fluorophores: syntheses and fluorescence properties of 3-azolyl-7-diethylaminocoumarin derivatives. *Chem. Pharm. Bull.* **48**, 1702–1710 (2000).
56. Mahuteau-Betzer, F. & Piguel, S. Synthesis and evaluation of photophysical properties of series of π-conjugated oxazole dyes. *Tetrahedron Lett.* **54**, 3188–3193 (2013).
57. Xing, Z.-H. et al. Novel oxazole-based emitters for high efficiency fluorescent OLEDs: synthesis, characterization, and optoelectronic properties. *Tetrahedron* **73**, 2036–2042 (2017).
58. Satyanarayana, I., Manjappa, K. B. & Yang, D.-Y. Nitromethane as a surrogate cyanating agent: 7-N,N-dimethylamino-4-hydroxycoumarin-catalyzed, metal-free synthesis of α-iminonitriles. *Green. Chem.* **22**, 8316–8322 (2020).
59. Vlachou, E.-E. N., Armatas, G. S. & Litinas, K. E. Synthesis of fused oxazolocoumarins from o-hydroxynitrocoumarins and benzyl alcohol under gold nanoparticles or $FeCl_3$ catalysis. *J. Heterocycl. Chem.* **54**, 2447–2453 (2017).
60. Balalas, T. D. et al. One-pot synthesis of 2-substituted 4H-chromeno[3,4-d]oxazol-4-ones from 4-hydroxy-3-nitrocoumarin and acids in the presence of triphenylphosphine and phosphorus pentoxide under microwave irradiation. *SynOpen* **2**, 105–113 (2018).
61. Qian, L. et al. Catalytic atroposelective dynamic kinetic resolution of biaryl lactones with activated isocyanides. *Org. Lett.* **23**, 5086–5091 (2021).
62. Tao, L.-F. et al. Diastereo- and enantioselective silver-catalyzed [3 + 3] cycloaddition and kinetic resolution of azomethine imines with activated isocyanides. *Angew. Chem. Int. Ed.* **61**, e202202679 (2022).
63. Luo, Z.-H. et al. Torsional strain-independent catalytic enantioselective synthesis of biaryl atropisomers. *Angew. Chem. Int. Ed.* **61**, e202211303 (2022).
64. Yu, J. et al. Organic compound synthetic accessibility prediction based on the graph attention mechanism. *J. Chem. Inf. Model.* **62**, 2973–2986 (2022).
65. Chen, S. & Jung, Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *J. Cheminform.* **16**, 83 (2024).
66. Seo, S., Lim, J. & Kim, W. Y. Molecular generative model via retrosynthetically prepared chemical building block assembly. *Adv. Sci.* **10**, 2206674 (2023).
67. Gao, W., Mercado, R. & Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. In *The Tenth International Conference on Learning Representations*. (ICLR, 2022).
68. Guo, J. & Schwaller, P. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. *Chem. Sci.* https://doi.org/10.1039/d5sc01476j (2025).
69. Saigiridharan, L. et al. AiZynthFinder 4.0: developments based on learnings from 3 years of industrial application. *J. Cheminform.* **16**, 57 (2024).
70. Chen, B., Li, C., Dai, H. & Song, L. Retro*: learning retrosynthetic planning with neural guided A* search. https://doi.org/10.48550/arXiv.2006.15820 (2024).
71. Tu, Z. et al. ASKCOS: an open source software suite for synthesis planning. Preprint at https://arxiv.org/abs/2501.01835 (2025).
72. Runge, E. & Gross, E. K. U. Density-functional theory for time-dependent systems. *Phys. Rev. Lett.* **52**, 997–1000 (1984).
73. Serratore, N. A. et al. Integrating metal-catalyzed C-H and C-O functionalization to achieve sterically controlled regioselectivity in arene acylation. *J. Am. Chem. Soc.* **140**, 10025–10033 (2018).
74. Zhu, Y. et al. A modular artificial intelligence framework to facilitate fluorophore design. *Figshare* https://doi.org/10.6084/m9.figshare.26317933 (2025).
75. Zhu, Y. ChemloverYuchen/FLAME: FLAME-1.0. *Zenodo* https://doi.org/10.5281/zenodo.14842448 (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82473881 and 82273887 to L.Q.), the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (LTZ22B020001 to J.L.), Key Research and Development Program of Zhejiang Province (2025C02087 to Z.M.), and Zhejiang University. We also thank Prof. Tingjun Hou for providing the server for computation and Prof. Chang-Yu Hsieh for helpful discussions throughout this work. The computational calculations were also supported by the HPC Center of Zhejiang University (Zhoushan Campus) and the High-performance Computing Platform of YZBSTCACC.

Author contributions

Y.Z. and J.F. contributed equally to this work. Y.Z. and J.F. performed the data collection, data processing, model development, figure preparation, and manuscript writing. Y.Z. and S.A.H.A. carried out the organic synthesis and experimental measurements. T.Z. guided the TD-DFT analysis. J.-Y.L., Z.M., and L.Q. designed the project and guided the experiments. L.Q. wrote the manuscript. T.Z., S.Z., J.-Y.L., and Z.M. revised the manuscript. All authors reviewed and had final approval of the submitted paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58881-5.

Correspondence and requests for materials should be addressed to Jia-Yu Liao, Zhongjun Ma or Linghui Qian.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.