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 300071, China.

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

*email: chengwei.ju99@gmail.com

These authors contributed equally: Cheng-Wei Ju, Hanzhi Bai, Bo Li, Rizhang Liu

Abstract

The predictions of photophysical parameters are of crucial practical importance for the development of functional organic fluorescent materials, whereas the expense of quantum mechanical calculations and the relatively low universality of QSAR models have challenged the task. New avenues opened up by machine learning (ML), we establish a database of solvated organic fluorescent dyes and develop highly efficient ML models for the predictions of maximum emission/absorption wavelength and photoluminescence quantum yields, providing a reliable and efficient approach to high-throughput screenings. Various combinations of ML algorithms and molecular fingerprints were investigated. For emission wavelengths, TD-DFT accuracy was achieved under real-world conditions. Reliable identification of strong fluorescent materials was also demonstrated. We show that the easily obtainable consensus fingerprint inputs combined with proper ML algorithms enables efficient re-training based on additional datapoints whereby systematic improvements of our ML models can be achieved.

Introduction

Organic fluorescent materials, especially small-molecule organic fluorescent dyes, have been used extensively not only as useful tools in biological research[1, 2, 3] but also as vital elements in material science[4, 5, 6, 7, 8, 9]. The last decades have seen the development of novel fluorescence-based applications such as electrically pumped organic laser (EPOL)[10] and stimulated emission depletion (STED) microscopy[11], attracting even more attention to the rationale design of organic materials with desired photophysical properties. To achieve this goal, several strategies have been proposed to predict the maximum absorption/emission wavelengths ($\lambda_{em}/\lambda_{abs}$) of fluorophores, including quantitative structure-activity relationships (QSAR) studies[12, 13] and computational quantum mechanical (QM) methods[14, 15, 16]. Time-dependent density functional theory (TD-DFT) has emerged as probably the most popular electronic structure method for such purpose. However, compared with the satisfactory accuracy for λ_{abs}[17], the accurate prediction of λ_{em} by TD-DFT remains a considerable challenge due to the various approximations embodied in the physical model[18, 19]. As an example, the ignorance of vibronic couplings by assuming vertical excitation is typically employed due to its considerably lower cost than 0-0 energies[20, 21]. Furthermore, the involved interplay between radiative and non-radiative processes and the resulting needs for detailed and costly explorations of potential energy surfaces have further complicated QM predictions for the photoluminescence quantum yields (PLQY, Φ_{PL}) of emissive organic molecules[22, 23, 24, 25]. Consequently, experimental chemists rarely rely on such method for the (pre-)screening of newly designed organic fluorescent materials. The development of a new approach with both satisfactory efficiency and high accuracy to the prediction of photophysical properties is thereby of great practical significance for the design and screening of novel organic fluorescent materials.

As a promising method to solve the contradiction between high-cost calculations and limited computational power, machine learning (ML) has exhibited enormous potential as a useful tool in medicinal chemistry[26], organic synthesis[27, 28], and material chemistry[29, 30, 31], and has been explored extensively in recent years. For organic materials, the ML-predicted properties explored so far can be roughly classified into two categories: (1) properties available from (TD-)DFT calculations[32, 33, 34], typically single-molecular properties (HOMO/LUMO energies, S_1-T_1 gaps, dipole moments, etc.); (2) macroscopic characteristic parameters that can only be obtained accurately from experimental measurements, including activity, strength, durability,

efficiency and so forth. For the first category, the performance of a variety of molecular descriptors and ML algorithms has been investigated for different properties. As a successful example of ML-assisted material design, Aspuru-Guzik et al.[32] achieved high-throughput pre-screening of thermally activated delayed fluorescence (TADF) organic light-emitting diodes (OLED) based on neural-network prediction of delayed fluorescence rate constant with data obtained from (TD-)DFT calculations. In contrast, only a few properties in the second category have been studied, examples being Power Conversion Efficiency (PCE)[35, 36, 37, 38, 39], gas absorption selectivity[40, 41, 42], and aggregation-induced emission (AIE) effect[43]. For these properties, the input expressions of existing ML models are usually obtained from expensive quantum mechanical calculations, limiting their application in large-scale fast virtual screening. More recently, Sun et al.[39] has achieved accurate prediction of PCE with molecular fingerprints, a type of input expression that can be generated without any quantum calculation. Nevertheless, the prediction of macroscopic characteristic parameters based on easily obtainable quantum-chemistry-free inputs remains challenging and largely unexplored.

In this work, we report the development of accurate and highly efficient ML models for the fast estimation of photophysical parameters (λ_{abs}, λ_{em}, and Φ_{PL}) of solvated organic fluorescent materials. The expense of quantum mechanical calculations for molecular descriptors is bypassed by the employment of fingerprints as input expressions. A database with more than 4,300 experimental samples and 11,000 data (λ_{abs}, λ_{em}, and Φ_{PL}) was established. With the optimal combination of regressors and molecular representations, the mean absolute errors (MAE) for λ_{em} and λ_{abs} were reduced to 14.30 nm/0.66 eV and 10.47 nm/0.70 eV, respectively. With molecular-based partition, high accuracy can still be maintained (MAE low to 17.36 nm/0.0802 eV). In addition, we have also developed ML regressors and classifiers for the prediction of PLQY for different demands. An MAE of 0.11 and accuracy of 0.86 was achieved by ML regressor and binary classifier, respectively, providing workable identification of strongly emissive organic materials. The universality of our ML models is supported by an MAE of 0.200 eV on a set of 116 emission energies gathered from TD-DFT benchmark studies. For unseen organic dyes that are less related to our database, we demonstrate that by including a small number (~15%) of similar molecules into the training set, the MAE of our ML model can be further reduced to ~0.1 eV. In practice, the improvement of our models is further facilitated by the low cost of re-training (< 5 minutes per model on a personal computer). We believe that our ML approach

will provide an efficient and reliable platform for large-scale screening of organic materials with different substituent groups under realistic conditions.

Results

Importance of Descriptors and ML Algorithms for the Prediction of Emission and Absorption Wavelengths. Fig. 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 (Fig. 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.

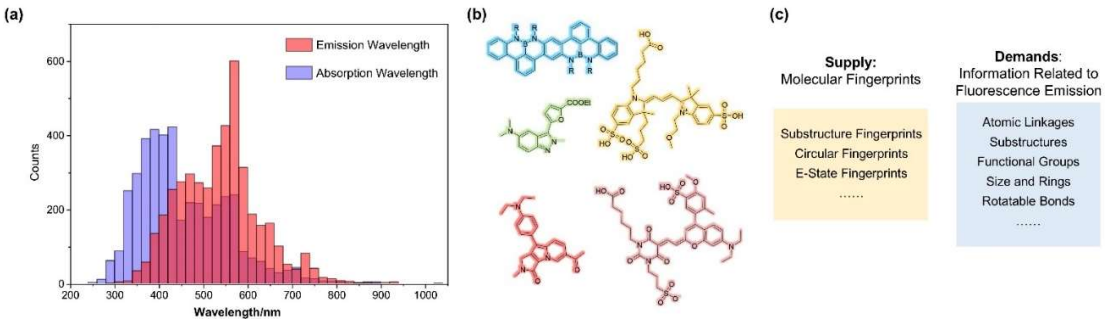

Fig. 1 a Distribution of maximum absorption and emission wavelengths of the solvated organic fluorescent materials in our database. **b** Selective organic dyes in our database. **c** Illustration for the motivation of using multiple fingerprints.

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 (Fig. 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, we use the combination of $E_T(30)$[44] and other four empirical scales[45] as solvent descriptors 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)[46], the most widely used ML algorithm, we also compared the performance and efficiency of other models including Support Vector Machine (SVM)[47], Kernel Ridge Regression (KRR)[48], Multi-Layer Perceptron (MLP)[49], k-Nearest Neighbors (kNN), Light Gradient Boosting Machine (LightGBM)[50] and Gradient Boost Regression Tree (GBRT)[51] to assess the relative merits of these approaches.

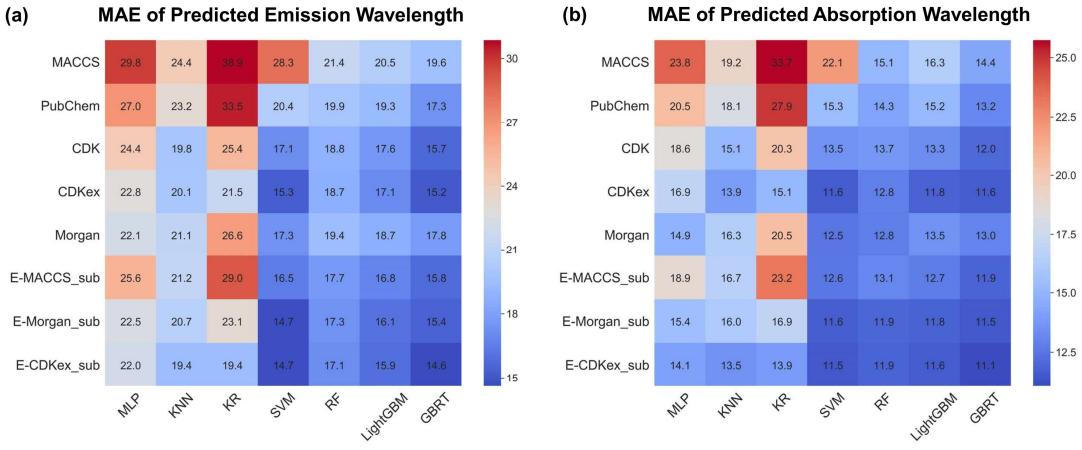

Fig. 2 Testing results of (a) emission wavelength and (b) absorption wavelength different combinations of ML models with different descriptors as inputs. The average MAE of ten tests are shown in the center of each colored block; For each test, we randomly select 10% of the data as the test set and use 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 consensus 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 (Fig. 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 molecular fingerprints[52], 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 Fig. 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.[53], 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 fewer molecular fingerprints describing features directly relevant to the phenomenon of interest might increase the efficiency of the expressions. Therefore, we combined two circular fingerprints (CDK fingerprints and Morgan fingerprints) with E-state fingerprints and substructure fingerprints (presence and count), giving rise to E-CDKex_sub and E-Morgan_sub (Fig. 2). 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 fingerprint features improve the performance of our ML models. The consensus fingerprint E-CDKex_sub 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
Absorption	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
	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 Fig. S2 to Fig. 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 Fig. 3a (see Fig. 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 (Fig. 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 (Fig. S11 and Fig. 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/E-CDKex_sub can be reasonably employed in further investigations to evaluate our ML approach.

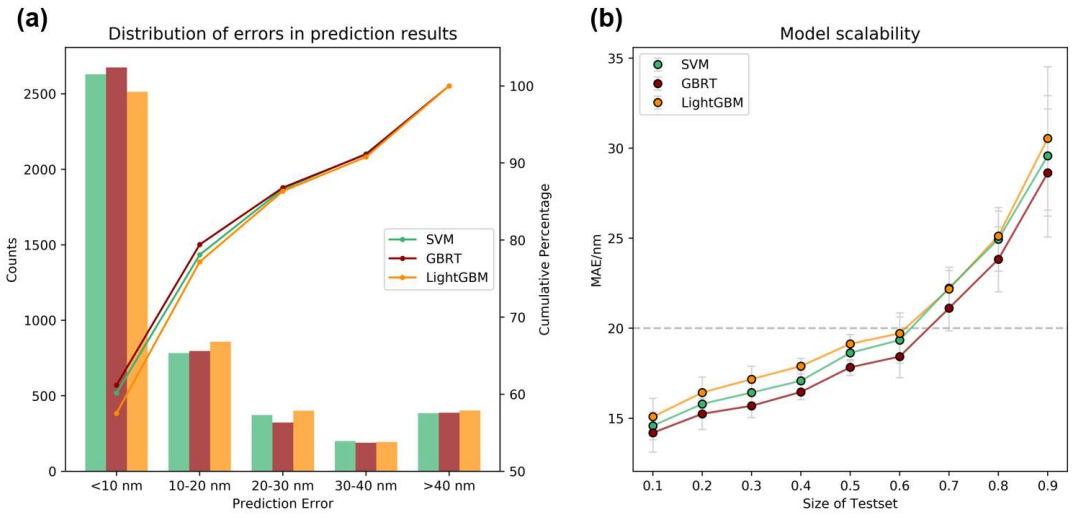

Fig. 3 (a) Error distribution and (b) change of mean absolute error with the increase of test set portion for SVM, GBRT and LightGBM. The E-CDKex_sub fingerprint is employed in all these assessments.

Due to the significance of solvent effects in organic photophysics, 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 model for unlearnt organic dyes in different solvents. 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/E-CDKex_sub are shown in Fig. 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 user to increase their own datapoint which can further increase its accuracy, 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, the E-CDKex_sub consensus fingerprint, and five solvent descriptors. 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, the E-CDKex_sub consensus fingerprint 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.

Exploration on prediction of PLQY with ML-models. 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/E-CDKex_sub 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 Fig. 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[24] even when only 10% of our database is used for training (Fig. 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 Fig. 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.

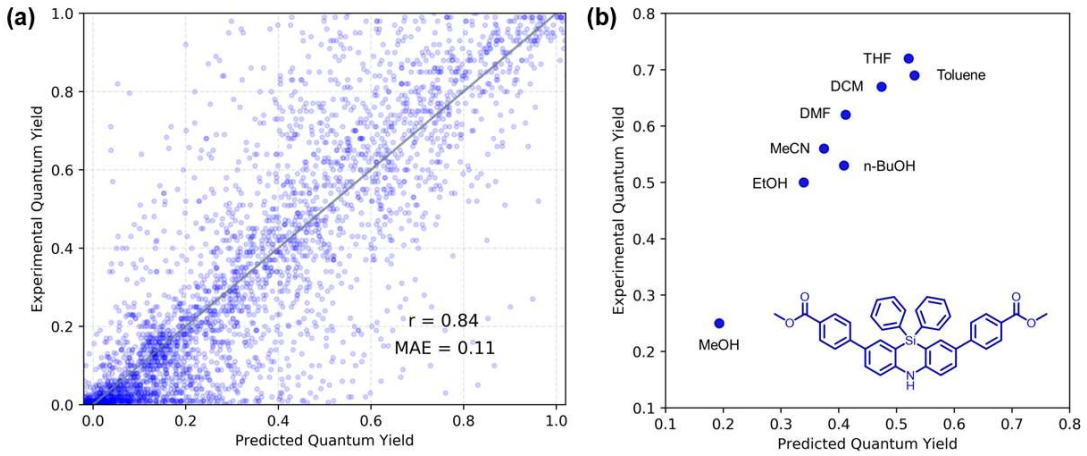

Fig. 4 Prediction of PLQY with ML regressor model. **a** Linear correlation between experimental PLQY and LightGBM-predicted values, along with the correlation coefficient (r = 0.84). Perfect positive correlation is depicted by the solid diagonal line. **b** Chemical structures and Quantum yield in different 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 (Fig. 4b and Fig. 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 solvent features

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.

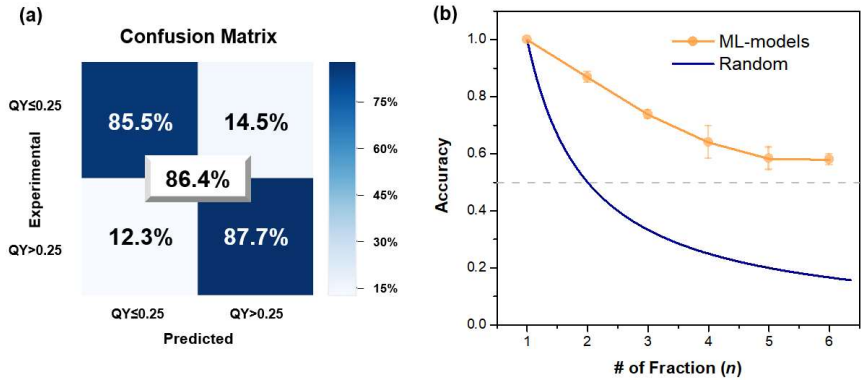

Fig. 5. Prediction of PLQY with ML classifier modl. **a** Performance of the LightGBM classifier on the test set (10% datapoints randomly selected from the database). **b** Accuracy versus the number of fractions (n) obtained by the LightGBM model with E-CDKex_sub.

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/E-CDKex_sub classifier is described by the confusion matrix in Fig. 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 (Fig. 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 Fig. 5b. When $n = 3$, the overall accuracy remains at a reasonable level (73.7%; see Fig. 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 (Fig. 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[54, 55, 56]. Unfortunately, an average result was obtained (accuracy = 72.7%) (Fig. 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 high-throughput virtual screening of novel organic fluorescent materials with high-quality ML predictions.

Comparison between ML Models and TD-DFT Calculations for Fluorescence Wavelength Predictions. 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[13]. 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[14, 15, 16, 18, 19, 22], 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[18]. Even double−hybrid density functionals, which explicitly include contribution from virtual orbitals, give large errors for these molecules[57]. 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]	TD-DFT Calculations		[Ref]
			MAE/eV	MAE/eV	Level of Theory[c]	
Large Fluorescent Dyes	**12** *BODIPY-Cyanines*	600-850 nm	*0.121 ± 0.006*	*0.350*	TD-M06-2X/6-311+G(2d,p)/LR-PCM// TD-M06-2X/6-31G(d)/LR-PCM	[18]
	11 *D-π-A Dyes*	470-650 nm		*0.100*	TD-ωB97X-D/6-31+G(d,p)/LR-PCM// TD-CAM-B3LYP/6-31G(d)/LR-PCM	[19]
	11 *Rhodamine Derivatives*	530-600 nm		*0.155*	TD-B3LYP-D/6-31+G(d,p)/CPCM	[20]
Small Fluorescent Dyes	**9** *Substituted Benzoxadiazolses*	370-500 nm	*0.197 ± 0.016*	*0.308*	TD-PBE0/6-31+G(d)	[14]
	*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	[15]
	*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	[16]
	*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/E-CDKex_sub. [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 (Fig. 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[14]. 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 package for both predictions and further expansion of database for re-training.

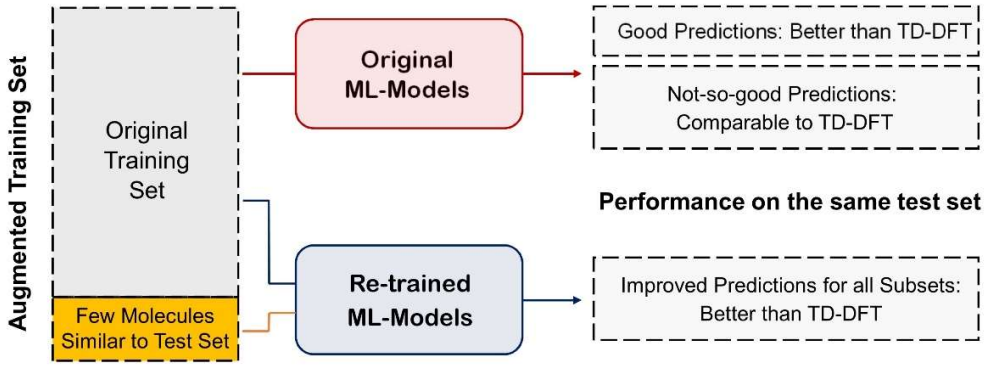

Fig. 6 Schematic illustration for the improvement of ML models.

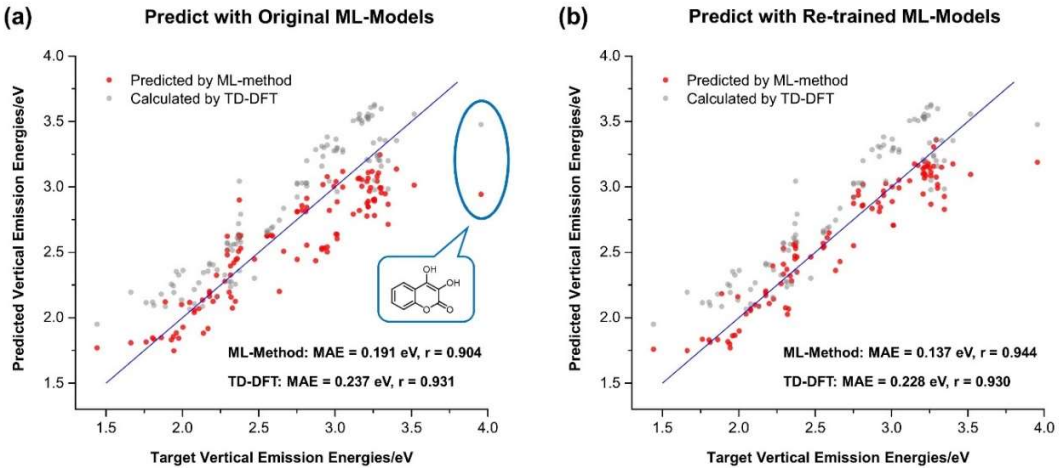

Fig. 7 Fluorescence emission energies predicted by ML methods (red points) and modelled by vertical emission with TD-DFT (grey points). Results obtained with the original dataset (**a**) and the augmented database (**b**) are given. Perfect positive correlation (r = 1) is given by the blue lines for 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.

Fig. 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 (Fig. 7b). 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 sizable amount of data about organic dyes similar to the targeted one(s).

Discussion

In summary, we introduced machine learning method into the prediction of photophysical parameters for organic fluorescent materials. A database with more than 4,000 solvated organic fluorescent dyes was established. After screening various fingerprints and algorithms, accurate prediction of emission wavelengths was achieved by learning experimental data. Consensus fingerprints with features relevant to the

phenomenon of interest (fluorescence emission in our case) were found to be efficacious input expressions. Our tests have suggested a level of accuracy similar to TD-DFT calculations in the face of real-world problems. Moreover, notable improvements of our ML models were achieved by including additional characterized molecules that share less similarities with the training set and more with the molecules to be predicted, a feature that we encourage the users to make use of.

Whereas emission wavelengths can be obtained from TD-DFT calculations, the prediction of quantum yields is more challenging. The quantum mechanical modelling of macroscopic characteristic parameters is difficult, often due to the multiscale nature of the system of interest as well as the physical complexity of various possible causes that might be coupled together in an involved manner. Bypassing the difficulties in physical modellings, we have shown that a simple ML-based binary classifier is already capable of offering reliable identification of strongly emissive organic dyes, showing good applicability to the (pre-)screening of organic fluorescent materials. The feasibility of achieving higher resolution (i.e. more than two groups) is demonstrated by the satisfactory performance of multiclass classifiers. ML regressors also display reasonably low MAEs. These results suggest the promising potential of ML models for quantum yield predictions of organic fluorescent dyes. We believe that more effective experimental PLQY data, especially negative ones, will greatly facilitate further improvements.

In addition to the applications mentioned above, the ML approaches presented here should also be transferable to the effective prediction of other important properties based on molecular fingerprints. We believe that the strategies demonstrated here will not only benefit the development of new ML models, but also promote the interaction between computer modelling and experimental explorations.

Method

ML Algorithms. Several supervised ML algorithms are used in this work, including Support Vector Machine (SVM), Kernel Ridge Regression (KRR), Multi-Layer Perceptron (MLP), k-Nearest Neighbors (kNN), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Gradient Boost Regression Tree (GBRT). All except LightGBM[50] can be found in Scikit-learn[58].

Database/Selection of Training and Test Sets. The database consists of 4371 λ_{em} data, 4237 λ_{abs} data and 3079 PLQY data of organic fluorescent dyes solvated in different solvents gathered from published works, open directories of Dyomics[59], and fluorophore[60] database. If multiple peaks were found, the peak with the longest wavelength/largest intensity was collected for absorption/emission data, respectively. Detailed information about the database can be accessed on our website. For individual tests in the presenting work, the

database is randomly partitioned into training set (90%) and test set (10%). The standard deviation in the 10-fold cross-validation is performed for the results of ten folds. Error bars in this work are drawn with standard error.

Fingerprints. Various fingerprints were investigated in this research. Most of them were obtained by PaDEL-Descriptor[61], including MACCS (166 bits), PubChem (881 bits), Substructure (presence and count of SMARTS patterns for Laggner functional group, 614 bits), Estate (E-State fragments, 79 bits), CDK (Chemistry Development Kit Fingerprints, 1024 bits), and CDKex (Chemistry Development Kit fingerprints and extended fingerprints, 2048 bits). Morgan circular fingerprints were generated with size 2048 bits and radius 2 by Rdkit[62]. E-CDKex_sub was generated directly by Padel, combining CDK fingerprints and extended fingerprints with E-States fingerprints and substructure fingerprints (both presence and count). E-MACCS_sub was the combination of MACCS, E-States fingerprints, and substructure fingerprints (both presence and count). E-Morgan_sub was the combination of Morgan fingerprints, E-States fingerprints, and substructure fingerprints (both presence and count).

Website/Package for Photophysical Properties Prediction. We have deployed a website where users can predict photophysical properties using our ML models (http://www.chemfluor.top). ML models with optimal accuracies (GBRT for λ_{em} and λ_{abs}, LightGBM for Φ_{PL}) are employed as back-end ML models of our online platform. Users can make predictions by inputting SMILES and solvent information. The outputs include maximum emission wavelength, maximum absorption wavelength, and photoluminescence quantum yield. Although the ML models per se are fast, the translation from SMILES to fingerprints by PaDEL costs more time. As a result, it takes ~4 seconds to finish one prediction. In addition, some larger molecules cannot be converted into molecular fingerprints through PaDEL in the server. Therefore, we also provide support for an (offline) python package where new compounds can be added into the training dataset for re-training. We have prepared a small patch for OLED (emission), used as a tutorial to teach users how to introduce their own data. The python package and the patch can be found in supporting data[63] or downloaded from our website.

Data availability

The data and codes used in this study are available on our website (http://www.chemfluor.top) or databases[63].

Acknowledgments

We are grateful to the Tianjin Training Programs of Innovation and Entrepreneurship for Undergraduates (No.201910055398). 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).

Author contributions

C.J. conceived the project, collected the data, optimized the ML-models, analyzed the data, and wrote the manuscript. H.B. conceived the project, constructed and optimized the ML-models, and analyzed the data. R.L. constructed and optimized the ML-models, analyzed the data, and prepared the web tool. B.L. analyzed the data and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

1. Vendrell M, Zhai D, Er JC, Chang Y-T. Combinatorial Strategies in Fluorescent Probe Development. *Chemical Reviews* **112**, 4391-4420 (2012).
2. Lee J-S*, et al.* Synthesis of a BODIPY Library and Its Application to the Development of Live Cell Glucagon Imaging Probe. *Journal of the American Chemical Society* **131**, 10077-10082 (2009).
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* **129**, 4510-4511 (2007).
4. Kondo Y*, et al.* Narrowband deep-blue organic light-emitting diode featuring an organoboron-based emitter. *Nature Photonics* **13**, 678-682 (2019).
5. Chen Q*, et al.* Synthesis and helical supramolecular organization of discotic liquid crystalline dibenzo hi,st ovalene. *Journal of Materials Chemistry C* **7**, 12898-12906 (2019).
6. Qiu Z*, et al.* Dynamic Visualization of Stress/Strain Distribution and Fatigue Crack Propagation by an Organic Mechanoresponsive AIE Luminogen. *Advanced Materials* **30**, 1803924 (2018).
7. Song F*, et al.* Highly Efficient Circularly Polarized Electroluminescence from Aggregation-Induced Emission Luminogens with Amplified Chirality and Delayed Fluorescence. *Advanced Functional Materials* **28**, 1800051 (2018).
8. Coles DM*, et al.* Strong Exciton-Photon Coupling in a Nanographene Filled Microcavity. *Nano Letters* **17**, 5521-5525 (2017).
9. Yang Z*, et al.* Recent advances in organic thermally activated delayed fluorescence materials. *Chemical Society Reviews* **46**, 915-1016 (2017).
10. Liu D*, et al.* Organic Laser Molecule with High Mobility, High Photoluminescence Quantum Yield, and Deep-Blue Lasing Characteristics. *Journal of the American Chemical Society* **142**, 6332-6339 (2020).
11. 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* **54**, 15213-15217 (2015).
12. 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* **28**, 695-706 (2018).
13. Schueller A, Goh GB, Kim H, Lee J-S, Chang Y-T. Quantitative Structure-Fluorescence Property Relationship Analysis of a Large BODIPY Library. *Molecular Informatics* **29**, 717-729 (2010).
14. Brown A, Ngai TY, Barnes MA, Key JA, Cairo CW. Substituted Benzoxadiazoles as Fluorogenic Probes: A Computational Study of Absorption and Fluorescence. *Journal of Physical Chemistry A* **116**, 46-54 (2012).
15. Jacquemin D*, et al.* Time-dependent density functional theory investigation of the absorption, fluorescence, and phosphorescence spectra of solvated coumarins. *Journal of Chemical Physics* **125**, 164324 (2006).
16. Jacquemin D, Perpete EA, Scalmani G, Ciofini I, Peltier C, Adamo C. Absorption and emission spectra of 1,8-naphthalimide fluorophores: A PCM-TD-DFT investigation. *Chemical Physics* **372**, 61-66 (2010).

17. Chibani S, Laurent AD, Le Guennic B, Jacquemin D. Improving the Accuracy of Excited-State Simulations of BODIPY and Aza-BODIPY Dyes with a Joint SOS-CIS(D) and TD-DFT Approach. *Journal of Chemical Theory and Computation* **10**, 4574-4582 (2014).
18. Charaf-Eddin A, Le Guennic B, Jacquemin D. Excited-states of BODIPY-cyanines: ultimate TD-DFT challenges? *Rsc Advances* **4**, 49449-49456 (2014).
19. Bernini C*, et al.* Excited State Geometries and Vertical Emission Energies of Solvated Dyes for DSSC: A PCM/TD-DFT Benchmark Study. *Journal of Chemical Theory and Computation* **10**, 3925-3933 (2014).
20. Jacquemin D, Duchemin I, Blase X. 0-0 Energies Using Hybrid Schemes: Benchmarks of TD-DFT, CIS(D), ADC(2), CC2, and BSE/GW formalisms for 80 Real-Life Compounds. *J Chem Theory Comput* **11**, 5340-5359 (2015).
21. Grimme S, Antony J, Ehrlich S, Krieg H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **132**, 154104 (2010).
22. Savarese M*, et al.* Fluorescence Lifetimes and Quantum Yields of Rhodamine Derivatives: New Insights from Theory and Experiment. *Journal of Physical Chemistry A* **116**, 7491-7497 (2012).
23. 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* **129**, 9333-9339 (2007).
24. Kohn AW, 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* **123**, 15394-15402 (2019).
25. Lin Z, Kohn AW, 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* **124**, 3925-3938 (2020).
26. Subramanian G, Ramsundar B, Pande V, Denny RA. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *Journal of Chemical Information and Modeling* **56**, 1936-1949 (2016).
27. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **360**, 186-190 (2018).
28. Granda JM, Donina L, Dragone V, Long D-L, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377-381 (2018).
29. Gu GH, Noh J, Kim I, Jung Y. Machine learning for renewable energy materials. *Journal of Materials Chemistry A* **7**, 17096-17117 (2019).
30. Raccuglia P*, et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73-76 (2016).
31. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature* **559**, 547-555 (2018).
32. Gomez-Bombarelli R*, et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **15**, 1120-1127 (2016).
33. Faber FA*, et al.* Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **13**, 5255-5264 (2017).

34. Pereira F, Xiao K, Latino DARS, 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* **57**, 11-21 (2017).
35. 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* **9**, 2639-2646 (2018).
36. Sahu H, Rao W, Troisi A, Ma H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials* **8**, 1801032 (2018).
37. 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* **9**, 1900891 (2019).
38. Sahu H, Ma H. Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *Journal of Physical Chemistry Letters* **10**, 7277-7284 (2019).
39. Sun W*, et al.* Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances* **5**, eaay4275 (2019).
40. Ma X, Li Z, Achenie LEK, Xin H. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. *Journal of Physical Chemistry Letters* **6**, 3528-3533 (2015).
41. Wang S, Zhang Z, Dai S, Jiang D-e. Insights into CO2/N2 Selectivity in Porous Carbons from Deep Learning. *ACS Materials Letters* **1**, 558-563 (2019).
42. Zhang Z*, et al.* Prediction of Carbon Dioxide Adsorption via Deep Learning. *Angewandte Chemie International Edition* **58**, 259-263 (2019).
43. Qiu J*, et al.* Prediction and understanding of AIE effect by quantum mechanics-aided machine-learning algorithm. *Chemical Communications* **54**, 7955-7958 (2018).
44. Reichardt C. Solvatochromic Dyes as Solvent Polarity Indicators. *Chemical Reviews* **94**, 2319-2358 (1994).
45. 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* **113**, 5951-5960 (2009).
46. Breiman L. Random forests. *Machine Learning* **45**, 5-32 (2001).
47. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology* **2**, Article 27 (2011).
48. Cortes C, Mohri M, Rostamizadeh A. Learning non-linear combinations of kernels.
49. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* **521**, 436-444 (2015).
50. Ke G*, et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 3146--3154 (2017).
51. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232 (2001).
52. 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* (eds Chackalamannil S, Rotella D, Ward SE). Elsevier (2017).
53. Sandfort F, Strieth-Kalthoff F, Kühnemund M, Beecks C, Glorius F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem*, doi: 10.1016/j.chempr.2020.02.017 (2020).

54. Li G, Hirano T, Yamada K. Bright near-infrared chemiluminescent dyes: Phthalhydrazides conjugated with fluorescent BODIPYs. *Dyes and Pigments* **178**, 108339 (2020).
55. Liu J, Zhu L, Wan W, Huang X. Gold-Catalyzed Oxidative Cascade Cyclization of 1,3-Diynamides: Polycyclic N-Heterocycle Synthesis via Construction of a Furopyridinyl Core. *Organic Letters*, doi: 10.1021/acs.orglett.0c01086 (2020).
56. Pei K*, et al.* Highly fluorescence emissive 5, 5′-distyryl-3, 3′-bithiophenes: Synthesis, crystal structure, optoelectronic and thermal properties. *Dyes and Pigments* **179**, 108396 (2020).
57. Grimme S, Neese F. Double-hybrid density functional theory for excited electronic states of molecules. *The Journal of Chemical Physics* **127**, 154116 (2007).
58. Pedregosa F*, et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825-2830 (2011).
59. Dyomics.(https://dyomics.com/fileadmin/uploads/Redakteur/EN/Downloads/Dyomics_2017.pdf)
60. Fluorophore. (http://www.fluorophores.tugraz.at/)
61. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry* **32**, 1466-1474 (2011).
62. Landrum G. RDKit: Open-source cheminformatics. (2006).
63. Ju C-W, Liu R, Bai H, Li B. ChemFluor *figshare* (https://dx.doi.org/10.6084/m9.figshare.12110619.v3)