Ensemble Machine Learning and Applicability Domain Estimation for Fluorescence Properties and its Application to Structural Design

Yuki Sugawara, Masaaki Kotera, Kenichi Tanaka, Kimito Funatsu*

Department of Chemical System Engineering, School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN

(Received April 9, 2019; Accepted May 29, 2019)

Fluorescent substances are used in a wide range of applications, and the method that effectively design molecules having desirable absorption and emission wavelength is required. In this study, we used boron-dipyrromethene (BODIPY) compounds as a case study, and constructed high precision wavelength prediction model using ensemble learning. Prediction accuracy improved in stacking model using RDKit descriptors and Morgan fingerprint. The variables related to the molecular skeleton and the conjugation length were shown to be important. We also proposed an applicability domain (AD) estimation model that directly use the descriptors based on Tanimoto distance. The performance of the AD models was shown better than the OCSVM-based model. Using our proposed stacking model and AD model, newly generated compounds were screened and we obtained 602 compounds which were estimated inside the AD in both absorption wavelength and emission wavelength.

Key Words: QSPR, fluorescent material, ensemble learning, applicability domain

Introduction

Fluorescent substances are used in a wide range of applications such as dye lasers, organic electro-luminescences, fluorescent labelings of cells. Development of novel fluorescent substances with desirable absorption and emission wavelengths for each application have been carried out by synthetic chemists, but trial and error by experiments is costly in terms of time and money. Attempts have been made to reduce costs by performing quantum chemistry computation before conducting experiments. However, it is difficult to apply it to a large number of candidate compounds because of high calculation cost. In addition, many quantum calculations use the solvent-free state, and also it is necessary to correct the deviation between measured and calculated values of the quantum calculation [1].

In recent years, with increasing compound data and the development of chemoinformatic techniques, quantitative structure-property relationship (QSPR) studies are widely used and expressed the relationship between chemical structures and physical properties. Machine learning models enable efficient QSPR, and
physical properties can be predicted at high speed even for unknown compounds.

QSPR studies for fluorescent materials are also being conducted. Chen et al. [2] predicted absorption and emission wavelength with high precision by adding quantum calculation results and solvent information as descriptors. Among the fluorescent substances, those having boron-dipyrromethene (BODIPY) in the molecular skeleton have a sharp spectrum [3] and stably high quantum yield [4], and the influence of solvent to the wavelength is small. Schüller et al. [5] constructed a QSPR model for compounds with BODIPY skeleton by combining multiple variable selection methods.

In this study, we constructed an ensemble QSPR model (Fig. 1) proposing new BODIPY compounds with desirable chemical properties within appropriate applicability domain (AD). AD refers to the region in the descriptor space where the prediction model can predict with high accuracy. We used RDKit descriptors [6] and Morgan fingerprints [7] as explanatory variables, and constructed a highly accurate wavelength prediction model using an ensemble learning [8,9] named stacking [10]. We compared the predictive performance with other machine learning methods and also with quantum computation. Furthermore, we developed the AD model to estimate the reliability of the newly generated compounds.

**Data and methods**

**Data**

We obtained BODIPY compounds reported in the previous paper [4]. The data contained compounds that had no basic BODIPY skeleton, which were excluded from the dataset. Finally, we obtained 143 and 141 compounds with measured values of absorption and emission wavelength, respectively.

In addition to the RDKit descriptors [6] and Morgan fingerprint [7] as described below, we designed a categorical variable representing the rough type of BODIPY skeletons and used it as the descriptor for the first layer of stacking. Most compounds (100 and 99 for absorption and emission wavelength, respectively) have the first type of BODIPY skeleton, and some others (16 for both absorption and emission wavelength) have the BODIPY skeleton fused to two rings (Fig. 2). Other compounds (27 and 26 for absorption and emission wavelength, respectively) have diverse structures. These skeleton information was represented as BODIPY_category descriptor.

![Figure 1. BODIPY compounds and our proposed prediction model](image-url)
RDKit descriptors and Morgan fingerprint

We used RDKit descriptors (version 2017.03.2) [6] that contain 196 descriptors such as the molecular weight and the calculated value of log P that are calculated from the entire molecule, as well as local structural information of the molecule such as the number of benzene rings and that of fused rings.

A fingerprint is regarded as a set of descriptors. A fingerprint typically represents presence or absence of chemical substructures in a compound as a bit (1 or 0) vector. A fingerprint can also represent the count number of chemical substructures in a compound as an integer vector. In this study, we used the latter type, i.e., a count-based fingerprint. Specifically, we used Morgan fingerprint [7], which dynamically determines the partial structure depending on the data set. Morgan fingerprint targets all substructures until the radius reaches from the center (distance 0) of each atom of the structure. Because of the large number of dimensions, Morgan fingerprint is often used after reducing the dimension such as 512, 1024, 2048 using the hash function. Hash function allots some different substructures to the same valuable, which is referred to as the hash collision. Since the hash collision is considered to have bad effect for prediction [11], we calculated unhashed Morgan fingerprint, as well as hashed Morgan fingerprint.

Ensemble methods: averaging and stacking

An ensemble learning [9,10] is a generic term for an integrated method consisting of more than two prediction models in its first layer, and its second layer uses the predicted values from the first layer and make the final decision. The simplest ensemble, referred to as “averaging”, outputs the average of the predicted values from the first layer. The more complex ensemble method, referred to as “stacking” [10], applies another prediction models in the second layer. We compared the predictive performances of averaging and stacking in this study.

A schematic diagram of the proposed stacking method is shown in Fig 1. The first layer took RDKit descriptors and count-based Morgan fingerprint (radius = 4, hashed or unhashed) as input, and conducted the following 10 prediction models:

- four linear models
- five decision tree methods
- Random Forest: RF [15]
- Extra Trees: ET [16]
- Bagging Regressor: BAG [17]
- Gradient Boosting: GBT [18,19]
- AdaBoost: ADA [18]
- k nearest neighbor (kNN) [20]

The second layer used the 20 predicted values (from 10 models for RDKit descriptors and count-based Morgan fingerprint) from the first layer, and applied a linear model (RI) and a nonlinear model (ET), outputting the average of those models.

Evaluation

For estimation of prediction accuracy, Mean Absolute Error (MAE) and coefficient of determination $R^2$ were used (equations 1 and 2, respectively). MAE has a value greater than or equal to 0, and indicate that the closer to 0 represents better prediction performance, $R^2$ takes a value less than or equal to 1, and indicates that closer to 1 better prediction performance.

$$\text{MAE} [\text{nm}] = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

Hyper-parameter tuning and validation

Table 1 lists the hyper-parameters tuned in this study. Other hyper-parameters were set to the default values as described in the user’s manual of the Python scikit-learn library. The hyper-parameter was determined using RandomizedSearchCV of scikit-learn.

When the data size is sufficiently large, external validation is available, where the data is divided into three (i.e. training, validation and test datasets). However, when the data size is small, the sizes of divided datasets become smaller. Small training datasets cause unstable models, and small test datasets cause unreliable evaluation. In this study, since the dataset was very small, we applied a cross model validation (CMV), which consists of a nest structure of two cross validation schemes. Specifically, we used leave-one-out cross model validation (LOO-CMV) to train the first layer, in which the outer and inner cross validation schemes apply LOO and five-fold cross validations, respectively. Hyper-parameter tuning was conducted in the inner cross validations.

The second layer did not apply any parameter tuning,
and we adjusted the n_estimators parameter to 300. We used LOO to validate the second layer.

**Performance comparison**

We compared several different conditions to clarify the effect on predictive performances. The first comparison was made between single and ensemble machine learning models. The second was between hashed and unhashed Morgan fingerprints. The third was between the separate use and the concatenated vector of RDKit descriptors and Morgan fingerprints. Finally, the precision of our proposed model for novel compounds was evaluated by comparing with the prediction using quantum calculation.

**Generation of new BODIPY compounds**

All known BODIPY compounds were divided into BODIPY skeleton and side chains. We obtained 15 different side chains. A BODIPY ring has seven bondable positions, among which two sites were selected and two side chains were joined by allowing overlapping from the 15 side chains. We conducted this process to generate all possible compounds, removed the structure if it is already included in the training data. As a result, we obtained 4410 new structures having BODIPY skeleton.

**Applicability Domain (AD)**

Applicability Domain (AD) represents the coverage of the prediction model. We investigated AD of our proposed model to evaluate the newly generated BODIPY compounds. A number of previous studies investigated quantitative AD: (1) range setting of each variable or latent variable after dimensional reduction [21] by Principal Component Analysis (PCA) [22], (2) distance from neighboring point [23] using kNN [20], (3) data density estimation [24] by OneClassSVM (OCSVM) [25], and (4) variation of predicted values [26] by ensemble learning [8,9].

| Table 1. Hyperparameters tuned in this study |
|---------------------------------------------|
| Method | Parameter names | Values |
| RF     | max_depth       | [None, 3, 5, 7, 9] |
|        | max_features    | ['auto', 0.5, 0.7, 0.9, 1.0] |
|        | max_leaf_nodes  | [None, 2, 4, 6, 10] |
|        | min_samples_split | [2, 4, 6, 10] |
|        | n_estimators    | [200] |
| GBT    | learning_rate   | [0.05, 0.08, 0.1, 0.12] |
|        | max_depth       | [None, 3, 5, 9] |
|        | max_leaf_nodes  | [None, 2, 6, 10] |
|        | min_samples_split | [2, 6, 10] |
|        | n_estimators    | [100, 200] |
| BAG    | bootstrap_features | [False, True] |
|        | max_features    | [0.5, 0.7, 0.9, 1.0] |
|        | max_samples     | [0.5, 0.75, 1.0] |
|        | n_estimators    | [200] |
| EX     | bootstrap_features | [False, True] |
|        | max_depth       | [None, 3, 5, 9] |
|        | max_leaf_nodes  | [None, 2, 6, 10] |
|        | min_samples_split | [2, 6, 10] |
|        | n_estimators    | [200] |
| ADA    | learning_rate   | [0.3, 0.5, 0.8, 1.0, 1.5] |
|        | n_estimators    | [300] |
| RI     | alpha           | [0.5, 1.0, 2, 4, 8, 16, 32, 64, 128, 256] |
| LA     | alpha           | [0.0078125, 0.015625, 0.03125, 0.0625, 0.125, 0.25, 0.5, 1.0, 2, 4] |
| kNN    | n_neighbors     | [2, 3, 5, 7, 10, 15] |
| PLS    | n_components    | [2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20] |
In this study, we proposed to use Tanimoto distance of binary unhashed Morgan fingerprint (radius = 4) with the training data as the AD estimation model. First, Tanimoto distances were calculated for each compound against all compounds in the training data. Consequently, the average distance of the N closest compounds was regarded as the AD score of the generated compound. For comparison, Tanimoto distance was calculated for all combination within the training data. The reference AD scores were sorted in ascending order, and the AD score at the top k% was used as the threshold. If the AD score of newly generated structure was equal to or less than the threshold value, the new structure was estimated to be inside AD.

In order to evaluate the effectiveness of the proposed AD estimation method, we compared it with data density estimation by Gaussian kernel OCSVM [25] that is commonly used in the field of chemoinformatics. In OCSVM, the prediction model is considered to be reliable in the region where the density of the training data is high in the descriptor space. Since the direct use of OCSVM was not considered to be effective, we applied dimension reduction by PCA, and the obtained principal components were subjected to OCSVM.

Results and Discussion

Performance comparison

We evaluated the prediction performance of the single models that combined two types of descriptors, i.e., 196 RDKit descriptors and count-based hashed Morgan fingerprint (radius = 4, 1024 bits), and the ensemble models (Table 2). It was observed that GBT performed the best in terms of the highest $R^2$ score in both RDKit descriptors ("R" in Table 2c) or hashed/unhashed Morgan fingerprint ("hM" and "uM" in Table 2c), for both analyses of absorption wavelength and emission wavelength. Decision tree-based models were shown to be higher prediction accuracy than linear models, implying that few variables in the descriptors had linear relationship with the objective wavelength. When using fingerprint, RF performed better than ET; RF does and ET does not select variables internally, indicating that some bits in the fingerprint are unnecessary for the prediction. More importantly, we confirmed that stacking performed the highest performance despite that it contains low precision single models, whilst averaging performed worse than some single models (Table 2b and 2d). This was supposed to be because all models were dealt with equally in the averaging model. As shown in the Fig. 3 of averaging and the best single model for both absorption wavelength and emission wavelength, the predicted value was smaller than the measured value in high wavelength region, and the predicted value was larger than the measured value in low wavelength region. In contrast, it was shown that stacking especially improved in the prediction in high and low wavelength region, as well as in medium wavelength region.

Next, we made a comparison between hashed and unhashed Morgan fingerprints. In contrast to the 1024 bits in the hashed Morgan fingerprint, the unhashed Morgan fingerprint obtained in this study consisted of 3579 bits, meaning that about 3.5 substructures were allotted to one bit in the hashed Morgan fingerprint on average. We found that unhashed Morgan fingerprint performed better than hashed Morgan fingerprint ("hM" and "uM" in Table 2a and 2c) in RF and ET but not in other decision tree-based models, although the difference was not significant. Unhashed Morgan fingerprint showed worse performance in kNN. The use of unhashed Morgan fingerprint improved the prediction accuracy in the linear model, especially the LA model.

The observed fact that unhashed Morgan fingerprint did not perform better than hashed Morgan fingerprint in RF was different from the result of the previous study [11]. This is probably because we did not use binary but count-based fingerprint in this study. It was not clearly written which type of fingerprint was used in the previous study [11], but binary fingerprints have been commonly used. We suspect this difference also comes from the difference of the compound sets, because we only considered BODIPY compounds and thus there should be some common bits in all datasets in this study. Since fingerprint is a sparse vector, a decision tree requires many nodes to divide the data set, easily causing overfitting, which may be the reason why unhashed Morgan fingerprint did not show significant improvement, such as the case in ET. We were able to adjust the tree division points for the best division in other decision tree-based models. In contrast, ET randomly decides the variable and the division points, therefore the probability that ET choses variables effective for prediction is very low when using unhashed Morgan fingerprint. The reason why the accuracy decreased with kNN is thought to be a curse of dimensionality; the distance between each data cannot be evaluated appropriately for high dimension vectors. Linear model directly estimates the relationship between explainable and objective variables, therefore the effect of removing the hash collision was very large, which was why unhashed Morgan fingerprint performed better in all linear models. Especially, LA is known to generally make the coefficients of many unnecessary explanatory variables 0, and it is suitable for sparse vectors such as fingerprints.

The final prediction of stacking performed the best in terms of $R^2$ in absorption wavelength, and in MAE and $R^2$ in emission wavelength (Table 2b and 2d, respectively). Similar to the result presented above, averaging did not perform well in the high and low wavelength regions. The LA model, which was the best single model, also showed similar tendency, but the prediction of some compounds were greatly out.
Consequent analysis compared the separate use of RDKit descriptors and Morgan fingerprint ("R+hM" and "R+uM" in Table 2b and 2d) with the concatenated vector of the two ("CRuM" in Table 2). Significant difference was not observed when using single models ("CRuM" in Table 2a and 2c), however, the concatenated vector performed worse when using stacking ("CRuM" in Table 2b and 2d). This result was understandable because the model diversity is known to be important in ensemble learning. Considering the diversity of models, it was suggested that higher accuracy could be achieved by building an ensemble model for each group of semantically consistent features within a descriptor.

As shown in Table 2, the predictive performances of RDKit descriptors and Morgan fingerprint were not significantly different when using single models. The former describes the chemical characteristics of the entire molecule and some substructure information, and the latter describes the comprehensive substructure information. This knowledge might be useful when the users want to select descriptors, however, the achievement of this study is to show the advantage of using the "stacking" ensemble model, rather than comparing RDKit descriptors and Morgan fingerprint.

### Table 2. Performance comparison of prediction models. R: RDKit descriptors, hM: hashed count-based Morgan fingerprint, uM: unhashed count-based Morgan fingerprint, CRuM: concatenated vector of RDKit descriptors and unhashed count-based Morgan fingerprint. Since we applied LOO because of the small dataset, standard deviations could not be calculated.

#### (a) MAE of single models

|          | Absorption wavelength | Emission wavelength |
|----------|-----------------------|---------------------|
|          | R       | hM     | uM     | CRuM     | R       | hM     | uM     | CRuM     |
| RF       | 23.1    | 18.8   | 20.2   | 21.7     | 27.6    | 20.6   | 21.1   | 21.7     |
| ET       | 23.9    | 23.8   | 26.3   | 23.7     | 25.8    | 23.7   | 27.4   | 23.7     |
| BAG      | 22.9    | 19.0   | 18.4   | 21.8     | 23.8    | 20.0   | 18.0   | 21.8     |
| ADA      | 25.3    | 28.8   | 27.3   | 25.4     | 27.8    | 26.6   | 28.3   | 25.4     |
| GBT      | 21.5    | 19.0   | 19.5   | 20.5     | 24.7    | 19.4   | 19.7   | 20.5     |
| kNN      | 26.0    | 24.4   | 29.5   | 29.8     | 27.4    | 27.3   | 32.3   | 29.8     |
| RI       | 33.9    | 28.3   | 25.6   | 24.2     | 34.5    | 28.6   | 26.2   | 24.2     |
| LA       | 25.3    | 31.2   | 18.2   | 18.4     | 27.1    | 25.4   | 17.2   | 18.4     |
| PLS      | 37.4    | 29.2   | 27.1   | 25.7     | 39.3    | 29.3   | 28.1   | 25.7     |
| BAY      | 32.1    | 28.5   | 25.2   | 23.8     | 31.6    | 28.5   | 25.6   | 23.8     |

#### (b) MAE of ensemble models

|          | Absorption wavelength | Emission wavelength |
|----------|-----------------------|---------------------|
|          | R+hM     | R+uM     | CRuM     | R+hM     | R+uM     | CRuM     |
| Averaging | 21.1     | 20.9     | 20.3     | 21.8     | 21.4     | 20.9     |
| Stacking | 17.8     | 17.2     | 18.8     | 18.3     | 16.6     | 18.4     |

#### (c) R² of single models

|          | Absorption wavelength | Emission wavelength |
|----------|-----------------------|---------------------|
|          | R       | hM     | uM     | CRuM     | R       | hM     | uM     | CRuM     |
| RF       | 0.83    | 0.87   | 0.85   | 0.83     | 0.81    | 0.85   | 0.86   | 0.81     |
| ET       | 0.84    | 0.84   | 0.79   | 0.82     | 0.83    | 0.84   | 0.79   | 0.83     |
| BAG      | 0.84    | 0.88   | 0.85   | 0.83     | 0.83    | 0.86   | 0.88   | 0.82     |
| ADA      | 0.85    | 0.81   | 0.80   | 0.83     | 0.83    | 0.85   | 0.83   | 0.82     |
| GBT      | 0.86    | 0.88   | 0.87   | 0.86     | 0.84    | 0.89   | 0.89   | 0.86     |
| kNN      | 0.75    | 0.75   | 0.66   | 0.65     | 0.74    | 0.74   | 0.67   | 0.67     |
| RI       | 0.70    | 0.73   | 0.73   | 0.77     | 0.72    | 0.76   | 0.77   | 0.80     |
| LA       | 0.80    | 0.71   | 0.85   | 0.85     | 0.79    | 0.81   | 0.88   | 0.88     |
| PLS      | 0.64    | 0.70   | 0.72   | 0.76     | 0.65    | 0.74   | 0.76   | 0.79     |
| BAY      | 0.69    | 0.73   | 0.74   | 0.77     | 0.72    | 0.77   | 0.78   | 0.80     |

#### (d) R² of ensemble models

|          | Absorption wavelength | Emission wavelength |
|----------|-----------------------|---------------------|
|          | R+hM     | R+uM     | CRuM     | R+hM     | R+uM     | CRuM     |
| Averaging | 0.87     | 0.86     | 0.86     | 0.87     | 0.87     | 0.87     |
| Stacking | 0.90     | 0.90     | 0.87     | 0.91     | 0.91     | 0.89     |
Feature importance

To ascertain what kind of grounds the model contributed to the prediction, feature importance was calculated. The concept of feature importance was defined in decision tree-based methods [15]. In this study, the feature importance values in our proposed stacking model were calculated by the multiplication of the feature importance in the first layer (for the descriptors) and those in the second layer (for the first layer models), followed by the normalization to make the sum 1.0.

We examined 10 variables that were the most important in RDKit descriptors. Most of the important variables were common in absorption and emission wavelength (Table 3). Especially up to 4 variables were in the same order. This is reasonable because the emission wavelength of the fluorescent substance causes peak shift from the absorption wavelength to the longer wavelength side. BODIPY_category, which we defined in this study, and RingCount, which is the number of ring structures, were important in both absorption and emission wavelength. NumAromaticRings, which is the number of aromatic rings, was important in the absorption wavelength. Variables related to the size of the molecule and the length of the conjugation length were thought to affect the long wavelength side. In addition, many variables concerning EState were estimated to be important. EState was an indicator showing the potential difference between compounds in the molecule, and it was considered that the electronic characteristics had an influence on the wavelength.

Fig. 4a shows the substructures corresponding top eight important variables of Morgan fingerprint. Fig. 4b is the box plot, showing that the compounds having the basic BODIPY skeleton (BODIPY_category 1) had a wide distribution of wavelengths. The compounds having a structure with two rings attached to the basic skeleton (BODIPY_category 2) also had a broad wavelength distribution but in the higher wavelength side than those having BODIPY_category 1.

Table 3. Top eleven important RDKit descriptors for absorption and emission wavelength

|                      | (a) Absorption wavelength |                      | (b) Emission wavelength |
|----------------------|---------------------------|----------------------|-------------------------|
| MaxEStateIndex       | 0.0596                    | MaxEStateIndex       | 0.0404                  |
| SlogP_VSA8           | 0.0580                    | SlogP_VSA8           | 0.0395                  |
| MaxAbsEStateIndex    | 0.0506                    | MaxAbsEStateIndex    | 0.0356                  |
| EState_VSA7          | 0.0306                    | EState_VSA7          | 0.0344                  |
| BODIPY_category      | 0.0248                    | VSA_EState8          | 0.0167                  |
| RingCount            | 0.0182                    | BODIPY_category      | 0.0137                  |
| BertzCT              | 0.0156                    | MinEStateIndex       | 0.0092                  |
| MinEStateIndex       | 0.0155                    | PEOE_VSA8            | 0.0081                  |
| VSA_EState8          | 0.0145                    | EState_VSA3          | 0.0076                  |
| BalabanJ             | 0.0124                    | BertzCT              | 0.0075                  |
| NumAromaticRings     | 0.0102                    | RingCount            | 0.0069                  |
(a) **Top eight important Morgan fingerprint**

- fp_3982076256
- fp_791239570
- fp_1599909365
- fp_1000548645
- fp_856514708
- fp_3527476261
- fp_422715066
- fp_951226070

(b) **Boxplots of BODIPY category and the top-ranked Morgan fingerprint**

*Figure 4. Top two important descriptors in unhashed Morgan fingerprint and BODIPY category.*
Validation using quantum calculation

In this study, we assumed that quantum chemical calculation values were correct, and used them in order to validate the predicted values by the proposed model.

First, in order to verify the accuracy of quantum calculation, their calculation values and the actual measurement values were compared for the 107 compounds whose actual absorption wavelength were known. Quantum calculation was performed using m062x/6-31g(d) in a solvent-free state, and the solvent content was corrected based on a deviation from the experimental value. We compared those calculation values with the experimental values by using the least squares method with LOO, and obtained $R^2 = 0.96$ and MAE = 13.18. It was confirmed that the error was sufficiently small, and estimation of the absorption wavelength was possible in a wide range.

Consequently, we compared the predicted values with the quantum calculation. The coefficient of determination $R^2 = 0.92$, MAE = 16.33, which proved the high correlation. We have not verified the accuracy in emission wavelength, because it is known that quantum computation on the luminescence wavelength is much more time-consuming than on absorption wavelength, and the calculation would not finish. Thus, we regarded quantum calculation can be used for obtaining correct absorption wavelength of the newly generated compounds, and used them in the following processes.

Comparison of AD estimation methods to generate new structures

We constructed our proposed AD estimation model based on the Tanimoto distance with the neighboring points of the training data. Our model was compared with the existing AD estimation model by Gaussian kernel OCSVM. The performances of the AD models were evaluated by calculating the absorption wavelength by quantum calculation as follows. We used 143 compounds for which absorption wavelength were experimentally verified. In the OCSVM-based AD estimation model, we did not use Morgan fingerprint because all newly generated structures were estimated to be inside AD and the model did not work properly.

Dimensions of RDKit descriptors were reduced using PCA to the first axis where the cumulative contribution rate becomes $n\%$ or more, followed by OCSVM where the covariance in the training data was set to maximize and $\nu$ was defined as the ratio of the outliers in the training data. Table 4a shows the obtained average error of predicted absorption wavelength of generated compounds inside/outside AD and the number of compounds inside AD calculated based on various $n$ and $\nu$.

In contrast, our proposed AD model required two parameters; $m$ samples and $k\%$ points in advance. Table 4b shows the obtained average error of predicted absorption wavelength of generated compounds inside/outside AD and the number of compounds inside AD calculated based on various $m$ and $k$.

As long as the OCSVM-based AD model functioned properly, as the ratio $\nu$ increases, the prediction error inside AD would become larger. Similarly, as long as our proposed AD model functioned properly, as the threshold $k$ increases, the prediction error inside AD would become larger. We observed that these were the case in our proposed AD model (Table 4b) but not always the case in the OCSVM-based AD model (Table 4a).

Our proposed AD model and the OCSVM-based AD model use different parameters, and the both result changes significantly depending on the threshold, it was difficult to directly compare those performances. Thus we made a comparison by using top 500 compounds in descending order of their AD scores. The purpose of this analysis was to verify whether the newly generated compound distant from the training data was estimated to be distant properly regardless of the setting of the

| (a) | $n$ = 70 | $n$ = 80 | $n$ = 90 |
|-----|---------|---------|---------|
| $v$=0.5 | 28.82 / 31.80 (2477) | 29.31 / 30.89 (2139) | 28.72 / 31.04 (1742) |
| $v$=0.16 | 29.70 / 31.38 (3305) | 29.69 / 30.72 (2547) | 29.70 / 31.38 (2574) |
| $v$=0.1 | 29.85 / 31.01 (3389) | 29.13 / 30.89 (2625) | 29.85 / 31.01 (2714) |
| $v$=0.05 | 30.88 / 26.39 (3656) | 29.41 / 31.25 (2703) | 30.88 / 26.39 (3186) |

| (b) | $m$ = 14 | $m$ = 28 | $m$ = 72 | $m$ = 142 |
|-----|---------|---------|---------|---------|
| $k$ = 0.83 | 5.4 / 30.1 ( 3) | 7.2 / 30.1 ( 5) | 17.7 / 30.3 (62) | 18.7 / 30.5 (152) |
| $k$ = 0.95 | 26.8 / 32.3 (1758) | 24.3 / 31.9 (1028) | 22.7 / 31.3 (601) | 23.5 / 31.8 (915) |
| $k$ = 1.00 | 30.1 / 25.5 (4402) | - | - | - |
threshold. Table 5 shows the average prediction errors of the structure inside and outside AD. It was confirmed that the prediction error of the inside AD was smaller than the prediction error of the outside AD in both models. Compared with the OCSVM-based model, the proposed model had a smaller prediction error of the structure of the inside AD, and it can be considered that the proposed model defined AD more correctly.

To further investigate the performance of our AD model, we sorted the newly generated compounds in the descending order of the AD scores, and observed the transition of the predictive error (Fig. 5). The error at each point is taken as the average value of errors of the front and back 251 structures. The parameters of the proposed model were $m = 72$, $k = 0.95$, and the parameters of the OCSVM-based model were $n = 90$, $\nu = 0.1$. The error in the proposed model has risen up to the 1,000 structure. The plot was flat from 1,000 structures to 4,000 structures, and it also risen after 4,000. In the OCSVM-based model, the error roughly risen up to the 700 structure. After 700, the prediction error was dispersed and hit the wave greatly. In the proposed model, it can be considered that the distances of the new candidate structures could be properly discriminated. In the OCSVM-based model, the prediction error of some structure which were estimated to be far from the training data is smaller, and it is considered that the missing of the structure which should be estimated to be inside AD. We thus concluded that the proposed AD model functioned more sufficiently than the Gaussian kernel OCSVM-based model.

The prediction model verified in the AD estimation model were applied to the newly generated 4,410 compounds. Fig. 6 represents the absorption and emission wavelength of the newly generated compounds before and after the AD screening. 602 structure passed through the screening. Fig. 7 are the examples of the new compounds inside AD in high and low absorption/emission wavelength.

|              | Inside AD | Outside AD |
|--------------|-----------|------------|
| OCSVM        | 26.31 nm  | 30.72 nm   |
| Tanimoto (proposed) | 20.98 nm  | 31.56 nm   |

Table 5. Average error of top 500 compounds inside AD and other compounds outside AD ranked by two AD estimation methods.

![Figure 5](image1.png)

**Figure 5. Comparison of average errors to select top-ranked compounds using two AD estimation methods.** Vertical axis represents average error [nm]. Horizontal axis represents the number of top-ranked compounds that are regarded as within AD.

![Figure 6](image2.png)

**Figure 6. Predicted absorption and emission wavelength of the newly generated compounds outside (top) and inside (bottom) AD.** Horizontal and vertical axis represent the predicted absorption and emission wavelength [nm], respectively.

![Figure 7](image3.png)

**Figure 7. Examples of newly generated compounds.** The first three (top) and the last three (bottom) were estimated to have absorption/emission in the high and low wavelength regions, respectively.
Conclusion

High precision wavelength prediction model and AD estimation model were constructed using ensemble learning for fluorescent substances with BODIPY skeleton. Screening was carried out using these models, and fluorescent substances were proposed. Prediction accuracy improved in stacking model using RDKit descriptors and count-based unhashed Morgan fingerprint. When using stacking, it was found that the diversity of the prediction model of the first layer had influence on accuracy. The variables related to the molecular skeleton and the conjugation length were shown to be important. The result agreed with known chemical findings. We also proposed the AD estimation model that directly use the descriptors based on Tanimoto distance. The performance of the AD models was shown better than the existing Gaussian kernel OCSVM-based model. Using our proposed stacking model and AD model, newly generated compounds were screened and we obtained 602 compounds which were estimated inside AD in both absorption wavelength and emission wavelength. The AD estimation could possibly improve by the use of Tanimoto kernel OCSVM [27], which would strengthen our stacking-based framework. Still, our AD estimation method would be interpretable because of the independence from specific non-linear regression models like SVM.

Acknowledgements

We thank Hiroshi Nakano, Masakazu Ukita, Raku Shirasawa, Shigetaka Tomiya in Sony Corporation, Japan, for helpful discussions.

References

[1] M. Adachi, S. Nakamura, Dye Pigment, 17, 287-296(1991).
[2] C.-H. Chen, K. Tanaka, K. Funatsu, J. Fluorescence, 28, 695-706 (2018).
[3] I. J. Arroyo, R. Hu, G. Merino, B.-Z. Tang, E. Peña-Cabrera, J. Org. Chem., 74, 5719-5722 (2009).
[4] A. Loudet, K. Burgess, Chem. Rev., 107, 4891-4932 (2007).
[5] A. Schüller, G. B. Goh, H. Kim, J. S. Lee, Y. T. Chang, Mol. Inform., 29, 717-729 (2010).
[6] RDKit. https://www.rdkit.org/. Accessed January 31, 2019.
[7] D. Rogers, M. Hahn, J. Chem. Inf. Model., 50, 742-754 (2010).
[8] T. G. Dietterich. Ensemble Methods in Machine Learning. http://www.cs.orst.edu/~tgd. Accessed January 31, 2019.
[9] Y. Liu, X. Yao, Neural Networks, 12,1399-1404 (1999).
[10] D. H. Wolpert, Neural Networks, 5, 241-259 (1992).
[11] A. Kenser, J. Alvarsson, U. Norinder, O. Spjuth O. J. Cheminform., 10, 49 (2018).
[12] D. J. C. MacKay, Bayesian Interpolation, Neural Computation, 4, 415-447 (1992).
[13] Regression Shrinkage and Selection via the Lasso on JSTOR. https://www.jstor.org/stable/2346178?seq=1#metadata_table_contents. Accessed January 31, 2019.
[14] P. Geladi, B. R. Kowalski, Anal Chim Acta., 185, 1-17 (1986).
[15] L. Breiman, Random Forest. 2001. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Accessed February 1, 2019.
[16] P. Geurts, D. Ernst, L. Wehenkel, Mach Learn., 63, 3-42 (2006).
[17] L. Breiman. Bagging predictors. Mach Learn, 24, 123-140 (1996).
[18] L. G. Valiant, Commun ACM., 27, 1134-1142 (1984).
[19] J. H. Friedman. Stochastic Gradient Boosting; 1999. http://www.salford-systems.com/treenet.html. Accessed January 31, 2019.
[20] T. Cover, P. Hart, IEEE Trans. Inf. Theory, 13, 21-27 (1967).
[21] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, Altern Lab Anim., 33, 445-459 (2005).
[22] K. LIII, Pearson, London, Edinburgh, Dublin Philos Mag J Sci., 2, 559-572 (1901).
[23] R. P. Sheridan, B. P. Feuston, V. N. Maiorov, S. K. Kearsley, J. Chem. Inf. Comput. Sci., 44, 1912-1928 (2004).
[24] N. Fechner, A. Jahn, G. Hinselmann, A. Zell, J Cheminform., 2, 2 (2010).
[25] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, Neural Comput., 13, 1443-1471 (2001).
[26] H. Kaneko, K. Funatsu, J. Chem. Inf. Model., 54, 2469-2482 (2014).
[27] L. Ralaivola, S. J. Swamidass, H. Saigo, P. Baldi, Neural Netw. 8, 1093-110 (2005).