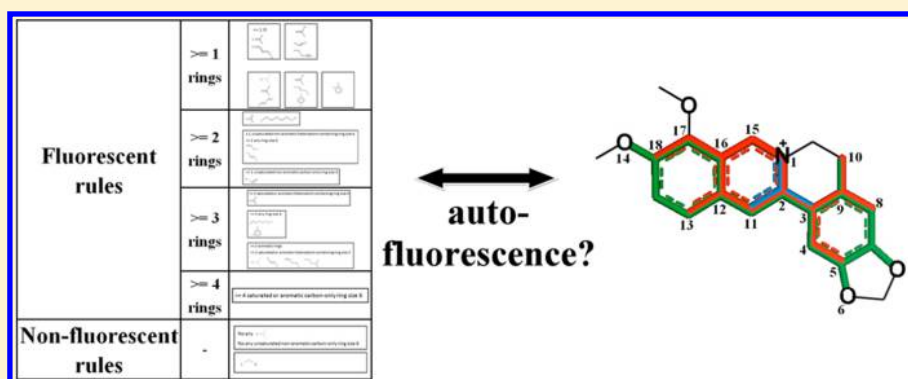


## Rule-Based Classification Models of Molecular Autofluorescence

Bo-Han Su,<sup>†,§</sup> Yi-Shu Tu,<sup>‡,§</sup> Olivia A. Lin,<sup>‡</sup> Yeu-Chern Harn,<sup>†,‡</sup> Meng-Yu Shen,<sup>†,‡</sup> and Yufeng J. Tseng<sup>\*,†,‡,||</sup><sup>†</sup>Department of Computer Science and Information Engineering and <sup>‡</sup>Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan 106<sup>||</sup>Drug Research Center, National Taiwan University School of Medicine, No. 1, Sec. 1, Ren-ai Road, Taipei, Taiwan 106

## S Supporting Information



**ABSTRACT:** Fluorescence-based detection has been commonly used in high-throughput screening (HTS) assays. Autofluorescent compounds, which can emit light in the absence of artificial fluorescent markers, often interfere with the detection of fluorophores and result in false positive signals in these assays. This interference presents a major issue in fluorescence-based screening techniques. In an effort to reduce the time and cost that will be spent on prescreening of autofluorescent compounds, *in silico* autofluorescence prediction models were developed for selected fluorescence-based assays in this study. Five prediction models were developed based on the respective fluorophores used in these HTS assays, which absorb and emit light at specific wavelengths (excitation/emission): Alexa Fluor 350 (A350) (340 nm/450 nm), 7-amino-4-trifluoromethyl-coumarin (AFC) (405 nm/520 nm), Alexa Fluor 488 (A488) (480 nm/540 nm), Rhodamine (547 nm/598 nm), and Texas Red (547 nm/618 nm). The C5.0 rule-based classification algorithm and PubChem 2D chemical structure fingerprints were used to develop prediction models. To optimize the accuracies of these prediction models despite the highly imbalanced ratio of fluorescent versus nonfluorescent compounds presented in the collected data sets, oversampling and undersampling strategies were applied. The average final accuracy achieved for the training set was 97%, and that for the testing set was 92%. In addition, five external data sets were used to further validate the models. Ultimately, 14 representative structural features (or rules) were determined to efficiently predict autofluorescence in data sets containing both fluorescent and nonfluorescent compounds. Several cases were illustrated in this study to demonstrate the applicability of these rules.

## ■ INTRODUCTION

Fluorescence-based detection has been increasingly developed in high-throughput screening (HTS) assays as a reliable tracking method<sup>1–3</sup> and widely employed due to its nonradioactivity, homogeneity, and high sensitivity advantages.<sup>1</sup> Fluorescence-based detection methods currently used in HTS<sup>2</sup> assays include: fluorescence intensity (FI),<sup>3</sup> fluorescence polarization (FP),<sup>4</sup> time-resolved fluorescence (TRF),<sup>5</sup> fluorescence resonance energy transfer (FRET),<sup>6</sup> fluorescence lifetime (FLT),<sup>7</sup> fluorescence correlation spectroscopy (FCS),<sup>8</sup> and fluorescent intensity distribution analysis (FIDA).<sup>9</sup> Unfortunately, all of the fluorescence-based methods have a few common limitations that could affect the analysis outcomes.<sup>10</sup> The most prominent limitation lies in the detection of “false-positive” fluorescence due to autofluorescence, or “false-negative” in fluorescence-quenching assays.<sup>1,11</sup> Autofluorescent compounds, which emit light in the absence of artificial fluorescent markers, often interfere with the

detection of fluorophores in HTS assays and contribute toward inaccurate screening.<sup>12</sup> It was estimated that up to 50% of false positive signals were due to fluorescence interference, and these inaccurate screenings could mask the active compounds.<sup>11</sup> Despite several attempts by previous groups to reduce fluorescence interferences, by means of adjusting concentrations and wavelengths of fluorophores,<sup>1</sup> or developing improved FLT-based assays,<sup>13</sup> numerous “false” enzyme activators still could not be excluded even after secondary validation was performed.<sup>14</sup> Therefore, to distinguish autofluorescence apart from artificially induced fluorescence from fluorophores is an ongoing challenge that needs to be addressed. For this reason, an additional method to screen for and identify compounds with autofluorescence may be necessary, in order to guarantee better

Received: December 16, 2014

Published: January 27, 2015

HTS assays outcome by preventing the detection of unwanted false positive signals.<sup>15</sup>

CE-Flu is a powerful tool used to detect autofluorescent compounds. The efficient separation capacity of capillary electrophoresis (CE) combined with fluorescence detection (Flu) offers higher sensitivity and improved selectivity. This is because CE enables the analysis of a wide range of biomolecular/pharmaceutical compounds, and Flu results in higher detection accuracy that is preferred over the UV detection method. CE-Flu is usually performed under low UV wavelengths due to the low optical transparency restriction of the chip materials. For this reason, de Kort and colleagues observed that when fluorescent compounds were excited in deep-UV, detection sensitivity of CE-Flu could be compromised.<sup>16</sup> Consequently, to rely solely on fluorescence-based assays to detect autofluorescence is insufficient.

The molecular mechanisms of fluorescence have been well studied.<sup>17,18</sup> Typically fluorescent molecules absorb light photons at an appropriate wavelength, become excited, and emit light in longer wavelength while transitioning back to ground state.<sup>18</sup> Hence, a fluorescent molecule should possess high absorbance, and this is the reason that currently well-defined fluorescent molecules include the following structural features: conjugated nonaromatic, aromatic, and heterocyclic groups.<sup>19</sup> However, the structural features that have been listed and categorized thus far<sup>20</sup> are not representative of or applicable to a wide range of fluorophores used in HTS, since they were derived from a small fraction of marketed fluorescence agents.<sup>15</sup> On the other hand, the National Institute of Health Chemical Genomics Center (NCGC)<sup>15</sup> and the Southern Research Molecular Libraries Screening Center (SRMLSC)<sup>21</sup> have screened and organized the autofluorescence profiles of compounds in a library. The fluorescence profiles and novel fluorophores of 71 391 compounds commonly used in HTS assays have been previously reported.<sup>15</sup> However, due to the limited correlation between structural features and fluorescence in the compound samples, there were insufficient representative rules or guidelines to reliably screen for autofluorescent compounds. To characterize fluorophores from more diverse libraries of compounds will improve detection of autofluorescence in HTS assays.

To reduce the time and cost that would normally be spent on additional screenings for autofluorescent compounds, *in silico* structure-based prediction models were developed to identify and remove these compounds before *in vitro* fluorescent-based HTS assays were performed. A recent work by Albert-Garcia et al. published an *in silico* autofluorescence prediction tool using linear discriminant analysis to distinguish fluorescent from nonfluorescent compounds.<sup>22</sup> A balanced data set consisting of 50 fluorescent and 50 nonfluorescent compounds were divided into a training set of 80 compounds and a testing set of 20 compounds. The fluorescence analysis was performed with the excitation wavelengths set between 220 and 440 nm, and the emission wavelengths set between 220 and 600 nm. This resulted in 170 molecular connectivity descriptors calculated to correlate structural features with fluorescence of molecules and a discriminant function containing 7 descriptors obtained to predict autofluorescence.

However, the accuracy and applicability of Albert-Garcia's system was restricted due to its insufficient end points, difficult-to-interpret descriptors, and unspecified excitation/emission profiles. First and foremost, 80 compounds in the training set insufficiently represent the structural diversity of these compounds to build a good prediction model. Second, solely relying on connectivity descriptor makes it much more difficult

to predict which components of a compound is associated with autofluorescence. Last but not least, not tailoring the system for a specific spectrum renders it less reliable because most HTS assays are screened at specific excitation/emission wavelengths with the appropriate fluorescent agents. In short, autofluorescence prediction models that are trained by fragment based descriptors and designed for specific excitation/emission wavelengths would more closely resemble HTS assays and better predict which components of a compound is likely to result in autofluorescence.

In this study, five autofluorescence prediction models were developed based on the five commonly used fluorophores, with different excitation/emission profiles: Alexa Fluor 350 (340 nm/450 nm), 7-amino-4-trifluoromethyl-coumarin (AFC) (405 nm/520 nm), Alexa Fluor 488 (480 nm/540 nm), Rhodamine (547 nm/598 nm), and Texas Red (547 nm/618 nm). A total of 5163 fluorescent and 309 212 nonfluorescent compounds were collected as our data set. To fine-tune our prediction models, undersampling and oversampling methodologies<sup>23</sup> were applied to overcome overfitting problem resulted from utilizing highly imbalanced data sets to build our models. The fuzzy *k*-means clustering algorithm was applied as undersampling strategy to reduce the majority set of training compounds. These prediction models were then optimized using C5.0 rule-based algorithm<sup>24</sup> combined with PubChem 2D fingerprint.<sup>25</sup> In addition, five external data sets were used to further validate our models. Last but not least, 14 novel rules were determined which could be applied to accurately predict autofluorescence in data sets containing both fluorescent and nonfluorescent compounds. Several molecules were described to illustrate the applicability of identified rules on autofluorescence prediction. Finally, constructed rule-set models and resources of data sets, including CID numbers and descriptor values, were provided online ([http://cmdm.tw/supp\\_files/fluor.rar](http://cmdm.tw/supp_files/fluor.rar)) to assist interested individuals who may wish to utilize this tool for autofluorescence predictions.

## MATERIAL AND METHODS

**Training and Testing Data Sets.** Training and testing data sets were obtained from the PubChem Assay Database. The assays were analyzed and published by NCGC<sup>15</sup> and SRMLSC.<sup>21</sup> The five fluorophores commonly used in fluorescence screening assays were selected: Alexa Fluor 350 (A350) (PubChem AID: 590/709, 4141 fluorescent compounds, 64 116 nonfluorescent compounds), 7-amino-4-trifluoromethyl-coumarin (AFC) (PubChem AID: 923, 780 fluorescent compounds, 71 923 nonfluorescent compounds), Alexa Fluor 488 (A488) (PubChem AID: 591, 60 fluorescent compounds, 58 241 nonfluorescent compounds), Rhodamine (PubChem AID: 594, 33 fluorescent compounds, 58 372 nonfluorescent compounds), and Texas Red (PubChem AID: 587, 35 fluorescent compounds, 47 620 nonfluorescent compounds).<sup>15,21,26</sup> The distinct excitation/emission profiles of these fluorophores were summarized in Table 1. Five prediction models were developed based on the five data sets collected. Each data set was randomly divided into training set and testing set, for development and evaluation of models performance.

Five external data sets were obtained from different literature sources<sup>16,19,27–29</sup> as testing data sets to further validate the five prediction models. Hundreds of natively fluorescent compounds were reported in these literatures, but only compounds with the appropriate excitation/emission profiles, closely corresponding to the profiles of compounds selected for training data sets, were considered.

**Table 1. Representative Fluorescence Agents, Excitation/Emission Wavelengths, Assay Sources, and the Numbers of Fluorescent/Nonfluorescent Compounds in Five Autofluorescence Dataset<sup>a</sup>**

representative fluorescence agent	spectrum wavelengths (nm)		PubChem assay ID (source)	fluorescent compounds	nonfluorescent compounds
	excitation	emission			
Alexa Fluor 350 (A350)	340	450	590 (NCGC) 709 (SRMLSC)	4141	64116
AFC	405	520	923 (NCGC)	780	71923
Alexa Fluor 488 (A488)	480	540	591 (NCGC)	60	58241
Rhodamine	547	598	594 (NCGC)	33	58372
Texas Red	547	618	587 (NCGC)	35	47620

<sup>a</sup>AFC, 7-amino-4-trifluoromethyl-coumarin. NCGC, NIH Chemical Genomics Center. SRMLSC, Southern Research Molecular Libraries Screening Center.

**PubChem Fingerprints.** PubChem 2D fingerprints were used to generate descriptors for the compounds in our data set.<sup>25</sup> In the PubChem fingerprints database, there are 881 bits of descriptors related to element counts, aromatic or nonaromatic ring counts, atom pairs, atom neighborhoods, and specific fragments.

**Ring-Based Filters.** The ratios between fluorescent and nonfluorescent compounds are highly imbalanced in the collected data sets, as shown in Table 1. Building prediction models from highly imbalanced data sets could result in overfitting, thus proper elimination of noisy data is a way to balance the training set.<sup>30</sup> In general, fluorescent compounds would possess at least an aromatic ring or a ring with heteroatoms.<sup>17</sup> Therefore, we applied 17 bits related to aromatic rings and heteroatom rings in PubChem fingerprints as our ring-based filter (Table 2) to remove compounds which contains no

**Table 2. Bits in PubChem Fingerprints Considered in the Ring-Based Filter**

bit no.	description
118	has 3-sized saturated or aromatic rings with heteroatoms
121	has 3-sized unsaturated nonaromatic rings with heteroatoms
132	has 4-sized saturated or aromatic rings with heteroatoms
135	has 4-sized unsaturated nonaromatic rings with heteroatoms
146	has 5-sized saturated or aromatic rings with heteroatoms
149	has 5-sized unsaturated nonaromatic rings with heteroatoms
181	has 6-sized saturated or aromatic rings with heteroatoms
184	has 6-sized unsaturated nonaromatic rings with heteroatoms
216	has 7-sized saturated or aromatic rings with heteroatoms
219	has 7-sized unsaturated nonaromatic rings with heteroatoms
230	has 8-sized saturated or aromatic rings with heteroatoms
233	has 8-sized unsaturated nonaromatic rings with heteroatoms
244	has 9-sized saturated or aromatic rings with heteroatoms
247	has 9-sized unsaturated nonaromatic rings with heteroatoms
251	has 10-sized saturated or aromatic rings with heteroatoms
254	has 10-sized unsaturated nonaromatic rings with heteroatoms
255	has aromatic rings

aromatic rings and heteroatom rings both in fluorescent and nonfluorescent data set. A compound was dropped from data set if all of these 17 bits were "0", or negative. The total numbers of fluorescent and nonfluorescent compounds in the training set and testing set after ring-based filtering was applied were reported (Table 3). In A350 and AFC data sets, two-thirds of the fluorescent compounds were randomly selected as training set, and the remaining fluorescent compounds were used as a testing set. However, in A488, Rhodamine, and Texas Red data sets, due to the limited samples of fluorescent compounds, all of the

**Table 3. Amount of Fluorescent and Nonfluorescent Compounds in Training and Testing sets, after Filtering with the Ring Filter<sup>a</sup>**

representative fluorescence agent	training set		testing set	
	fluorescent	nonfluorescent	fluorescent	nonfluorescent
Alexa Fluor 350 (A350)	2751	41870	1376	20935
AFC	520	46575	260	23288
Alexa Fluor 488 (A488) <sup>a</sup>	60	38057	60	19029
Rhodamine <sup>a</sup>	33	38144	33	19072
Texas Red <sup>a</sup>	35	31121	35	31121

<sup>a</sup>The active compounds in A488, Rhodamine, and Texas Red datasets were all in both training set and testing set due to the small amount of fluorescent compounds.

fluorescent compounds were included in both training and testing set. Nonfluorescent compounds in all five data set were also randomly separated into training sets and testing sets of 2:1 ratio. In the fluorescent class, 0.3% compounds were eliminated by ring-based filter and these compounds only belong to A350 data set. It assures that most fluorescent compounds in our data set have at least one aromatic ring or a ring with heteroatoms. In the nonfluorescent class, 1324 compounds on average were filtered out. The compounds that have been filtered out will be predicted as nonfluorescent in our system. Although it would produce a few false negatives, eliminating a portion of training compounds which has no fluorescence-related chemical properties (aromatic rings or rings with heteroatoms) might reduce disturbance for the performance of system when development of fluorescent prediction model. It is surprising that nearly 98% nonfluorescent compounds contain the known fluorescent properties. In other words, whatever the compounds are fluorescent or nonfluorescent almost satisfy the known fluorescent properties. It is shown that the well-known fluorescent properties are insufficient to identify the autofluorescence.

**C5.0 Rule-Based Classification Methods.** We built prediction models of each fluorescent data set by RuleQuest C5.0 data mining method in R package.<sup>24,31</sup> C5.0 was not only an improved version of well-known C4.5 decision tree algorithm, but also able to produce rule-based models. A rule-based model is composed of a set of rules, and users can interpret the models based on these rules. Therefore, the prediction models we built may give us some insight of how chemical structures related to their fluorescence properties.

In the C5.0 software, there were two types of prediction models provided: decision tree models and rule set models. A tree-based model provides only one classification tree, whereas a



rule-based model provides multiple rulesets for classification. Compared to decision tree models, rule set models were more accurate in prediction power and easier for model explanation.<sup>32</sup> We selected the rule-based mode to generate the fluorescence prediction models. The rule set models were made from branches from the decision trees by deleting sub-branches that will not reduce too much prediction power. The algorithm would produce fluorescent and nonfluorescent rulesets separately. Each ruleset contains a set of conditions and a “confidence score” which represents its prediction power. While classifying, all rulesets will be tested; the confidence score for each ruleset will be summed up; then the predicted class will be the one which has the highest voting score. We can utilize the rule-set based prediction models for easier model explanation and linking the correlation between molecule characteristics and fluorescence.

**Oversampling Strategy.** For evaluation of models performance, five data sets were divided into training sets and testing sets. However, the ratios between fluorescent and nonfluorescent compounds are highly imbalanced. To prevent overfitting, an oversampling technique was applied to the training set to minimize the large difference between the numbers of fluorescent and nonfluorescent compounds in each of the five data sets.<sup>23</sup> We repeated the fluorescent data in the training set to make the ratio of fluorescent compounds and nonfluorescent compounds equal to 2 to 3 in each fluorescence data set. For example, when the number of nonfluorescent compounds in Rhodamine training set is 38 144 and the number of fluorescent compounds is 33 only (Table 3), we repeat the 33 fluorescent compounds 771 times. Thus, the produced 25 443 ( $33 \times 771$ ) fluorescent compounds and 38 144 nonfluorescent compounds are used as our new training data set to build the prediction model. The oversampling strategy can enlarge the significance of the minority class of data sets that contributes to the prediction model. However, too much redundant data might cause the overfitting on those fluorescent training compounds for the prediction models, and thus, the effectiveness of oversampling strategy is limited. We then considered the undersampling strategy to decrease the number of nonfluorescent training compounds.

#### Fuzzy *k*-Means Clustering As Downsampling Strategy.

Fuzzy *k*-means clustering was the undersampling strategy adopted to select the representative compounds from our nonfluorescent training data sets to resolve the highly imbalanced data set problem in our study. The clustering methodology is commonly applied to select for representative chemical structures from a large data sets for screening.<sup>33–35</sup> We can retrieve one key compound from each cluster as the representative chemicals. Fuzzy *k*-means clustering is an appropriate method for classifying huge numbers of chemical compounds on account of its low computational complexity.<sup>36</sup> *k* is a prior parameter representing the expected number of clusters. The algorithm of *k*-means initialized *k* partitions of input data set, and then iteratively reallocates the centroids of each partition to refine the partitions until an optimal criterion is reached. Each partition produced by *k*-means method is a *crisp* cluster since each compound only can be assigned to one single cluster. Alternatively, in a *fuzzy* clustering approach, each compound is allowed to belong to more than one cluster (called *fuzzy* cluster).<sup>37,38</sup> The concept of fuzzy cluster is implemented by a defined membership function to determine degree of each compound that could be classified to each cluster. Fuzzy *k*-means clustering so far is the most widely used method among all of the fuzzy clustering approaches and is first developed by Dunn<sup>39</sup> and Bezdek.<sup>37</sup>

Holliday et al.<sup>40</sup> first employed the fuzzy *k*-means clustering method for clustering files of chemical structures. The simulated prediction results demonstrated that the fuzzy *k*-means approach is superior to the conventional *k*-means method based on 2D-fingerprints when the parameters are configured appropriately. In our study, we adopted the same fuzzy *k*-means procedure to select the representative compounds from the nonfluorescent data set. The algorithm of fuzzy *k*-means is similar to the conventional *k*-means approach other than the additional membership function indicating what degree of a compound belongs to each cluster. The procedure of fuzzy *k*-means will complete computation when an objective function is minimized, and each compound can then be assigned to one cluster with the corresponding maximum membership value. The detailed methods have been summarized in ref 40.

However, the greater the expected number of cluster *k*, the more the computational time especially for large compound data sets. In our cases, when *k* is set to 100, the computational time approaches 2 days. To make sure that the program can be halted in a reasonable execution time, the expected number of clusters *k* is considered to be lower than 100 in our system. If we only select one key nonfluorescent compound from each cluster as our training nonfluorescent data set, the number of training fluorescent compounds after performing oversampling methods would be much higher than the nonfluorescent training compounds. In order to balance the ratio of fluorescent to nonfluorescent compounds, we retrieve *N* number of representative nonfluorescent compounds from each cluster according to the following procedure. All of the nonfluorescent compounds classified in same cluster are first ranked based on Euclidean distance measurement between the centroid of the cluster and its containing compounds. Assume total number of compounds in one cluster is *C*, the compounds which are ordered in ranks,  $C \times 1/N$ ,  $C \times 2/N$ ,  $C \times 3/N$ , ..., and  $C \times N/N$ , are then selected as the representative nonfluorescent compounds in this cluster. That is, the total number of selected nonfluorescent training compounds is  $N \times k$ . We attempt different values of *N* in this study to adjust the ratio of fluorescent to nonfluorescent compounds for optimizing system performance.

Besides to the parameter *N*, there are three additional parameters that need to be configured in fuzzy *k*-means clustering algorithm. The first parameter is the number of clusters, *k*, which is given a priori or heuristically determined by user. In our study, *k* influences the time performance and the structural diversity of compounds, as we have discussed above. The second parameter is the fuzziness index, *m*, which defines the degree of cluster fuzziness. The value of *m* is between 1 and  $\infty$ . A larger fuzziness index causes a smaller membership function values. The value of *m* in 1 results in a traditionally crisp *k*-means clustering, and instead, when the value of *m* approaches  $\infty$ , the membership function value is close to  $1/k$ . A too large value of *m* will cause inappropriate partitions while a too small value will lose the preponderance of fuzziness. Therefore, the setting of *m* should be careful. According to the studies of Dunn and Bezdek and previous works,<sup>37,39,41</sup> only small values of *m* that are lower than 2 lead to suitable results in many applications and adopted in our system. The last parameter is the termination criterion,  $\epsilon$ , which influences the convergence of the algorithm.  $\epsilon$  is set in a sufficient value of 0.01 in our study. How the predicted performance of the model is influenced by these different setting of the parameters is discussed in the Results and Discussion.

**Statistical Evaluation of Prediction Power in Models.** For measuring and evaluating the prediction power of our

**Table 4. Statistical Evaluation of the Five Optimized Fluorescence Prediction Models Using the Nonfluorescent Training Dataset Built by the Fuzzy *k*-Means Clustering Methods, after Evaluations Based on Different Parameter Settings<sup>a</sup>**

model		fluorescent:nonfluorescent compounds	accuracy	sensitivity	specificity	G-mean
A350	training	2751:4620	90.2	83.5	94.2	88.7
	testing	1376:58214	82.2	78.1	82.3	80.2
AFC	training	520:2088	96.5	96.4	96.6	96.5
	testing	260:67768	88.0	87.3	88.0	87.7
A488 <sup>b</sup>	training	60:1492	99.3	100	99.3	99.6
	testing	60:55589	94.9	100	94.9	97.4
Rhodamine <sup>b</sup>	training	33:1007	99.8	100	99.8	99.9
	testing	33:56203	97	100	97	98.5
Texas Red <sup>b</sup>	training	35:990	99.2	100	99.2	99.6
	testing	35:45680	96.4	100	96.4	98.2

<sup>a</sup>The rest of the nonfluorescent compounds not selected by fuzzy *k*-means clustering methods were used as testing compounds. An oversampling strategy was applied to the fluorescent training dataset to adjust the ratio of fluorescent and nonfluorescent compounds to about 2:3, but the numbers of repeated fluorescent compounds were excluded from the table. <sup>b</sup>In the A488, Rhodamine, and Texas Red models, all of the fluorescent compounds were included in the training as well as the testing datasets, due to the extremely low number of fluorescent compounds in these datasets.

**Table 5. Accuracy of Optimized Fluorescent Models for Five Fluorescent Agents According to Our Collected External Datasets**

	A350	AFC	A488	Rhodamine	Texas Red
accuracy (number correct/total number)	100% (9/9)	75% (3/4)	62.5 (5/8)	100% (2/2)	50% (1/2)

models, the following statistical measurements were calculated: *accuracy* (rate of correctly predicted compounds in both fluorescent and nonfluorescent compounds; eq 1), *sensitivity* (the rate of correctly predicted fluorescent compounds; also referred as *recall*; eq 2), *specificity* (the rate of correctly predicted nonfluorescent compounds; eq 3), and *geometric mean* (*G-mean*; the square-root of the multiplied product of sensitivity and specificity; eq 4). In the equations below, “positive” represents the numbers of fluorescent compounds, “negative” represents the numbers of nonfluorescent compounds, and “true” indicates correct prediction.

$$\text{accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{overall data}} \quad (1)$$

$$\text{sensitivity} = \frac{\text{true positive}}{\text{all positive data}} \quad (2)$$

$$\text{specificity} = \frac{\text{true negative}}{\text{all negative data}} \quad (3)$$

$$\text{geometric mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (4)$$

## RESULTS AND DISCUSSION

**Rule-Based Model Developed Using Training Data Sets.** Five fluorescence-based prediction models were naively built and trained using training data sets without using sampling strategies, and the performance of these models were then evaluated using the compounds in testing data sets. The performance of each of the five models, determined by accuracy, sensitivity, specificity, and G-mean, was summarized in Supplementary Table 2. Despite the overall high specificities, the sensitivities of these models were all below 60%.

To optimize the accuracies of our five prediction models, oversampling strategy was applied to amplify the numbers of fluorescent compounds in each data set, to achieve a ratio of 2:3 fluorescent to nonfluorescent compounds in all the data sets. The results of statistical evaluation were summarized in

Supplementary Table 3. In the A350 and AFC models, the sensitivity and specificity of the training set were almost 100%, yet in the testing set, only approximately 60% of fluorescent compounds were correctly predicted. This demonstrated that although the oversampling strategy can moderately improve the overall prediction power of the models, however, utilizing the strategy alone may still possibly lead to overfitting problem on the testing data sets.

To reduce the difference between the numbers of fluorescent and nonfluorescent training compounds, we utilized fuzzy *k*-means clustering as the undersampling strategy, to select representative nonfluorescent compounds as our new nonfluorescent training data set. The remaining nonfluorescent compounds not selected were used as our new nonfluorescent testing data set. We considered different configurations for the following parameters: *m*, *k*, and *N* in fuzzy *k*-means clustering, to adjust the sampling of nonfluorescent compounds. The detailed discussion on models optimization is described in the supplementary file. The performance of the five optimized models was summarized in Table 4. The optimized fluorescence models yielded the average accuracy, sensitivity, specificity, and G-mean values of 97.0%, 96.0%, 97.8%, and 96.7% for training sets, respectively. The corresponding average accuracy, sensitivity, specificity, and G-mean values for testing sets were 91.7%, 93.1%, 91.7%, and 92.4%. This demonstrated that the coapplication of oversampling strategy with undersampling strategy, combined with the appropriate selection of representative compounds in the training data sets, produced excellent fluorescent prediction models.

**External Data Sets for Validation of Optimized Prediction Models.** Five external data sets, containing 9, 4, 8, 2, and 2 compounds, respectively (Supplementary Table 1), were utilized to further validate our five optimized fluorescence classification models. The accuracies of each prediction model are listed in Table 5. The fluorescence classification model of A350 and Rhodamine data set achieved 100% accuracies.

Sialic acid from the AFC external data set, a fluorescent compound, was misclassified (Table S1). In the AFC rule-based classification model, three rules that matched the

Table 6. Identified Significant Fluorescent and Nonfluorescent Rules in Each Optimized Prediction Model<sup>a</sup>

model	conditions	fingerprint description	model	conditions	fingerprint description
A350	<b>Fluorescent rule 1</b>		Rhodamine	<b>Fluorescent rule 1</b>	
	PubchemFP18 val=1	≥1 O		PubchemFP192 val=1	≥3 any ring size 6
	PubchemFP431 val=1	C(-C)(-C)(=N)		PubchemFP704 val=1	O=C-C-C-C-C-C
	PubchemFP621 val=1	N-C:C:C:N		PubchemFP737 val=1	Cc1cc(N)ccc1
	FR: 8%			FR: 24%	
	NR: 0.4%			NR: 0.7%	
	D: 3.10			D: 2.10	
	<b>Fluorescent rule 2</b>			<b>Fluorescent rule 2</b>	
	PubchemFP577 val=1	C:C-N-C:C		PubchemFP200 val=1	≥4 saturated or aromatic carbon-only ring size 6
	PubchemFP621 val=1	N-C:C:C:N		FR: 21%	
AFC	PubchemFP770 val=1	Nc1c(N)cccc1		NR: 0.1%	
	FR: 6%			D: 2.48	
	NR: 0.3%			<b>Fluorescent rule 3</b>	
	D: 3.09			PubchemFP758 val=1	Cc1c(N)cccc1
	<b>Fluorescent rule 1</b>			FR: 18%	
	PubchemFP184 val=1	≥1 unsaturated nonaromatic heteroatom-containing ring size 6		NR: 1%	
	PubchemFP185 val=1	≥2 any ring size 6		D: 2.21	
	PubchemFP490 val=1	C-C-C=C	Texas Red	<b>Fluorescent rule 1</b>	
	PubchemFP502 val=1	N-C=C-[#1]		PubchemFP182 val=1	≥1 unsaturated nonaromatic carbon-only ring size 6
	FR: 51%			PubchemFP376 val=1	C(~N)(:C)
	NR: 3%			FR: 17%	
	D: 2.75			NR: 0.5%	
A488	<b>Fluorescent rule 1</b>			D: 2.56	
	PubchemFP434 val=1	C(-C)(-H)(=C)		<b>Fluorescent rule 2</b>	
	PubchemFP443 val=1	C(-C)(=O)		PubchemFP153 val=1	≥2 saturated or aromatic heteroatom-containing ring size 5
	PubchemFP676 val=1	N#C-C-C-C		PubchemFP259 val=1	≥3 aromatic rings
	FR: 21%			PubchemFP340 val=1	C(~C)(~C)(~N)
	NR: 0.7%			PubchemFP502 val=1	N-C=C-[#1]
	D: 2.73			PubchemFP641 val=1	O-C-C-C=C
	<b>Fluorescent rule 2</b>			PubchemFP709 val=1	C-C(C)-C-C-C
	PubchemFP340 val=1	C(~C)(~C)(~N)		FR: 26%	
	PubchemFP434 val=1	C(-C)(-H)(=C)		NR: 0.9%	
	PubchemFP671 val=1	O=C-C=C-C		D: 2.50	
	FR: 25%			<b>Fluorescent rule 3</b>	
	NR: 4%			PubchemFP200 val=1	≥4 saturated or aromatic carbon-only ring size 6
	D: 2.00			FR: 17%	
	<b>Fluorescent rule 3</b>			NR: 0.4%	
	PubchemFP340 val=1	C(~C)(~C)(~N)		D: 2.66	
	PubchemFP633 val=1	N-C-C:C-C		<b>Nonfluorescent rule 1</b>	
	PubchemFP714 val=1	Cc1ccc(O)cc1		PubchemFP182 val=0	≥1 unsaturated nonaromatic carbon-only ring size 6
	FR: 13%			PubchemFP340 val=0	C(~C)(~C)(~N)
	NR: 2%			FR: 0%	
	D: 2.07			NR: 23%	
	<b>Nonfluorescent rule 1</b>			D: 4.06	
	PubchemFP368 val=1	C(~H)(~S)			
	FR: 0%				
	NR: 26%				
	D: 4.89				

<sup>a</sup>The rules are composed of different PubChem substructure fingerprints. “val=1” and “val=0” denote presence and absence of the PubChem fingerprints, respectively. Only the presence of PubChem fingerprints are shown in the table unless the absences of PubChem fingerprints are the dominant descriptors in the rules. In the PubChem fingerprints between PubchemFP327 and PubchemFP415, “~” denotes any bond orders and “:” denotes bond aromaticity. In the PubChem fingerprints between PubchemFP460 and PubchemFP712, the descriptors are denoted by SMARTS patterns and bond aromaticity matches both single and double bonds. The matched rate of fluorescent compounds, the matched rate of nonfluorescent compounds, and our defined contributed degrees are also represented followed by each rules.

chemical properties of sialic acid, misclassify sialic acid as a nonfluorescent molecule with high confidence score. However, natural fluorescence of sialic acid involves  $\text{IO}_4^-$ -lutidine derivatization,<sup>42</sup> for this reason, the misclassification of sialic acid was disregarded.

Among the A488 external data set, three autofluorescent isoquinoline alkaloids, including berberine, jatrorrhizine, and palmatine were also misclassified as nonfluorescent molecules. Further inspection revealed that all of matched rules which guide the three alkaloids to be misclassified in our prediction model

were in only 0.625 of confidence score. This could imply that the PubChem fingerprints of the selected compounds for A488 fluorescent training set did not adequately contain the structural properties of these three alkaloids, and thus, caused them to the misclassification. To promote the precision of A488 fluorescence classification model, berberine, jatrorrhizine, and palmatine were included in our A488 fluorescent training data set. After the re-establishment of A488 fluorescence classification model for the new training data set including 63 fluorescent and 1492 nonfluorescent compounds, the accuracy, sensitivity, specificity, and G-mean were improved, 98.3%, 100%, 98.3%, and 99.1%, respectively, for the training set and 93.7%, 100%, 93.7%, and 96.8%, respectively, for the testing set (data not shown).

Since the excitation/emission wavelengths of Rhodamine and Texas Red fluorophores are within the same range, the same external data set was used for the evaluation of both prediction models. Interestingly, Propidium Iodide was correctly classified in the Rhodamine model, but misclassified in the Texas Red model. In our system, if a compound was identified by one of the five prediction models as being fluorescent, it is regarded as a autofluorescent compound. Therefore, the misclassification of Propidium Iodide by Texas Red model will not affect the overall prediction power.

A diverse collection of fluorescent and nonfluorescent compounds has been considered and incorporated into the development and training of our prediction models. Although the five models are applicable to compounds satisfying the specific excitation/emission profiles chosen, and may not be efficient for compounds outside the considered wavelengths, nevertheless these models are novel and convenient tools for precise detection of autofluorescence. More importantly, a series of comprehensive fluorescent and nonfluorescent rules were identified, they will be discussed in the following sections.

**Characterization of the Contributions of Fluorescent Rules to the Classification Models.** So far, there are insufficient structural properties that can be used to identify natural fluorescent compounds. Our five optimized classification models have deduced 14 reliable new rules for the detection of natural fluorescence. On top of utilizing these models directly for prediction of natural fluorescence, these new rulesets can further served as a guideline for designing drug candidates or autofluorescent molecules in the early drug development processes.

The rulesets for autofluorescence prediction were derived from an empirical equation ( $D$ ), which determines the degree to which any given rules contributed to the optimized fluorescence models:

$$D = \log\left(N_F \times \frac{FR}{NR + 0.01}\right) \text{ for fluorescent rule,} \\ \text{or } \log\left(N_N \times \frac{NR}{FR + 0.01}\right) \text{ for nonfluorescent rule} \quad (5)$$

where  $N_F$  is the total number of fluorescent rules in the corresponding fluorescence data set,  $N_N$  is the total number of nonfluorescent rules in the corresponding fluorescence data set, FR is the percentage of fluorescent compounds satisfying a rule condition, and NR is the percentage of nonfluorescent compounds satisfying a rule condition. A good discriminating fluorescent rule should contain as many FR as possible, as few NR as possible, and vice versa for a good discriminating non-fluorescent rule. Since our fluorescence prediction models are composed of  $N_F$  fluorescent and  $N_N$  nonfluorescent rules, more number of rules decreases the degree of contribution of each rule

to the prediction model.  $N_F$  and  $N_N$  are then considered as weighting factors in eq 5 so that the values of  $D$  can be standardized between different prediction models. To ensure that the errors from division by zero does not occur, 0.01 is added to the denominator in equation  $D$ . The average number of rules in 5 optimized fluorescent models is 19. Assume we expected the ratio of FR to NR is 10 at least, the resultant  $D$  is near to 2.27. We approximately adopt 2.0 of  $D$  as our criterion to select the significant fluorescent or nonfluorescent rules from the prediction models. In Table 6, the identified significant fluorescent and nonfluorescent rules whose values of  $D$  are higher than 2.0 were listed in each model. The rules are composed of different PubChem substructure fingerprints where “val=1” and “val=0” denote presence and absence of the PubChem fingerprints, respectively. Because the absences of descriptors are not easily observed and directly expressed on the molecules, only the presence of PubChem fingerprints were shown in the table unless the absences of PubChem fingerprints were the dominant descriptors in the rule.

In A350 model, there are two significant fluorescent rules whose FRs are near to 7%. Actually, the A350 model is totally constituted by 64 rules. A compound which is fluorescent or nonfluorescent is codetermined by all of the 64 rules. Therefore, we can observe that even the two significant rules have lower FRs, they still produce higher  $D$  than most of rules in other models. According to the identified fluorescent rule 1 in the A350 model, a compound possessing “one or more oxygen atoms”, “2-Propanimine”, and “1-Propen-1-amine, 3-imino-” could be predicted as an autofluorescent molecule. If a compound contains substructures of “ethenamine, N-ethenyl-”, “1-propen-1-amine, 3-imino-”, and “1,2-diaminobenzene”, the compound could be also predicted as an autofluorescent molecule. The descriptor of “1-propen-1-amine, 3-imino-” is a quite important fluorescent property as it existed in both fluorescent rules of A350 model. In the AFC model, we discerned one significant fluorescent rule constituted by structural characteristics of “one or more unsaturated non-aromatic heteroatom-containing ring size 6”, “more than one any ring size 6”, “1-butene”, and “ethenamine whose C terminal only can be connected to one hydrogen atom”. In the A488 model, there are three identified fluorescent rules. The descriptors in first fluorescent rule of A488 model are composed of “propylene”, “ethanol”, and “butyronitrile”. The descriptors in second fluorescent rule of A488 model contain “2-propylamine”, “propylene”, and “butenal”. The descriptors in third fluorescent rule of A488 model include “2-propylamine”, “(E)-but-2-en-1-amine”, and “4-cresol”. We also found one representative nonfluorescent rule from the A488 model. If a structure contains a carbon atom which is adjacent to one hydrogen atom and one sulfur atom with any bond orders, the compound might have no fluorescence in high possibilities. In Rhodamine fluorescent model, we also recognized three significant fluorescent rules. “More than two any ring size 6”, “heptanal”, and “2-propanimine” are three important descriptors depicted in the first fluorescent rule. The descriptors in the fluorescent rules 2 and 3 are, respectively, “more than three saturated or aromatic carbon-only ring size 6” and “2-toluidine”. Most of the representative PubChem fingerprints in Rhodamine model are related to the multiple or conjugated rings, and the finding is consistent with the well-known fluorescent properties. In Texas Red model, three important fluorescent and one nonfluorescent rules possessing high value of  $D$  are determined. The first fluorescent rule of Texas Red refers to “one or more unsaturated non-aromatic

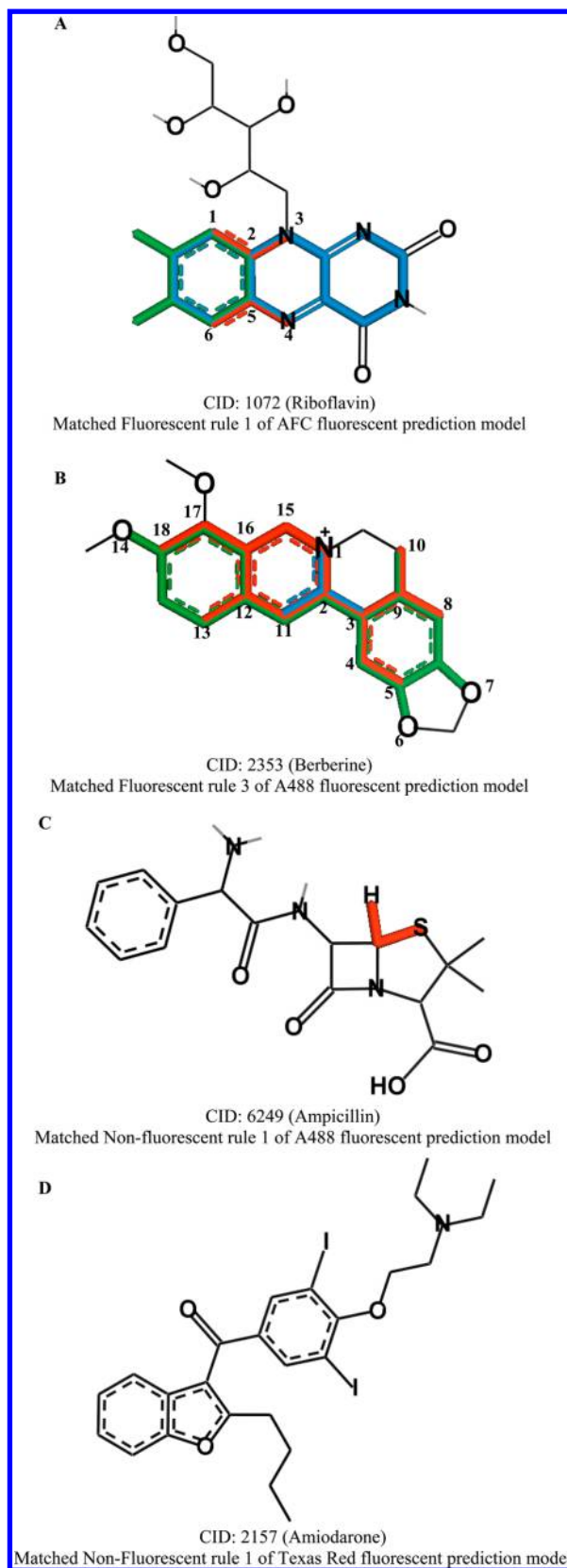


carbon-only ring size 6" and "one carbon atom connected to one nitrogen atom regardless of bond orders and another aromatic carbon atom". The second important fluorescent rule of Texas Red model includes "more than one saturated or aromatic heteroatom-containing ring size 5", "more than two aromatic rings", "2-propylamine", "ethenamine", "3-buten-1-ol", and "2-methylpentane". The third representative fluorescent rule of Texas Red model refers to "more than three saturated or aromatic carbon-only ring size 6". The nonfluorescent rule in Texas Red model is composed of two absence of descriptors including "one or more unsaturated non-aromatic carbon-only ring size 6", and "2-methylpentane". It is noted that terminal atoms defined in all of the above depicted descriptors of five models can be connected to atoms in any bond orders. Among the 14 rules, the descriptor PubchemFP340 is a much important descriptors as it appeared four times in different rules, three presences for fluorescence and one absence for nonfluorescence. The PubchemFP340 is depicted as "2-propylamine" above but exactly it refers to "a carbon atom connected to two carbon and one nitrogen atoms in any bond orders". In the next section we will takes different fluorescent or nonfluorescent compounds as examples to express and discuss each significant rules.

**Interpretation of Rules for Prediction of Molecular Autofluorescence.** A bonus contribution that comes from the rule sets in our fluorescence models is that the representative 2D substructural properties can be identified from the rule-based model. Fourteen autofluorescent and nonfluorescent compounds from the training, testing, and external data sets were selected to demonstrate the applicability of the 14 identified vital structural rules listed in Table 6, respectively. However, in this section, only two well-known fluorescent and two non-fluorescent compounds (Figure 1) were chosen for further discussion. The remaining ten compounds and their corresponding discussion are included as supplement (Supplementary Figures 4–7).

Riboflavin (vitamin B<sub>2</sub>), a yellow fluorescent compound, was chosen as an example from the AFC prediction model (Figure 1A). There are two unsaturated nonaromatic heteroatom-containing rings of size 6 (PubchemFP184) and totally three any rings of size 6 (PubchemFP185). We combined the features, PubchemFP184 and PubchemFP185, shown in blue. PubchemFP490 intends four carbons chain having one double bond in terminal end. Again, since bonds in aromatic ring can be regarded as single or double bonds, and there is one aromatic ring in Riboflavin, any paths in length of three within the aromatic ring of Riboflavin and two carbon atoms adjacent to the aromatic ring forms the PubchemFP490 descriptor, which was illustrated in green. The last key structural feature, described by the PubchemFP502 descriptor, defines the structures similar to Ethenamine of which C terminal can only be connected to a hydrogen atom, were shown in red paths (1, 2, and 3) and (4, 5, and 6). Riboflavin contains an isoalloxazine ring joined to a ribityl group. Our suggested fluorescent characteristics were all located on the isoalloxazine ring. The contribution of isoalloxazine ring in autofluorescence has been examined and reported in another study.<sup>43</sup> From this, we can confirm our identified fluorescent rules in the AFC model are sound.

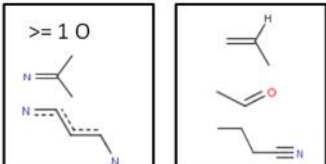
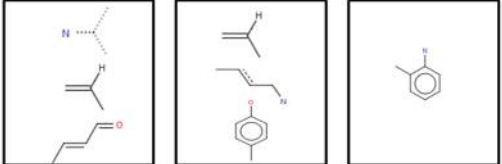
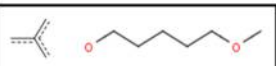



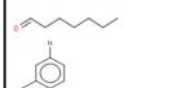
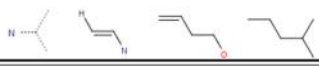



In Figure 1B, Berberine was chosen as an example to illustrate the representative structural features of fluorescent rule 3 in the A488 model. Comparable to fluorescent rule 2, PubchemFP340 is also a key fingerprint in fluorescent rule 3. PubchemFP340 appeared once in the center of Berberine, as shown in blue. The second feature of fluorescent rule 3, PubchemFP633, denotes a nitrogen atom joined by four carbons where the second and third



**Figure 1.** Examples of fluorescent compounds (A and B) and nonfluorescent compounds (C and D) which are matched to our identified rules in optimized fluorescent prediction model.

carbon atoms have to be aromatic. Examples of PubchemFP633 are highlighted in red and include the following paths: (1, 2, 3, 4, and 5),



Fluorescent rules	$\geq 1$ rings	<div> <math>\geq 1</math> O  </div> <div>  </div>
	$\geq 2$ rings	<div>  </div> <div> <math>\geq 1</math> unsaturated non-aromatic heteroatom-containing ring size 6  <math>\geq 2</math> any ring size 6  </div> <div> <math>\geq 1</math> unsaturated non-aromatic carbon-only ring size 6  </div>
	$\geq 3$ rings	<div> <math>\geq 3</math> saturated or aromatic heteroatom-containing ring size 6  </div> <div> <math>\geq 3</math> any ring size 6  </div> <div> <math>\geq 3</math> aromatic rings  <math>\geq 2</math> saturated or aromatic heteroatom-containing ring size 5  </div>
	$\geq 4$ rings	$\geq 4$ saturated or aromatic carbon-only ring size 6 
Non-fluorescent rules	-	<div> No any  </div> <div> No any unsaturated non-aromatic carbon-only ring size 6  </div>

**Figure 2.** Identified significant fluorescent and nonfluorescent rules according to different numbers of ring.

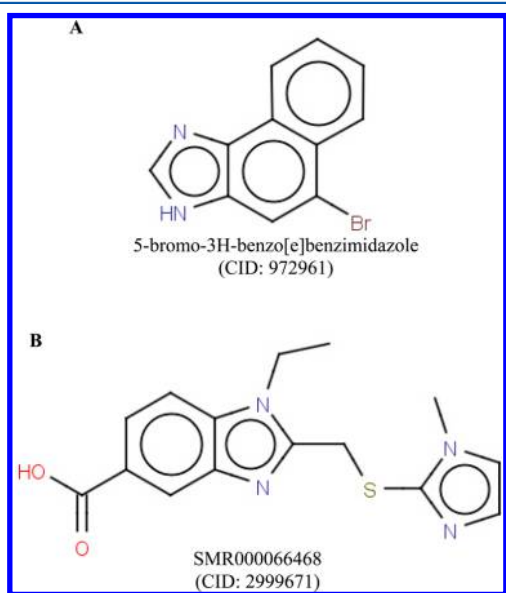
(1, 2, 3, 9, and 8), (1, 2, 3, 9, and 10), (1, 2, 11, 12, and 13), (1, 2, 11, 12, and 16), (1, 15, 16, 12, and 13), and (1, 15, 16, 17, and 18). The last fingerprint, PubchemFP714 (4-cresol), was highlighted in green. Benzenes between atom 6 and 10, atom 2 and atom 7, and atom 11 and atom 14 forms the substructure of 4-cresol. Berberine contains many structural features important for compound fluorescence, and indeed, they are known strong fluorescent compounds.

An important nonfluorescent rule for A488 prediction model is listed in Table 6, and illustrated in the Ampicillin example, in Figure 1C. PubchemFP368, highlighted in red, denotes a carbon atom adjacent to one hydrogen and one sulfur atoms. Any compound with this feature will be classified as nonfluorescent.

In Figure 1D, Amiodarone was the chosen representative nonfluorescent compound to demonstrate the applicability of non-fluorescent rule 1 from the Texas Red prediction model. Non-fluorescent rule 1 listed under the Texas Red prediction model contains two fingerprints (Table 6), PubchemFP182 and PubchemFP340. Interestingly, none of the two fingerprints were observed in Amiodarone. It is worth noting that the absence of nonfluorescent fingerprints, does not correlate to compound fluorescence. Also, none of the key fingerprints for autofluorescence is observed. Amiodarone is a known nonfluorescent compound. We can apply the above identified rules to any compounds for detection of fluorescence during the early stage of drug design.

**Exploration of Novel Extended Rules for Molecular Autofluorescence.** To determine whether the aforementioned rulesets can be utilized for the rapid prediction of autofluorescence, the fluorescent rules (exclude nonfluorescent rules) were divided into four categories (Figure 2), based on the rings counts. The four categories of fluorescent rules are  $\geq 1$  ring,  $\geq 2$  rings,  $\geq 3$  rings, and  $\geq 4$  rings. Under the first category, there are five rules represented by five boxes. The same was presented in the latter three categories.

Most fluorescent compounds are highly conjugated, but highly conjugated compounds may not be fluorescent in nature. Two nonfluorescent compounds, 5-bromo-3*H*-benzo[*e*]benzimidazole and SMR000066468, with highly conjugated structures were illustrated in Figure 3. The structures presented in Figure 3 can



**Figure 3.** Example of highly conjugated compounds which are nonfluorescent.

be correctly classified as nonfluorescent compounds, using the inductive rules listed in Figure 2. First, determine whether the structures satisfy any of the nonfluorescent rules. Since nonfluorescent fingerprints were not present in the structure of 5-bromo-3*H*-benzo[*e*]benzimidazole (Figure 3A), the next step involves the identification of fluorescent fingerprints within the structure. Of all the fluorescent fingerprints presented in Figure 2, the seven fingerprints related to oxygen atom could be disregarded, because the structure of 5-bromo-3*H*-benzo[*e*]benzimidazole does not contain an oxygen atom. The remaining four possibilities are the only rule with “2-toluidine” in the first category, the rules associated with nonaromatic rings in the second category, the only rule associated with three or more saturated or aromatic heteroatom-containing six-membered rings in the third category, or the only rule presented in the fourth category associated with four or more saturated or aromatic carbon-only six-membered rings. The structure of 5-bromo-3*H*-benzo[*e*]benzimidazole does not contain “2-toluidine”, and out of the three aromatic rings, only the five-membered ring consists of heteroatom. It can be concluded from the deductive reasoning above that 5-bromo-3*H*-benzo[*e*]benzimidazole does not satisfy any fluorescent rules; therefore, it should be classified as a nonfluorescent compound according to the rules listed in Table 6. In fact, 5-bromo-3*H*-benzo[*e*]-

benzimidazole is a nonfluorescent compound. In the second example (Figure 3B), SMR000066468 contains a sulfur atom joined by one saturated carbon atom. This is the determining nonfluorescent property, as shown in Table 6. The illustrative examples in Figure 3 demonstrated the applicability of our novel expanded fluorescent/nonfluorescent rules, which could be easily and rapidly applied to distinguish the fluorescence properties of a compound, with the appropriate excitation/emission profile.

## CONCLUSION

In this study, five *in silico* rule-based autofluorescence prediction models were generated. These models were each developed, tested and validated for a commonly used fluorescent agent listed below: Alexa Fluor 350, 7-amino-4-trifluoromethyl-coumarin (AFC), Alexa Fluor 488, Rhodamine, and Texas Red. Using the highly skewed data set to build prediction models for autofluorescence is inevitable. Oversampling of fluorescent data and undersampling of nonfluorescent data effectively optimized the prediction power of these models, which yielded an average accuracy, sensitivity, and specificity of 97.0%, 96.0%, and 97.8% among the five models for the training set and 91.7%, 93.1%, and 91.7% among the five models for the testing set. Our systems can be utilized to predict fluorescence in compounds with the indicated excitation/emission profiles.

Another important contribution in our study is the identification as well as organization of structural characteristics for fluorescent and nonfluorescent compounds. A total of 14 novel rules (fingerprints or structural features) have been discussed in detail, and demonstrated in various fluorescence predictions. Etazolate, Daunorubicin, Riboflavin, Berberine, and Propidium Iodide are all well-known fluorescent compounds which have been correctly classified using the 14 novel rules. To allow for rapid determination of autofluorescence, these rules were organized into four categories accompanied by graphical representations of the rules or fingerprints important for fluorescence. This enables the rapid classification of compounds.

In overall, our molecular autofluorescence prediction models can be used as a tool to find out compounds with autofluorescence properties in specific excitation/emission spectra. By *in silico* prescreening based on our prediction models, the time and cost of *in vitro* fluorescence detection before performing HTS can be reduced. Furthermore, our identified novel fluorescent and nonfluorescent rules can be used as a new guidance to easily detect autofluorescence or modify the structures during the early stage of drug development processes.

## ASSOCIATED CONTENT

### Supporting Information

This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*Tel.: +886.2.3366.4888 #529. Fax: +886.2.2362.8167. E-mail: [yjtseng@csie.ntu.edu.tw](mailto:yjtseng@csie.ntu.edu.tw).

### Author Contributions

<sup>§</sup>B.-H.S. and Y.-S.T. share equal contribution in this work.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was funded by the Ministry of Science and Technology, Taiwan, grants number 103-2325-B-002-048. Resources of the Laboratory of Computational Molecular Design and Detection and Department of Computer Science and Information Engineering of National Taiwan University were used in performing these studies. We are grateful to the National Center for High-performance Computing for computer time and facilities.

## REFERENCES

- (1) Turek-Etienne, T. C.; Small, E. C.; Soh, S. C.; Xin, T. A.; Gaitonde, P. V.; Barrabee, E. B.; Hart, R. F.; Bryant, R. W. Evaluation of fluorescent compound interference in 4 fluorescence polarization assays: 2 kinases, 1 protease, and 1 phosphatase. *J. Biomol. Screening* **2003**, *8* (2), 176–184.
- (2) Bannwarth, W.; Hinzen, B. *Combinatorial Chemistry: From Theory to Application*, 2nd ed.; Wiley VCH, 2006.
- (3) Chen, B.-H.; Wang, C.-C.; Lu, L.-Y.; Hung, K.-S.; Yang, Y.-S. Fluorescence assay for protein post-translational tyrosine sulfation. *Anal. Bioanal. Chem.* **2013**, *405* (4), 1425–1429.
- (4) Owicki, J. C. Fluorescence polarization and anisotropy in high throughput screening: perspectives and primer. *J. Biomol. Screening* **2000**, *5* (5), 297–306.
- (5) Hemmälä, I.; Webb, S. Time-resolved fluorometry: an overview of the labels and core technologies for drug screening applications. *Drug Discovery Today* **1997**, *2* (9), 373–381.
- (6) Selvin, P. R. The renaissance of fluorescence resonance energy transfer. *Nat. Struct. Biol.* **2000**, *7* (9), 730–734.
- (7) Pritz, S.; Meder, G.; Doering, K.; Drueckes, P.; Woelcke, J.; Mayr, L. M.; Hassiepen, U. A fluorescence lifetime-based assay for Abelson kinase. *J. Biomol. Screening* **2011**, *16* (1), 65–72.
- (8) Auer, M.; Moore, K. J.; Meyer-Almes, F. J.; Guenther, R.; Pope, A. J.; Stoeckli, K. A. Fluorescence correlation spectroscopy: lead discovery by miniaturized HTS. *Drug Discovery Today* **1998**, *3* (10), 457–465.
- (9) Rüdiger, M.; Haupts, U.; Moore, K. J.; Pope, A. J. Single-molecule detection technologies in miniaturized high throughput screening: binding assays for G protein-coupled receptors using fluorescence intensity distribution analysis and fluorescence anisotropy. *J. Biomol. Screening* **2001**, *6* (1), 29–37.
- (10) Ma, H.; Deacon, S.; Horiuchi, K. The challenge of selecting protein kinase assays for lead discovery optimization. *Expert Opin. Drug Discovery* **2008**, *3* (6), 607–621.
- (11) Natasha, T.; Douglas, S. A.; James, I. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Curr. Opin. Chem. Biol.* **2010**, *14* (3), 315–324.
- (12) Sink, R.; Gobec, S.; Pecar, S.; Zega, A. False Positives in the Early Stages of Drug Discovery. *Curr. Med. Chem.* **2010**, *17* (34), 4231–4255.
- (13) Meyners, C.; Wawrzinek, R.; Krämer, A.; Hinz, S.; Wessig, P.; Meyer-Almes, F. J. A fluorescence lifetime-based binding assay for acetylpolymine amidohydrolases from *Pseudomonas aeruginosa* using a [1,3]dioxolo[4,5-f][1,3]benzodioxole (DBD) ligand probe. *Anal. Bioanal. Chem.* **2014**, *406* (20), 4889–4897.
- (14) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740.
- (15) Simeonov, A.; Jadhav, A.; Thomas, C. J.; Wang, Y.; Huang, R.; Southall, N. T.; Shinn, P.; Smith, J.; Austin, C. P.; Auld, D. S.; Inglese, J. Fluorescence Spectroscopic Profiling of Compound Libraries. *J. Med. Chem.* **2008**, *51* (8), 2363–2371.
- (16) de Kort, B. J.; de Jong, G. J.; Somsen, G. W. Native fluorescence detection of biomolecular and pharmaceutical compounds in capillary electrophoresis: detector designs, performance and applications: a review. *Anal. Chim. Acta* **2013**, *766*, 13–33.
- (17) Valeur, B. Characteristics of fluorescence emission. In *Molecular Fluorescence: Principles and Applications*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2002; pp 34–71.
- (18) Valeur, B. Introduction. In *Molecular Fluorescence*; Wiley-VCH Verlag GmbH, 2001; pp 3–19.
- (19) Guilbault, G. G. *Practical Fluorescence*, 2nd ed.; Marcel Dekker, Inc.: New York, 1990.
- (20) Haugland, R. P. *The handbook: a guide to fluorescent probes and labeling technologies*; Molecular probes: Eugene, OR, 2005.
- (21) Southern Research Molecular Libraries Screening Center (SRMLSC). *PubChem Assays AID 709*; 2007.
- (22) Albert-Garcia, J. R.; Antón-Fos, G. M.; Duarte, M. J.; Lahuerta Zamora, L.; Martínez Calatayud, J. Theoretical prediction of the native fluorescence of pharmaceuticals. *Talanta* **2009**, *79* (2), 412–418.
- (23) Chang, C.-Y.; Hsu, M.-T.; Esposito, E. X.; Tseng, Y. J. Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition, Molecular Descriptors, and Predictive Modeling Methods. *J. Chem. Inf. Model.* **2013**, *53* (4), 958–971.
- (24) Kuhn, M.; Weston, S.; Coulter, N.; Quinlan, R. C50: C5.0 Decision Trees and Rule-Based Models. <http://CRAN.R-project.org/package=C50> (accessed March 4, 2014).
- (25) Modi, S.; Li, J.; Malcomber, S.; Moore, C.; Scott, A.; White, A.; Carmichael, P. Integrated in silico approaches for the prediction of Ames test mutagenicity. *J. Comput.-Aided Mol. Des.* **2012**, *26* (9), 1017–1033.
- (26) Lin, Z.-H. Computational classification molecular fluorescence models. Master's Thesis, National Taiwan University, Taipei, Taiwan, 2012.
- (27) Bestvater, F.; Spiess, E.; Stobrawa, G.; Hacker, M.; Feurer, T.; Porwol, T.; Berchner-Pfannschmidt, U.; Wotzlaw, C.; Acker, H. Two-photon fluorescence absorption and emission spectra of dyes relevant for cell imaging. *J. Microsc.* **2002**, *208* (2), 108–115.
- (28) Möller, L.; Krause, A.; Bartsch, I.; Kirschning, A.; Witte, F.; Dräger, G. Preparation and In Vivo Imaging of Lucifer Yellow Tagged Hydrogels. *Macromol. Symp.* **2011**, *309–310* (1), 222–228.
- (29) Watson, D. G. *Pharmaceutical Analysis*, 2nd ed.; Elsevier Health Sciences: UK, 2005.
- (30) Shen, M.-Y.; Su, B.-H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A Comprehensive Support Vector Machine Binary hERG Classification Model Based on Extensive but Biased End Point hERG Data Sets. *Chem. Res. Toxicol.* **2011**, *24* (6), 934–949.
- (31) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.
- (32) Ross Quinlan, J. Is See5/C5.0 Better Than C4.5? <http://rulequest.com/see5-comparison.html> (accessed March 4, 2014).
- (33) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (3), 572–584.
- (34) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered. *J. Chemom.* **2000**, *14* (5–6), 599–616.
- (35) Haranczyk, M.; Holliday, J. Comparison of Similarity Coefficients for Clustering and Compound Selection. *J. Chem. Inf. Model.* **2008**, *48* (3), 498–508.
- (36) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1–40.
- (37) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981.
- (38) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; John Wiley: New York, 2001.
- (39) Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3* (3), 32–57.
- (40) Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M. Y.; Mahfouf, M.; Lawson, K.; Mullier, G. Clustering files of chemical structures using the fuzzy k-means clustering method. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 894–902.
- (41) Babuška, R.; Alic, L.; Louren, M. S.; Verbaak, A. F. M.; Bogaard, J. Estimation of respiratory parameters via fuzzy clustering. *Artif. Intell. Med.* **2001**, *21* (1–3), 91–105.

- (42) Shukla, A. K.; Schauer, R. Fluorimetric determination of unsubstituted and 9(8)-O-acetylated sialic acids in erythrocyte membranes. *Hoppe-Seyler's Z. Physiol. Chem.* **1982**, 363 (3), 255–262.
- (43) Visser, A. J. W. G.; Müller, F. Absorption and Fluorescence Studies on Neutral and Cationic Isoalloxazines. *Helv. Chim. Acta* **1979**, 62 (2), 593–608.