



Advancing promiscuous aggregating inhibitor analysis with intelligent machine learning classification

Luxuan Wang, Beihong Ji, Jingchen Zhai , Junmei Wang *

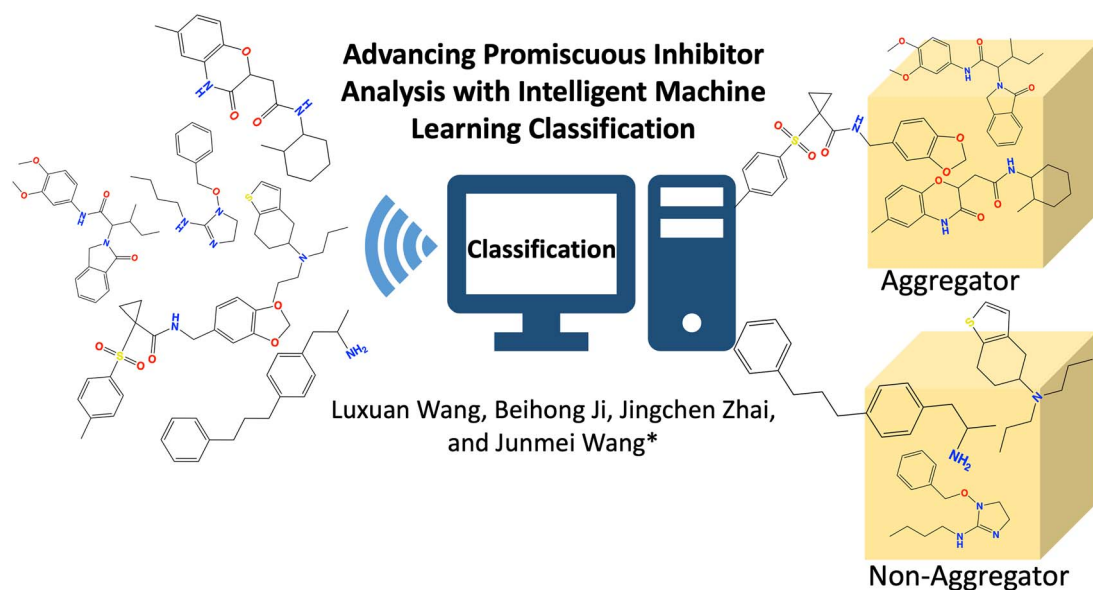
Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy, University of Pittsburgh, 3501 Terrace St., Pittsburgh, PA 15261, United States

*Corresponding author. Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy, University of Pittsburgh, 3501 Terrace St., Pittsburgh, PA 15261, United States. E-mail: junmei.wang@pitt.edu

Abstract

Small molecules have been playing a crucial role in drug discovery; however, some exhibit nonspecific inhibitory effects during hit screening due to the formation of colloidal aggregators. Such false positives often lead to significant research costs and time investment. Therefore, to identify potential aggregating compounds efficiently and accurately at an early stage of drug discovery, we employed several machine learning techniques to develop classification models for identifying promiscuous aggregating inhibitors. Using a training dataset of 10 000 aggregators and 10 000 nonaggregators, models were trained by combining four different molecular representations with various machine learning algorithms. We found that the best-performing model is the one that employs path-based FP2 fingerprints in conjunction with the cubic support vector machine algorithm, which achieved the highest accuracy and area under the receiver operating characteristic curve values for both the validation and test datasets while maintaining high sensitivity and specificity levels (>0.93). Additionally, we have proposed a new model interpretation method, global sensitivity analysis (GSA), to complement the well-recognized SHapley Additive exPlanations analysis. Several comparative studies have shown that GSA is a time-efficient and accurate approach for identifying crucial descriptors that contribute to model prediction, especially in the scenario where the dataset contains a substantial number of data entries with a limited set of descriptors. Our models as well as GSA findings can provide useful guidance on screening library design to minimize false positives.

Graphical Abstract



Keywords: colloidal aggregator; machine learning; global sensitivity analysis; drug screening; compound library design

Received: January 5, 2025. Revised: March 9, 2025. Accepted: March 28, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

The field of high-throughput screening (HTS) has emerged as a powerful technological approach to drug discovery and development. Its popular objective is to enable the rapid screening of vast compound libraries against specific targets, facilitating the identification of promising hit compounds that can be further explored for their therapeutic potential [1]. Enzymes have gained increasing attention as attractive targets for HTS in recent years. Their significance lies not only in their established role as drug targets for numerous diseases but also in their catalytic properties that make them particularly suitable for HTS applications. While HTS has undoubtedly revolutionized lead compound discovery, it is confronted with a persistent challenge known as false positives, which can even account for up to 95% of the initially identified active molecules obtained through HTS [2], leading to inefficient allocation of resources and potential misinterpretation of research outcomes. Extensive data analysis revealed that these false positives do not bind to a specific target as initially expected but instead exhibit promiscuous inhibition by binding to multiple targets. One of the main factors contributing to false positives is the formation of colloidal aggregators, which arise from poorly water-soluble compounds. These aggregators can lead to nonspecific interactions with various targets, further complicating the screening process [2–4]. Similar to the process of micelle formation, colloidal aggregators can form when the concentration of monomers surpasses the critical aggregation concentration within a buffer solution. This triggers the self-aggregation of monomers, leading to the formation of colloidal particles spanning a radius range of 50–500 nm. These aggregates can sequester 10^5 – 10^6 enzymes, resulting in partial denaturation and inactivation, thus providing time-dependent and noncompetitive inhibition of enzymes [5–7]. Colloidal aggregation not only affects target-based assays but also has implications for cell-based assays. The formation of stable colloids leads to a decline in the concentration of a free drug, resulting in reduced intracellular levels and diminished effectiveness [6]. Therefore, early identification of colloidal aggregators is critical to ensure accurate target-based and cell-based assay results, preventing the unnecessary expenditure of resources and time.

Various experimental methods have emerged for the detection of aggregators, which can be divided into two main categories: biochemistry techniques and biophysical techniques. In the biochemistry approach, the detergent-based assay is utilized, relying on the property of detergent reversibility to solubilize and disperse colloidal aggregators [8]. On the other hand, biophysical techniques such as resonant waveguide gravity [9], dynamic light scattering [10], nuclear magnetic resonance [11], and fluorescence-based assays [12] leverage diverse physical principles to investigate and characterize aggregators. However, each of those methods has certain limitations [2].

Consequently, computational machine learning models come into the stage to complement the shortage of experimental methods. Based on the filtered physicochemical criteria, Seidler et al. [13] applied a recursive partitioning method to generate a model using 111 compounds and 260 physicochemical properties, obtaining an accuracy of 93.7%. However, the limited amount of training data raised concerns about the model's reliability when applied to large-scale screening. Rao et al. [14] created a support vector machine (SVM) model for aggregator identification, utilizing 1319 aggregators and 128 325 nonaggregators. Although the overall prediction accuracy was high, the model displayed low precision (77.8%) in detecting aggregators, resulting in a

notable rate of false negatives. Aggregator Advisor 2.0 [15], a web tool for aggregator identification, takes a model-free approach based on molecular similarity. It predicts the likelihood of a compound becoming an aggregator by considering factors such as lipophilicity, affinity, and similarity to known aggregators. However, due to the limited reference data, its generalizability cannot be guaranteed. In another study, Yang et al. [16] developed a classification model using a dataset comprising 12 119 aggregators and 24 172 nonaggregators. They explored the relationship between molecular features and aggregators, leading to the creation of ChemAGG (<http://admet.scbdd.com/ChemAGG/index>), a free online web tool for aggregator detection. However, their choice of training data did not consider matching the logP values between aggregators and nonaggregators, which might potentially introduce a bias toward compounds with specific lipophilic properties, as the model may assign disproportionate importance to logP as a feature.

Therefore, to achieve more accurate and efficient screening of aggregators, we developed the compound aggregation classification models in MATLAB (2022b) utilizing four different descriptors. Also, we incorporated the recently well-recognized popular graph neural network, Attentive FP [17], to compare the classification performance. Simultaneously, the importance of understanding the rationale underlying a model's prediction is as significant as the accuracy of the prediction [18]. Therefore, we brought up a new protocol called Global Sensitivity Analysis (GSA), which identifies important features and assesses their impact on the model's predictions. This method not only achieves a similar level of accuracy as the widely recognized SHapley Additive exPlanations (SHAP) analysis but also offers a substantial improvement in speed. Based on this analysis, we listed several substructures and physicochemical properties that are considered crucial in aggregator identification, ultimately contributing to the initial screening of potential hits in drug discovery.

Materials and methods

Data preparation

The data used in this study includes an aggregator dataset obtained from Yang et al. [16] and a nonaggregator dataset obtained from the ChEMBL database [19]. To ensure the dataset's quality, compounds from the ChEMBL database underwent preliminary checks for duplicate molecules and molecules previously reported as aggregators. Next, those compounds were filtered based on having an inhibition constant lower than 100 nmol, indicating high target affinity rather than nonspecific aggregation, and falling within a molecular weight range of 25–1600 Da, as small molecules are less prone to aggregation compared to larger compounds. To establish a fair comparison, we further selected nonaggregators that closely matched the aggregators in terms of their properties, primarily focusing on molecular weights (MWs) and the calculated logarithm of partition coefficients (clogP), since these two properties measure the molecular size/van der Waals interaction and electrostatic interaction, respectively, which are pertinent to compound aggregation. To be more specific, we randomly selected an aggregation compound and identified the ChEMBL compound that has the most similar MW and clogP values. This procedure was repeated 29 247 times to compile the final negative dataset. As a result, we obtained the nonaggregator dataset containing 29 247 compounds, while the aggregator dataset contains 12 116 compounds. It is worth noting that the nonaggregator dataset

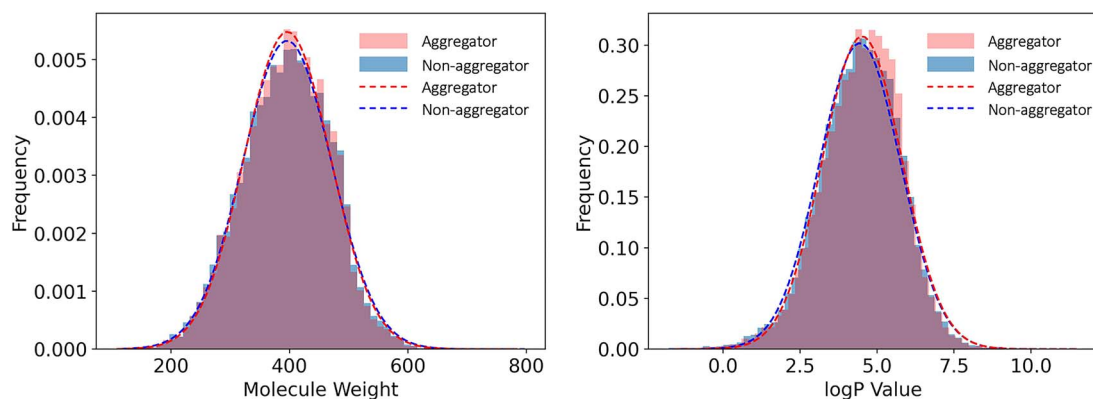


Figure 1. Distributions of molecule weight and clogP values for aggregators and nonaggregators.

demonstrated almost the same distributions as the aggregator dataset for two molecular properties, MW and clogP (Fig. 1). Other properties including topological polar surface area, number of hydrogen bond donors, number of hydrogen bond acceptors, atom numbers, bond numbers, and molar refractivity also exhibit similar patterns between nonaggregators and aggregators as expected (Fig. S1). To critically evaluate the performance of the descriptor-based models, only 10 000 aggregator compounds and 10 000 nonaggregator compounds were randomly selected to participate in model construction, and the rest served as the external test set compounds for evaluation. The SMILES strings of the training and test set molecules are listed in Table S1. The graph-based model, Attentive FP, utilized datasets that were constructed by excluding compounds with atom numbers >200 from the aforementioned datasets. The details on these refined datasets for Attentive FP can be found in Table S2.

Molecular representations

We used five different molecular representations as the readable input for aggregator classification including FP2 fingerprints [20], MACCS (Molecular ACCEss System) keys [21], GAFF (General AMBER Force Field) descriptors [22], and RDKit descriptors [23]. More details on those molecular representations are provided in the Supplementary Information (SI).

Model construction

In this study, we used the available algorithms that support training data with the Classification Learner Tool in MATLAB (2022b) to construct descriptor-based models, which include 47 algorithms in the domains of Decision Trees, Logistic Regression, Naïve Bayes, SVM, K-nearest Neighbor (KNN), Neural Network, Discriminant analysis, Kernel Approximation, and Ensemble classifiers [24]. Each type of classifier contributes to the variance in model flexibility and interpretability. Even for the choice of algorithms in the same classifier, the different algorithms show various effects on the model performance. For example, SVM, which aims to create a hyperplane between two classes so that it is as far away as possible from the closest data point in each class, is mainly used to construct a binary classifier. Although this algorithm was originally used to build a linear model, introducing the kernel function, such as the radial basis function, can effectively project data into high dimensions. So, the nonlinear data set that does not contain exact pairs of objects with opposite labels could be separated successfully. However, for finding the optimal kernel function to seek the best separation, the only method is employing multiple trials [25–27].

Apart from the descriptor-based models, we have also adopted a well-constructed graph-based model, Attentive FP, to compare the classification performance. Attentive FP is a state-of-the-art graph neural network, originally proposed by Xiong *et al.* [17]. It turned the atom connectivity of a molecule into a graph with vertexes and edges representing atoms and bonds, respectively. To gather comprehensive insights from neighboring to distant molecule graphs, the model employs recursive neural networks. Furthermore, it leverages graph attention mechanisms to extract nonlocal effects at the intramolecular level, allowing it to focus on the most relevant parts of the input. These techniques significantly enhance the model's learning and predictive capabilities, enabling accurate predictions for various molecular properties [17, 28]. In this study, we employed the random search method to fine-tune the hyperparameters, which has proven to be more effective than the traditional grid search approach [29]. Through the random search, we systematically explored various combinations within predetermined ranges: learning rates ranging from $10^{-5.5}$ to $10^{-1.5}$, batch sizes ranging from 100 to 300, and epochs ranging from 0 to 500, trying to identify the best possible set of hyperparameters for this specific case.

Model evaluation

We validated models by performing the 10-fold cross-validation and further tested using the test dataset. For the cross-validation, the training set was randomly divided into 10 subsets of roughly equal size. During each iteration, the model was trained using nine subsets while one subset was held out for validation. This process was repeated 10 times, ensuring that every subset had a chance to participate in the validation stage. By calculating the average accuracy across all subsets, the risks of overfitting and underfitting could be largely reduced and the model selection could become more rational. For the model performance evaluation, we adopt the accuracy (ACC), sensitivity (SE), specificity (SP), receiver operating characteristic curve (ROC), and the area under the ROC curve (AUC) as the metrics. The definitions of those metrics are provided in the SI.

Model interpretation

The SHAP analysis is a widely used method for interpreting model prediction results. However, in this study, SHAP analysis was found to be more appropriate for datasets with an extensive number of descriptors far outnumbering the data entries. In contrast, our specific case, which involves a large number of data entries with a few descriptors, revealed SHAP's computational inefficiencies. To address this limitation, we introduced a novel approach

called GSA to effectively and efficiently assess the contribution of each descriptor/feature and examine the impact of individual descriptors/features on the model's prediction outcomes.

GSA systematically evaluates the influence of each descriptor on the classification model by screening through all descriptors to identify if any specific ones can significantly impact model performance under a defined model algorithm while keeping the remaining descriptors unchanged. Initially, it's necessary to obtain a stable test result on an additional test dataset using the well-performing model trained with the full-size training set. This serves as a baseline for subsequent identification of significant descriptors. Then, we generated descriptor groups by calculating the Pearson's correlation coefficient (R) between each pair of descriptors, identifying those with the R value >0.5 as a group. Specifically, if descriptor A exhibits strong correlations with both descriptor B and descriptor C, we consider descriptors A, B, and C as a unified group. Conversely, descriptors that lack correlations above 0.5 with others will form their own group. Using descriptor groups rather than individual descriptors greatly accounts for potential correlations among descriptors, thereby avoiding the risk of overlooking important descriptors that may have synergistic effects. Once this preparation is complete, we proceed to conduct GSA. The first step involves systematically removing one group at a time from the training set, i.e. descriptors that stand alone due to weak correlations with others are removed individually, and descriptors forming a correlated group are collectively removed. Subsequently, the model undergoes retraining, and its performance is evaluated using the test set to obtain the corresponding test accuracy. This iterative process continues until every single group has been omitted once from the training set. We then compared the test accuracies obtained from these iterations with the test accuracy of the model trained with the full-size training set (baseline). By ranking the iterations in descending order according to the differences in test accuracy, we identify the iterations with the largest disparities. These iterations correspond to the descriptor groups that have been removed and are considered significant descriptors deserving further investigation. In contrast to deterministic algorithms, which require a single execution of the GSA, stochastic machine learning algorithms necessitate multiple runs of the entire GSA procedure to guarantee reliable outcomes, accounting for the influence of random variations on algorithm parameters. This treatment is essential because under the influence of random effects, a descriptor might be accidentally deemed significant in one execution of GSA while remaining unselected in subsequent runs. Given that a single GSA yields 15 potential descriptors in our study, conducting n runs of GSA leads to a maximum of $15 \times n$ potential descriptors, of note, indispensable descriptors likely appear multiple times. Next, we conducted an occurrence analysis on these potential candidates to select those that appeared five or more times and considered them as important descriptors that contribute significantly to the classification model. To finally analyze the influence of these significant features on the model's prediction output, we performed logistic regression by comparing the descriptor values with the corresponding true labels (0 or 1). When the resulting coefficient is positive, it indicates that the presence of the descriptor or a larger value for this descriptor increases the likelihood of predicting positive cases. The overall workflow of the GSA is shown in Fig. 2.

Shapley value is a mathematical concept derived from cooperative game theory and is utilized to quantify the individual contributions of players in a game. It offers a fair and equitable approach to distributing the value generated through collaborative efforts among the players. In the domain of machine learning, SHAP is a framework that encompasses a collection of algorithms

applying the principles of Shapley values to provide explanations for predictions made by machine learning models. By computing Shapley values for each feature across all possible feature sets and then averaging the absolute values of each feature's Shapley scores across all samples, SHAP analysis identifies key features and provides deeper insights into their impact on specific predictions. We performed the SHAP analysis with MATLAB (2022b) and adopted the extension to the Kernel SHAP algorithm with a conditional value function, which is more flexible and adaptable, allowing us to effectively handle scenarios with interdependent features [18, 30].

Results

Model selection and performance

We constructed the model by integrating four different molecular representations with the available algorithms in MATLAB (2022b). The overall workflow is shown in Fig. S2. Each representation captures unique structural and physicochemical features, enabling a diverse molecular characterization to assess their effectiveness in aggregation prediction. To ensure the reliability and generalizability of the models, we utilized a 10-fold cross-validation approach during the model construction phase. By examining the average validation accuracy across the 10 folds, we were able to effectively decrease the probability of overfitting or underfitting and identified the top two performing models for each of the four molecular representation methods, resulting in a total of eight top classifiers for aggregation prediction. To further assess the performance of those eight models, we conducted tests using a separate test set, and the results are summarized in Table 1 and ROC curves are plotted in Fig. 3. From Table 1, we can observe that the SVM, KNN, and Ensemble classifiers more frequently stick out as they exhibit excellent performance in conjunction with the four molecular representations for both the training set and the test set. The average ACC across all eight models reaches 0.91, while the average AUC is 0.95. Furthermore, the SE and SP values are consistently above 0.90. Among the eight models, the cubic SVM algorithm achieves the best performance when built with FP2 fingerprints, RDKit descriptors, and MACCS keys. However, when paired with GAFF descriptors, it is not the optimal choice. The one using the FP2 fingerprints in conjunction with the cubic SVM algorithm achieves the best results. It obtains a validation ACC of 0.938 and the highest validation AUC of 0.980. Moreover, the model achieves a comparable performance with a test accuracy of 0.943 and AUC of 0.980, demonstrating the strong robustness of the model when applied to predict unseen data. Both the validation and test SE are above 0.93, while the SP values are above 0.94. The runner-up model was constructed using the cubic SVM algorithm with RDKit descriptors as the molecular representation. Its performance follows closely to that of the best model. However, two models utilizing GAFF as the molecular representation exhibit the poorest overall performance, with an ACC of only around 0.89 and SE and SP values mostly below 0.90. Based on the validation results presented in Table 1, we chose the top two performing models, namely, the ones trained with cubic SVM algorithms utilizing FP2 fingerprints and RDKit descriptors as molecular representations, respectively, for comparison with the performance of the introduced graph-based Attentive FP model. A summary of the results can be found in Table 2, and the ROC curves are shown in Fig. 4. Additionally, the precision-recall curves are also provided in Fig. S3 in the SI. The analysis of Table 2 reveals that the Attentive FP model achieved good performance in validation through hyperparameter tuning, demonstrating an accuracy of 0.924 and an AUC of 0.971. However, when it was applied to an

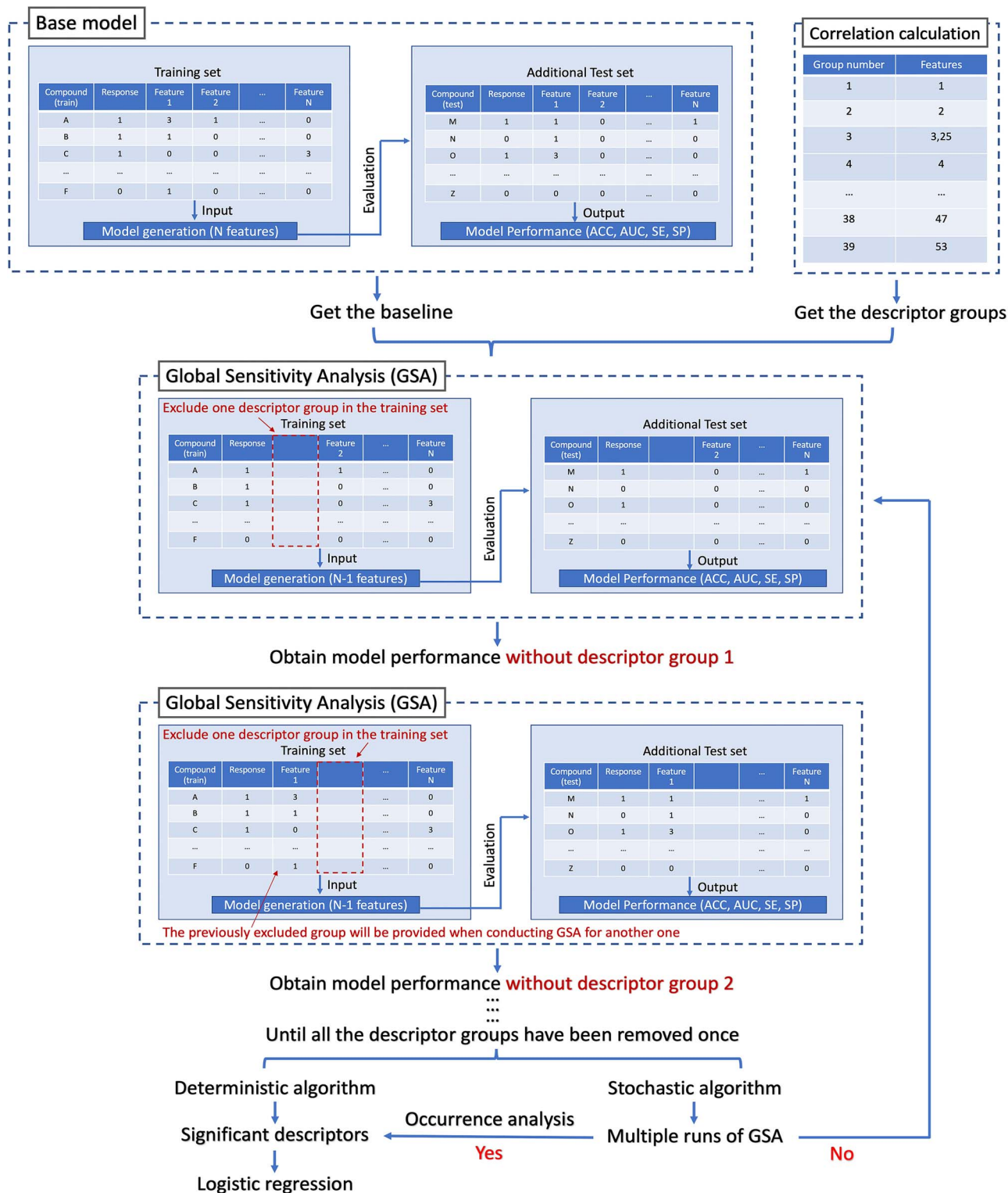


Figure 2. Overall workflow of GSA. For GSA, descriptor groups are removed one by one, retraining the model and evaluating each time, until every group has been removed once. If a deterministic algorithm is used for training, significant descriptors can be directly identified by comparing the differences in model performance after each group removal with the baseline. Otherwise, multiple runs of GSA and occurrence analysis are necessary. After identifying significant descriptors, logistic regression can be used to further understand the impact of descriptor values on the model's predictions.

unseen test dataset, the prediction accuracy dropped to 0.896. This decrease in accuracy might indicate that the model was overfitted to the training data, and its generalization ability to new data was compromised. More discussion on the performance of the Attentive FP model was provided in the SI.

Methods comparison for model interpretation

To validate our newly proposed method, coined as GSA, we first conducted the SHAP analysis to assess the contributions of individual descriptors. Given the significant computational time required to compute Shapley values for every descriptor of

Table 1. Performance of top two models on 10-fold cross-validation and additional test set for each descriptor

Descriptor	Algorithms	10-fold cross-validation				Test set			
		ACC	AUC	SE	SP	ACC	AUC	SE	SP
FP2	SVM (Cubic SVM)	0.938	0.980	0.933	0.943	0.943	0.980	0.930	0.945
	Ensemble (Subspace KNN)	0.929	0.972	0.924	0.934	0.936	0.974	0.937	0.936
RDKit	SVM (Cubic SVM)	0.930	0.972	0.940	0.920	0.929	0.975	0.944	0.928
	KNN (Fine KNN)	0.914	0.914	0.916	0.911	0.916	0.918	0.921	0.916
MACCS	SVM (Cubic SVM)	0.917	0.967	0.918	0.916	0.919	0.970	0.921	0.918
	KNN (Fine KNN)	0.903	0.903	0.908	0.899	0.906	0.902	0.897	0.907
GAFF	KNN (Weighted KNN)	0.889	0.950	0.909	0.870	0.883	0.954	0.910	0.880
	Ensemble (Bagged trees)	0.885	0.949	0.884	0.886	0.894	0.951	0.886	0.895

Table 2. Performance of three models on validation and additional test set

Model	Validation				Test			
	ACC	AUC	SE	SP	ACC	AUC	SE	SP
FP2 + Cubic SVM	0.938	0.980	0.933	0.943	0.943	0.980	0.930	0.945
RDKit + Cubic SVM	0.930	0.972	0.940	0.920	0.929	0.975	0.944	0.928
Attentive FP	0.924	0.971	0.931	0.917	0.896	0.965	0.923	0.893

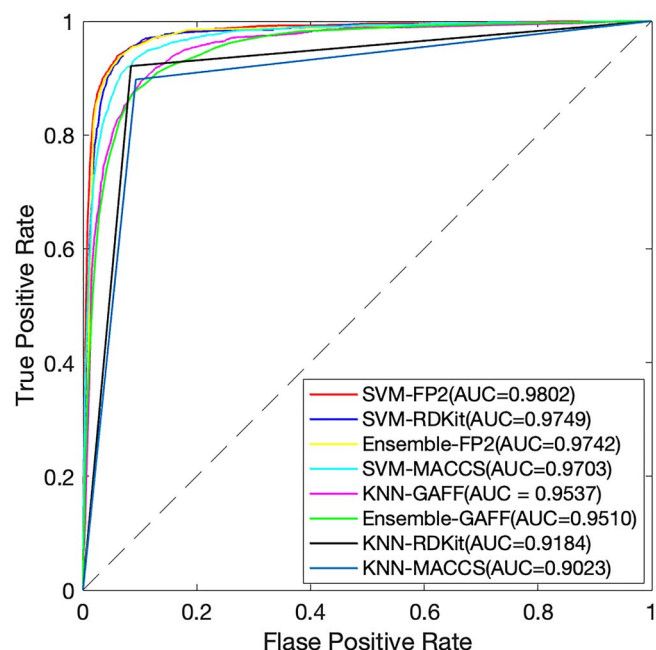


Figure 3. ROC curves of top two models for each descriptor on the additional test set.

each compound, we opted to analyze the model utilizing GAFF molecular representation, which has the fewest descriptors. To achieve this, we selected the previously developed model based on the bagged trees algorithm and computed the corresponding Shapley values for every descriptor of each compound in the additional test set. By summing the absolute value of the descriptor's Shapley values and calculating their average, we identified the 10 descriptors with the most significant contributions to the model performance. Moreover, we delved further into the impact of descriptors on the likelihood of a compound being qualified as an aggregator by exploring the distribution of descriptor

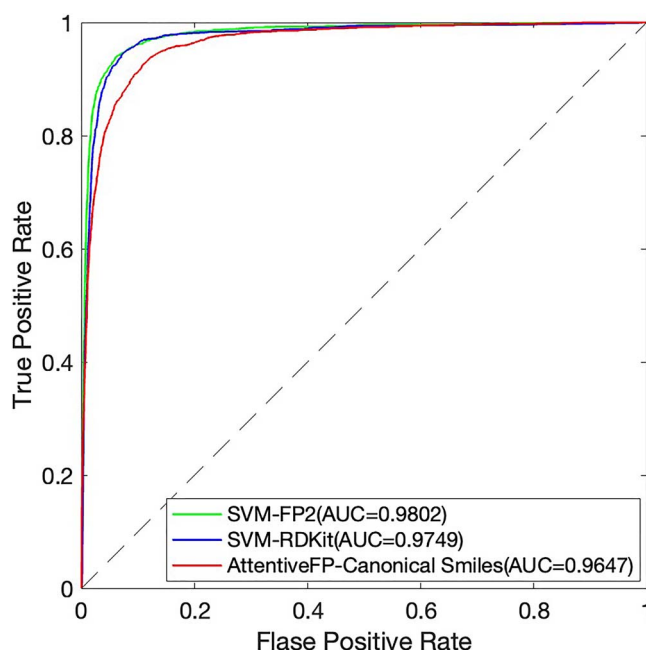


Figure 4. Comparative analysis of ROC curves for three models on the additional test set: Cubic SVM algorithm with FP2 fingerprints, Cubic SVM algorithm with RDKit descriptors, and Attentive FP model.

Shapley values in relation to their corresponding descriptor values.

From the SHAP dependence plot shown in Fig. 5, we observed the top 10 important descriptors ranked in descending order based on their contributions. Each feature's contribution to every compound is represented by plotted points in the figure. The SHAP values are arranged along the X-axis, where positive values indicate an increased probability of the model predicting the compound as an aggregator, while negative values indicate an

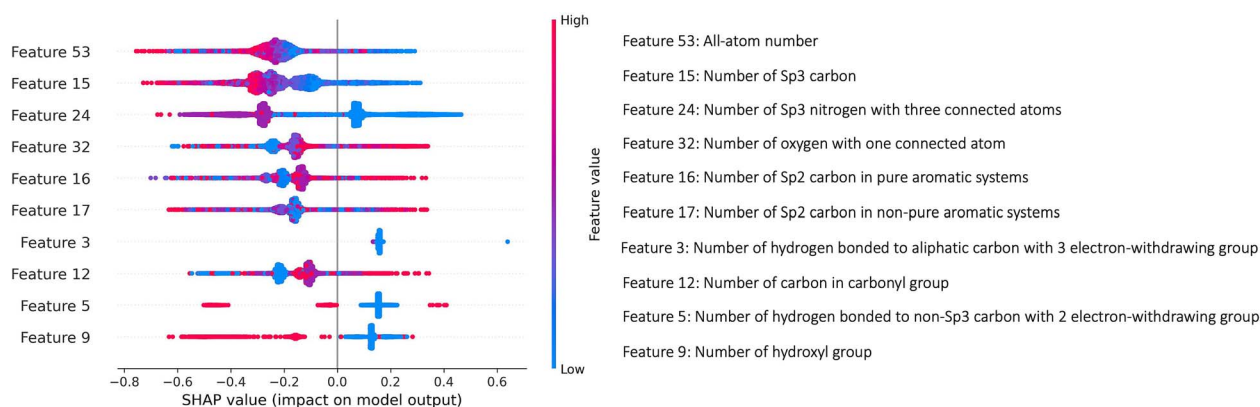


Figure 5. SHAP dependence plot of the top 10 key features for the GAFF model.

increased probability of the model predicting it as a nonaggregator. The magnitude of the feature values is depicted from blue to red, signifying small to large values. Considering feature 15 as an example. The corresponding SHAP value becomes more negative as the feature value increases (indicated by a reddish color). This implies that a larger feature value leads to a higher likelihood of the model predicting the compound as a nonaggregator. In other words, as the number of sp^3 carbon atoms increases, the compound is less likely to be classified as an aggregator. Conversely, when the feature value decreases (indicated by a bluish color), the corresponding SHAP value becomes more positive. This indicates that a smaller number of sp^3 carbon atoms makes the compound more likely to be classified as an aggregator. By analyzing the SHAP dependence plot, we can easily identify the highly contributing features and understand their impact on whether a compound becomes an aggregator.

Next, we proceeded with the GSA. To ensure comparability between the results obtained from this method and SHAP analysis, we combined the inputs generated by GAFF descriptors with the bagged trees algorithm to retrain the model and also used all the compounds in the test set as the test data. It's important to note that the bagged trees algorithm is a stochastic algorithm, meaning it introduces random variations in algorithm parameters. Therefore, when initially training the model with the complete training data and determining the test accuracy to set up the baseline, we conducted this process 10 times. Averaging the results from these repetitions ensures the baseline's robustness and reliability. The results of these repeated rounds are displayed in Table 3, where the average test accuracies, with a value of 0.8957, were chosen as the baselines for GSA. Then, we assessed the impact of the individual descriptor group by systematically excluding each one from the training set to train the model, analyzing the deviation from the corresponding model's test performance compared to the baseline. Similarly, we also repeated this procedure 10 times to counteract the impact of randomness on the results. Subsequently, the occurrence analysis was conducted to identify the significant descriptors. Besides, to understand the impact of those descriptors on whether a compound is an aggregator, we performed logistic regression by comparing the true labels of all compounds in the test data with the values corresponding to each identified important descriptor. The results are presented in Table 4. The symbol "+" means that with the descriptor value increases, the compound is more likely to be an aggregator. And the symbol "-" means that with the descriptor value decreases, the compound is more likely to be an aggregator. From Table 4, it is evident that Descriptor 8 is

considered the most important, as it appeared as a potential candidate in all ten repetitions, with its exclusion resulting in the largest deviation from the baseline performance in six repetitions. Note that Descriptor 8 indicates the number of hydrogen atoms bonded to nitrogen atoms. As Descriptor 8 has a negative logistic regression coefficient, we can infer that a compound with none or few numbers of hydrogen atoms bonded to nitrogen is more likely to be an aggregator. Further details and feature meanings can be found in Table S3.

By comparing the results obtained from these two different model interpretation protocols, we observed a 60% overlap in the results, which includes the prominently contributing descriptors and their impact on model prediction. Specifically, Descriptors 9, 24, 15, 12, 16, and 53 were found to be consistent (highlighted in Table 4). However, it is important to note that the interpretative process for this model with the SHAP analysis is much more time-demanding than GSA, as the latter does not need to estimate the influence of each descriptor for individual compounds, while SHAP analysis does. A scenario characterized by a small number of descriptors alongside a large number of data entries, as in this study, is frequently encountered in molecular property prediction. Under this scenario, GSA is a more appropriate choice since it not only achieves comparable results to SHAP analysis but also dramatically reduces computation time.

GSA was then applied to the second-ranked model, which utilizes RDKit as molecular representation and was trained with the cubic SVM algorithm. After establishing a robust and dependable baseline, GSA was repeated 10 times to identify truly important descriptors (referring to SI for more details). These descriptors are listed in Table 5, with further details provided in Table S4.

Discussion

This study aimed to develop a robust classification model for compound aggregation prediction using both descriptor-based and graph-based approaches. Among the models evaluated, the best-performing classifier was built using the cubic SVM algorithm with FP2 fingerprints, achieving ~0.94 accuracy and 0.98 AUC on both validation and test datasets, with high sensitivity and specificity (>0.93). Such superior performances can be attributed to FP2 fingerprints' richer structural representation compared to GAFF and MACCS, as well as their greater homogeneity relative to RDKit descriptors. These characteristics enable FP2 to stand out, particularly when paired with the cubic SVM algorithm, which is well suited for processing high-dimensional discrete data. While our final model demonstrated strong performance, there is still

Table 3. Model performance on additional test set with GAFF descriptors across multiple iterations

	Round1	Round2	Round3	Round4	Round5	Round6	Round7	Round8	Round9	Round10	Average	Standard deviation
ACC	0.8958	0.8982	0.8943	0.8954	0.8952	0.8940	0.8954	0.8978	0.8960	0.8945	0.8957	0.0015
AUC	0.9522	0.9531	0.9516	0.9536	0.9518	0.9522	0.9517	0.9517	0.9513	0.9520	0.9521	0.0007
SE	0.8847	0.8875	0.8866	0.8837	0.8866	0.8894	0.8823	0.8875	0.8875	0.8889	0.8865	0.0024
SP	0.8970	0.8994	0.8952	0.8967	0.8962	0.8945	0.8969	0.8989	0.8970	0.8952	0.8967	0.0017

Table 4. Global sensitivity analysis (GSA) results for significant features ranked by identification frequency in descending order for GAFF model. Highlighted features in bold refer to features consistent with SHAP analysis results

Feature	Meaning	Logistic regression coefficient
8	hn: Hydrogen bonded to nitrogen atoms	−
9,33	ho: Hydroxyl group/oh: Oxygen in hydroxyl group	−, +
39	ss: Sp3 sulfur in thio-ester or thio-ether	+
4	h4: Hydrogen bonded to non-sp3 carbon with one electron-withdrawing group	−
20	cp: Head Sp2 carbon that connect two rings in biphenyl system	−
24	n3: Sp3 nitrogen with three connected atoms	−
27	nb: Sp2 nitrogen in pure aromatic systems	−
7,15	hc: Hydrogen bonded to aliphatic carbon without electron-withdrawing group/c3: Sp3 carbon	+, −
28	nc: Sp2 nitrogen in nonpure aromatic systems	+
12,21	c: Carbon in carbonyl group/n: Sp2 nitrogen in amide groups	+, −
6,16	ha: Hydrogen bonded to aromatic carbon/ca: Sp2 carbon in pure aromatic systems	+, −
53	Number of atoms	−
18	ce: Inner Sp2 carbons in conjugated systems	+
43	f: Atom number for fluorine	−

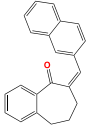
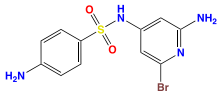
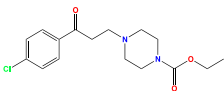
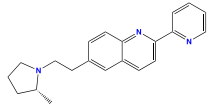
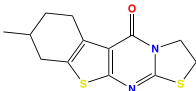
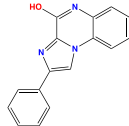
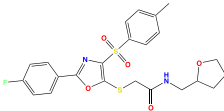
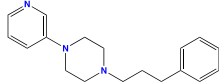
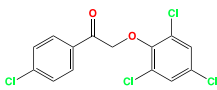
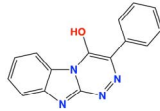
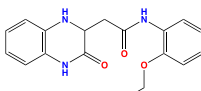
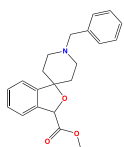
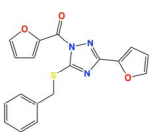
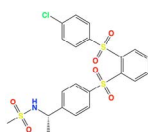
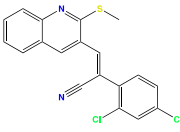
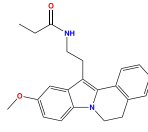
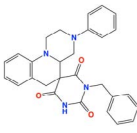
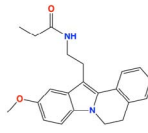
room for improvement. Exploring alternative molecular representations such as the Extended-connectivity fingerprints (ECFPs) could provide additional structural insights. Additionally, incorporating feature selection techniques like principal component analysis might enhance model robustness and performance by reducing dimensionality. Meanwhile, the deep learning model introduced in this study, Attentive FP, exhibited a performance drop when tested on unseen data, suggesting potential overfitting and compromised generalization ability. Future work could focus on more extensive hyperparameter optimization or exploring alternative deep learning architectures, which may enable a fairer comparison with descriptor-based models and potentially yield superior results for aggregation prediction.

Our new model interpretation method, GSA, offers significant advantages in terms of time efficiency and accuracy when interpreting model outcomes. GSA focuses on individually evaluating the impact of a descriptor group by systematically excluding it and analyzing the resulting model performance. While MATLAB software includes built-in feature selection functions such as the chi-squared test, analysis of variance, and Kruskal–Wallis methods, these approaches primarily focus on individual features and their statistical association with the outcome variable. They are not able to capture the interactions between features as achieved by SHAP analysis and GSA. Hence, we did not discuss the feature selection results obtained from MATLAB, even though there may be overlaps with the outcomes of the aforementioned two methods. The Kernel SHAP algorithm, employing a conditional value function, explicitly considers feature interactions and provides a more accurate assessment of descriptor importance. However, the computation of Shapley values takes a large amount of time when dealing with large datasets like the one utilized in our study. On the contrary, GSA better reveals the collaborative relationships among descriptors by evaluating the influence of each descriptor

group on the overall model. It exposes the indispensability of a descriptor group among all groups. Essentially, if the remaining descriptors can effectively collaborate and compensate for the excluded group, the model's performance remains relatively unchanged, indicating the minimal contribution of that specific group to the overall model. Both SHAP analysis and GSA provide valuable insights into model interpretation. In the scenario represented by our study, where the dataset comprises a substantially large number of data entries with a small set of descriptors, GSA is notably more efficient, significantly reducing computation time while accurately identifying key descriptors. However, as the number of descriptors greatly surpasses the number of data entries, this efficiency advantage diminishes. In such high-dimensional settings, SHAP analysis becomes more suitable, particularly for deep learning models that rely on a vast number of descriptors. Nonetheless, even when its computational advantage is lost, GSA remains highly effective in accurately identifying key descriptors, as it can better account for correlations between features.

In this classification study, we compared the interpretation results of the GAFF model from GSA and SHAP analysis, identifying a 60% overlap between the two methods. Additionally, we demonstrated that our GSA method is significantly faster than SHAP analysis, with GSA taking just a few minutes compared to 10 days for SHAP analysis. In fact, we also attempted SHAP analysis on another GAFF model that trained with the weighted KNN algorithm. However, it seems to take over five times longer time for interpretation. After demonstrating the accuracy and efficiency of GSA, we successfully applied it to analyze the second-ranked descriptor-based model, revealing significant structural and physicochemical properties for hit screening. Since the second-ranked descriptor-based model used the SVM algorithm, which introduces a heuristic procedure to determine the optimal value of the scale parameter instead of relying on a fixed

Table 5. GSA results for significant features ranked by identification frequency in descending order for the RDKit model

Feature	Meaning	Logistic regression coefficient	Example	
			Aggregator	Nonaggregator
127 141	fr_ArN: Number of nitrogen functional groups attached to aromatics/fr_NH2: Number of primary amines	-, -		
197	fr_pyridine: Number of occurrences of the pyridine fragment	-		
171	fr_imidazole: Number of imidazole rings	-		
4	MinAbsEStateIndex: Minimum absolute EState index	-		
175 176	fr_ketone: Number of ketones/fr_ketone_Topliiss: Number of ketones excluding diaryl, a,b-unsat. Dienones, heteroatom on Calpha	+, +		
188	fr_para_hydroxylation: Number of para-hydroxylation sites	+		
201	fr_sulfone: Number of sulfone groups	-		
75	SlogP_VSA3: MOE-type descriptors using LogP contributions and surface area contributions	-		
172	fr_imide: Number of imide groups	+		

MOE, Molecular Operating Environment.

value, randomness is introduced to the model training process. Therefore, similar to the GAFF model used in the comparison between GSA and SHAP analysis, we also need multiple runs to obtain more reliable results not only for the baseline generation but also for conducting GSA. Besides, we established the accuracy of the test set as the baseline measure, as changes in the AUC are generally less sensitive compared to accuracy in this specific case.

While aggregation is used to showcase the application of GSA in machine learning model interpretation, we further made two comparative studies to demonstrate the excellent performance of GSA. In the first study, we first constructed the bagged trees classification model for blood–brain barrier (BBB) prediction using a large dataset (Table S6). The model achieved a satisfactory performance with a prediction accuracy of 87.7% for the validation subset and 90.0% for the external test set. Details of dataset preparation and model construction were provided in SI. As shown in Table S7, 90% of key features identified by SHAP (Fig. S5) were also recognized by GSA. It is reasonable to investigate the impact of features using the feature group if features in that group are highly correlated. Additional information can be derived by investigating the pattern of coefficients of individual features in logistic regression. For example, the hc/c3/NATOM group has positive coefficients for hc and c3 but a negative coefficient for NATOM, suggesting that a molecule with a large Sp3 carbon to total atom number ratio is likely to penetrate BBB.

In the second study, we constructed computational models to predict protein–ligand binding affinities for six protein receptors. The models were trained using ligand–residue interaction profiles (LRIPs) [31, 32]. The details on model construction are presented in SI. It is not difficult to imagine that many features of LRIP are highly correlated. As shown in Table S8, the LRIP models trained by MATLAB's Regression Learner achieved a performance comparable with that in the literature [31]. Table S9 lists the total number of key features, i.e. hotspot residues for protein–ligand binding, by SHAP and GSA. Encouragingly, GSA recognized 0.7–2.2 folds more hotspot residues than SHAP, consistent with the fact that LRIP features are highly correlated. We then illustrated those hotspot residues that are also neighbors of ligands in 3D structures (Fig. S6). Obviously, there are much more neighboring hotspot residues surrounding the ligand identified by GSA (red and green sticks) than by SHAP (blue and green sticks). GSA-identified hotspot residues are more reasonable as they are more evenly distributed around the ligand for all six drug targets. Thus, GSA is a flexible approach that can efficiently identify important descriptors and facilitate the interpretation of classification and regression models. A more reasonable interpretation is expected with GSA when the descriptors are highly correlated.

Conclusions

In this study, we employed machine learning techniques to train and test classification models for compound aggregation prediction using a training set of 20 000 compounds and a test set of 21 363 compounds. By combining four different descriptors with a variety of machine learning algorithms, we successfully developed the top two accurate and robust models, which also outperformed a graph-based model constructed by Attentive FP. These models can be seamlessly integrated into early-stage drug discovery pipelines to screen out potential aggregators before HTS or biochemical assays, thereby reducing false positives, improving hit identification reliability, and optimizing resource

allocation in downstream experiments. Additionally, our novel model interpretation method, GSA, proved to be successful in identifying descriptors that made significant contributions to the model performance and shedding light on how these descriptors influence the prediction outcome. With GSA, we have identified important substructures and physicochemical properties, thereby aiding us in identifying potential colloidal aggregators in drug screening and guiding chemical structure modification in drug lead optimization.

Associated content

The Supplementary Information includes six Excel files: the dataset for training and testing descriptor-based models (Table S1), the dataset for training and testing the graph-based model (Table S2), detailed GSA results and descriptor meanings for GAFF (Table S3) and RDKit descriptors (Table S4), and the dataset for training and testing models for BBB (Table S6). It also contains a Word file, which includes details on molecular representations (Part 1), model evaluation (Part 2), model performance of the Attentive FP model (Part 3), GSA on the RDKit/cubic SVM model (Part 4), construction of BBB models (Part 5) and ligand–residue interaction profile models for six protein drug targets (Part 6). Additionally, this file contains supplementary tables and figures: model performance on the test set with RDKit descriptors across multiple iterations (Table S5), key descriptors of the BBB model identified by GSA (Table S7), summary statistics on the performance of LRIP models (Table S8), and summary statistics of key LRIP features identified by SHAP and GSA (Table S9). Figures include distributions of properties such as topological polar surface area, hydrogen bond donors/acceptors, atom and bond numbers, and molar refractivity for aggregators and non-aggregators (Fig. S1), the overall study workflow (Fig. S2), precision–recall curves of the top two models for each descriptor on the additional test set (Fig. S3), epoch-wise validation accuracy curve for the Attentive FP model (Fig. S4), SHAP dependence plot for the BBB model (Fig. S5), and 3D mapping of identified neighboring hotspot residues for the ligand–residue interaction profile models (Fig. S6).

Key Points

- Constructed a highly accurate and robust classification model to predict compounds' tendency to form aggregators from their chemical structures and physicochemical properties.
- Proposed a new method, Global Sensitivity Analysis (GSA), to identify key descriptors of a machine learning model. GSA can efficiently deal with correlated data and provides additional insights into model interpretation.
- GSA achieved a similar or better performance of SHAP analysis but with a much-reduced computational cost, especially for machine learning models trained with the number of data entries far exceeding the number of descriptors.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

The authors are grateful for the financial support provided by the National Institutes of Health (R01GM147673 and R01GM149705) and the National Science Foundation (1955260). The authors also thank the computing resources provided by the Center for Research Computing (CRC) at the University of Pittsburgh.

Author contributions

L.W. conducted the research and wrote the manuscript under the guidance of J.W. J.W. conceived the ideas, prepared the data, and revised the manuscript. B.J. improved the study design and revised the manuscript. J.Z. improved the figure and revised the manuscript.

Funding

None declared.

Data availability

All the data used for model construction were provided in [Tables S1, S2, and S6](#) of the SI. The models were constructed using MATLAB R2022b software. The best two models and codes developed for GSA are publicly accessible on GitHub (<https://github.com/ClickFF/GSA>).

References

1. Szymanski P, Markowicz M, Mikiciuk-Olasik E. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *Int J Mol Sci* 2012;**13**:427–52. <https://doi.org/10.3390/ijms13010427>.
2. Reker D, Bernardes GJL, Rodrigues T. Computational advances in combating colloidal aggregation in drug discovery. *Nat Chem* 2019;**11**:402–18. <https://doi.org/10.1038/s41557-019-0234-9>.
3. Sassano MF, Doak AK, Roth BL. et al. Colloidal aggregation causes inhibition of G protein-coupled receptors. *J Med Chem* 2013;**56**:2406–14. <https://doi.org/10.1021/jm301749y>.
4. Feng BY, Shelat A, Doman TN. et al. High-throughput assays for promiscuous inhibitors. *Nat Chem Biol* 2005;**1**:146–8. <https://doi.org/10.1038/nchembio718>.
5. Lak P, O'Donnell H, Du X. et al. A crowding barrier to protein inhibition in colloidal aggregates. *J Med Chem* 2021;**64**:4109–16. <https://doi.org/10.1021/acs.jmedchem.0c02253>.
6. Owen SC, Doak AK, Wassam P. et al. Colloidal aggregation affects the efficacy of anticancer drugs in cell culture. *ACS Chem Biol* 2012;**7**:1429–35. <https://doi.org/10.1021/cb300189b>.
7. Ghattas MA, Al Rawashdeh S, Atatreh N. et al. How do small molecule aggregates inhibit enzyme activity? A molecular dynamics study. *J Chem Inf Model* 2020;**60**:3901–9. <https://doi.org/10.1021/acs.jcim.0c00540>.
8. Feng BY, Shoichet BK. A detergent-based assay for the detection of promiscuous inhibitors. *Nat Protoc* 2006;**1**:550–3. <https://doi.org/10.1038/nprot.2006.77>.
9. Wang R, Wang J, Liu Y. et al. Resonant waveguide grating based assays for colloidal aggregate detection and promiscuity characterization in natural products. *RSC Adv* 2019;**9**:38055–64. <https://doi.org/10.1039/C9RA06466D>.
10. Allen SJ, Dower CM, Liu AX. et al. Detection of small-molecule aggregation with high-throughput microplate biophysical methods. *Curr Protoc Chem Biol* 2020;**12**:e78. <https://doi.org/10.1002/cpch.78>.
11. LaPlante SR, Carson R, Gillard J. et al. Compound aggregation in drug discovery: Implementing a practical NMR assay for medicinal chemists. *J Med Chem* 2013;**56**:5142–50. <https://doi.org/10.1021/jm400535b>.
12. Lifeng C, Gochin M. Colloidal aggregate detection by rapid fluorescence measurement of liquid surface curvature changes in multiwell plates. *J Biomol Screen* 2007;**12**:966–71. <https://doi.org/10.1177/1087057107306503>.
13. Seidler J, McGovern SL, Doman TN. et al. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem* 2003;**46**:4477–86. <https://doi.org/10.1021/jm030191r>.
14. Rao H, Li Z, Li X. et al. Identification of small molecule aggregators from large compound libraries by support vector machines. *J Comput Chem* 2010;**31**:752–63. <https://doi.org/10.1002/jcc.21347>.
15. Irwin JJ, Duan D, Torosyan H. et al. An aggregation advisor for ligand discovery. *J Med Chem* 2015;**58**:7076–87. <https://doi.org/10.1021/acs.jmedchem.5b01105>.
16. Yang ZY, Yang ZJ, Dong J. et al. Structural analysis and identification of colloidal aggregators in drug discovery. *J Chem Inf Model* 2019;**59**:3714–26. <https://doi.org/10.1021/acs.jcim.9b00541>.
17. Xiong Z, Wang D, Liu X. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2020;**63**:8749–60. <https://doi.org/10.1021/acs.jmedchem.9b00959>.
18. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 2017;**30**:4768–77.
19. Mendez D, Gaulton A, Bento AP. et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:D930–40. <https://doi.org/10.1093/nar/gky1075>.
20. O'Boyle NM, Banck M, James CA. et al. Open babel: An open chemical toolbox. *J Chem* 2011;**3**:33. <https://doi.org/10.1186/1758-2946-3-33>.
21. Gao K, Nguyen DD, Sresht V. et al. Are 2D fingerprints still valuable for drug discovery? *Phys Chem Chem Phys* 2020;**22**:8373–90. <https://doi.org/10.1039/DOCP00305K>.
22. Wang JM, Wolf RM, Caldwell JW. et al. Development and testing of a general amber force field. *J Comput Chem* 2004;**25**:1157–74. <https://doi.org/10.1002/jcc.20035>.
23. Landrum G. RDKit: Open-source cheminformatics, <https://www.rdkit.org> 2006.
24. Mathworks: Choose Classifier Options, <https://www.mathworks.com/help/stats/choose-a-classifier.html> 2024.
25. Huang S, Cai N, Pacheco PP. et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* 2018;**15**:41–51.
26. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006;**24**:1565–7. <https://doi.org/10.1038/nbt1206-1565>.
27. Rodriguez-Perez R, Vogt M, Bajorath J. Support vector machine classification and regression prioritize different structural features for binary compound activity and potency value prediction. *Acs Omega* 2017;**2**:6371–9. <https://doi.org/10.1021/acsomega.7b01079>.
28. Jiang D, Wu Z, Hsieh CY. et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Chem* 2021;**13**:12. <https://doi.org/10.1186/s13321-020-00479-8>.
29. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 2012;**13**:281–305.

30. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence Chemistry* 2021;**298**:103502. <https://doi.org/10.1016/j.artint.2021.103502>.
31. Ji B, He X, Zhai J. et al. Machine learning on ligand-residue interaction profiles to significantly improve binding affinity prediction. *Brief Bioinform* 2021;**22**:bbab054. <https://doi.org/10.1093/bib/bbab054>.
32. Niu T, Wang N, Wang J. Machine learning and deep learning based scoring functions in deciphering ligand-receptor binding: An application in drug design for GPCRs. *Annual Report Computational Chemistry* 2024;**20**:189–224. <https://doi.org/10.1016/bs.arcc.2024.10.001>.