Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields

Cheng-Wei Ju [1] †, Hanzhi Bai [2‡], Bo Li [3‡], Rizhang Liu [4‡]*

[1] College of Chemistry, Nankai University, Tianjin 300071, China.

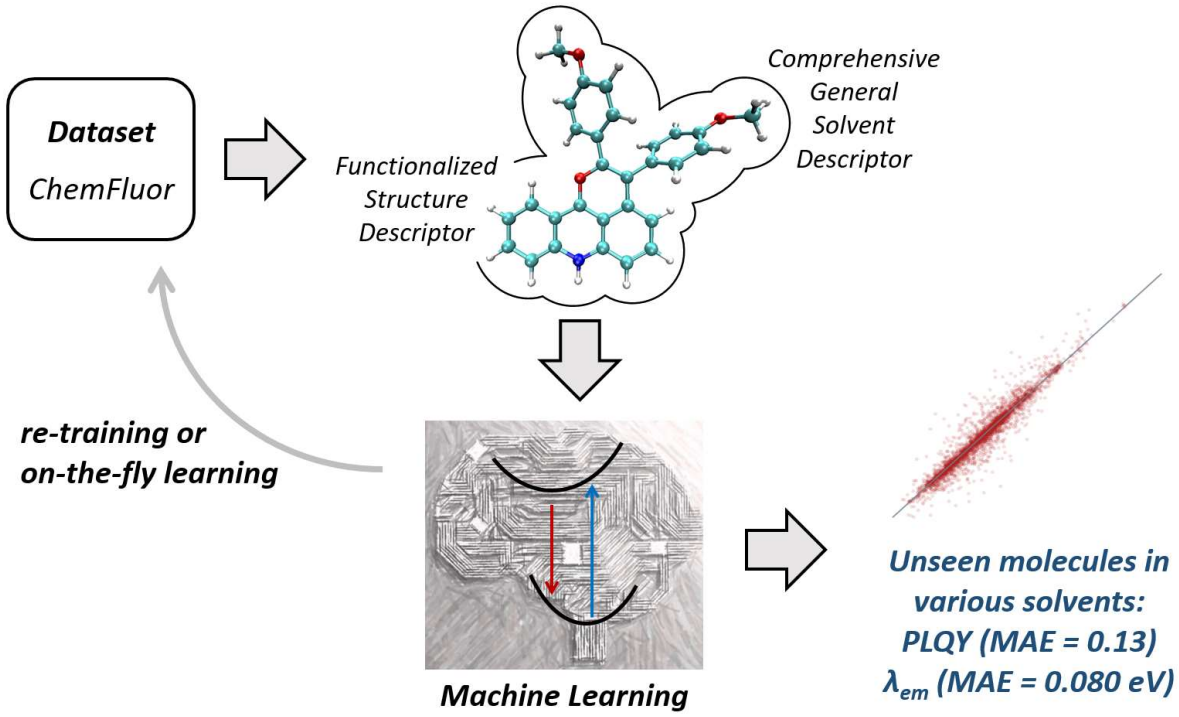
[2] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.

[3] Department of Chemistry, College of Science, Tianjin University, Tianjin 300072, China.

[4] College of Software Engineering, Sichuan University, Chengdu, Sichuan, 610064, China.

Keywords: Photoluminescence Quantum Yield • Machine Learning • Excited States • Chromophores • Emission Wavelengths

SYNOPSIS



BRIEFS: Machine learning has for the first time overcome the holistic and rapid prediction of PLQY of solvated organic molecules in various solvents with a mean absolute error of 0.13. Multi-faceted investigations with TD-DFT have proved that machine learning can achieve comparable accuracy and far superior speed. Retraining makes learning on-the-fly become possible, prove the generalizability of the model.

Abstract

The development of functional organic fluorescent materials calls for fast and accurate predictions of photophysical parameters for processes such as high-throughput virtual screening, while the task is challenged by the limitations of quantum mechanical calculations. We establish a database covering >4,300 solvated organic fluorescent dyes and develop new machine learning (ML) approach aimed at efficient and accurate predictions of emission wavelength and photoluminescence quantum yield (PLQY). Our feature engineering has given rise to Functionalized Structure Descriptor (FSD) and Comprehensive General Solvent Descriptor (CGSD), whereby a highly black-box computational framework is realized with consistently good accuracy across different dye families, ability of describing substitution effects and solvent effects, efficiency for large-scale predictions and workability with on-the-fly learning. Evaluations with unseen molecules suggests a remarkable MAE of 0.13 for PLQY and 0.080 eV for emission energy, the latter comparable to time-dependent density functional theory (TD-DFT) calculations. An online prediction platform was constructed based on the ensemble model to make prediction in various solvents. Our statistical learning methodology will complement quantum mechanical calculations as an efficient alternative approach for the prediction of these parameters.

Introduction

Organic fluorescent materials, especially small-molecule organic fluorescent dyes, have been used extensively not only as useful tools in biological research [1-4] but also as vital elements in material science[5-10]. The last decades have seen the development of novel fluorescence-based applications including electrically pumped organic laser (EPOL)[11] ,stimulated emission depletion (STED) microscopy[12, 13], thermally activated delayed fluorescence (TADF) organic light-emitting diode (OLED)[14, 15], etc., attracting great attention to the rationale design of organic materials with high photoluminescence quantum yields (PLQY, Φ_{PL}) and precisely controlled maximum absorption and/or emission wavelengths (λ_{abs}, λ_{em})[16, 17]. Nevertheless, it remains a great challenge to predict Φ_{PL} with first-principle calculations[18-20]. The high expense of excited-state calculations, combined with the involved interplay between radiative and non-radiative processes, has made it exceedingly costly to fully explore excited-state potential energy surfaces (PESs) without prior information[21, 22]. The situation is further compounded by the involvement of triplet excited states via intersystem crossing, whose modelling relies on accurate singlet-triplet gaps, spin-orbit coupling (SOC) strengths, etc[23]. For solvated organic dyes, the modelling of solvent, implicitly or explicitly, can lead to an array of additional issues for solvent response, hydrogen bonding effects, etc., albeit dramatic dependence of Φ_{PL} on solvent is not any rare phenomenon[24]. As a neat quantum mechanical treatment, the thermal vibrational correlation function (TVCF) formalism developed by Shuai *et al.* has been applied to the Φ_{PL} predictions of bodipy [25] and rationalization of aggregation induced emission (AIE) [18, 26] and room-temperature phosphorescence (RTP) [27, 28]. However, successful TVCF calculations are still within a limited scope, and the efficiency is far from satisfactory for large-scale screening. Despite the efforts by Van Voorhis *et al*. that develop semi-empirical methods to improve the efficiency (and accuracy) of Φ_{PL} predictions, the specific backbones and moderate accuracy (MAE ≈ 0.2) again reflects the great challenge of Φ_{PL} predictions [23, 29]. Besides, the application of all these approaches requires details about the major photophysical processes. For the high-throughput screening of organic materials with high Φ_{PL}, it is still strongly desired to develop a black-box computational framework with simple inputs, no requirement for pre-knowledge in photophysical processes, consistently good accuracy across different dye families, capability of describing substitution effects and necessary solvent effects, and efficiency for large-scale predictions.

Although the predictions of λ_{abs} and λ_{em} are much more tractable than Φ_{PL}, the commonly adopted computational methods, especially linear response time-dependent density functional theory (TD-DFT), is still in urgent need of

improvement in various aspects[30-32]. Seemingly a black-box method, the level of theory being used can have serious influences on the performance of TD-DFT[33]. In particular, the percentage of Hartree-Fock exchange greatly affects the description of charge-transfer excited states, the prediction of excited-state geometries (as in biaryl compounds), and, in many cases, the systematic overestimation of excitation/emission energies[34, 35]. Strategies like optimal tuning can partly alleviate these issues, but also creating a notable increase in computational expense that is inviable in the context large-scale predictions[36, 37]. Moreover, TD-DFT is unavoidably biased towards certain backbones (e.g. even double hybrid functionals fail for cyanines)[38]. Finally, the efficiency of TD-DFT calculations, especially taking its $O(N^4)$ scaling into account, can hardly meet the requirement for the large-scale screening of organic materials. Analogous to PLQY predictions, the fast, accurate, black-box prediction of λ_{abs} and λ_{em} with unbiased generality across different dyes is sought after in this work.

The difficulties in first-principle photophysical modellings have motivated us to explore a fundamentally different, top-down data-driven approach. In recent years, machine learning (ML) has exhibited enormous potential as a useful tool in medicinal chemistry[39], organic synthesis[40, 41], and material chemistry[42-44]. For organic materials, although ML models have been established for various single-molecular properties available from (TD-)DFT calculations[45-47], predicting macroscopic characteristic parameters (activity, strength, durability, efficiency, etc.) based on molecular-level structural information is still a great challenge. So far, first principles can hardly predict these parameters. Reported ML predictions are limited to Power Conversion Efficiency (PCE)[48-52], gas absorption selectivity[53-55] and AIE effect[56], most relying on expensive quantum mechanical calculations to generate input expressions. For solvated molecules, expression of solvent features is critical but rarely studied in detail[57]. To achieve large-scale predictions for emission wavelength and PLQY with low/no sacrifice in accuracy, new strategies for feature engineering as well as the selection/designing of ML algorithms must be explored.

Herein, we report the development of highly accurate ML models for the fast estimation of photophysical parameters (λ_{abs}, λ_{em}, and Φ_{PL}) for solvated organic fluorescent materials. A database with more than 4,300 experimental samples and 11,000 data (λ_{abs}, λ_{em}, and Φ_{PL}) was established. Functionalized Structure Descriptor (FSD) and Comprehensive General Solvent Descriptor (CGSD) were developed and shown to be efficacious quantum-chemistry-free input expressions, enabling high-speed ML predictions. Remarkably, our optimal ML models predict Φ_{PL} with MAE = 0.11 and λ_{em} with MAE = 14.30 nm (0.066 eV in energy scale). This data-driven approach exhibits satisfactory universality towards unseen molecules and is systematically improvable via re-training and on-the-fly

learning. Using our new solvent descript (CGSD), the pronounced change of PLQY in different solvents can be predicted. Detailed comparison of our approach with TD-DFT calculations suggests dramatically less time cost for λ_{em} predictions (ML <1 s versus TD-DFT ~50 CPU hours for each molecule) with minor difference in accuracy. Ensemble model was incorporated into a Web version freely available at http://www.chemfluor.com, which hoped to be a useful tool for pre-screening and rational designing of organic fluorescent molecules. We believe that our black-box ML approach will serve as an efficient and reliable strategy that complements quantum mechanical calculations for the estimation of PYQY as well as the high-speed prediction of emission wavelengths.

Results and Discussion

Development of Machine Learning Models.

Figure 1a shows the statistics of absorption/emission wavelengths (>4,000 molecules with >8,000 wavelength data) collected from the literature. The data consist mainly of commercial fluorescent dyes and novel organic molecules with fluorescent activity reported in recent years (Figure 1b), including various skeletons with different functional groups. Most of the emission wavelengths are distributed in the range of 400 – 700 nm (blue to near-infrared). One reason is that fluorescent dyes with longer emission wavelengths are believed conducive to the applications in biological imaging, and are synthesized extensively in recent years.

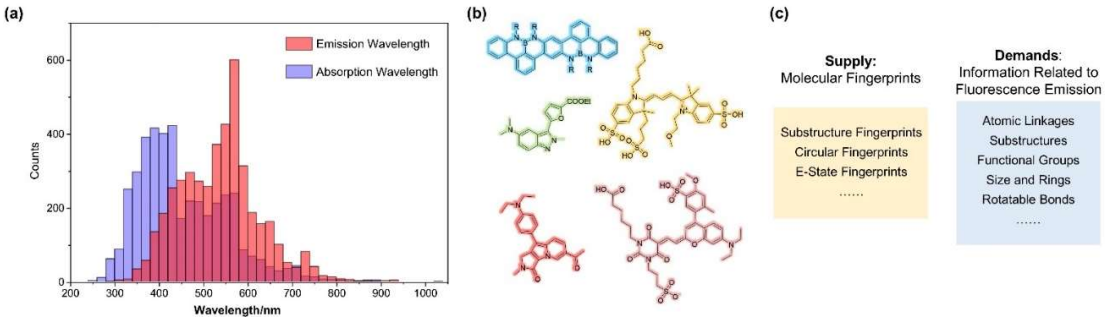


Figure 1. (a) Distribution of maximum absorption and emission wavelengths of the solvated organic fluorescent materials in our database. (b) Selective organic dyes in our database. (c) Illustration for the motivation of using multiple fingerprints.

In order to develop ML models, we started by the choice of molecular and solvent descriptors. Molecular descriptors serve as the basis for machine learning, for it transforms molecular information into computer-readable data. Molecular fingerprints, a subclass of molecular descriptors available without any quantum mechanical calculation, are used in our study due to the high potential in high-throughput screening of materials. A potential challenge originates

from the multifold molecular features involved in fluorescence emission, but a single molecular fingerprint hardly covers all of them (Figure 1c). For this reason, several kinds of fingerprints such as substructure key-based fingerprints and circular fingerprints as well as a handful of consensus fingerprints are investigated and compared. Because fluorescence properties are also sensitive to solvents especially for molecules with intramolecular charge transfer (ICT) features, Comprehensive General Solvent Descriptor (CGSD), which combines $E_T(30)$[58] with other four empirical scales[59], is proposed here in order to discern a wide spectrum of solvents,

The choice of ML algorithm is key to precise prediction. In addition to Random Forest (RF)[60], the most widely used ML algorithm, we also compared the performance and efficiency of other models including Support Vector Machine (SVM)[61], Kernel Ridge Regression (KRR)[62], Multi-Layer Perceptron (MLP)[63], k-Nearest Neighbors (kNN), Light Gradient Boosting Machine (LightGBM)[64] and Gradient Boost Regression Tree (GBRT)[65] to assess the relative merits of these approaches.

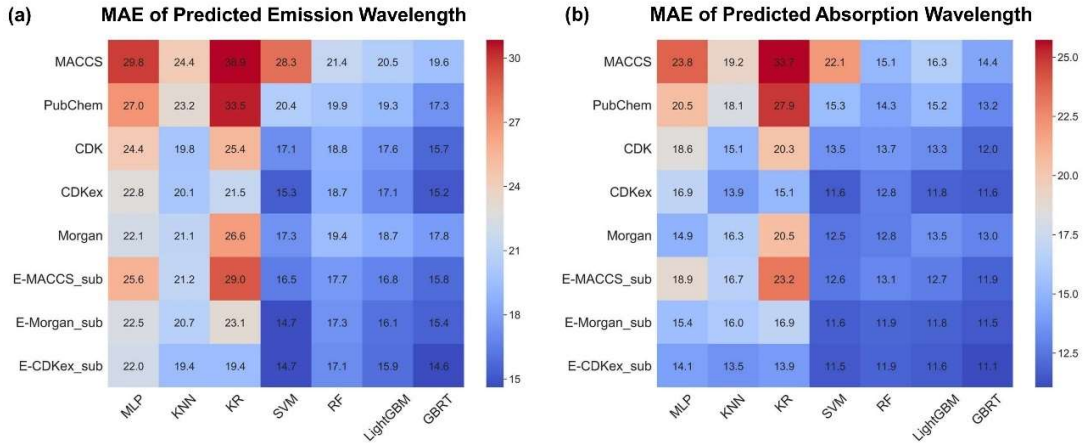


Figure 2. Testing results of (a) emission wavelength and (b) absorption wavelength different combinations of ML models with different structure-based descriptors as inputs. The average MAE of ten tests are shown in the center of each colored block; For each test, we randomly select 10% of the data as the test set and use the rest as the training set. Other metrics (R^2, RMSE and their confidence intervals) could be found in Table S1 (Supporting Information). Details about the abbreviation of the FSD and fingerprints can be found in the Method section.

To gain preliminary insights into the predictive powers of these ML models in conjunction with various molecular fingerprints, we first compared their mean absolute errors (MAE) for predicted absorption and emission wavelengths (Figure 2). In terms of tendency, shorter inputs show better performance with tree-based algorithms (RF, LightGBM and GBRT) while kernel-based algorithms (KRR and SVM) become comparable to tree-based ones with long input

features. MLP and kNN only show average results in our model, possibly because molecular fingerprints are sparse high-dimensional vectors. LightGBM, SVM and GBRT regressors exhibit the lowest MAEs, and are used for assessing fingerprints before further differentiation.

With regard to the efficacy of molecule descriptors[66], substructure key-based fingerprints (MACCS, PubChem), which are based on the presence of certain substructures in a limited structure list, exhibit poor performance according to Figure 2. By comparison, circular fingerprints including Chemistry Development Kit (CDK) fingerprints and Morgan fingerprints show better performance, which implies that the representation of molecular structures by atom neighborhoods might be better for our purpose. In the recent study by Glorious et al.[67], the benefits of combining multiple fingerprints features (MFFs) as a composite input molecular descriptor was demonstrated. However, due to the extreme lengths of MFFs (more than 70,000 bits), the resultant increase of computation cost limits its application. We propose that the combination of several descriptor describing features directly relevant to the phenomenon of interest might increase the efficiency of the expressions. Following this proposal, we have designed Functionalized Structure Descriptor (FSD), which combines two circular fingerprints (CDK fingerprints and Morgan fingerprints) with E-state fingerprints and substructure fingerprints (presence and count), giving rise to FSD_CDK (E-CDKex_sub) and FSD_Morgan (E-Morgan_sub) (Figure 2). One key underlying motivation is that substructure fingerprints provide a differentiation for an array of functional groups while structure descriptors (circular fingerprints) such as Morgan and CDKex are better expressions of the molecular backbones. Meeting our expectations, such strategy does increase the performance for all algorithms considered here. We also applied this method to MACCS, the smallest fingerprint, and the resulting E-MACCS_sub also exhibits improved performance. These results indicate that composite inputs with multiple relevant features improve the performance of our ML models. The FSD_CDK gives the lowest MAEs in reproducing both emission and absorption wavelengths, and are thereby used throughout the further assessments of ML algorithms.

Table 1. Performance of selected algorithms [a].

Prediction object	Algorithms	*r*	R^2	MAE/nm	RMSE/nm	MAE/eV	RMSE/eV
Emission	SVM	0.959 ± 0.009	0.918 ± 0.018	14.419 ± 0.683	25.736 ± 2.531	0.067 ± 0.003	0.126 ± 0.012
	LightGBM	0.957 ± 0.008	0.916 ± 0.016	15.295 ± 0.839	26.192 ± 2.044	0.071 ± 0.005	0.126 ± 0.013
	GBRT	0.962 ± 0.007	0.925 ± 0.014	14.307 ± 1.118	24.768 ± 2.238	0.066 ± 0.005	0.119 ± 0.012

	SVM	0.975 ± 0.005	0.951 ± 0.010	11.187 ± 0.984	22.217 ± 2.625	0.076 ± 0.006	0.157 ± 0.015
Absorption	LightGBM	0.973 ± 0.005	0.946 ± 0.009	11.614 ± 0.548	23.177 ± 1.845	0.077 ± 0.005	0.156 ± 0.019
	GBRT	0.977 ± 0.005	0.954 ± 0.010	10.471 ± 1.023	21.459 ± 2.565	0.070 ± 0.006	0.146 ± 0.019

[a] The presented results for each algorithm are achieved by 10-fold cross validation. The standard deviation is obtained by the difference of the prediction of each fold.

To further differentiate SVM, LightGBM and GBRT to find the optimal prediction model, we further analyzed their performance with more performance metrics over our database with 10-fold cross-validation (Table 1; see Table S2 for other algorithms and Figure S2 to Figure S9 for scatter plots). Since in the TD-DFT studies, the MAE of eV is a more commonly used evaluation standard, so we transformed the test result through the equation $E = 1240 / \lambda$ to show the MAE of our models under eV. The superior performance of the GBRT regressor is consistently suggested by the lowest MAEs (10.47 nm and 0.70 eV for absorption, 14.31 nm and 0.66 eV for emission) as well as the highest coefficients of determination (R^2 = 0.954 for λ_{abs} and 0.925 for λ_{em}). In the prediction results of the absorption wavelength, the MAE is lower in the case of wavelength (nm), but higher in the case of energy (eV). This is completely acceptable and mainly due to the illusion brought about by the unit conversion. Due to the higher decision coefficients (R^2) and correlation coefficients (r), we argue that the ML models perform more reliably for absorptions. It is worth noting that although the prediction of absorption shows a higher accuracy (by R^2 and r; plausibly due to the more direct structure-property relationship), more attention should be paid on emission due to the greater challenge of accurate prediction and the significance in fluorescence-based applications.

As described by Figure 3a (see Figure S1 for the rest of the algorithms), the advantage of GBRT over SVM and LightGBM is further supported by error distribution. The errors of more than 80% of the GBRT-predicted results are smaller than 20 nm, demonstrating the high accuracy of our approach for predicting molecules with similar backbones. Furthermore, it can be seen that GBRT has consistently larger cumulative percentage of error than the SVM and LightGBM. In order to further evaluate GBRT, SVM and LightGBM by their upgradeability and universality, the dependence of MAE on the partition ratio of training/test sets was examined (Figure 3b). When the test set makes up increasingly higher portions, the MAE of all three regressors increases accordingly. Following this tendency, it can be inferred that our model can perform even better with more available training data, and the same conclusion has been suggested by the learning curve for the fixed dataset (Figure S11 and Figure S12). The GBRT regressor, whose MAE remains smaller than 20 nm even when the training set is reduced to 40% of the entire database, shows smaller

MAE than the other two models at all tested partition ratios. Therefore, with the analysis on performance metrics, error distribution and model upgradeability, GBRT/FSD_CDK can be reasonably employed in further investigations to evaluate our ML approach.

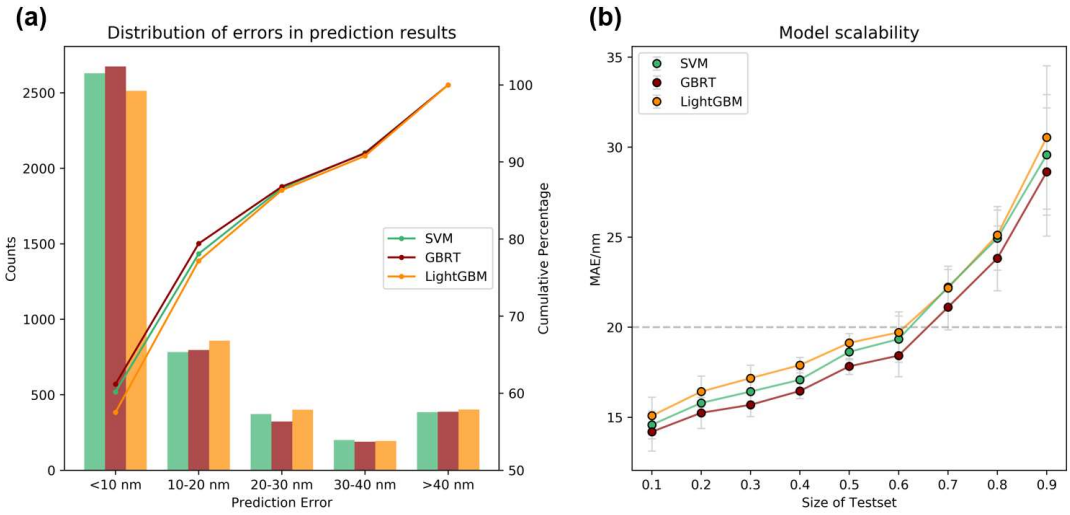


Figure 3. (a) Error distribution and (b) change of mean absolute error with the increase of test set portion for SVM, GBRT and LightGBM. FSD_CDK is employed in all these assessments.

Due to the significance of solvent effects in organic photophysics [58, 68], a successful model should be able to make predictions in the face of both new molecules and different solvents. Therefore, we have assessed our ML models for unlearnt organic dyes in different solvents to test the performance of CGSD. This is achieved by re-partitioning the database into training/test sets based on molecules, that is, the datapoints of the same molecules in different solvents will only appear in either training or test sets. In practice, we discriminate between molecules appearing only once (*part 1*) and molecules appearing for multiple times in different solvents (*part 2*). Then, we randomly and separately chosen 20% of the datapoints from *part 1* and *part 2* to form the test set. The performance of several algorithms following this approach is described in Table S3, which suggests that GBRT is the most suitable model for our purpose among selected algorithms. The predictions by GBRT/FSD_CDK are shown in Figure S13. The overall MAE (17.36 nm, 0.0802 eV) is only slightly less accurate than randomly sampling 20% of the entire dataset (MAE: 15.25 nm, 0.0700 eV). Although *part 2* shows less satisfactory performance (MAE: 20.83 nm, 0.0933 eV for this part of the test set), such accuracy is still noticeable. To alleviate the error of *part 2*, we have devised and trained a stacking model using four ML models as basic learners and the linear regressor as meta learner (details and discussions can be found

in the supporting information,). This ensemble model has reduced the overall error to 17.20 nm, 0.800 eV and the *part 2* error to 19.79 nm, 0.0887 eV. The benefit of ensemble model adds to the improvability of the ML approach. Nevertheless, we have continued to use a single GBRT model due to the following two reasons: (1) its high training efficiency (< 5 min) promote the potential on-the-fly learning, while the ensemble model needs comparably much longer training time; (2) acceptable accuracy can be achieved by single model, since errors at the level of 1 nm/0.005 eV is not so obvious in practical applications.

Summarizing this section, we have assessed an array of ML algorithms and molecular fingerprints for the prediction of absorption/emission wavelengths of solvated organic dyes, leading to the development of a ML regressor combining the GBRT algorithm, FSD_CDK and CGSD. In the course of our evaluations, the GBRT algorithm shows optimal performance on our database according to multiple indicators, error analysis, and upgradeability comparisons. Regarding to feature engineering, FSD_CDK has been developed by combining fingerprints describing features that are directly relevant to absorption and emission, and have proven effective within the scope of our investigations. Furthermore, it has been demonstrated that our ML approach is improvable by the expansion of database and the introduction of ensemble models. These results suggest the merits of our ML models for practical applications.

Machine Learning Predictions for PLQY.

Photoluminescence quantum yield is one of the most critical factors affecting the fluorescence intensity of organic fluorescent materials, but attempts to its prediction is still limited. Oriented towards high-throughput screening of emissive organic materials, we hope to achieve the ML prediction of PLQY with efficient quantum-chemistry-free molecular representations. In our database, around 3,000 PLQY data measured in various solvents have been collected. Screening over several fingerprints and algorithms indicate that the LightGBM/FSD_CDK regressor has optimal accuracy in our database (Table S4 and Table S5). Reasonable accuracy is achieved with this regressor (r = 0.84, MAE = 0.11; see Figure 4a), which is sufficient for applications such as pre-screening of fluorophore candidates. In addition, if we only focused on the samples that are a bit bright (defined as QY > 0.10 here), the MAE value is still 0.12, indicating the high performance of the ML model (Table S6). Moreover, the accuracy remains better than reported estimations with TD-DFT calculations[23] even when only 10% of our database is used for training (Figure S14 and Table S7), showing the superiority of our approach on this specific problem.

In attempts to reduce the error of our model, we noticed that experimental QY can have a large error bar. The best measurement method (integration sphere) may still have an error of about 10%, the relative method even higher. For

this reason, we have investigated the effect of using only the high-quality data (~45% of the dataset) by relative measurement. As expected, the resultant accuracy (r = 0.86; see Figure S15 for details) is slightly improved even though the dataset is considerably smaller. According to this result, it is believed that our model can be further improved with more available high-quality QY data.

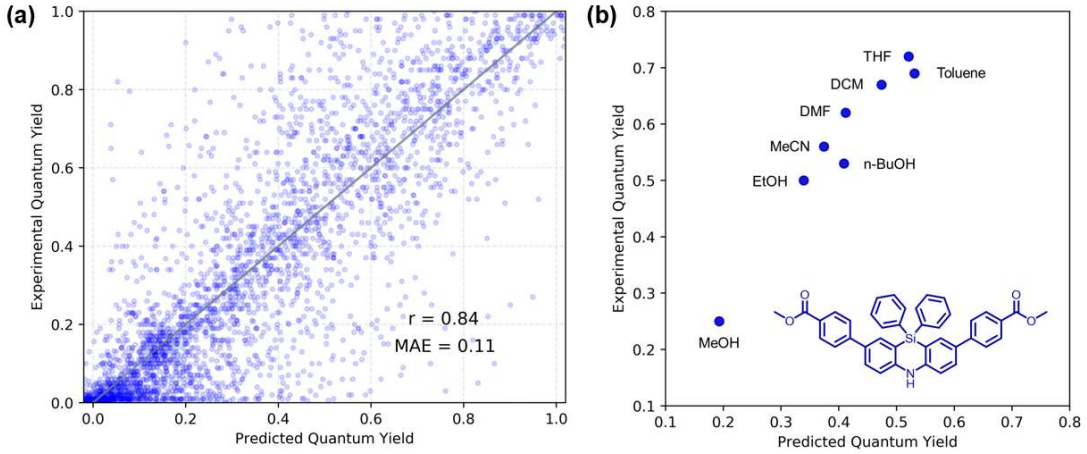


Figure 4. Prediction of PLQY with ML regressor model. a Linear correlation between experimental PLQY and LightGBM-predicted values, along with the correlation coefficient (r = 0.84). Perfect positive correlation is depicted by the solid diagonal line. b Chemical structures and Quantum yield in different solvents of typical compounds which can be accurate predicted.

Analogous to absorption/emission wavelengths, we have also evaluated the impact of molecule-based partition on QY predictions to show the predictive power of our models in the face of solvent effects. The reasonably higher MAE (0.131) compared with the datapoint-based approach (0.120) suggests insignificant overfitting in our models. However, solvent effects have a more involved influence on QY than emission wavelengths – even the same molecule can display distinct QYs in different solvents. Questioning whether our models can discriminate between large solvent effects in the same compound, we have selected several organic dyes whose emission shows notable solvent dependence (Figure 4b and Figure S18 to S20). It is shown that the dramatic solvent effects have been well reproduced for these examples, which is at least indicating the ability of our model for capturing the necessary solvent features for these molecules and suggesting the potential transferability to other cases. Further analysis suggests that our models can also differentiate the importance of solvent for different photophysical parameters. The overall importance of CGSD follows the order of QY (LightGBM: 14.68%, GBRT: 11.84%) > emission (GBRT: 5.84%) > absorption (GBRT: 0.69%) (see Table S10 for details), which meets with our cognition on solvent effects.

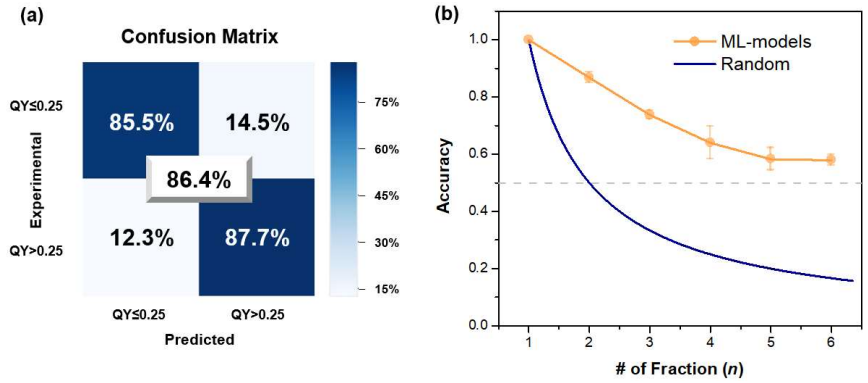


Figure 5. Prediction of PLQY with ML classifier model. a Performance of the LightGBM classifier on the test set (10% datapoints randomly selected from the database). b Accuracy versus the number of fractions (n) obtained by the LightGBM model with FSD_CDK.

Seeking for higher reliabilities than the regressors, we have also evaluated the performance of classifier models. To develop binary classifiers, the median of experimental PLQY (0.25) was used as the threshold to equally divide the database into two groups. This threshold is also suitable in realistic applications. The performance of the LightGBM/FSD_CDK classifier is described by the confusion matrix in Figure 5a. The accuracies of the best-performing models for the first ($\Phi_{PL} < 0.25$) and second ($\Phi_{PL} > 0.25$) groups are 85.5% and 87.7%, respectively, giving rise to a satisfactory overall accuracy (86.8%). Further assessment suggests that the accuracy remains greater than 80% when the training set shrinks to only 40% of the dataset (Figure S21). Hence, functional organic materials with strong fluorescence ($\Phi_{PL} > 0.25$) can be identified by the ML binary classifier even when a relatively small training set is available.

With the binary classifier in hand, we hope to increase the resolution of our classifier by introducing multiclass classifier models. The dependence of accuracy on the number of groups (n) is given in Figure 5b. When $n = 3$, the overall accuracy remains at a reasonable level (73.7%; see Figure S22 for confusion matrix). As n increases, the accuracy tends to decrease, but is significantly superior to random classifier. For $n = 6$, we can still obtain a 57.9% accuracy, which is around 3.5 times that of random classification. In fact, 68% of the incorrect predictions lie in intervals adjacent to the correct one (Figure S23), adding to the usability of our classifier. It can be inferred from the results here that ML classifier models are capable of providing reasonable predictions to PLQYs.

Because the binary classifier can already be applied to large-scale pre-screening of strong light-emission materials, we use it as example to test the accuracy of QY prediction on 22 molecules collected from three recent papers[69-71].

Unfortunately, an average result was obtained (accuracy = 72.7%) (Figure S24 and Table S12). One of the underlying reasons might be the lack of negative data, that is, materials with weak/no fluorescence are often reported without quantum yields. But still, the recall of strong fluorescent materials can be achieved 86.7%, which means that most tested molecules with strong fluorescence emission have been recognized by the binary classifier.

To conclude, we believe that our ML models, including regressors and classifiers, display reasonable accuracy in the tests presented above. The expansion of database is likely to enable further improvements that facilitate the design and virtual screening of novel organic fluorescent materials with high-quality ML predictions.

Comparison between Machine Learning and TD-DFT.

Whereas in principle quantum mechanical methods are efficacious as long as the physical approximations remains reasonable, empirical models such as QSAR and ML typically relies heavily on the scope of the training set and thereby lacks universality. For example, the published QSAR studies on the relationship between molecular structures and photophysical properties are usually limited to a maximum of hundreds of molecules[72, 73]. It is therefore important to assess the scope of our ML model (hence the potential in real-world applications) for the prediction of emission wavelengths. Accordingly, we have collected 116 molecules from TD-DFT studies on vertical emission energies[38, 74-78], mostly benchmark studies. The best levels of theories in each benchmark study were used to compare with our ML models. The ML-predicted emission wavelengths were translated into emission energies (eV) to be directly compared with TD-DFT. Note that the same level of error in wavelengths (nm) appears to be different when converted into energies (eV) due to the inverse proportionality ($E = 1240 / \lambda$). To alleviate such effect, the set of 116 molecules are divided into two categories, namely large fluorescent dyes whose emission wavelengths range from orange to red, and smaller ones with blue-to-green fluorescence emissions.

The results of the assessment are summarized in Table 2. In terms of overall performance, our ML model displays a lower MAE than TD-DFT (0.200 eV for ML vs. 0.237 eV for TD-DFT). The ML prediction of large fluorescent dyes seems excellent (MAE = 0.121 eV), superior to TD-DFT for BODIPY cyanines and rhodamine derivatives. In fact, these cyanines represent a particular challenge for TD-DFT calculations, which has been ascribed to the failure of TD-DFT for not correctly describing the difference of dynamic correlation between the two electronic states[38]. Even double-hybrid density functionals, which explicitly include contribution from virtual orbitals, give large errors for these molecules[79]. In contrast, our approach does not encounter such issue due to the direct statistical learning of experimental data, demonstrating the advantage of bypassing physical framework. Although the MAEs of our models

are generally larger for small fluorescent dyes, the performance is still comparable with TD-DFT for benzodiazoles (MAE = 0.197 eV) and coumarins (MAE = 0.234 eV) and is application to realistic problems. Since most small dyes collected in our dataset are novel heterocyclic dyes synthesized in the last decade, thus share fewer common features with this test set, and the relatively worse results on these molecules can be understood accordingly.

Table 2. Comparison between ML Models and TD-DFT Calculations for the Prediction of Emission Wavelengths[a].

Datasets	Skeletons	Range of λ_{em}	ML Predictions[b] MAE/eV	TD-DFT Calculations MAE/eV	Level of Theory[c]	[Ref]
Large Fluorescent Dyes	12 *BODIPY-Cyanines*	600-850 nm		*0.350*	TD-M06-2X/6-311+G(2d,p)/LR-PCM// TD-M06-2X/6-31G(d)/LR-PCM	[38]
	11 *D-π-A Dyes*	470-650 nm	*0.121 ± 0.006*	*0.100*	TD-ωB97X-D/6-31+G(d,p)/LR-PCM// TD-CAM-B3LYP/6-31G(d)/LR-PCM	[74]
	11 *Rhodamine Derivatives*	530-600 nm		*0.155*	TD-B3LYP-D/6-31+G(d,p)/CPCM	[75]
Small Fluorescent Dyes	9 *Substituted Benzoxadiazolses*	370-500 nm	*0.197 ± 0.016*	*0.308*	TD-PBE0/6-31+G(d)	[76]
	With 12 related molecules included into our dataset.		*0.141 ± 0.020*			
	49 *Coumarins*	350-500 nm	*0.234 ± 0.017*	*0.280*	TD-PBE0/6-31+G(d)/LR-PCM	[77]
	With 8 coumarins randomly moved from test set to training set.		*0.142 ± 0.005*			
	24 *1,8-Naphthalimides*	350-550 nm	*0.220 ± 0.018*	*0.160*	TD-PBE0/6-31+G(d)/LR-PCM	[78]
	With 4 naphthalimides randomly moved from test set to training set.		*0.149 ± 0.010*			
Overall	116 *Organic Fluorescent Materials* (Original Training Set)		*0.200 ± 0.005*	*0.237*		
	104 *Organic Fluorescent Materials* (Augmented Training Set)		*0.144 ± 0.006*	*0.228*		

[a] See Table S14 for details. [b] The ML-models are constructed with GBRT/FSD_CDK. [c] Best levels are chosen for each skeleton.

Although a prediction power comparable to TD-DFT is observed on the tested examples, there are still chances for the ML model to exhibit larger errors for more generalized cases. To demonstrate the applicability of our approach under such circumstances, we have investigated the improvability of our ML models for molecules with lower similarity to the training set, especially newly designed ones with unprecedented backbone structures. Note that aside from the original training set, learnable structural features might also be shared by certain subset(s) outside the training set (Figure 6). Inspired by this idea, we tested the impact of including a certain number of molecules analogous to the targeted ones into the training set. Benzoxadiazole dyes were used for preliminary explorations because 12 characterized molecules with similar backbones were provided in the TD-DFT paper[76]. The effect of including the 12 datapoints was notable (MAE reduced to 0.141 eV), which meets with our expectation. For coumarins and

naphthalimides, a different yet similar approach was investigated. We tried to move a small portion (< 17%; randomly selected) of the test set into our training set. Again, the updated ML models show excellent performance (MAE = 0.142 eV and 0.149 eV, respectively). According to these results, we infer that the improvement of our ML models for less-learned backbones can be readily achieved by utilizing similar molecules as effective training data. The low cost of the (re-)training step (less than 5 minutes) is considerably lower than TD-DFT computations. These results have also motivated us to provide a python script for both predictions and further expansion of database for learning on-the-fly, which is viable utilizing either more TD-DFT calculations or experimental feedbacks.

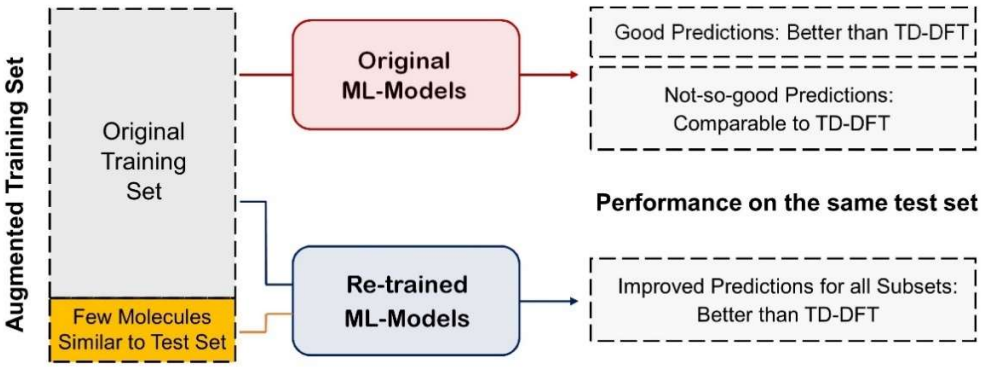


Figure 6. Schematic illustration for the improvement of ML models.

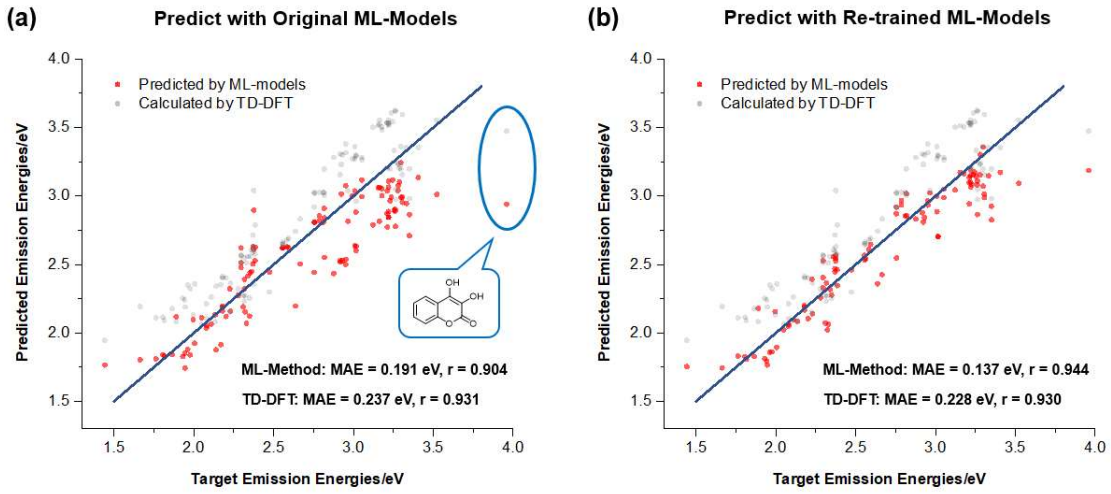


Figure 7. Fluorescence emission energies predicted by ML methods (red points) and modelled by vertical emission with TD-DFT (grey points). Results obtained with the original dataset (a) and the augmented database (b) are given. Perfect positive correlation ($r = 1$) is given by the blue lines for reference. The results of TD-DFT are slightly different from Table 2 because 12 molecules are moved from the test set to the training set.

Figure 7a shows the correlation between experimentally measured emission energies and vertical emission energies calculated by TD-DFT and predicted by ML-model directly of the 116 molecules shown in Table 2. Compared with TD-DFT, ML shows a smaller MAE but worse correlation coefficient. The compound with the largest error is 3,4-dihydroxy-2H-chromen-2-one, for which the Lewis structure might deviate from the real one due to tautomerization. When the training set is augmented with a few molecules that are structurally related to the test set, the ML model exhibits better performance than TD-DFT computations (Figure 7b).

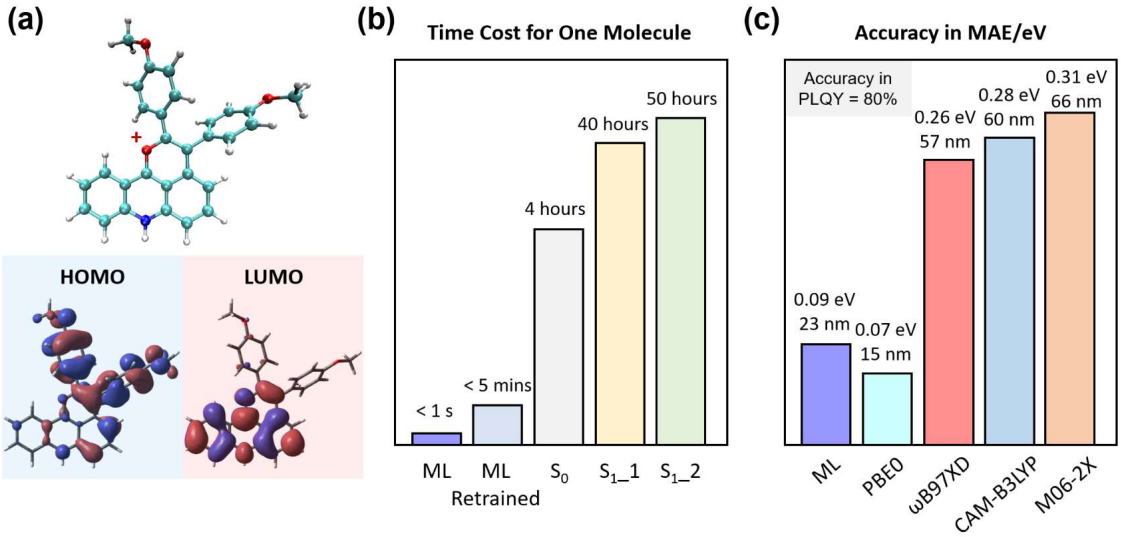


Figure 8. (a) Structure, HOMO and LUMO of representative molecule from the external dataset (30 molecules). (b) Comparison of the CPU time cost between ML method and first principle method (E5-2650v3). ML Retrained means the time cost with re-trained process, which will not change with the increase of the molecule number. S_0 is only for optimize the structure in ground state; S_1_1 means optimize the excited structure in S1 with CAM-B3LYP/6-31G*/LR-PCM(DCM); S_1_2 means after optimizing the excited structure in S_1 with CAM-B3LYP/6-31G*/LR-PCM(DCM), single point based on CAM-B3LYP/6-311+G*/LR-PCM(DCM) have been calculated. (c) Prediction accuracy of emission wavelength for this external dataset with ML model and various functional on 6-311+G*; Prediction accuracy of PLQY is shown in the grey color.

In order to further assess the performance of our ML methods, we collected 30 heterocyclic fluorescent dyes published recently as an additional test set[4]. These dyes are ionic, a category that is missing from TD-DFT benchmark studies. Using experimental reference data, PLQYs are predicted with MAE = 0.21 for ML regressor and 80% accuracy (i.e. 24 out of 30 correct answers) for ML classifiers. These are considered good results given the difficulty of QY predictions. For emission wavelengths, we make use of this extra dataset to make another comparison between TD-DFT and our ML models. For part of the selected dyes, the FMO diagrams imply charge-transfer character (Figure 8a). Our ML approach is able to make predictions costing less than 1 s for each molecule, reaching an overall MAE

of 23 nm (0.09 eV). To contrast this with TD-DFT, we optimized all molecules in the S_1 state with TD-DFT at the level of CAM-B3LYP/6-31G*/LR-PCM(DCM) (Reference can be found in supporting information). Then, emission energies were computed at the optimized geometries with 6-311+G*, a more extended Pople triple-zeta basis set equipped with polarization and diffuse functions for heavy atoms, in conjunction with the LR-PCM(DCM) solvent model and a series of typically recommended exchange-correlation functionals including hybrid functionals PBE0 and M06-2X as well as range-separated functionals ωB97X-D and CAM-B3LYP. These levels of theory are believed representative for realistic TD-DFT predictions. The accuracy and time cost with TD-DFT are compared with our ML models in Fig 8b-8c (see Fig SX for details). It is suggested that PBE0 is the only functional showing lower MAE than ML on this dataset (15 nm, 0.07 eV for PBE0 vs 23 nm, 0.09 eV for ML). The MAEs with other functionals are in the range of 57 ~ 66 nm (0.26 ~ 0.31 eV), which are generally larger than those with ML models. Meanwhile, our ML approach shows a clear advantage in time cost, since most TD-DFT predictions require 40 ~ 50 CPU hours.

From the comparison between TD-DFT and ML models, the following conclusions can be drawn: (1) By directly learning experimental data, ML models can predict emission wavelength with similar level of accuracy as TD-DFT in our tests, (2) Improvement of ML models can be readily achieved by the introduction of a certain amount of data about organic dyes similar to the targeted one(s), implying potential for on-the-fly learning.

Conclusion

Machine learning methodology was introduced to the predictions of photophysical parameters for organic fluorescent dyes. With a database of >4,300 solvated organic dyes established, ML models were developed. Molecular structures and solvent properties were efficiently expressed by the newly designed Functionalized Structure Descriptor (FSD) and Comprehensive General Solvent Descriptor (CGSD), respectively, and the underlying design rules were demonstrated. Combined with algorithm selection, ML prediction was realized for PLQYs with a good differentiation power in the solvent's degrees of freedom. For emission wavelengths, thorough comparison between our ML approach and TD-DFT calculations was drawn, showing comparable accuracy and considerably lower time cost in the quantum-chemistry-free data-driven approach. Moreover, the improvability of our ML models was shown by the re-training process with additional datapoints. Online platform (http://www.chemfluor.top) will be a useful tool in the pre-screening by experimenters for discovery of new materials. Our work demonstrates how challenges in excited-state

modelling especially those with highly involved physical nature (e.g. quantum yields, and potentially lifetime, bandwidths and so forth) can be effectively solved by simple machine-learning models.

Data availability

Detailed The data and codes used in this study are available on our website (http://www.chemfluor.top) or databases[80].

ASSOCIATED CONTENT

The following files are available free of charge.

Detailed information about method, ensemble learning model, solvent effect, model performance, Figure S1 to Figure S26 and Table S1 to S16 can be found in Supporting Information. (PDF)

AUTHOR INFORMATION

Corresponding Author

Cheng-Wei Ju - College of Chemistry, Nankai University, Tianjin 300071, China; orcid.org/0000-0002-2250-8548; Email: nkuchemjcw@mail.nankai.edu.cn

Author

Hanzhi Bai - Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.

Bo Li - Department of Chemistry, College of Science, Tianjin University, Tianjin 300072, China.

Rizhang Liu - College of Software Engineering, Sichuan University, Chengdu, Sichuan, 610064, China.

Author Contributions

‡These authors contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We thank Yumiao Ma (BSJ Institute, Beijing) for providing computation facilities. We also thank Dr. Zijie Qiu (Max Planck Institute for Polymer Research), Yu-Zhi Xu (South China University of Technology), Hao-Zhe Wang, Qian-Zhen Shao (both Nankai University) and Ziwei Fan for valuable discussions. We thank all researchers who reported the data collected in this study, epically Dr. Andreas Schüller (Pontificia Universidad Católica de Chile) and Prof. Dr. Young-Tae Chang (Pohang University of Science and Technology).

REFERENCES

1. Vendrell, M.; Zhai, D.; Er, J. C.; Chang, Y.-T., Combinatorial Strategies in Fluorescent Probe Development. Chemical Reviews 2012, 112, 4391-4420.
2. Lee, J.-S.; Kang, N.-y.; Kim, Y. K.; Samanta, A.; Feng, S.; Kim, H. K.; Vendrell, M.; Park, J. H.; Chang, Y.-T., Synthesis of a BODIPY Library and Its Application to the Development of Live Cell Glucagon Imaging Probe. Journal of the American Chemical Society 2009, 131, 10077-10082.
3. Ahn, Y.-H.; Lee, J.-S.; Chang, Y.-T., Combinatorial rosamine library and application to in vivo glutathione probe. Journal of the American Chemical Society 2007, 129, 4510-4511.
4. Ma, W.; Zhang, L.; Shi, Y.; Ran, Y.; Liu, Y.; You, J., Molecular Engineering to Access Fluorescent Trackers of Organelles by Cyclization: Chemical Environment of Nitrogen Atom-Modulated Targets. Advanced Functional Materials 2020, 30, 2004511.
5. Kondo, Y.; Yoshiura, K.; Kitera, S.; Nishi, H.; Oda, S.; Gotoh, H.; Sasada, Y.; Yanai, M.; Hatakeyama, T., Narrowband deep-blue organic light-emitting diode featuring an organoboron-based emitter. Nature Photonics 2019, 13, 678-682.
6. Chen, Q.; Zajaczkowski, W.; Seibel, J.; De Feyter, S.; Pisula, W.; Muellen, K.; Narita, A., Synthesis and helical supramolecular organization of discotic liquid crystalline dibenzo hi,st ovalene. Journal of Materials Chemistry C 2019, 7, 12898-12906.
7. Qiu, Z.; Zhao, W.; Cao, M.; Wang, Y.; Lam, J. W. Y.; Zhang, Z.; Chen, X.; Tang, B. Z., Dynamic Visualization of Stress/Strain Distribution and Fatigue Crack Propagation by an Organic Mechanoresponsive AIE Luminogen. Advanced Materials 2018, 30, 1803924.
8. Song, F.; Xu, Z.; Zhang, Q.; Zhao, Z.; Zhang, H.; Zhao, W.; Qiu, Z.; Qi, C.; Zhang, H.; Sung, H. H. Y.; Williams, I. D.; Lam, J. W. Y.; Zhao, Z.; Qin, A.; Ma, D.; Tang, B. Z., Highly Efficient Circularly Polarized Electroluminescence from Aggregation-Induced Emission Luminogens with Amplified Chirality and Delayed Fluorescence. Advanced Functional Materials 2018, 28, 1800051.
9. Coles, D. M.; Chen, Q.; Flatten, L. C.; Smith, J. M.; Muellen, K.; Narita, A.; Lidzey, D. G., Strong Exciton-Photon Coupling in a Nanographene Filled Microcavity. Nano Letters 2017, 17, 5521-5525.
10. Yang, Z.; Mao, Z.; Xie, Z.; Zhang, Y.; Liu, S.; Zhao, J.; Xu, J.; Chi, Z.; Aldred, M. P., Recent advances in organic thermally activated delayed fluorescence materials. Chemical Society Reviews 2017, 46, 915-1016.
11. Liu, D.; De, J.; Gao, H.; Ma, S.; Ou, Q.; Li, S.; Qin, Z.; Dong, H.; Liao, Q.; Xu, B.; Peng, Q.; Shuai, Z.; Tian, W.; Fu, H.; Zhang, X.; Zhen, Y.; Hu, W., Organic Laser Molecule with High Mobility, High Photoluminescence Quantum Yield, and Deep-Blue Lasing Characteristics. Journal of the American Chemical Society 2020, 142, 6332-6339.
12. Wang, C.; Fukazawa, A.; Taki, M.; Sato, Y.; Higashiyama, T.; Yamaguchi, S., A Phosphole Oxide Based Fluorescent Dye with Exceptional Resistance to Photobleaching: A Practical Tool for Continuous Imaging in STED Microscopy. Angewandte Chemie-International Edition 2015, 54, 15213-15217.
13. Wang, C.; Taki, M.; Kajiwara, K.; Wang, J.; Yamaguchi, S., Phosphole-oxide-based Fluorescent Probe for Super-resolution Stimulated Emission Depletion (STED) Live Imaging of the Lysosome Membrane. ACS Materials Letters 2020, 2, 705-711.
14. Zhang, Q.; Li, B.; Huang, S.; Nomura, H.; Tanaka, H.; Adachi, C., Efficient blue organic light-emitting diodes employing thermally activated delayed fluorescence. Nature Photonics 2014, 8, 326-332.

15. Lee, J.-H.; Chen, C.-H.; Lee, P.-H.; Lin, H.-Y.; Leung, M.-k.; Chiu, T.-L.; Lin, C.-F., Blue organic light-emitting diodes: current status, challenges, and future outlook. Journal of Materials Chemistry C 2019, 7, 5874-5888.

16. Kim, E.; Koh, M.; Lim, B. J.; Park, S. B., Emission wavelength prediction of a full-color-tunable fluorescent core skeleton, 9-aryl-1, 2-dihydropyrrolo [3, 4-b] indolizin-3-one. Journal of the American Chemical Society 2011, 133, 6642-6649.

17. Carter, E. A., Challenges in modeling materials properties without experimental input. Science 2008, 321, 800-803.

18. Peng, Q.; Yi, Y.; Shuai, Z.; Shao, J., Toward quantitative prediction of molecular fluorescence quantum efficiency: Role of Duschinsky rotation. Journal of the American Chemical Society 2007, 129, 9333-9339.

19. Shuai, Z.; Wang, D.; Peng, Q.; Geng, H., Computational Evaluation of Optoelectronic Properties for Organic/Carbon Materials. Accounts of Chemical Research 2014, 47, 3301-3309.

20. Humeniuk, A.; Bužančić, M.; Hoche, J.; Cerezo, J.; Mitrić, R.; Santoro, F.; Bonačić-Koutecký, V., Predicting fluorescence quantum yields for molecules in solution: A critical assessment of the harmonic approximation and the choice of the lineshape function. The Journal of Chemical Physics 2020, 152, 054107.

21. Polyak, I.; Hutton, L.; Crespo-Otero, R.; Barbatti, M.; Knowles, P. J., Ultrafast photoinduced dynamics of 1, 3-cyclohexadiene using XMS-CASPT2 surface hopping. Journal of chemical theory and computation 2019, 15, 3929-3940.

22. Li, X.; Chung, L. W.; Morokuma, K., Photodynamics of all-trans retinal protonated schiff base in bacteriorhodopsin and methanol solution. Journal of chemical theory and computation 2011, 7, 2694-2698.

23. Kohn, A. W.; Lin, Z.; Van Voorhis, T., Toward Prediction of Nonradiative Decay Pathways in Organic Compounds I: The Case of Naphthalene Quantum Yields. The Journal of Physical Chemistry C 2019, 123, 15394-15402.

24. Chi, W.; Chen, J.; Liu, W.; Wang, C.; Qi, Q.; Qiao, Q.; Tan, T. M.; Xiong, K.; Liu, X.; Kang, K., A General Descriptor Δ E Enables the Quantitative Development of Luminescent Materials Based on Photoinduced Electron Transfer. Journal of the American Chemical Society 2020, 142, 6777-6785.

25. Ou, Q.; Peng, Q.; Shuai, Z., Toward Quantitative Prediction of Fluorescence Quantum Efficiency by Combining Direct Vibrational Conversion and Surface Crossing: BODIPYs as an Example. The Journal of Physical Chemistry Letters 2020, 11, 7790-7797.

26. Zhao, W.; He, Z.; Lam, Jacky W. Y.; Peng, Q.; Ma, H.; Shuai, Z.; Bai, G.; Hao, J.; Tang, Ben Z., Rational Molecular Design for Achieving Persistent and Efficient Pure Organic Room-Temperature Phosphorescence. Chem 2016, 1, 592-602.

27. Ma, H.; Yu, H.; Peng, Q.; An, Z.; Wang, D.; Shuai, Z., Hydrogen Bonding-Induced Morphology Dependence of Long-Lived Organic Room-Temperature Phosphorescence: A Computational Study. The Journal of Physical Chemistry Letters 2019, 10, 6948-6954.

28. Ma, H.; Peng, Q.; An, Z.; Huang, W.; Shuai, Z., Efficient and Long-Lived Room-Temperature Organic Phosphorescence: Theoretical Descriptors for Molecular Designs. Journal of the American Chemical Society 2019, 141, 1010-1015.

29. Lin, Z.; Kohn, A. W.; Van Voorhis, T., Toward Prediction of Nonradiative Decay Pathways in Organic Compounds II: Two Internal Conversion Channels in BODIPYs. The Journal of Physical Chemistry C 2020, 124, 3925-3938.

30. Loos, P.-F.; Scemama, A.; Blondel, A.; Garniron, Y.; Caffarel, M.; Jacquemin, D., A mountaineering strategy to excited states: Highly accurate reference energies and benchmarks. Journal of Chemical Theory and Computation 2018, 14, 4360-4379.

31. Grimme, S., A simplified Tamm-Dancoff density functional approach for the electronic excitation spectra of very large molecules. The Journal of Chemical Physics 2013, 138, 244104.

32. Seibert, J.; Bannwarth, C.; Grimme, S., Biomolecular structure information from high-speed quantum mechanical electronic spectra calculation. Journal of the American Chemical Society 2017, 139, 11682-11685.

33. Laurent, A. D.; Jacquemin, D., TD-DFT benchmarks: a review. International Journal of Quantum Chemistry 2013, 113, 2019-2039.

34. Jacquemin, D.; Mennucci, B.; Adamo, C., Excited-state calculations with TD-DFT: from benchmarks to simulations in complex environments. Physical Chemistry Chemical Physics 2011, 13, 16987-16998.

35. Jacquemin, D.; Planchat, A.; Adamo, C.; Mennucci, B., TD-DFT assessment of functionals for optical 0–0 transitions in solvated dyes. Journal of Chemical Theory and Computation 2012, 8, 2359-2372.

36. Refaely-Abramson, S.; Baer, R.; Kronik, L., Fundamental and excitation gaps in molecules of relevance for organic photovoltaics from an optimally tuned range-separated hybrid functional. Physical Review B 2011, 84, 075144.

37. Rubešová, M.; Muchová, E.; Slavíček, P., Optimal Tuning of Range-Separated Hybrids for Solvated Molecules with Time-Dependent Density Functional Theory. Journal of Chemical Theory and Computation 2017, 13, 4972-4983.

38. Charaf-Eddin, A.; Le Guennic, B.; Jacquemin, D., Excited-states of BODIPY-cyanines: ultimate TD-DFT challenges? Rsc Advances 2014, 4, 49449-49456.

39. Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A., Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. Journal of Chemical Information and Modeling 2016, 56, 1936-1949.

40. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting reaction performance in C-N cross-coupling using machine learning. Science 2018, 360, 186-190.

41. Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L., Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 2018, 559, 377-381.

42. Gu, G. H.; Noh, J.; Kim, I.; Jung, Y., Machine learning for renewable energy materials. Journal of Materials Chemistry A 2019, 7, 17096-17117.

43. Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-learning-assisted materials discovery using failed experiments. Nature 2016, 533, 73-76.

44. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., Machine learning for molecular and materials science. Nature 2018, 559, 547-555.

45. Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A., Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nature Materials 2016, 15, 1120-1127.

46. Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A., Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. Journal of Chemical Theory and Computation 2017, 13, 5255-5264.

47. Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-de-Sousa, J., Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. Journal of Chemical Information and Modeling 2017, 57, 11-21.

48. Nagasawa, S.; Al-Naamani, E.; Saeki, A., Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. Journal of Physical Chemistry Letters 2018, 9, 2639-2646.

49. Sahu, H.; Rao, W.; Troisi, A.; Ma, H., Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. Advanced Energy Materials 2018, 8, 1801032.

50. Lee, M.-H., Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design. Advanced Energy Materials 2019, 9, 1900891.

51. Sahu, H.; Ma, H., Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. Journal of Physical Chemistry Letters 2019, 10, 7277-7284.

52. Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; Lu, S.; Li, Y.; Sun, K., Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Science Advances 2019, 5, eaay4275.

53. Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H., Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. Journal of Physical Chemistry Letters 2015, 6, 3528-3533.

54. Wang, S.; Zhang, Z.; Dai, S.; Jiang, D.-e., Insights into CO2/N2 Selectivity in Porous Carbons from Deep Learning. ACS Materials Letters 2019, 1, 558-563.

55. Zhang, Z.; Schott, J. A.; Liu, M.; Chen, H.; Lu, X.; Sumpter, B. G.; Fu, J.; Dai, S., Prediction of Carbon Dioxide Adsorption via Deep Learning. Angewandte Chemie International Edition 2019, 58, 259-263.

56. Qiu, J.; Wang, K.; Lian, Z.; Yang, X.; Huang, W.; Qin, A.; Wang, Q.; Tian, J.; Tang, B.; Zhang, S., Prediction and understanding of AIE effect by quantum mechanics-aided machine-learning algorithm. Chemical Communications 2018, 54, 7955-7958.

57. Yang, Q.; Li, Y.; Yang, J. D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J. P., Holistic Prediction of the pKa in Diverse Solvents Based on a Machine-Learning Approach. Angewandte Chemie 2020, doi:10.1002/anie.202008528.

58. Reichardt, C., Solvatochromic Dyes as Solvent Polarity Indicators. Chemical Reviews 1994, 94, 2319-2358.

59. Catalán, J., Toward a Generalized Treatment of the Solvent Effect Based on Four Empirical Scales: Dipolarity (SdP, a New Scale), Polarizability (SP), Acidity (SA), and Basicity (SB) of the Medium. The Journal of Physical Chemistry B 2009, 113, 5951-5960.

60. Breiman, L., Random forests. Machine Learning 2001, 45, 5-32.

61. Chang, C.-C.; Lin, C.-J., LIBSVM: A Library for Support Vector Machines. Acm Transactions on Intelligent Systems and Technology 2011, 2, Article 27.

62. Cortes, C.; Mohri, M.; Rostamizadeh, A. Learning non-linear combinations of kernels. 2009; pp 396-404.

63. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. Nature 2015, 521, 436-444.

64. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y., LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 2017, 3146-3154.

65. Friedman, J. H., Greedy function approximation: a gradient boosting machine. Annals of statistics 2001, 1189-1232.

66. Bajusz, D.; Rácz, A.; Héberger, K. 3.14 - Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In Comprehensive Medicinal Chemistry III, Chackalamannil, S.; Rotella, D.; Ward, S. E., Eds.; Elsevier: Oxford, 2017, pp 329-378.

67. Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F., A Structure-Based Platform for Predicting Chemical Reactivity. Chem 2020, 6, 1379-1390.

68. Sun, W.; Li, S.; Hu, R.; Qian, Y.; Wang, S.; Yang, G., Understanding solvent effects on luminescent properties of a triple fluorescent ESIPT compound and application for white light emission. The Journal of Physical Chemistry A 2009, 113, 5888-5895.

69. Li, G.; Hirano, T.; Yamada, K., Bright near-infrared chemiluminescent dyes: Phthalhydrazides conjugated with fluorescent BODIPYs. Dyes and Pigments 2020, 178, 108339.

70. Liu, J.; Zhu, L.; Wan, W.; Huang, X., Gold-Catalyzed Oxidatve Cascade Cyclization of 1,3-Diynamides: Polycyclic N-Heterocycle Synthesis via Construction of a Furopyridinyl Core. Organic Letters 2020 22, 3279–3285.

71. Pei, K.; Zhou, H.; Yin, Y.; Zhang, G.; Pan, W.; Zhang, Q.; Guo, H., Highly fluorescence emissive 5, 5′-distyryl-3, 3′-bithiophenes: Synthesis, crystal structure, optoelectronic and thermal properties. Dyes and Pigments 2020, 179, 108396.

72. Schueller, A.; Goh, G. B.; Kim, H.; Lee, J.-S.; Chang, Y.-T., Quantitative Structure-Fluorescence Property Relationship Analysis of a Large BODIPY Library. Molecular Informatics 2010, 29, 717-729.

73. Chen, C.-H.; Tanaka, K.; Funatsu, K., Random forest approach to QSPR study of fluorescence properties combining quantum chemical descriptors and solvent conditions. Journal of fluorescence 2018, 28, 695-706.

74. Bernini, C.; Zani, L.; Calamante, M.; Reginato, G.; Mordini, A.; Taddei, M.; Basosi, R.; Sinicropi, A., Excited State Geometries and Vertical Emission Energies of Solvated Dyes for DSSC: A PCM/TD-DFT Benchmark Study. Journal of Chemical Theory and Computation 2014, 10, 3925-3933.

75. Savarese, M.; Aliberti, A.; De Santo, I.; Battista, E.; Causa, F.; Netti, P. A.; Rega, N., Fluorescence Lifetimes and Quantum Yields of Rhodamine Derivatives: New Insights from Theory and Experiment. Journal of Physical Chemistry A 2012, 116, 7491-7497.

76. Brown, A.; Ngai, T. Y.; Barnes, M. A.; Key, J. A.; Cairo, C. W., Substituted Benzoxadiazoles as Fluorogenic Probes: A Computational Study of Absorption and Fluorescence. Journal of Physical Chemistry A 2012, 116, 46-54.

77. Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Frisch, M. J.; Assfeld, X.; Ciofini, I.; Adamo, C., Time-dependent density functional theory investigation of the absorption, fluorescence, and phosphorescence spectra of solvated coumarins. Journal of Chemical Physics 2006, 125, 164324.

78. Jacquemin, D.; Perpete, E. A.; Scalmani, G.; Ciofini, I.; Peltier, C.; Adamo, C., Absorption and emission spectra of 1,8-naphthalimide fluorophores: A PCM-TD-DFT investigation. Chemical Physics 2010, 372, 61-66.

79. Grimme, S.; Neese, F., Double-hybrid density functional theory for excited electronic states of molecules. The Journal of Chemical Physics 2007, 127, 154116.

80. Ju, C-W.; Liu, R.; Bai, H.; Li, B.; ChemFluor figshare (https://dx.doi.org/10.6084/m9.figshare.12110619.v3)