Prediction of molecular electronic transitions using random forests

Beomchang Kang,[†] Chaok Seok,[†] and Juyong Lee[*,‡]

[†]*Department of Chemistry, Seoul National University, 08826, Seoul, Republic of Korea*

[‡]*Division of Chemistry and Biochemistry, Department of Chemistry, Kangwon National University, 24341, Chuncheon, Republic of Korea*

E-mail: juyong.lee@kangwon.ac.kr

Abstract

Fluorescent molecules, fluorophores, play essential roles in bioimaging. Attachment of fluorophores to proteins enables observation of the detailed structure and dynamics of biological reactions occurring in the cell. Effective bioimaging requires fluorophores with high quantum yields to detect weak signals. Besides, fluorophores with various emission frequencies are necessary to extract richer information. An essential computational component to discover novel functional molecules is to predict molecular properties. Here, we present statistical machines that predict excitation energies and associated oscillator strengths of a given molecule using a random forest algorithm. Excitation energies and oscillator strengths are directly related to the emission spectrum and the quantum yields of fluorophores, respectively. We discovered specific molecular substructures and fragments that determine the oscillator strengths of molecules from the feature importance analysis of our random forest machine. This discovery is expected to serve as a new design principle for novel fluorophores.

Introduction

Fluorophores play essential roles in diverse fields such as biochemistry, analytical chemistry, medicine, and spectroscopy.[1–4] For example, they allow us to observe protein-protein interactions and to screen toxic compounds at the molecular level.[5,6] However, only a limited number of fluorophores are commonly used. The discovery of novel fluorophores may open new possibilities to observe unseen biological reactions. To find novel and favorable fluorophores, it is essential to identify their photophysical properties accurately, such as absorption, emission wavelength, absorption coefficient, and quantum yield.[7]

Conventional approaches to developing fluorophores have limitations. Novel fluorophores have been discovered by modifying existing ones based on human experts' chemical intuition. Despite decades of endeavors by experimental chemists, only a few fluorophores are commonly used, such as fluorescein,[8] BODIPY,[9] cyanine,[10] and Seoul-Fluor.[7,11] For example, various BODIPY derivatives are discovered by modifying the s-indacene core structure.[9] Seoul-Fluor is developed based on the indolizine core.[7,11] These molecules are designed to have distinct emission frequencies by attaching diverse functional groups to their core structure.[7,9,11] With traditional approaches, finding a novel molecular scaffold takes considerable time and resources to synthesize and verify candidate compounds.[12]

Theoretical approaches such as quantum mechanical (QM) calculation or machine learning (ML) may solve this problem more efficiently than conventional approaches. Diverse QM approaches have been developed to predict the photophysical properties of molecules.[13–18] Most previous studies focused on predicting the photophysical properties of inorganic solids. They do not guarantee accurate property prediction of fluorophores used in biomedical researches.[13–17] Sumita and coworkers used density-functional theory (DFT) with the B3LYP hybrid functional and the 3-21G* basis set to evaluate the photophysical properties of molecules.[18] One critical limitation of QM approaches is that they need more considerable computation resources than ML-based approaches.[19]

Although there are a few studies to predict the electronic properties of small organic molecules with ML approaches, their prediction accuracy was inferior to that of QM approaches.[19–21] Ghosh and coworkers utilized multiple deep-learning (DL) methods to train a machine that predicts the electronic properties of molecules from their structures as inputs.[19] They used multi-layer perceptron (MLP),[21] convolutional neural network (CNN),[22] and deep tensor neural network (DTNN)[17] to predict the density of states from the DFT results of the QM7[23,24] and QM9[25,26] data set.[19] Nakata and coworkers utilized the support vector machine (SVM) and the ridge regression to predict the HOMO-LUMO gaps of molecules.[20] The root mean squared error (RMSE) between QM values and prediction value of their models spans from 0.36 to 0.48 eV.[20] Note that prediction errors of ML studies are similar to that of the time-dependent density functional theory (TD-DFT) benchmarks against experiments.[27,28] To the best of our knowledge, previous ML approaches predicted excitation energies of a molecule, but not oscillator strengths.[19,20]

Montavon and coworkers predicted the maximum absorption intensity and corresponding excitation energy from time-dependent Hartree-Fock results using a deep neural network.[21,29] The RMSE of excitation energy of maximum intensity predicted with the existing deep-learning model (1.71 eV) exceeds the range (1.45 eV) of visible light (1.65-3.10 eV). This amount of prediction error appears to be rather large for practical molecular design tasks.[21]

Here, we developed statistical models that predict the electronic transitions of a molecule using random forests (RF) algorithm[30] from SMILES input.[31] RF is known to have low overfitting risk and high accuracy.[32] Another advantage of RF over DL is that it facilitates the analysis of relative feature importance.[30,33] In this study, we identified specific fragments that play essential roles in determining oscillator strengths of electronic transitions of molecules. The performance of our model is encouraging for the discovery of novel fluorophores.

Our model was trained with the TD-DFT results of about half a million molecules. The highest oscillator strength and associated excitation energy among ten excitation states of molecules are used to train the model. Molecular fingerprints and various molecular property descriptors are used as input features for our model. From the RF results, molecular fragments that play essential roles in determining the molecular electronic properties are identified. Chemical insights into critical factors determining the electronic properties of molecules will be discussed.

Method

Overview of the Workflow

In this study, the target properties for predictions are the excitation energy and oscillator strength of the transition with the maximum oscillator strength among the first ten excited states. The training and test sets contain 450,000 and 50,000 molecules, respectively. A schematic diagram of the workflow is shown in Figure 1.

Two independent RF machines were trained to predict the maximum oscillator strength, and the other RF machine was trained to predict the excitation energy coupled with the maximum oscillator strength. We utilized the RF method implemented in the Scikit-learn 0.21.3 library.[34] The feature vector of a molecule consists of the extended connectivity fingerprints (ECFP),[35] molecular access system (MACCS) keys,[36] and various molecular descriptors provided by RDkit.[37] The obtained feature vectors were used as the input of the RF models.

Molecular Data Set

All data used in this study are from the PubChemQC database.[20] PubChemQC contains the quantum mechanical calculation results of almost four million molecules in PubChem[38] using the TD-DFT methods implemented in GAMESS.[39,40] PubChemQC provides the ground-state electronic molecular structures optimized by DFT at the B3LYP/6-31G* level and the ten low-lying excited state information calculated by TD-DFT with B3LYP/6-31+G*. A half-million molecules were randomly selected from a subset of molecules that consist only of H, B, C, N, O, F, P, S, Cl atoms with no net charge. Among the molecules in the selected set, 450,000 molecules were used as a training set, and the rest as a test set.

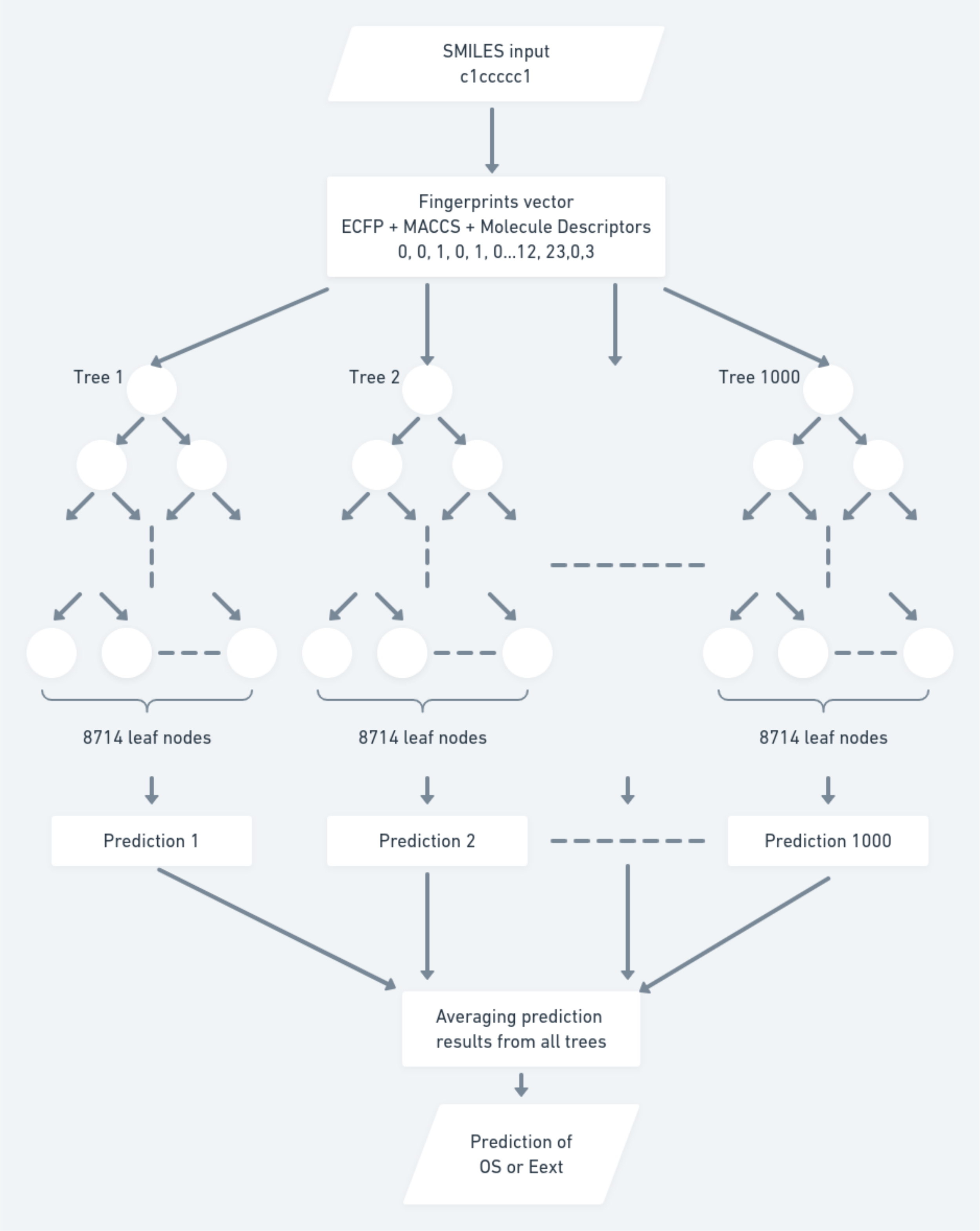


Figure 1: Overview of the algorithm. First, SMILES representations are converted to fingerprints vector (combination of ECFP, MACCS keys, molecular descriptors calculated with RDkit).

Target Properties and Molecular Features

We use the dimensionless quantity oscillator strength (f_{LU}) that represents the rate of absorption to deal with the excitation intensity.[41] We focus on the transition of the highest oscillator strength because it corresponds to the peak of absorption spectra. Oscillator strength is defined as follows:[42]

$$f_{LU} = \frac{2m_e}{3\hbar^2}(E_U - E_L)|\langle \psi_L d_L|\mathbf{r}|\psi_U d_U\rangle|^2 \tag{1}$$

where m_e is the mass of an electron, L and U denote the lower and upper state of transition, and d_k and ψ_k represent the degeneracy and wave function of state k.

Two fingerprint methods—ECFP[35] and MACCS keys[36]—were used to generate a feature vector from a molecule. In addition to the fingerprints, various molecular descriptors calculated by RDkit were included[37] as well. These molecular features were generated from the SMILES representations of molecules.[31,43] The used molecular descriptors are listed in supplementary information. ECFP is a topological fingerprint for molecular characterization that accounts for the relationships between molecular substructure efficiently. The generation procedure of ECFP systematically records the neighborhood of each non-hydrogen atom into multiple circular layers up to a given diameter as integer identifiers.[35] All these atom-centered substructural identifiers are then mapped into a fixed-length binary representation using a hashing function.[35] The MACCS keys are binary codes that identify the presence of 166 pre-determined molecular sub-structures. Additional molecular descriptors related to the overall topology or the physicochemical properties of a molecule, such as molecular weight or the fraction of sp3 carbon, were included in input features.

Two types of feature vectors were tested in this study. The first vector includes 4096 bits from ECFP4 (of diameter 4), 166 MACCS keys and, 39 molecular descriptors from RDkit, resulting in a 4031-dimensional vector. The second vector includes 16384 bits solely from ECFP6 (of diameter 6) to discover bigger fragments, which determine the electronic properties of molecules. We call the first vector ”Vector 1” and the second vector ”Vector 2”.

Random Forest Regression

To predict the oscillator strength and the excitation energy, random forests regression (RFR), an ensemble learning method was applied as schematically shown in Figure 1. It operates by constructing multiple decision trees during training and averaging the prediction values of all trees for final prediction.[44,45] Each tree starts from the root node encompassing all data points, and nodes are divided into two new sub-nodes. Splits are decided to minimize mean square error (MSE). We performed extensive searches of hyperparameters (the number of trees, leaves, and the ratio of features). Extension of each tree ended when the number of used features for the decision reached one-third of the dimension of the input vector. Our final RF model consists of 1000 trees and 8714 leaf nodes. Four models were constructed in total. Two models were trained with Vector 1—ECFP4 4096 bits, MACCS keys, and molecular descriptors—to predict the excitation energy and oscillator strength of a molecule independently. The other two models are trained with ECFP6 16386 bits to identify larger fragments.

Feature Importance Analysis

From the trained RF models, the importance of each feature in performing prediction was extracted. Molecular fragments that have high contributions to the oscillator strength of a molecule were identified from the ECFP. Also, critical molecular substructures defined in the MACCS keys and molecular descriptors were identified. To analyze the feature importance, a fingerprint with a longer diameter (ECFP6) and more vector bits (16384) were used to search for huger and more diverse fragments. The four bits with the highest feature importance were selected to discover important fragments that induce high oscillator strength. All fragments corresponding to each selected bit were obtained by the reverse process of the generation. All molecules that include these fragments were then searched from the PubChemQC database. Finally, eight fragments were selected.

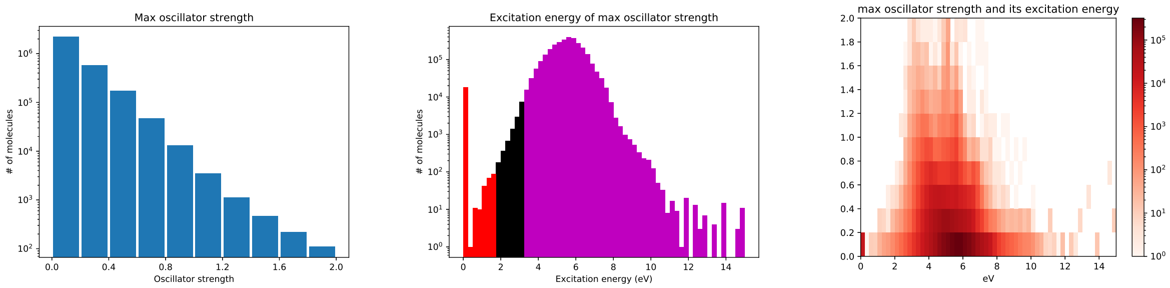
Results and discussion

Distribution of the Highest Intensity Transitions from PubChemQC

Analysis of the PubchemQC database shows that it is not trivial to discover favorable fluorophores with desired properties (Figure 2). Favorable fluorophores are required to have high oscillator strength, which may lead to a strong emission spectrum, and diverse emission colors. About a half of PubChemQC molecules have the maximum oscillator strength less of than 0.1 (Figure 2a). Only 4.12 and 0.18 percent of molecules have oscillator strength over 0.5 and 1.0, respectively. In terms of excitation energy, most molecules have their maximum intensity absorption in the ultraviolet region (Figure 2b). Only 0.4 percent of molecules in the database have their absorption peak in the visible region. Taken together, the number of molecules whose maximum intensity absorption in the visible range and oscillator strength over 0.5 is only 880, corresponding to 0.03% of the dataset. This result demonstrates that highly accurate electronic property prediction methods are essential for designing a novel fluorophore efficiently.

Hyperparameter Optimization

Five hyperparameters of the model were optimized. Two hyperparameters are related to input vector generation: the diameter and the number of bits of ECFP. Three hyperparameters are related to the random forest model: the number of trees, the number of data points in each leaf node, and the ratio of tested features to find the best split. Except for the number of data points per leaf nodes, increasing the size of the model did not improve the performances of models (Table S1 and Table S2). After extensive hyperparameter optimization, ECFP4 with 4096 bits, 1000 trees, 1/3 feature ratio, and 1000 data points per leaf



(a) Oscillator strength

(b) Excitation energy

(c) Heat map in log-scale

Figure 2: Distribution of the maximum oscillator strength transition of molecules in PubChemQC. (a) Distribution of (a) the maximum oscillator strength, (b) the excitation energy of max oscillator strength are colored in red, black, and violet if they correspond to the IR, visible and UV region, respectively. (c) The heat map showing the distribution of excitation energy and oscillator strength in log-scale.

nodes were used to train the model.

Prediction Accuracy of the Highest Intensity Transition

The target properties of our model are maximum oscillator strength and the corresponding excitation energy of a molecule. One of the most important properties to be a favorable fluorophore is high-intensity emission in the visible light range. However, the number of emission spectrum data is still not enough for applying deep-learning or other state-of-the-art approaches.[46,47] Additionally, the quantum calculation of emission spectrum is computationally more complicated than that of the absorption spectrum because the geometry optimization of excited states is less accurate than that of the ground state.[47,48] Because the wavelength of absorption and emission are correlated in general,[43] we used the absorption spectrum as a proxy of the emission spectrum.

The correlation coefficient between our prediction and TD-DFT calculation exceeds 0.8 for both oscillator strength and excitation energy(Table 1). The RMSE of oscillator strength and excitation energy predictions with vector1 are 0.088 and 0.448 eV, respectively (Table 1). Figures 3b and 4b show that most molecules are distributed near the diagonal line, representing the perfect prediction. For 96 percent of molecules, the error of oscillator strength between prediction and TD-DFT is less than 0.2 (yellow lines in Figure 3a). In terms of excitation energy prediction, 80 percent of predictions have errors less than 0.5 eV (yellow lines in Figure 4).

To the best of our knowledge, our models are the most accurate models in predicting the oscillator strength and the excitation energy at the highest intensity transition of organic compounds. Two existing approaches predicted excitation spectra based on the HOMO-LUMO gap,[19,20] not the highest intensity transition, which makes a direct comparison with our model hard. Only Montavon and coworkers developed a model that predicts the highest intensity transition by using fully connected hidden layers with a multi-task neural network.[21] They optimized the 3D geometry of molecules using a DFT method starting from SMILES. The atom-pair matrix of Coulomb interaction was used as input. The training was carried out with 5000 molecules, and the number of the hidden layers of the model was four. The RMSE values of their model were 0.12 and 1.76 eV for the maximum oscillator strength and corresponding excitation energy, respectively. The accuracy of our model outperforms their model in terms of both oscillator strength and excitation energy prediction. Notably, the RMSE of our model in excitation energy prediction is only one-third of their model.

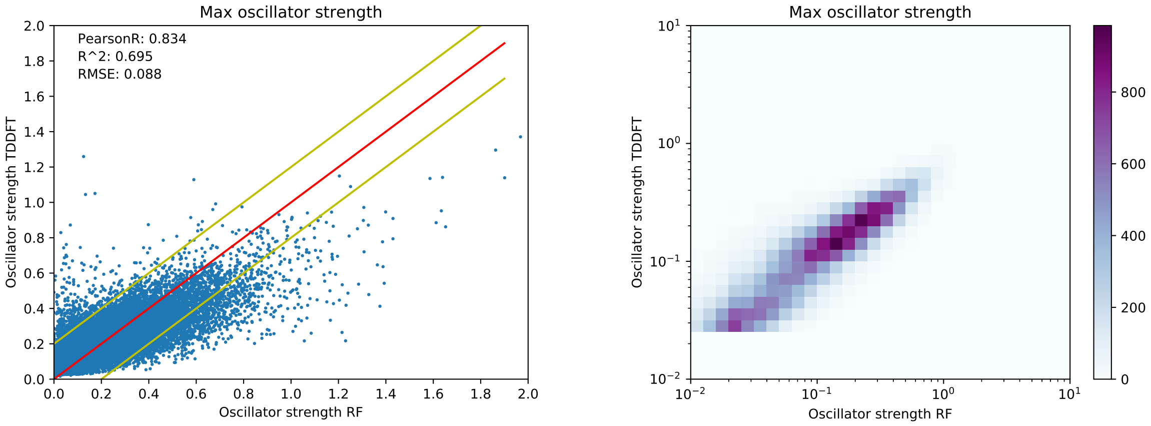
There are three possible explanations for the high performance of our model. First, the size of our training set is 100 times larger than that of the previous study. It may decrease the risk of overfitting to a small set of molecules. Second, non-Coulomb interactions that were not considered by an atom-pair Coulomb interaction matrices may carry important information. The ECFP, MACCS keys, and molecular descriptors are able to capture local environments including the effects of non-Coulomb interactions. Third, the depth of layers in their model was four, which may be too shallow to learn the electronic properties accurately. In contrast, the number of leaf nodes and the number of trees in our model are saturated (Table S1 and Table S2).

It is worth noting that the error of our models is comparable to that of TD-DFT quantum calculation against the experiment.[27,28,49] Fabian and coworkers reported that the mean

Table 1: Root mean square error (RMSE) and Pearson R between TD-DFT and random forest (RF) regression comparing to Motavon's DL (Deep Learning) model.[21]

Model	Feature	Quantum property	RMSE	Pearson R	Computation time[e]
RF	Vector 1[a]	Excitation energy (eV)	0.448	0.881	27 min
RF	Vector 1[a]	Oscillator strength	0.088	0.834	23 min
DL	Matrix[b]	Excitation energy (eV)	1.76[d]		
DL	Matrix[b]	Oscillator strength	0.12[d]		
RF	Vector 2[c]	Excitation energy	0.447	0.882	14h 2min
RF	Vector 2[c]	Oscillator strength	0.092	0.814	13h 18min

[a] ECFP4 4096 bits + 166 MACCS keys + 39 descriptors;
[b] Atom-pair Coulomb potential matrix; [c] ECFP6 16384 bits;
[d] Results come from Motavon's paper;[21]
[e] Intel Xeon CPU E5-2650 v4 2.20 GHz 24 core, 48 processors, 128 GB memory.



(a) Scatter plot.

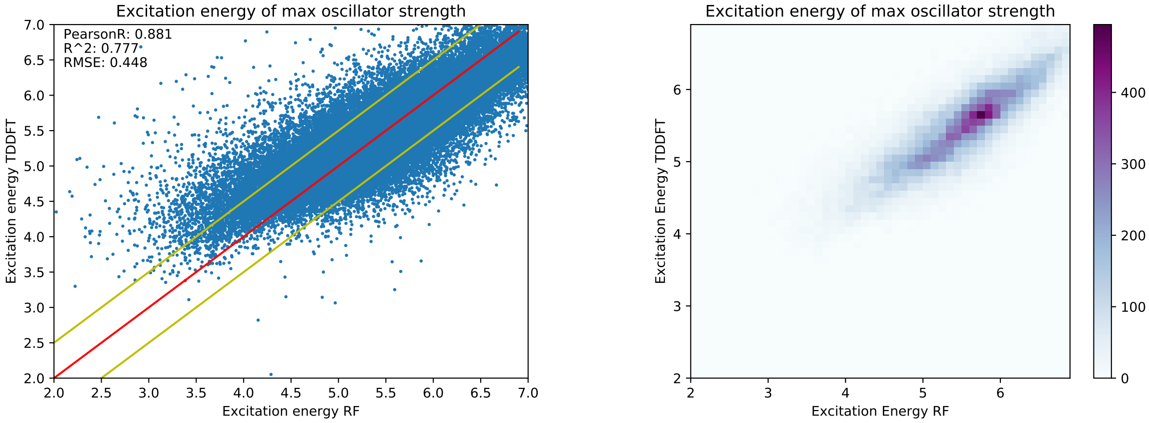
(b) Distribution of molecules.

Figure 3: Prediction results of the maximum oscillator strength. The X-axis represents our RF prediction results. The Y-axis represents TD-DFT results. (a) Scatter plot. (b) Heat map. Only 0.6% of molecules are predicted to have oscillator strength below 0.01

absolute deviation of excitation energy between TD-DFT predictions with B3LYP functional and the 6-31+G* basis set compared to the experiment is 0.27 eV.

Fragments contributing to high oscillator strength

We identified molecular fragments that determine high maximum oscillator strength. To discover bigger and more diverse molecular fragments, we extracted information from the RF results using ECFP6 with 16384 bits. To confirm that the oscillator strength distribu-



(a) Scatter plot.

(b) Distribution of molecules.

Figure 4: Accuracy of excitation energy (eV) of maximum oscillator strength prediction. The X-axis represents our RF prediction results. The Y-axis represents TD-DFT results. (a) Scatter plot. (b) Heat map. Only 0.6% and 2.3% of molecules are predicted to have excitation energy of maximum intensity lower than 2.0 eV and higher than 7.0 eV, respectively.

tions of the identified fragments are statistically higher than the average, the p-value of the distribution was calculated by the Wilcoxon rank sum test. The fragments satisfying the following conditions were selected for further analysis; 1) the p-value is lower than 0.0001, 2) the mean oscillator strength of molecules containing the fragment is over 0.3, and 3) the number of molecules containing the fragment is over 10. The eight fragments satisfying the above criteria were retrieved from the PubchemQC database. These fragments are presented in the ascending order of p-value, descending order of statistical significance (Figure 5).

Overall, it is identified that nitrogen-containing heterocycles are playing important roles in inducing high oscillator strength of a molecule. The first six statistically significant fragments have one or more nitrogen atoms inside a ring structure. The most statistically significant fragment contains an indazole-like structure, the fragment c(cc)(c(c)c)n(C)n (Figure 5a). The 69 molecules containing the indazole-like structure have the lowest p-value of 1.158*10^{-28} and the second highest mean oscillator strength, 0.415 (Figure 5). 1-Methylbenzo[f]indazole has the highest oscillator strength among the compounds that contain fragment c(cc)(c(c)c)n(C)n (Figure 6a). The maximum oscillator strength and excitation energy of the molecule are 1.430 and 5.381 eV, respectively. The highest intensity electron transition arises from the second-highest occupied MO (Figure 6c) to LUMO (Figure 6d). The little overlap between the two orbitals suggests that a significant amount of electron density change is associated with the transition, which leads to strong absorption of light.

The second statistically significant fragment is a pyzazine-like fragment, c1nc(C)cnc1NC (Figure 5c). The fragment is found in 45 molecules, and their mean oscillator strength is 0.471, which is the highest among the eight identified fragments. The third and fourth significant fragments are a triazole-like structure (4H-Triazol-4-ylmethanamine), and 1-acetylpyrrolidine-like structure, respectively. They both are contained in 26 molecules in the dataset and have oscillator strength over 0.35.

The rest four fragments are relatively less significant and included in a smaller number of molecules, less than 20. The fifth and sixth fragments are quinoxaline-like and 1-phenylpyrazole-like fragments with the mean oscillator strength of 0.371 and 0.344, respectively. Unlike the other fragments, the last two fragments, mesitylene and 2-methylhex-1-en-4-yne, do not have nitrogen inside its scaffold.

Molecular Properties inducing High Oscillator Strength

Our RF models provide which molecular descriptors are significantly related to the electronic property of a molecule. The most critical, highest feature importance, eleven molecular descriptors are illustrated in Figure 7.

In the previous section, it is identified that all fragments inducing high oscillator strength have many π conjugation bonds. It is consistent with the fact that many dye molecules have a large portion of the conjugated π-system. The most critical molecular descriptor determining oscillator strength is the fraction of sp^3 carbon atoms. It is in good agreement with the fact that many known fluorophores have π conjugation systems. The number of stereo-centers and the existence of carbon-carbon double bonds are closely related to the fraction of sp^3

carbon atoms. In other words, it implies that a higher sp^3 carbon fraction leads to a low oscillator strength. The $^1\kappa$ shape index, the molecular cyclicity, is ranked in the second place. The index is defined as follows:

$$^1\kappa = A(A-1)^2/(B)^2 \tag{2}$$

where A is the number of atoms, and B is the number of bonds. The cyclicity indicates the relative portion of cyclic/branched structures inside a molecule[50,51] The index becomes maximum if a molecule is in a linear shape and minimum if a molecular graph forms a complete graph. The fact that molecular cyclicity is the second most crucial feature indicates that a large portion of aromatic rings is essential for high oscillator strength.

The next critical molecular features are as follows: the maximum partial charge, the minimum absolute partial charge, the existence of -CH_2N, and the existence of nitrogen connected aromatic rings. These properties appear to contribute to the high oscillator strength through charge separation of a molecule because nitrogen atoms donate lone pair electrons. It is known that charge separation is closely correlated with light-induced electron transfer.[52]

Conclusion

In this work, we developed random forest models that predict the maximum oscillator strength and corresponding excitation energy of a molecule. Our models trained with almost one million DFT results yielded highly accurate predictions. In addition to the high prediction accuracy compared to the existing prediction models, this study presents its importance at one of the few theoretical studies on the prediction of the oscillator strength of a molecule. We further identified molecular fragments and molecular descriptors important for determining the oscillator strength of a molecule. We believe that these findings will serve as a useful guideline for the design of novel fluorophores.

Acknowledgement

Supporting Information Available

Hyperparameter Optimization

Table S1: Hyperparameter optimization results for oscillator strength prediction

ECFP		Random Forest			Result		
Diameter	Bits	Trees	Data/leaf	Feature ratio	RMSE	Pearson R	Training time
4	2048	1000	50	1/3	0.089	0.83	0d 1hr 34min
4	2048	1000	50	2/3	0.089	0.83	0d 2hr 53min
4	2048	1000	100	1/3	0.092	0.82	0d 1hr 32min
4	2048	1000	100	2/3	0.091	0.82	0d 2hr 40min
4	2048	10000	50	1/3	0.089	0.83	0d 10hr 50min
4	2048	10000	50	2/3	0.089	0.83	1d 0hr 34min
4	2048	10000	100	1/3	0.091	0.82	0d 11hr 24min
4	2048	10000	100	2/3	0.091	0.82	1d 13hr 12min
4	4096	1000	50	1/3	0.088	0.83	0d 3hr 11min
4	4096	1000	50	2/3	0.088	0.83	0d 5hr 33min
4	4096	1000	100	1/3	0.091	0.82	0d 2hr 54min
4	4096	1000	100	2/3	0.091	0.82	0d 6hr 36min
4	4096	10000	50	1/3	0.088	0.83	1d 2hr 47min
4	4096	10000	50	2/3	0.088	0.83	2d 1hr 3min
4	4096	10000	100	1/3	0.091	0.82	0d 22hr 18min
4	4096	10000	100	2/3	0.091	0.82	1d 8hr 48min
4	8192	1000	50	1/3	0.088	0.84	0d 6hr 59min

continued from previous page

ECFP		Random Forest			Result		
Diameter	Bits	Trees	Data/leaf	Feature ratio	RMSE	Pearson R	Training time
4	8192	1000	50	2/3	0.088	0.83	0d 9hr 24min
4	8192	1000	100	1/3	0.090	0.82	0d 6hr 21min
4	8192	1000	100	2/3	0.090	0.82	0d 9hr 18min
4	8192	10000	50	1/3	0.088	0.84	2d 5hr 56min
4	8192	10000	50	2/3	0.088	0.83	9d 0hr 7min
4	8192	10000	100	1/3	0.090	0.82	2d 19hr 35min
4	8192	10000	100	2/3	0.090	0.82	4d 3hr 3min
6	2048	1000	50	1/3	0.090	0.83	0d 1hr 37min
6	2048	1000	50	2/3	0.090	0.83	0d 2hr 46min
6	2048	1000	100	1/3	0.092	0.82	0d 1hr 15min
6	2048	1000	100	2/3	0.092	0.82	0d 2hr 36min
6	2048	10000	50	1/3	0.090	0.83	0d 12hr 34min
6	2048	10000	50	2/3	0.089	0.83	0d 19hr 48min
6	2048	10000	100	1/3	0.092	0.82	0d 11hr 8min
6	2048	10000	100	2/3	0.092	0.82	0d 21hr 31min
6	4096	1000	50	1/3	0.089	0.83	0d 3hr 9min
6	4096	1000	50	2/3	0.089	0.83	0d 5hr 46min
6	4096	1000	100	1/3	0.092	0.82	0d 2hr 54min
6	4096	1000	100	2/3	0.091	0.82	0d 7hr 42min
6	4096	10000	50	1/3	0.089	0.83	1d 5hr 33min
6	4096	10000	50	2/3	0.089	0.83	3d 7hr 3min
6	4096	10000	100	1/3	0.092	0.82	1d 0hr 39min
6	4096	10000	100	2/3	0.091	0.82	2d 2hr 3min
6	8192	1000	50	1/3	0.088	0.83	0d 5hr 59min

continued from previous page

ECFP		Random Forest			Result		
Diameter	Bits	Trees	Data/leaf	Feature ratio	RMSE	Pearson R	Training time
6	8192	1000	50	2/3	0.088	0.83	0d 13hr 56min
6	8192	1000	100	1/3	0.091	0.82	0d 4hr 58min
6	8192	1000	100	2/3	0.091	0.82	0d 18hr 6min
6	8192	10000	50	1/3	0.088	0.83	1d 23hr 56min
6	8192	10000	50	2/3	0.088	0.83	4d 11hr 10min
6	8192	10000	100	1/3	0.091	0.82	3d 3hr 19min
6	8192	10000	100	2/3	0.091	0.82	5d 23hr 48min

Table S2: Hyperparameter training of excitation energy

ECFP		Random Forest			Result		
Diameter	Bits	Trees	Data/leaf	Feature ratio	RMSE	Pearson R	Training time
4	2048	1000	50	1/3	0.45	0.88	0d 1hr 29min
4	2048	1000	50	2/3	0.45	0.88	0d 2hr 31min
4	2048	1000	100	1/3	0.46	0.87	0d 1hr 40min
4	2048	1000	100	2/3	0.46	0.87	0d 4hr 32min
4	2048	10000	50	1/3	0.45	0.88	0d 12hr 22min
4	2048	10000	50	2/3	0.45	0.88	0d 22hr 24min
4	2048	10000	100	1/3	0.46	0.87	0d 12hr 53min
4	2048	10000	100	2/3	0.46	0.87	1d 1hr 3min
4	4096	1000	50	1/3	0.45	0.88	0d 2hr 55min
4	4096	1000	50	2/3	0.45	0.88	0d 5hr 40min
4	4096	1000	100	1/3	0.46	0.87	0d 3hr 3min

continued from previous page

ECFP		Random Forest			Result		
Diameter	Bits	Trees	Data/leaf	Feature ratio	RMSE	Pearson R	Training time
4	4096	1000	100	2/3	0.46	0.87	0d 5hr 21min
4	4096	10000	50	1/3	0.45	0.88	1d 12hr 44min
4	4096	10000	50	2/3	0.45	0.88	1d 23hr 41min
4	4096	10000	100	1/3	0.46	0.87	1d 1hr 50min
4	4096	10000	100	2/3	0.46	0.87	2d 2hr 47min
4	8192	1000	50	1/3	0.45	0.88	0d 7hr 13min
4	8192	1000	50	2/3	0.45	0.88	0d 10hr 14min
4	8192	1000	100	1/3	0.46	0.88	0d 9hr 48min
4	8192	1000	100	2/3	0.46	0.88	0d 13hr 30min
4	8192	10000	50	1/3	0.45	0.88	2d 12hr 58min
4	8192	10000	50	2/3	0.45	0.88	6d 18hr 48min
4	8192	10000	100	1/3	0.46	0.88	2d 0hr 11min
4	8192	10000	100	2/3	0.46	0.88	5d 4hr 5min
6	2048	1000	50	1/3	0.45	0.88	0d 1hr 34min
6	2048	1000	50	2/3	0.45	0.88	0d 4hr 27min
6	2048	1000	100	1/3	0.46	0.87	0d 2hr 19min
6	2048	1000	100	2/3	0.46	0.87	0d 2hr 54min
6	2048	10000	50	1/3	0.45	0.88	0d 14hr 3min
6	2048	10000	50	2/3	0.45	0.88	0d 21hr 15min
6	2048	10000	100	1/3	0.46	0.87	0d 12hr 25min
6	2048	10000	100	2/3	0.46	0.87	0d 21hr 53min
6	4096	1000	50	1/3	0.45	0.88	0d 3hr 10min
6	4096	1000	50	2/3	0.45	0.88	0d 5hr 11min
6	4096	1000	100	1/3	0.46	0.87	0d 3hr 2min

continued from previous page

ECFP		Random Forest			Result		
Diameter	Bits	Trees	Data/leaf	Feature ratio	RMSE	Pearson R	Training time
6	4096	1000	100	2/3	0.46	0.87	0d 5hr 50min
6	4096	10000	50	1/3	0.45	0.88	1d 9hr 58min
6	4096	10000	50	2/3	0.45	0.88	2d 14hr 18min
6	4096	10000	100	1/3	0.46	0.87	0d 20hr 44min
6	4096	10000	100	2/3	0.46	0.87	2d 0hr 48min
6	8192	1000	50	1/3	0.45	0.88	0d 7hr 25min
6	8192	1000	50	2/3	0.45	0.88	0d 11hr 18min
6	8192	1000	100	1/3	0.46	0.87	0d 6hr 26min
6	8192	1000	100	2/3	0.46	0.87	0d 23hr 34min
6	8192	10000	50	1/3	0.45	0.88	2d 2hr 28min
6	8192	10000	50	2/3	0.45	0.88	4d 18hr 19min
6	8192	10000	100	1/3	0.46	0.87	2d 3hr 42min
6	8192	10000	100	2/3	0.46	0.87	3d 4hr 54min

MACCS keys used for RF training

MACCS 166 keys information.[53]

Atom symbols

A : Any valid periodic table element symbol,

Q : Hetero-atoms; any non-C or non-H atom

X : Halogens; F, Cl, Br, I

Z : Others; other than H, C, N, O, Si, P, S, F, Cl, Br, I

Bond types - : Single, = : Double, T : Triple, # : Triple, : Single or double query bond, % : An aromatic query bond, $: Ring bond; $ before a bond type specifies ring bond, ! :

Chain or non-ring bond; ! before a bond type specifies chain bond

@ : A ring linkage and the number following it specifies the atoms position in the line, thus @1 means linked back to the first atom in the list.

ISOTOPE, 103 < ATOMIC NO. < 256, GROUP IVA,VA,VIA PERIODS 4-6 (Ge...), ACTINIDE, GROUP IIIB,IVB (Sc...), LANTHANIDE, GROUP VB,VIB,VIIB (V...), QAAA@1, GROUP VIII (Fe...), GROUP IIA (ALKALINE EARTH), 4M RING, GROUP IB,IIB (Cu...), ON(C)C, S-S, OC(O)O, QAA@1, CTC, GROUP IIIA (B...), 7M RING, SI, C=C(Q)Q, 3M RING, NC(O)O, N-O, NC(N)N, C$=C($A)$A, I, QCH2Q, P, CQ(C)(C)A, QX, CSN, NS, CH2=A, GROUP IA (ALKALI METAL), S HETEROCYCLE, NC(O)N, NC(C)N, OS(O)O, S-O, CTN, F, QHAQH, OTHER, C=CN, BR, SAN, OQ(O)O, CHARGE, C=C(C)C, CSO, NN, QHAAAQH, QHAAQH, OSO, ON(O)C, O HETEROCYCLE, QSQ, Snot%A%A, S=O, AS(A)A, A$A!A$A, N=O, A$A!S, C%N, CC(C)(C)A, QS, QHQH (&...), QQH, QNQ, NO, OAAO, S=A, CH3ACH3, A!N$A, C=C(A)A, NAN, C=N, NAAN, NAAAN, SA(A)A, ACH2QH, QAAAA@1, NH2, CN(C)C, CH2QCH2, X!A$A, S, OAAAO, QHAACH2A, QHAAACH2A, OC(N)C, QCH3, QN, NAAO, 5M RING, NAAAO, QAAAAA@1, C=C, ACH2N, 8M RING, QO, CL, QHACH2A, A$A($A)$A, QA(Q)Q, XA(A)A, CH3AAACH2A, ACH2O, NCO, NACH2A, AA(A)(A)A, Onot%A%A, CH3CH2A, CH3ACH2A, CH3AACH2A, NAO, ACH2CH2A > 1, N=A, HETEROCYCLIC ATOM > 1 (&...), N HETEROCYCLE, AN(A)A, OCO, QQ, AROMATIC RING > 1, A!O!A, A$A!O > 1 (&...), ACH2AAACH2A, ACH2AACH2A, QQ > 1 (&...), QH > 1, OACH2A, A$A!N, X (HALOGEN), Nnot%A%A, O=A > 1, HETEROCYCLE, QCH2A > 1 (&...), OH, O > 3 (&...), CH3 > 2 (&...), N > 1, A$A!O, Anot%A%Anot%A, 6M RING > 1, O > 2, ACH2CH2A, AQ(A)A, CH3 > 1, A!A$A!A, NH, OC(C)C, QCH2A, C=O, A!CH2!A, NA(A)A, C-O, C-N, O > 1, CH3, N, AROMATIC, 6M RING, O, RING, FRAGMENTS

Molecular Descriptors used for RF training

Morgan fingerprint density 1, 2, 3, Molecular weight, Heavy atom weight, Max absolute partial charge, Max partial charge, Min absolute partial charge, Min partial charge, Number of radical electrons, Number of valence electrons, Fraction of SP3 carbon, kappa shape index 1, 2, 3, Labute ASA, Number of aliphatic carbocycles, Number of aliphatic heterocycles, Number of aliphatic rings, Number of amide bonds, Number of aromatic carbocycles, Number of aromatic heterocycles, Number of aromatic rings, Number of stereocenters, Number of bridgehead atoms, Number of HB acceptors, Number of HB donor, Number of heteroatoms, Number of heterocycles, Number of Lipinski HB acceptors, Number of Lipinski HB donors, Number of rings, Number of rotatable bonds, Number of saturated carbocycles, Number of saturated heterocycles, Number of saturated rings, Number of spiroatoms, Number of unspecified atom stereocenters, Topological SA.

50 Molecular Properties inducing High Oscillator Strength

Table S3: The input features with the fifty highest importance values among the MACCSkeys and molecular descriptors provided by RDkit.

Descriptors	Feature importance
Fraction of SP3 carbon	0.14312
Kappa1	0.04294
Exist of aromatic ring	0.03969
Number of aromatic ring > 1	0.02887
Mumber of stereocenters	0.02294
Member of unspecified atom stereocenter	0.02269
Exist of C=C	0.01568
Max partial charge	0.01486

continued from previous page

Descriptors	Feature importance
Min absolute partial charge	0.01303
Exist of CH2N	0.01233
Exist of nitrogen connected to aromatic ring	0.01046
Exist of C-N	0.00949
Min partial charge	0.0081
Max absolute partial charge	0.00751
Exist of C that has a double bond with C and 2 single bonds	0.0074
Morgan density 1	0.00739
Morgan density 3	0.00735
Morgan density 2	0.007
Kappa2	0.00633
Kappa3	0.00632
Exist of two rings connected by bond	0.00603
Topological surface area	0.00544
Labute accessible surface area	0.00497
Exist of aromatic rings with 2 functional group(1,2)	0.00463
Exist of -CH2-	0.00455
Number of radical electrons	0.00448
Heavy atom weight	0.00419
Number of aromatic rings	0.00413
Molecular weight	0.00398
ACH2AAACH2A[a]	0.00382
NACH2A[a]	0.0038
Exist of F	0.00373

continued from previous page

Descriptors	Feature importance
Exist of CH2 connected to herero atom	0.00367
Number of Hydrogen bond acceptors	0.00352
Exist of 8-membered ring	0.00293
Exist of C=C(C)C	0.00293
Number of valence electrons	0.00292
Exist of nitrogen connected to aromatic ring with double bond	0.00285
QAAAA@1[a]	0.00281
Number of rotatable bonds	0.00261
Exist of 5-membered ring	0.00256
6-membered ring > 1	0.00249
Exist of rings with 2 functional group(1,2)	0.00219
Number of Lipinski hydorgen bond acceptors	0.00211
Exist of 2rings sharing 1 bond	0.00211
Number of Lipinski hydorgen bond donors	0.00207
Exist of nitrogen connected to ring	0.00205
Number of aromatic heterocycles	0.00199
Number of aromatic carbocycles	0.00182
Number of heteroatoms	0.0018

[a] A : Any valid periodic table element symbol, Q : Hetero-atoms,

@ : A ring linkage and the number n following it specifies the atoms position in the line;

References

(1) Cartlidge, E. The light fantastic. *Science* **2018**, *359*, 382–385.

(2) Zinchuk, V.; Grossenbacher-Zinchuk, O. Recent advances in quantitative colocalization

analysis: Focus on neuroscience. *Progress in Histochemistry and Cytochemistry* **2009**, *44*, 125–172.

(3) Murphy, K. R.; Stedmon, C. A.; Wenig, P.; Bro, R. OpenFluor– an online spectral library of auto-fluorescence by organic compounds in the environment. *Anal. Methods* **2014**, *6*, 658–661.

(4) Evanko, D. A flaky but useful fluorophore. *Nature Methods* **2005**, *2*, 160–161.

(5) Moczko, E.; Mirkes, E. M.; Cáceres, C.; Gorban, A. N.; Piletsky, S. Fluorescence-based assay as a new screening tool for toxic chemicals. *Scientific Reports* **2016**, *6*.

(6) Martin, S. F.; Tatham, M. H.; Hay, R. T.; Samuel, I. D. Quantitative analysis of multi-protein interactions using FRET: Application to the SUMO pathway. *Protein Science* **2008**, *17*, 777–784.

(7) Kim, E.; Lee, Y.; Lee, S.; Park, S. B. Discovery, Understanding, and Bioapplication of Organic Fluorophore: A Case Study with an Indolizine-Based Novel Fluorophore, Seoul-Fluor. *Accounts of Chemical Research* **2015**, *48*, 538–547.

(8) Kobayashi, H.; Ogawa, M.; Alford, R.; Choyke, P. L.; Urano, Y. New Strategies for Fluorescent Probe Design in Medical Diagnostic Imaging. *Chemical Reviews* **2010**, *110*, 2620–2640.

(9) Loudet, A.; Burgess, K. BODIPY Dyes and Their Derivatives: Syntheses and Spectroscopic Properties. *Chemical Reviews* **2007**, *107*, 4891–4932.

(10) Mishra, A.; Behera, R. K.; Behera, P. K.; Mishra, B. K.; Behera, G. B. Cyanines during the 1990s: A Review. *Chemical Reviews* **2000**, *100*, 1973–2012.

(11) Choi, E. J.; Kim, E.; Lee, Y.; Jo, A.; Park, S. B. Rational Perturbation of the Fluorescence Quantum Yield in Emission-Tunable and Predictable Fluorophores (Seoul-

Fluors) by a Facile Synthetic Method Involving CH Activation. *Angewandte Chemie International Edition* **2014**, *53*, 1346–1350.

(12) Song, H.-O.; Lee, B.; Bhusal, R. P.; Park, B.; Yu, K.; Chong, C.-K.; Cho, P.; Kim, S. Y.; Kim, H. S.; Park, H. Development of a Novel Fluorophore for Real-Time Biomonitoring System. *PLoS ONE* **2012**, *7*, e48459.

(13) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine learning bandgaps of double perovskites. *Scientific Reports* **2016**, *6*.

(14) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*.

(15) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B* **2016**, *93*.

(16) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **2017**, *8*.

(17) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **2014**, *89*.

(18) Sumita, M.; Yang, X.; Ishihara, S.; Tamura, R.; Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Central Science* **2018**, *4*, 1126–1133.

(19) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Advanced Science* **2019**, *6*, 1801367.

(20) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling* **2017**, *57*, 1300–1308.

(21) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. c. *New Journal of Physics* **2013**, *15*, 095003.

(22) Antonellis, G.; Gavras, A. G.; Panagiotou, M.; Kutter, B. L.; Guerrini, G.; Sander, A. C.; Fox, P. J. Shake Table Test of Large-Scale Bridge Columns Supported on Rocking Shallow Foundations. *Journal of Geotechnical and Geoenvironmental Engineering* **2015**, *141*, 04015009.

(23) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters* **2012**, *108*, 058301.

(24) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732.

(25) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.

(26) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*.

(27) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. Extensive TD-DFT Benchmark: Singlet-Excited States of Organic Molecules. *Journal of Chemical Theory and Computation* **2009**, *5*, 2420–2435.

(28) Laurent, A. D.; Jacquemin, D. TD-DFT benchmarks: A review. *International Journal of Quantum Chemistry* **2013**, *113*, 2019–2039.

(29) López, C. S.; Faza, O. N.; Estévez, S. L.; de Lera, A. R. Computation of vertical excitation energies of retinal and analogs: Scope and limitations. *Journal of Computational Chemistry* **2006**, *27*, 116–123.

(30) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(31) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.

(32) Davies, A.; Ghahramani, Z. The Random Forest Kernel and other kernels for big data from random partitions. 2014.

(33) Kim, S.; Jun, S. AI Technology Analysis using Variable Importance of Deep Learning. *Journal of Korean Institute of Intelligent Systems* **2019**, *29*, 70–75.

(34) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

(36) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.

(37) Landrum, G. RDKit: Open-Source Cheminformatics Software. **2016**,

(38) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Research* **2015**, *44*, D1202–D1213.

(39) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular electronic structure system. *Journal of Computational Chemistry* **1993**, *14*, 1347–1363.

(40) Gordon, M. S.; Schmidt, M. W. *Theory and Applications of Computational Chemistry*; Elsevier, 2005; pp 1167–1189.

(41) Hilborn, R. C. Einstein coefficients, cross sections, f values, dipole moments, and all that. *American Journal of Physics* **1982**, *50*, 982–986.

(42) Zheng, L.; Polizzi, N. F.; Dave, A. R.; Migliore, A.; Beratan, D. N. Where Is the Electronic Oscillator Strength? Mapping Oscillator Strength across Molecular Absorption Spectra. *The Journal of Physical Chemistry A* **2016**, *120*, 1933–1943.

(43) Band, Y. B.; Heller, D. F. Relationships between the absorption and emission of light in multilevel systems. *Physical Review A* **1988**, *38*, 1885–1895.

(44) Ho, T. K. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995; pp 278–282 vol.1.

(45) Ho, T. K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1998**, *20*, 832–844.

(46) Mewes, S. A.; Plasser, F.; Krylov, A.; Dreuw, A. Benchmarking Excited-State Calculations Using Exciton Properties. *Journal of Chemical Theory and Computation* **2018**, *14*, 710–725.

(47) Brémond, E.; Savarese, M.; Adamo, C.; Jacquemin, D. Accuracy of TD-DFT Geometries: A Fresh Look. *Journal of Chemical Theory and Computation* **2018**, *14*, 3715–3727.

(48) Adamo, C.; Jacquemin, D. The calculations of excited-state properties with Time-Dependent Density Functional Theory. *Chem. Soc. Rev.* **2013**, *42*, 845–856.

(49) Fabian, J.; Diaz, L.; Seifert, G.; Niehaus, T. Calculation of excitation energies of organic chromophores: a critical evaluation. *Journal of Molecular Structure: THEOCHEM* **2002**, *594*, 41–53.

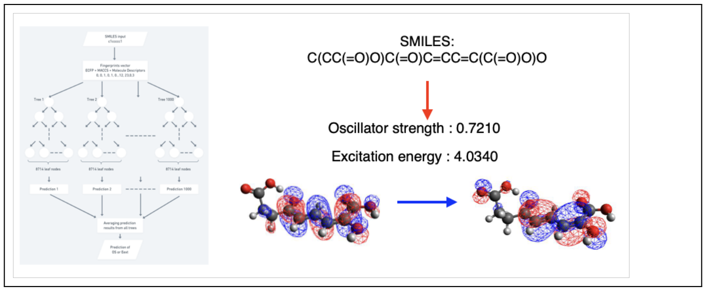
(50) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quantitative Structure-Activity Relationships* **1986**, *5*, 1–7.

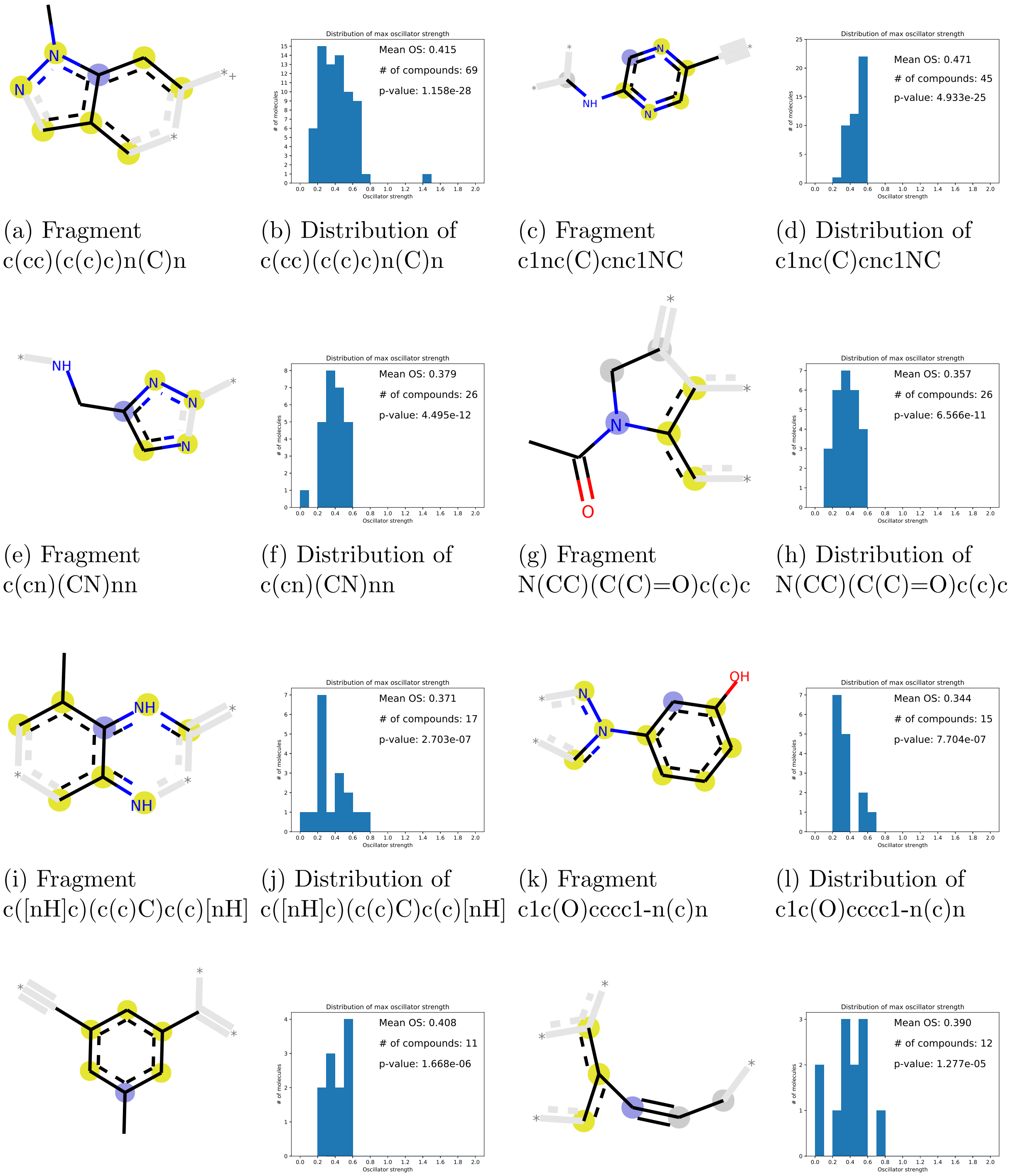
(51) Hall, L. H.; Kier, L. B. *Reviews in Computational Chemistry*; John Wiley & Sons, Inc., 2007; pp 367–422.

(52) Mochizuki, K.; Shi, L.; Mizukami, S.; Yamanaka, M.; Tanabe, M.; Gong, W.-T.; Palonpon, A. F.; Kawano, S.; Kawata, S.; Kikuchi, K.; Fujita, K. Nonlinear fluorescence imaging by photoinduced charge separation. *Japanese Journal of Applied Physics* **2015**, *54*, 042403.

(53) Sud, M. MayaChemTools: An Open Source Package for Computational Drug Discovery. *Journal of Chemical Information and Modeling* **2016**, *56*, 2292–2297.

Graphical TOC Entry





(a) Fragment c(cc)(c(c)c)n(C)n

(b) Distribution of c(cc)(c(c)c)n(C)n

(c) Fragment c1nc(C)cnc1NC

(d) Distribution of c1nc(C)cnc1NC

(e) Fragment c(cn)(CN)nn

(f) Distribution of c(cn)(CN)nn

(g) Fragment N(CC)(C(C)=O)c(c)c

(h) Distribution of N(CC)(C(C)=O)c(c)c

(i) Fragment c([nH]c)(c(c)C)c(c)[nH]

(j) Distribution of c([nH]c)(c(c)C)c(c)[nH]

(k) Fragment c1c(O)cccc1-n(c)n

(l) Distribution of c1c(O)cccc1-n(c)n

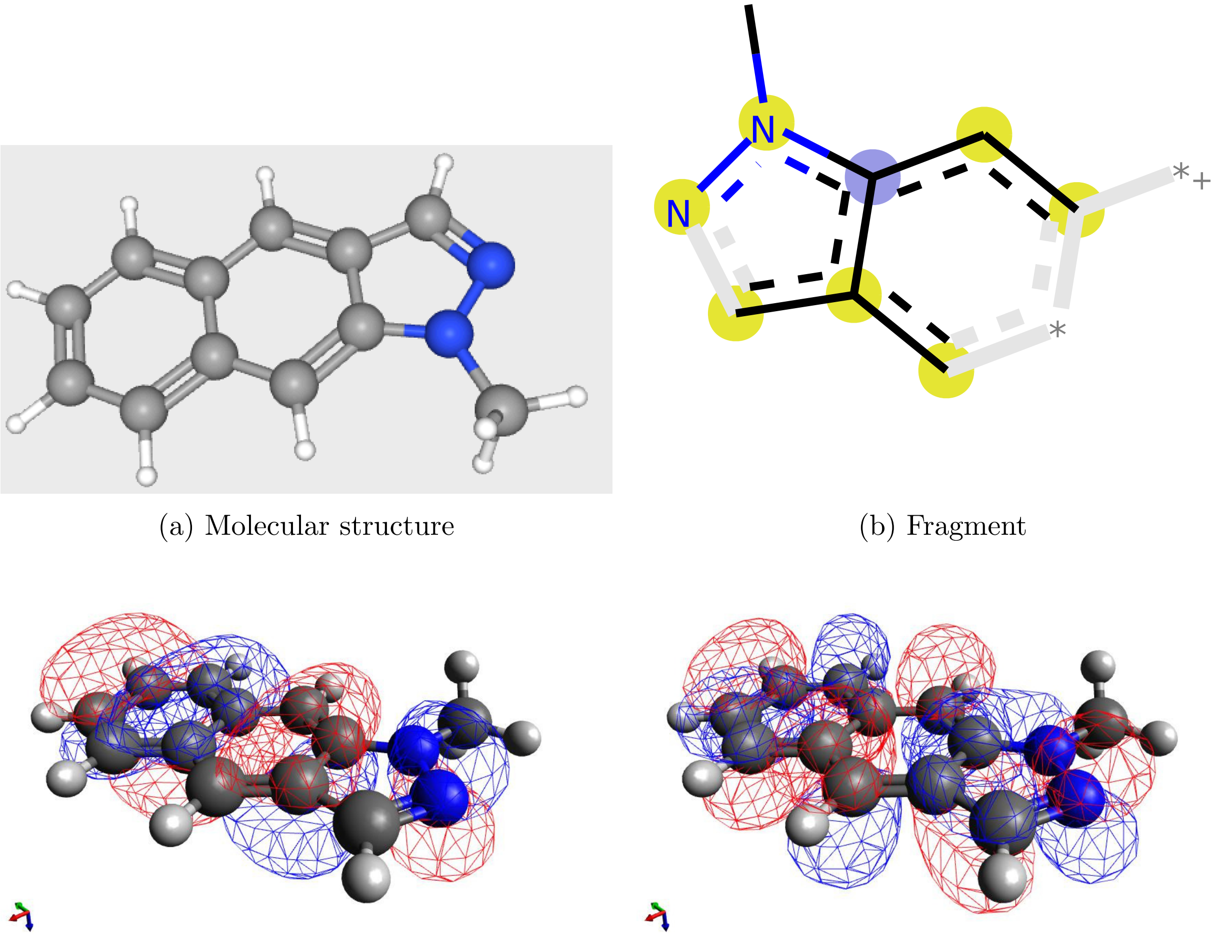
(m) Fragment c1(C)cc(C)cc(C)c1

(n) Distribution of c1(C)cc(C)cc(C)c1

(o) Fragment C(#CC)c(c)c

(p) Distribution of C(#CC)c(c)c

Figure 5: Eight molecular fragments that induce high oscillator strength. The first and third columns present the structures of molecular fragments. The second and fourth columns present the histogram of oscillator strength of molecules containing the fragment on the left side.



(a) Molecular structure

(b) Fragment

(c) Lower state (HOMO-1)

(d) Upper state (LUMO)

Figure 6: An example of a molecule containing the indazole-like fragment: 1-methyl-benzo[f]indazole. (a) Molecular structure (b) The fragment that is expected to induce high oscillator strength. (c) Lower state (the second highest occupied MO) and (d) upper state (LUMO) of the maximum oscillator strength excitation.

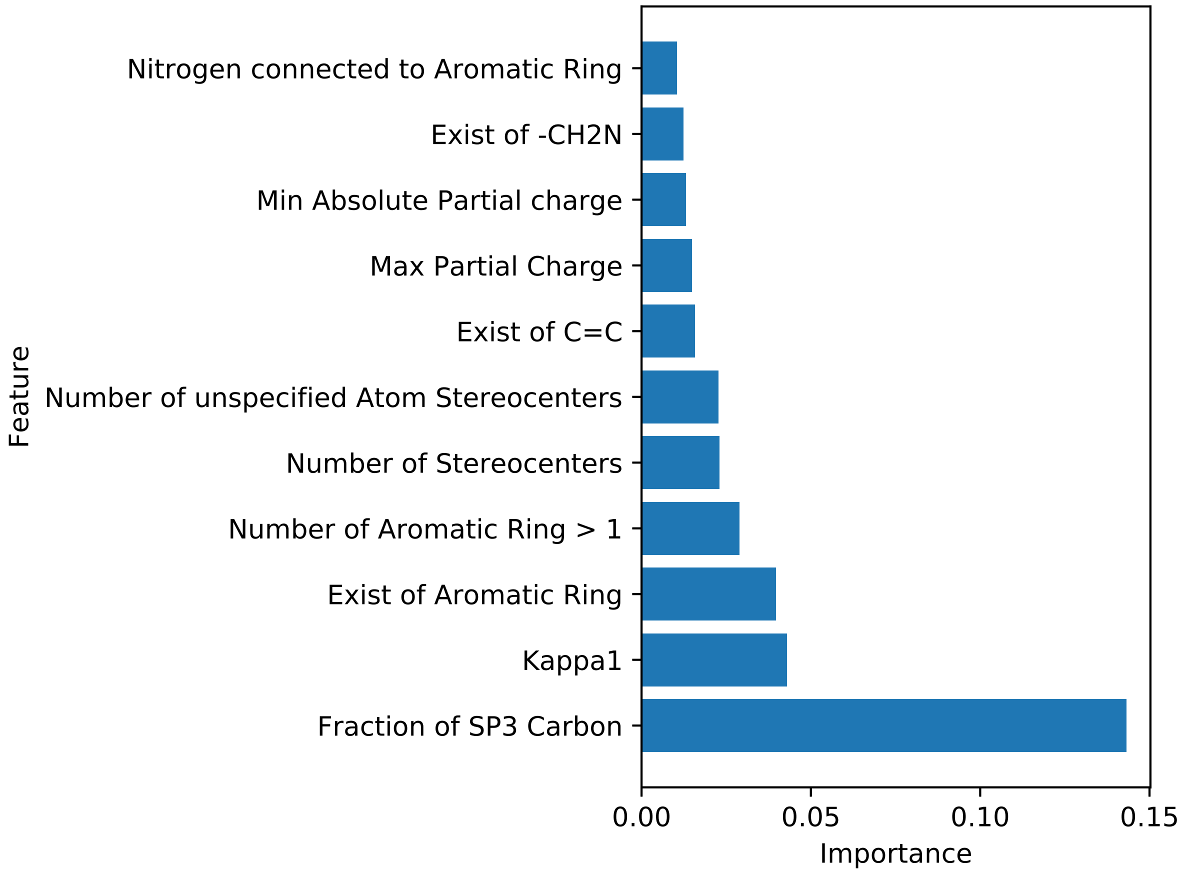


Figure 7: The input features with the eleven highest importance values among the MACCS keys and molecular descriptors provided by RDkit. Only the features whose importance is over 0.1 are plotted.