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. 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 hypotheses 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.

Restriction of intramolecular motion (RIM) mechanism is the milestone of research about AIE behaviour. 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, 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). 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.

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, 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 equilibration structure at A, it will go to an excited state and then re-
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 conformation 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,\textsuperscript{15} we optimized the structure at the ground and S1 states under BLYP functional with def2-SVP basis set and RJCOSX-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.\textsuperscript{16,17}

The oscillator strength change (Δf) could be calculated by simple DFT and TD-DFT method:

\[
\Delta f = |f_{\text{eq}}^{S} - f_{\text{eq}}^{A}| \tag{1}
\]

In which \(f_{\text{eq}}^{A}\) denote the oscillator strength at equilibrium conformation at ground state (A) and \(f_{\text{eq}}^{S}\) 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.\textsuperscript{18} After optimization, maximum deviance reduction was interpreted as split criterion and the maximum number of splits is 11. The validated accuracy is 84.1%; Supports-vector machine (SVM) was used for classification for decades.\textsuperscript{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.\textsuperscript{20} Triangle function was optimized as kernel function. The validated accuracy is 84.1%, too; We also build a diagonal quadratic model for reference.\textsuperscript{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.\textsuperscript{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.\textsuperscript{24} The purpose of the weighted function is 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.
The performance of a model could be evaluated by such factors: TP denotes true positives, TN denotes true negatives, FP denotes false positives, FN denotes false negatives. The AUC value was calculated as 0.93, which was larger than reported models, implying that weighted KNN performed the best.

Figure 3B displayed the decision region for ACQ and AIE lumogens 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 threshold of false positive rate and m and Δf.

All performances of the different classifier were displayed 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 Δf 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:

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{4}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{5}
\]

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 Δf 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
7 P. Zhou, P. Li, Y. Zhao and K. Han, Journal of Physical Chemistry Letters, 2019, 10, 6929–6935.
15 F. Nese, Wiley Interdisciplinary Reviews: Computational Molecular Science, 2018, 8, year.
23 I. Petersen, Scholarpedia, 2009, 4, 1883.