Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

# Quantitative criterion for AIEgens†

Junyi Gong,[a] Jacky W. Y. Lam[a] and Ben Zhong Tang*[a,b,c,d]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

**We defined two novel descriptors to demonstrate the flexibility of both chemical and electronic structures of organic fluorescence compounds upon excitation. Classification algorithms were introduced to predict the aggregation-induced emission behavior from the chemical structures based on the new descriptors. A dataset was built to train the classifier, which is optimized to 87.3% accuracy finally.**

Aggregation-induced emission (AIE) phenomenon was reported and termed in 2001 and made a new epoch of light-emitting materials field.[1–3] AIE luminogens, with strong emission in aggregates and solid state but weak emission in dilute solution, brings new possibilities for development of fluorescence imaging agents, photovoltaic devices, chemosensors and actuators. Many hypothesizes of the relationship between AIE property and chemical structures were claimed when few years after the first AIE paper published but most of them are abstract and tremendously relying on experimental verification.[4,5]

Restriction of intramolecular motion (RIM) mechanism is the milestone of research about AIE behaviour.[6,7] RIM claimed that the intramolecular motion (rotation and vibration) from flexible elements in chemical structure in AIE luminogens after excitation would lead them to a conical intersection or a weak emission point,[8] which behave no or weak emission in dilute form but will be restricted by aggregates formation or in the solid-state, displaying strong luminescence in such forms. So far, RIM mechanism is the essential criterion for researchers to determine if compounds are AIE or aggregation-caused quenching (ACQ).[9] However, the RIM criterion is highly abstract and dependent on the experiences of researchers. Many confusing cases were reported, implying high demand for a quantitative and transferable criterion.

Machine learning brings fresh air to chemical research with a powerful statistical solution to classification or regression problems.[10,11] It is highly modifiable and extendable, could be easily transferred and implemented, and could process high throughput data flow automatically. Our contributors proposed a classification model for AIE systems based on triphenylamine fused donor-acceptor structures with high accuracy of 84%.[12] However, this model is not universally applicable, which means the general criterion for different types of AIEgens is still in high demand. In this contribution, we are using popular machine learning algorithms to classify the information from quantum mechanics (QM) calculation of chemical structures with AIE or ACQ labels. For digitizing raw data from chemical structures, we derived two novel factors to describe the flexibility in conformation and electronic structure of any given systems. The final model could classify digitized information from chemical structures to AIE or ACQ with not only generally suitable potential but also high accuracy of 87.3% and high sensitivity of 94.1%.

Realizing the AIE phenomenon is the first step to quantify it. The chemical structure should include all information of a small molecular compound and determine the AIE/ACQ behaviour. From the preliminary results,[13,14] we know that aggregation-induced emission systems are almost non-emissive or weakly emissive in dilute solutions due to the energy dissipation during molecular motion after excitation. We may conclude that photo-flexibility is the first feature that AIE systems have. Photo-flexibility means the geometry of organic compounds are potential to move drastically after excitation. In addition, this flexibility would always lead the system to a state with weaker or no emission, indicating that oscillator strength of corresponding transition would also change a lot. For a simple AIEgen with equilibrium structure at A, it will go to an excited state and then re-

[a] *Department of Chemistry, Hong Kong Branch of Chinese National Engineering Research Center for Tissue Restoration and Reconstruction, Institute for Advanced Study, State Key Laboratory of Molecular Nanoscience, Division of Life Science and Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China; E-mail: tangbenz@ust.hk*
[b] *Center for AIE Research, Shenzhen Key Laboratory of Polymer Science and Technology, Guangdong Research Center for Interfacial Engineering of Functional Materials, College of Material Science and Engineering, Shenzhen University, Shenzhen 518061, China;*
[c] *Center for Aggregation-Induced Emission, SCUT-HKUST Joint Research Institute, State Key Laboratory of Luminescent Materials and Devices, South China University of Technology, Guangzhou 510640, China;*
[d] *AIE Institute, Guangzhou Development District, Huangpu, Guangzhou 510530, China.*

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 00.0000/00000000.

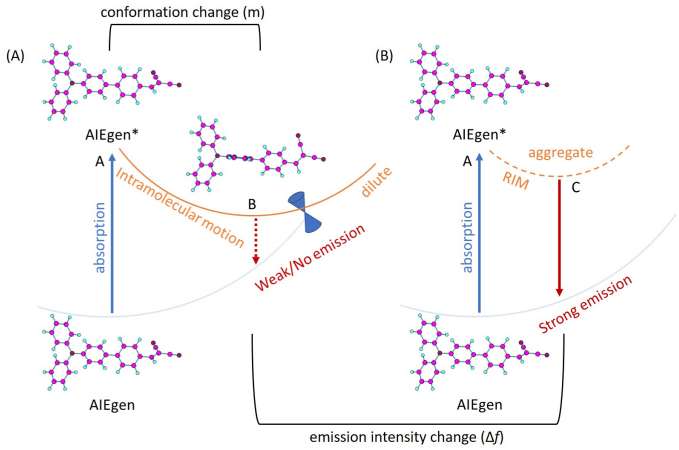


**Fig. 1** Photo-physical process of AIE system in dilute form (A) and aggregates (B).

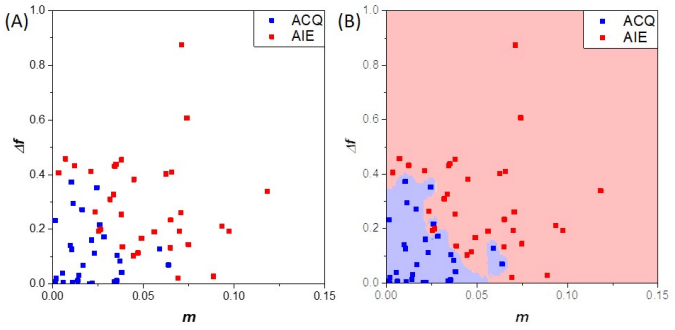


**Fig. 2** (A) Scatter of m and $\Delta f$ value of AIE and ACQ compounds in dataset; (B) Classified region of weighted KNN model with $k = 10$ and $1/d^2$ as weighted function.

laxed to a weak or non-emission conformation B through molecular motion after excitation as shown in figure 1A. However, in aggregates or solid-state, the molecular motion would be inhibited, and the relaxation process ended in a conformation C with stronger emission and less coordination change, as shown in figure 1B. Because we cannot get exact conformation of B from simple DFT method without packing information, extremely speaking, we can assume that confirmation of B is very close to A instead of C. So if a compound is AIE, it is evident that the change of oscillator strength and conformation from A to C need to be as large as possible. The structures information at conformation A and B could be calculated through quantum mechanical methods. By using ORCA 4.2.1 quantum mechanical calculation suite, [15] we optimized the structure at the ground and S1 states under BLYP functional with def2-SVP basis set and RI-approximation switch on. The oscillator strengths values were obtained with TD-DFT calculation under PBE0 hybrid functional with def2-SVP basis set and RIJCOSX-approximation switch on. [16,17]

The oscillator strength change ($\Delta f$) could be calculated by simple DFT and TD-DFT method:

$$\Delta f = |f_{GS}^{eq} - f_{S1}^{eq}| \tag{1}$$

In which $f_{GS}^{eq}$ denote the oscillator strength at equilibrium conformation at ground state (A) and $f_{S1}^{eq}$ denote the oscillator strength at equilibrium conformation at S1 state (B). To quantify conformation change, we defined the mean absolute deviation of atom coordination (after aligning) as $m$ and have:

$$m = \frac{1}{n}\sum\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \tag{2}$$

In which $n$ denotes the atom number, and $\Delta x$, $\Delta y$, $\Delta z$ denote the deviations of coordinators before and after a molecular motion from A to B of an atom. For any given chemical structure, the value of $\Delta f$ and $m$ could be calculated by QM optimization and TD-DFT calculation. A Python script for automatic calculation of the factors is attached. An original dataset with 29 typical ACQ cores and 34 typical AIEgens were inputted into the process and get all the value of factors, which are plotted in figure 2A as scatter point. Since the dataset is relatively small because the exceptions of structures with same AIE cores, high-folds cross-validations are required for better performance and reliable results. It could be indicated obviously that ACQ cores have generally smaller $\Delta f$ and $m$ than AIEgens, implying that the two new descriptors m and $\Delta f$ have the general potential for predicting AIE/ACQ behaviour of a random compound with chemical structure.

The machine learning process was performed on MATLAB R2019a software with its own classification learner suite and 20-fold cross-validation. Decision tree is a basic method for classification. [18] After optimization, maximum deviance reduction was interpreted as split criterion and he maximum number of splits is 11. The validated accuracy is 84.1%; Supports-vector machine (SVM) was used for classification for decades. [19] SVM provided powerful performance and highly modifiable expression. By adjusting the kernel function and parameter, SVM provided flexibility in different types of data. After optimization, gaussian function was chosen as kernel on raw data without standardization. The validated accuracy is 85.7%; Naive Bayes classifier is another classifier which could provide not only class but also possibility of a data point belongs to a specific category. [20] Triangle function was optimized as kernel function. The validated accuracy is 84.1%, too; We also build a diagonal quadratic model for reference. [21] The validated accuracy is also 84.1%. The decision surfaces of above four models with original scatter dots were shown in figure S1.

K nearest neighbours (KNN) method is one of the strongest and easy-handling algorithms among available machine-learning-based methods, which is non-parametric and distance-based. [22,23] In the KNN model, an object should be assigned to the class most common among its k nearest neighbours. For better performance, we introduced weighted function in KNN model. [24] The purpose of the weighted function is to to give more weight to the points which are nearby and less weight to the points which are farther away. We use $1/d^2$ as weighted function in which $d$ denotes the distance between data points and object waiting for assigning. A Python script for predicting the AIE/ACQ behaviour of new datapoint based on KNN algorithm without cross-validation was attached.

**Table 1** Performance data of all used models (AIE as positive).

| Models | True positive | True negative | False positive | False negative | Sensitivity | Specificity | Precision | Accuracy |
|---|---|---|---|---|---|---|---|---|
| KNN | 32 | 23 | 6 | 2 | 0.941 | 0.793 | 0.842 | 0.873 |
| SVM | 30 | 24 | 5 | 4 | 0.882 | 0.828 | 0.857 | 0.857 |
| Quadratic | 28 | 25 | 4 | 6 | 0.824 | 0.862 | 0.875 | 0.841 |
| Naive bayes | 29 | 24 | 5 | 5 | 0.853 | 0.828 | 0.853 | 0.841 |
| Decision trees | 32 | 21 | 8 | 2 | 0.941 | 0.724 | 0.800 | 0.841 |

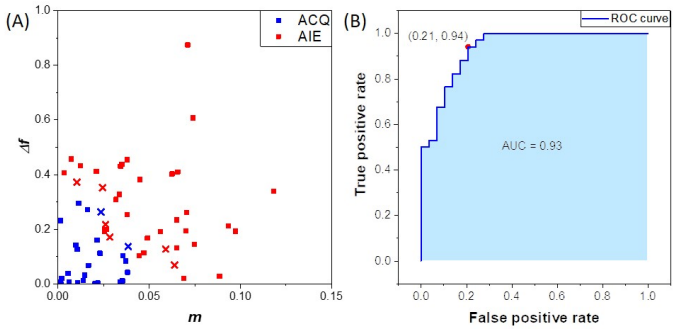


**Fig. 3** (A) Scatter diagram of model predictions; (B) ROC curve of AIE as positive class and AUC value. Orange dot corresponding to the best threshold of $m$ and $\Delta f$.

Figure 2B displayed the decision region for ACQ and AIE luminogens based on the raw method and data. It's evident that over-fitting was appeared and made the decision region tattered and incompatible to the physical picture. To avoid over-fitting, cross-validation was introduced. When K equals from 1 to 8, the accuracy increased stepwise. When K equals 8 or beyond, the accuracy gets to its maximum of 87.3%. Finally we choose ten as k value for better robustness. The final model was plotted in figure 3A. The ROC curve of the weighted KNN model were shown in figure 3B. AUC value was calculated as 0.93, which was larger than reported models, implying that $m$ and $\Delta f$ are more reliable descriptors to classify AIEgens and ACQ fluorescence materials. Assuming that TP denotes true positives, TN denotes true negatives, FP denotes false positives, FN denotes false negatives. The performance of a model could be evaluated by such factors:

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

All performances of the different classifier were displayed in figure S2 and Table 1 in which we could be indicated that in general, weighted KNN performed the best.

In this contribution, we discovered a new quantitative criterion from classic restriction of intramolecular motion theory: $m$ and $\Delta f$ to describe the photo-flexibility and optical instability in excited state dynamics of a specific compound, respectively. For classification of unknown materials based on known examples of AIE and non-AIE materials, supervised learning algorithms including supports-vector machine, k nearest neighbors, decision trees, quadratic discrimination, and naive Bayes were implemented. All these models finally outputted high accuracy beyond 80% and k nearest neighbors algorithm performed best at 87.3 % of accuracy. This model was proved to have strong potential for further development of applicable in silicon high throughput prediction of AIE materials among huge structures base.

## Conflicts of interest

There are no conflicts to declare.

## Notes and references


1 B. Z. Tang, X. Zhan, G. Yu, P. P. Sze Lee, Y. Liu and D. Zhu, *Journal of Materials Chemistry*, 2001, **11**, 2974–2978.
2 Y. Hong, J. W. Lam and B. Z. Tang, *Chemical Society Reviews*, 2011, **40**, 5361–5388.
3 J. Mei, N. L. Leung, R. T. Kwok, J. W. Lam and B. Z. Tang, *Chemical Reviews*, 2015, **115**, 11718–11940.
4 J. Mei, Y. Hong, J. W. Y. Lam, A. Qin, Y. Tang and B. Z. Tang, *Advanced Materials*, 2014, **26**, 5429–5479.
5 J. Luo, K. Song, F. L. Gu and Q. Miao, *Chemical Science*, 2011, **2**, 2029–2034.
6 N. L. Leung, N. Xie, W. Yuan, Y. Liu, Q. Wu, Q. Peng, Q. Miao, J. W. Lam and B. Z. Tang, *Chemistry - A European Journal*, 2014, **20**, 15349–15353.
7 P. Zhou, P. Li, Y. Zhao and K. Han, *Journal of Physical Chemistry Letters*, 2019, **10**, 6929–6935.
8 B. F. Curchod and F. Agostini, *Journal of Physical Chemistry Letters*, 2017, **8**, 831–837.
9 T. Förster and K. Kasper, *Zeitschrift fur Physikalische Chemie*, 1954, **1**, 275–277.
10 B. Fulkerson, D. Michie, D. J. Spiegelhalter and C. C. Taylor, *Technometrics*, 1995, **37**, 459.
11 Y. Baştanlar and M. Özuysal, *Methods in Molecular Biology*, 2014, **1107**, 105–128.
12 J. Qiu, K. Wang, Z. Lian, X. Yang, W. Huang, A. Qin, Q. Wang, J. Tian, B. Tang and S. Zhang, *Chemical Communications*, 2018, **54**, 7955–7958.
13 Q. Peng, Y. Niu, Q. Wu, X. Gao and Z. Shuai, *Aggregation-Induced Emission: Fundamentals*, John Wiley and Sons Ltd, Chichester, United Kingdom, 2013, vol. 1-2, pp. 357–398.
14 Q. Wu, C. Deng, Q. Peng, Y. Niu and Z. Shuai, *Journal of Computational Chemistry*, 2012, **33**, 1862–1869.
15 F. Neese, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2018, **8**, year.
16 A. J. Garza and G. E. Scuseria, *Journal of Physical Chemistry Letters*, 2016, **7**, 4165–4170.
17 J. Gong, J. W. Lam and B. Z. Tang, *Physical Chemistry Chemical Physics*, 2020, **22**, 18035–18039.
18 X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand and D. Steinberg, *Knowledge and Information Systems*, 2008, **14**, 1–37.
19 C. Cortes and V. Vapnik, *Machine Learning*, 1995, **20**, 273–297.
20 T. Hastie, *The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations*, Springer, New York, 2001.
21 A. Tharwat, *International Journal of Applied Pattern Recognition*, 2016, **3**, 145.
22 N. S. Altman, *The American Statistician*, 1992, **46**, 175–185.
23 L. Peterson, *Scholarpedia*, 2009, **4**, 1883.
24 S. A. Dudani, *IEEE Transactions on Systems, Man and Cybernetics*, 1976, **SMC-6**, 325–327.