ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Prediction of the Aqueous Solubility of Compounds Based on Light Gradient Boosting Machines with Molecular Fingerprints and the Cuckoo Search Algorithm

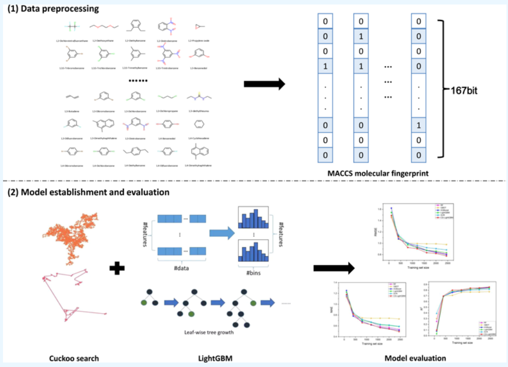
Mengshan Li,* Huijie Chen, Hang Zhang, Ming Zeng, Bingsheng Chen, and Lixin Guan

Cite This: *ACS Omega* 2022, 7, 42027−42035

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Aqueous solubility is one of the most important physicochemical properties in drug discovery. At present, the prediction of aqueous solubility of compounds is still a challenging problem. Machine learning has shown great potential in solubility prediction. Most machine learning models largely rely on the setting of hyperparameters, and their performance can be improved by setting the hyperparameters in a better way. In this paper, we used MACCS fingerprints to represent the structural features and optimized the hyperparameters of the light gradient boosting machine (LightGBM) with the cuckoo search algorithm (CS). Based on the above representation and optimization, the CS-LightGBM model was established to predict the aqueous solubility of 2446 organic compounds and the obtained prediction results were compared with those obtained with the other six different machine learning models (RF, GBDT, XGBoost, LightGBM, SVR, and BO-LightGBM). The comparison results showed that the CS-LightGBM model had a better prediction performance than the other six different models. RMSE, MAE, and R^2 of the CS-LightGBM model were, respectively, 0.7785, 0.5117, and 0.8575. In addition, this model has good scalability and can be used to solve solubility prediction problems in other fields such as solvent selection and drug screening.



1. INTRODUCTION

Aqueous solubility of compounds is a key physicochemical property in drug development because it affects drug absorption, distribution, metabolism, excretion, and toxicity (ADMET properties).[1−3] Therefore, the accurate and efficient prediction of the aqueous solubility of compounds is significant in reducing drug development costs and avoiding development failures.

Since the last century, a series of methods based on mechanistic models have been proposed to predict the aqueous solubility of compounds,[4,5] including general solubility equations (GSEs), Monte Carlo (MC) simulation, and COSMO-RS.[6−10] However, these methods rely on mathematical equations or physical constants, so they have certain limitations and require a great deal of calculation. Due to the high diversity of compound drugs, the poor variability of fitting equations, and the high fitting cost in terms of time and manpower, these methods are not ideal or efficient.

To replace traditional mechanistic models, many researchers turned to the quantitative structure−property relationship (QSPR) model.[11−13] In QSPR, the quantitative relationship among the physicochemical properties, biological properties, and molecular structures of compounds is explored with various statistical methods and mathematical models.[14,15] Usually, the molecular descriptors were selected as inputs of the models.[16] In previous studies, the main methods used in the QSPR model include multivariable linear regression (MLR), artificial neural network (ANN), Gaussian process (GP), and support vector machine (SVM).[17,18] However, the complex correlation between molecular descriptors and high-dimensional nonlinear data required for dissolution prediction poses great difficulties in traditional machine learning methods.[19]

Advanced machine learning methods expand the application scope of the QSPR model.[20−23] In recent years, ensemble learning methods, especially random forest (RF)[24] and light gradient boosting machine (LightGBM), have yielded satisfactory results in dissolution prediction.[25,26] In 2021, Ye et al.[27] predicted the solubility of compounds in organic solvents with the LightGBM algorithm, which showed better generalization ability compared to deep learning and other traditional machine learning algorithms.

In previous studies, a single model was generally used and the hyperparameters were set through exhaustive search. It is worth noting that the performance of the model is directly related to

Received: June 21, 2022
Accepted: October 18, 2022
Published: November 8, 2022

https://doi.org/10.1021/acsomega.2c03885

the setting of hyperparameters. In addition, the exhaustive search efficiency is low, especially under the conditions of too many hyperparameters. The swarm intelligence optimization algorithm has great advantages in search. The Cuckoo search (CS) algorithm is an excellent swarm intelligence optimization algorithm with few parameters and a strong search ability. Combined with advanced ensemble learning methods, it can solve the problem of hyperparameter setting well.

In this paper, the CS algorithm was used to optimize the setting of hyperparameters for LightGBM and molecular fingerprints were used as molecular descriptors to express the structure of compounds. Then, the CS-LightGBM model was established to predict the aqueous solubility of different compounds and the results were compared with those obtained with other models. In addition, the performance of each model was analyzed based on prediction results.

The main significance of this paper is as follows. (1) A new idea based on CS was proposed to optimize the setting of hyperparameters for LightGBM. (2) The proposed novel solubility prediction model showed great advantages in prediction accuracy, stability, correlation, etc. (3) The model has good scalability and great application potential in the fields of chemistry, materials, biology, and medicine.

2. MODEL ESTABLISHMENT

2.1. Theory. LightGBM is a Boosting framework based on decision tree proposed in 2017[28] and shows greatly optimized training speed and memory. LightGBM realizes the distributed and efficient framework mainly through gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS is an algorithm that balances the amount of data and the calculation accuracy. GOSS can filter out samples with a smaller gradient and calculate the information gain through the remaining data so as to reduce the amount of data and improve the efficiency. EFB is a dimensionality reduction technology used to bundle mutually exclusive features. Through feature bundling, EFB can reduce feature dimensions and improve the computing efficiency. If the features are completely exclusive (one feature value is 0, the other feature value is not 0), the features are directly bundled to prevent the loss of key information. If the features are not completely exclusive, based on the degree of nonexclusion of features (conflict ratio), the features with a low conflict ratio are selected and bundled to reduce the impact on accuracy. Different from most of the current models based on gradient boosting decision tree (GBDT), LightGBM adopts the leaf-wise strategy with depth restrictions other than the level-wise decision tree growth strategy. Therefore, as shown in Figure 1, on the same leaf node, compared with the level-wise strategy, the leaf-wise strategy can reduce information loss and memory consumption.[29]

The Cuckoo search algorithm proposed by Xin-She Yang and Suash Deb[30] is a swarm intelligence search technology integrating cuckoo nest parasitism with Levy flight. According to the long-term observation and research of entomologists, some cuckoos raise their young ones in a parasitic way. They do not build nests but lay their eggs in the nests of other birds and get their young ones hatched and reared by other birds. However, if the foreign eggs are found by the host, the host may discard them or build a new nest. For the algorithm, the nest represents the solution, and the process of the cuckoo looking for a nest to lay eggs is the process of looking for the solution in the *n*-dimensional space. The Levy flight is a non-Gaussian random process. During the Levy flight, the short-distance walk with a small step alternates with the long-distance walk with a large step. The long step is used to expand the search space and prevent falling into a local optimum, whereas the small step makes the population converge to the global optimal solution in a small space. The alternation of long and short steps enhances the global search ability.

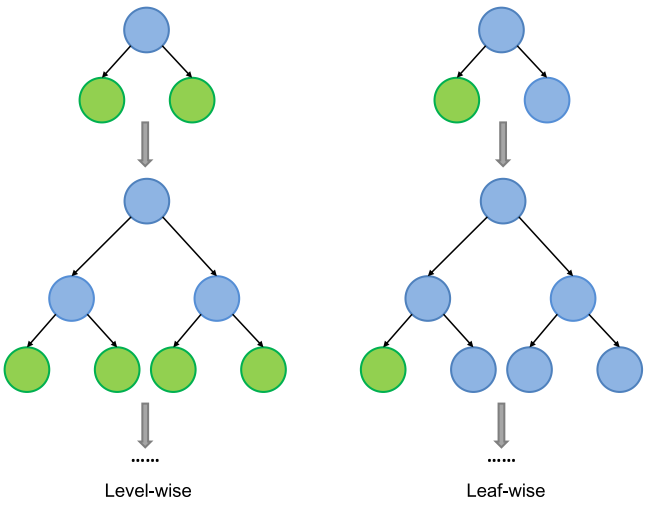


Figure 1. Tree generation strategy of LightGBM.

To better study the parasitism of a cuckoo nest, Yang et al. hypothesized a cuckoo's oviposition behavior into three ideal states: (1) each cuckoo lays one egg at a time and dumps its egg in a randomly chosen nest; (2) the best nests with a high quality of eggs will carry over to the next generations; and (3) the number of available host nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability $p_a \in [0, 1]$. Thus, the updating formula of the CS algorithm[31] is provided as follows

$$x_i^{t+1} = x_i^t + \alpha \oplus \mathrm{Levy}(\beta) \tag{1}$$

where $\alpha > 0$ is used to control the step. In most cases, $\alpha = 1$. The operator $\oplus$ refers to entrywise multiplications.

The Levy flight strategy is expressed as

$$\mathrm{Levy}(\beta) = \frac{\varphi \cdot u}{|v|^{1/\beta}} \tag{2}$$

$$\varphi = \left(\frac{\Gamma(1+\beta) \cdot \sin(\pi \cdot \beta/2)}{\Gamma\left(\left(\frac{1+\beta}{2} \right) \times \beta \times 2^{(\beta-1)/2} \right)} \right)^{1/\beta} \tag{3}$$

where $v \sim \mathrm{N}(0, 1)$, $\mu \sim \mathrm{N}(0, 1)$, β is a constant on [1, 2], and $\Gamma(\bullet)$ is the γ function.

The performance of LightGBM depends on the setting of hyperparameters. Generally, three methods are used to set the hyperparameters of the model: grid search, random search, and Bayesian optimization (BO). Grid search is an exhaustive search method and has a low search efficiency because it finds all of the solutions in the search space for permutations and combinations. Random search improves the search efficiency by randomly selecting parameter combinations in the hyperparameter space, but the improvement effect is not satisfactory when there are too many hyperparameters. In general, BO is more efficient than random search and grid search, but BO more easily falls into the local optimum, indicating that the improvement effect is not stable. Through the Levy flight, CS

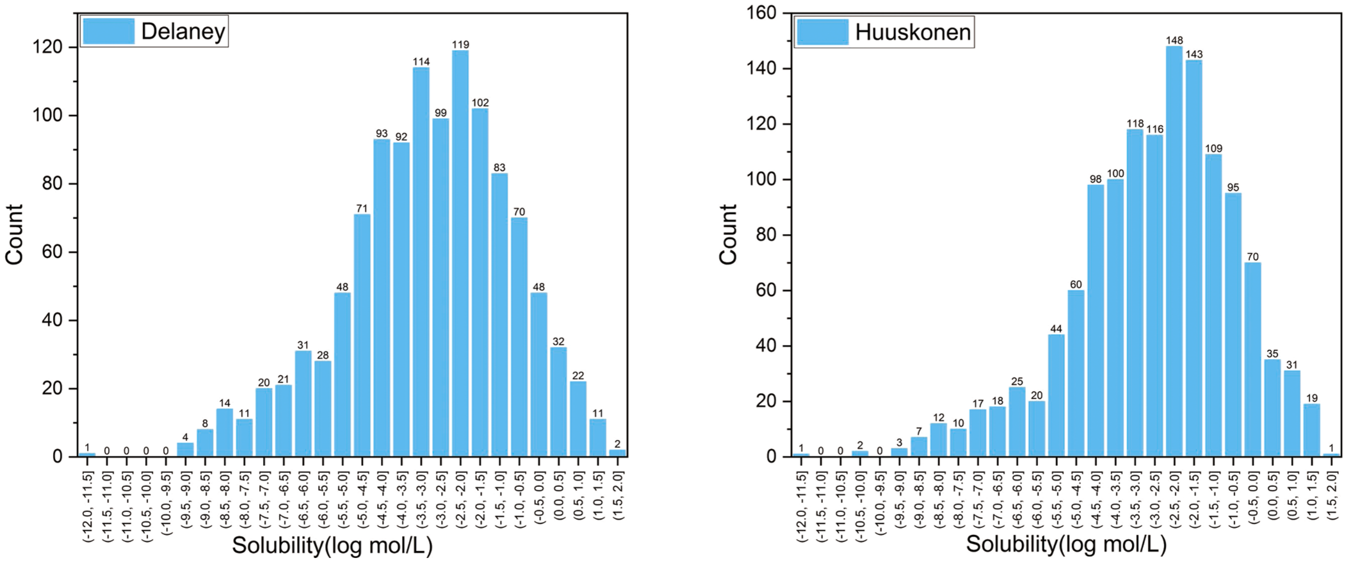


Figure 2. Distribution of the aqueous solubility of 2446 compounds.

can randomly walk and constantly update and iterate to find the optimal nest (that is, the global optimal solution) for incubating eggs so as to achieve an efficient search mode with few parameters, simple operation, and strong optimization capability, which has good performance in parameter optimization.[30]

Therefore, through the combination of CS and LightGBM, the excellent search ability of CS plays a key role in the hyperparameter setting of LightGBM, so that LightGBM can improve the search performance on the basis of low memory and high speed. In this paper, the five main hyperparameters of learning_rate, num_leaves, max_depth, subsample, and colsample_bytree were optimized through CS. The fitness function in this paper is root-mean-square error (RMSE) as the fitness function value, CS searches for a set of hyperparameters with minimum RMSE, the hyperparameters required by LightGBM. The pseudocode of the CS-LightGBM is provided as follows:

Algorithm: CS-LightGBM

```
nest = [learning_rate, num_leaves, max_depth, subsample, colsample_bytree]
Initialize nests = [nest 1, nest 2, …, nest n]
fitness = calc_fitness(fit_func, nests)
best_fitness = min(fitness), best_nest = nests[best_fitness_index]
for i=1 to iter_num do
    update_nests[nests], abandon_nests[nests]
    fitness = calc_fitness(fit_func, nests)
    min_fitness = min(fitness), min_nest = nests[min_fitness_index]
    if min_fitness < best_fitness
        best_nest ← min_nest
        best_fitness ← min_fitness
end for
```

Output: best_fitness ; best_nest

fit_func: a function including LightGBM training and evaluation, evaluation criteria is root mean square error (RMSE), and the return value is RMSE.

update_nests: a function based on Levy flight to find the better nest.

abandon_nests: a function to abandon some nests.

2.2. Data Collection. The experimental data were collected from previous studies and consisted of two datasets: Delaney and Huuskonen.[32,33] These data contained the names, SMILES (simplified molecular input line entry system) expressions, CAS (chemical abstracts service) numbers, and aqueous solubility (at 20–25 °C, and the unit of solubility is log mol/L) of 2446 organic compounds. As shown in Figure 2, the solubility data are unevenly distributed in the interval of (−12.0, 2.0) and mainly concentrated between −5.5 and 0. Due to the large span of solubility data, solubility prediction is a great challenge.

2.3. Data Preprocessing. SMILES is a specification that can guide the definite description of all of the details of a molecular structure with ASCII strings, including the basic information contained in a molecular system.[34] The SMILES expressions of compounds can be converted into molecular fingerprints with the open source toolkit RDKit. The MACCS molecular fingerprint is a fingerprint derived from the chemical structure database developed by MDL (Medical Discovery Leader), known as cheminformatics. It has 167 bits, including 166 bits representing substructures and 1 bit for saving information in RDKit. According to the experimental results of Chevillard et al.,[21] MACCS had a better prediction performance than ECFP4 and ECFP6. Therefore, in this paper, we chose the MACCS molecular fingerprint as the input of the model. To realize better performance, the whole dataset was randomly divided into two datasets, the training set and the test set, and verified by tenfold cross-validation. The whole prediction process is shown in Figure 3.

2.4. Evaluation Criterion. In the regression models of machine learning, root-mean-square error (RMSE) and mean absolute error (MAE) are commonly used to evaluate the degree of approximation between predicted results and real data. In addition, the coefficient of determination (R^2) is often used to characterize the fitting degree of the regression line to the observed value. RMSE, MAE, and R^2 are, respectively, defined as follows

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i^{\mathrm{pre}} - y_i^{\mathrm{exp}})^2}{n}} \tag{4}$$

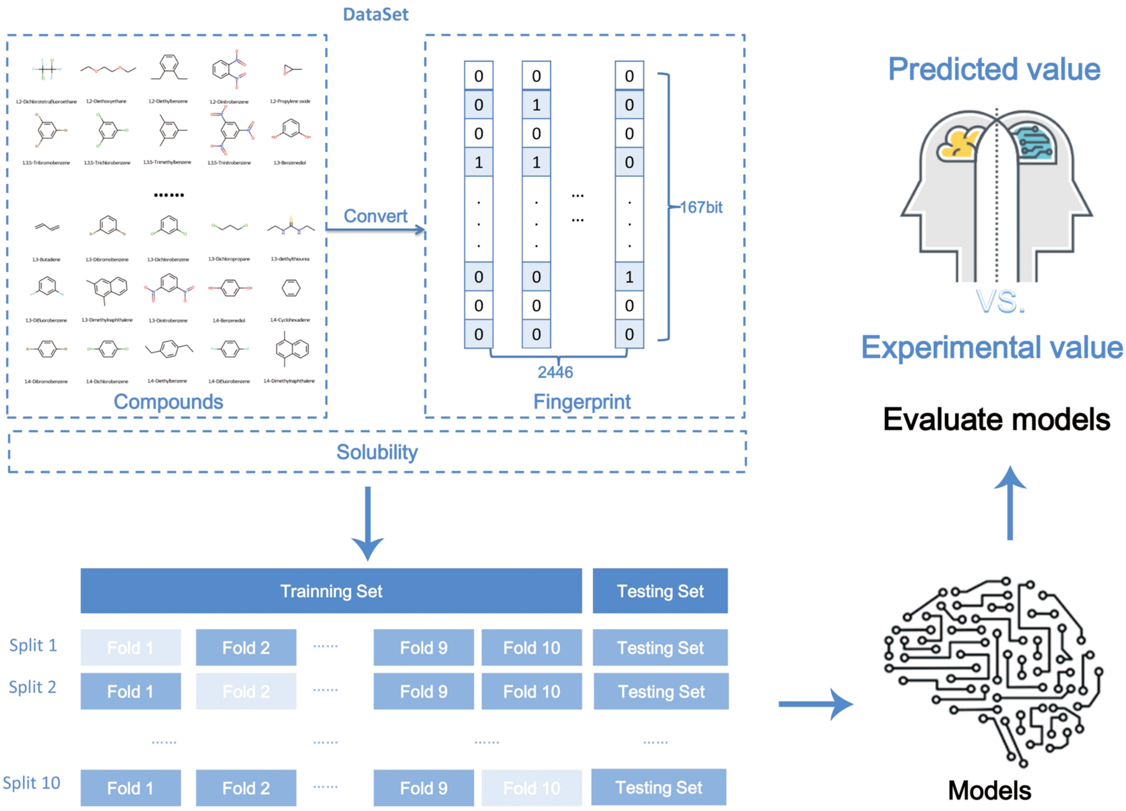


Figure 3. Solubility prediction process based on machine learning models.

Table 1. Hyperparameter Settings of the CS-LightGBM Model

parameter	learning_rate	num_leaves	max_depth	subsample	colsample_bytree
value	0.3	43	20	0.7	0.7

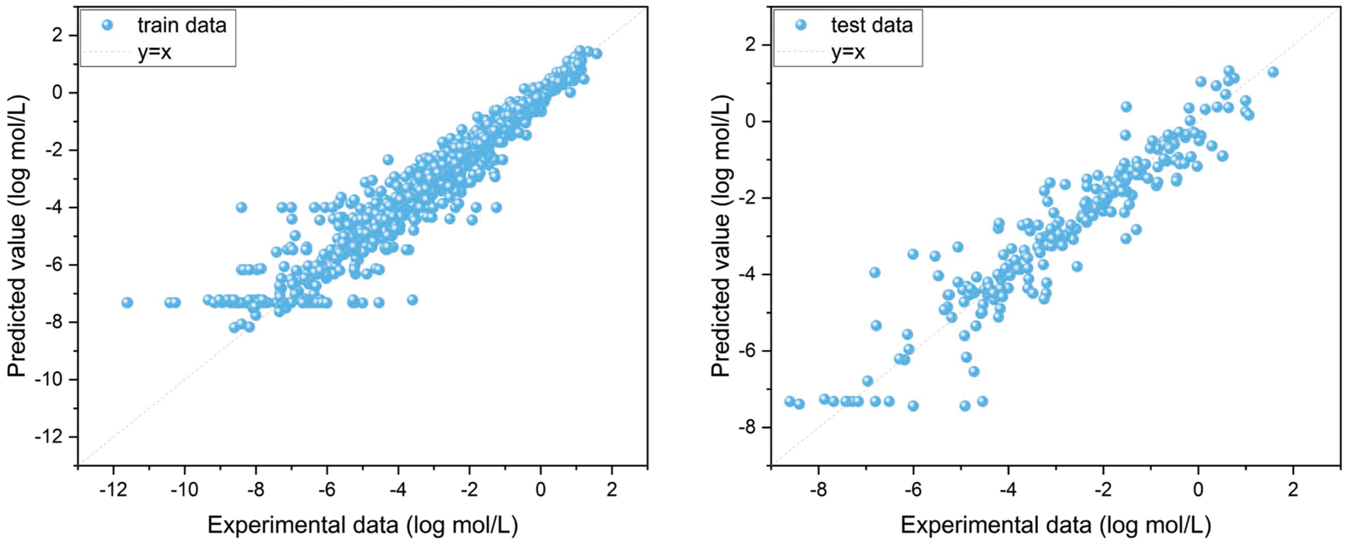


Figure 4. Experimental data and predicted values in the training set and testing set of CS-LightGBM.

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}|y_i^{\mathrm{pre}}-y_i^{\mathrm{exp}}|}{n} \tag{5}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i^{\mathrm{pre}}-y_i^{\mathrm{exp}})^2}{\sum_{i=1}^{n}(y_i^{\mathrm{exp}}-\overline{y})^2} \tag{6}$$

where n is the number of samples, y_i^{pre} is the predicted value, y_i^{exp} is the real value, and $\overline{y}$ is the average value.

3. RESULTS AND DISCUSSION

To increase the reliability of experimental results, all models were trained with the same dataset and division rules. The experiment was based on a Windows 7 64-bit operating system

https://doi.org/10.1021/acsomega.2c03885

(8.00 GB RAM, Intel(R) Core(TM) i5-9400F CPU) and implemented by Python.

3.1. Results of the Proposed Model.

In this paper, we established a reliable model with the CS-LightGBM algorithm to predict the aqueous solubility of different compounds and analyzed its predictive performance with some evaluation criteria.

Table 1 lists the hyperparameters of CS-LightGBM obtained with the CS algorithm. During the search process, with minimum RMSE as the fitness function, the optimal solution was obtained through continuous update iterations. However, the time cost of search also increases due to increasing iterations. Therefore, through several experiments, the number of iterations was set as 30 so as to reduce search time.

Figure 4 shows the fitting effect between predicted values and experimental values in the training set and test set. Less solubility data were between −12 and −4. However, the amount of data affected the prediction effect of the model. In the interval of [−9, −4] in the test set, most of the sample points significantly deviated from $y = x$ (the predicted value = the experimental value), indicating that the predicted solubility values in this interval had a large deviation from experimental values. However, the prediction effect of the model was better and the prediction accuracy was higher in the interval of [−4, 0] because the sample points were closer to the straight line. In addition, the amount of data between −4 and 0 was relatively large. Similarly, in the training set, the data points in the interval of [−12, −6] were in the discrete distribution. In short, the uneven distribution of data had a certain influence on the prediction results of the model.

Figure 5 shows the prediction error curve and can be used to further analyze the prediction performance of the model. In Figure 5, the horizontal axis is the number of samples, the vertical axis indicates the means of the predicted values and experimental values, and the error bar indicates their standard deviations. Most of the error bars were within a certain width range of 0 to 1.5, and about 25 data points had abnormal errors, namely, large errors. Obviously, except for some points that were affected by the uneven distribution of the original data, the prediction errors of this model were all within the acceptable range, indicating that this model had good stability.

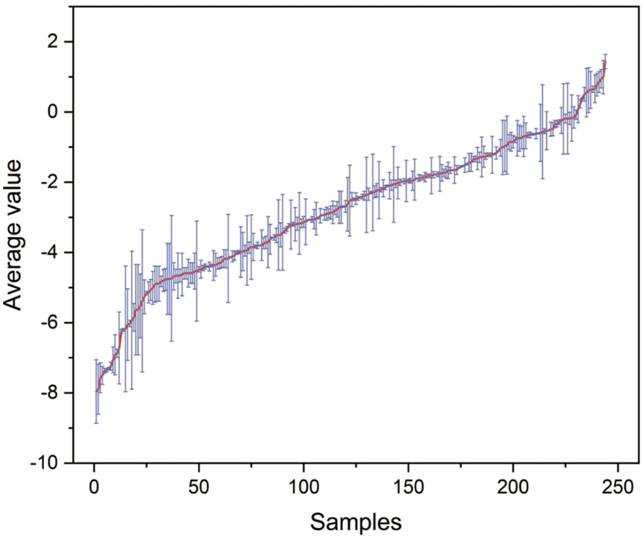


Figure 5. Error bar of the CS-LightGBM model in the testing set.

Table 2 shows the tenfold cross-validation results of this model. The purpose of cross-validation is to obtain a reliable and stable model. Although the uneven distribution of data had a certain impact on the prediction results of this model, the ten times cross-validation results showed a relatively stable trend without abnormal data. In addition, in the test set, the RMSE, MAE, and R^2 of the model were, respectively, 0.7785, 0.5117, and 0.8575 after averaging the data ten times, indicating that this model had a small prediction error and high correlation. In general, the stability and accuracy of this model were further verified by cross-validation and the model also showed good generalization ability.

Table 2. Performances of the CS-LightGBM Model in the Training Set and the Test Set

	RMSE		MAE		R^2	
	train	test	train	test	train	test
1	0.5291	0.7095	0.3021	0.4719	0.9329	0.8700
2	0.5455	0.7294	0.3049	0.4784	0.9310	0.8637
3	0.5411	0.7527	0.3030	0.4870	0.9314	0.8693
4	0.5413	0.7332	0.2996	0.4727	0.9314	0.8719
5	0.5374	0.8146	0.2980	0.5649	0.9331	0.8448
6	0.5402	0.8460	0.2982	0.5583	0.9321	0.8263
7	0.5282	0.7364	0.2902	0.4832	0.9342	0.8758
8	0.5480	0.7343	0.3060	0.4850	0.9295	0.8732
9	0.5528	0.8746	0.3068	0.5665	0.9292	0.8359
10	0.5517	0.8543	0.3061	0.5492	0.9290	0.8440
average	0.5415	0.7785	0.3015	0.5117	0.9314	0.8575

3.2. Comparison and Analysis of Various Models.

To further demonstrate the comprehensive performance of the proposed model, we compared the proposed model with other prediction models and discussed their performance differences, which proved that the proposed model had a significant advantage in predicting aqueous solubility. We separately implemented six comparison models of RF, GBDT, XGBoost, LightGBM, SVR, and BO-LightGBM and collected the experimental results of each model. LightGBM and XGBoost models were implemented, respectively, with Microsoft's LightGBM package and the XGBoost package, and other models were implemented with the sklearn package (Table 3).

Table 3. Comparison Models in This Study

model	description	reference
RF	random forest	35
GBDT	gradient boosting decision tree	36
XGBoost	eXtreme gradient boosting	37
LightGBM	light gradient boosting machine	28
BO-LightGBM	Bayesian optimization−light gradient boosting machine	38
SVR	support vector regression	39

Figure 6 shows the distributions of absolute errors of six prediction models. On the whole, RF, XGBoost, and CS-LightGBM models have similar error intervals between 0 and 1.5, while the other three models have relatively large error intervals. Under the condition of similar error interval, the data points exceeding the upper limit (i.e., outliers) in the results of CS-LightGBM were distributed within 3 and less than those in the results of RF and XGBoost models. In addition, most of the error points in the results of the CS-LightGBM model were

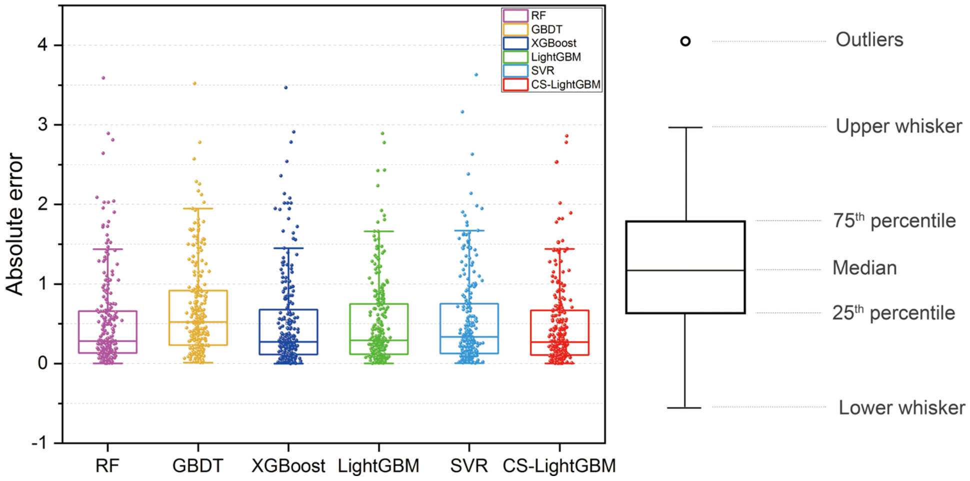


Figure 6. Distribution of absolute errors of RF, GBDT, XGBoost, LightGBM, SVR, and CS-LightGBM.

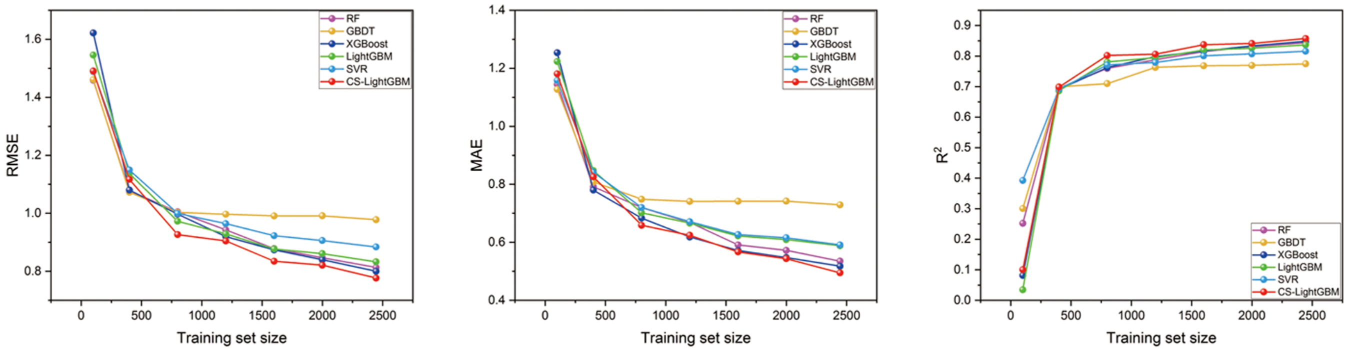


Figure 7. Prediction performances of RF, GBDT, XGBoost, LightGBM, SVR, and CS-LightGBM.

Table 4. Performance Comparison of the Six Models

	RF	GBDT	XGBoost	LightGBM	SVR	CS-lightGBM
RMSE	0.8132	0.9779	0.7999	0.8328	0.8838	0.7785
MAE	0.5349	0.7291	0.5177	0.5881	0.5908	0.5117
R^2	0.8439	0.7745	0.8485	0.8366	0.8159	0.8575
time(s)	16.3930	5.6900	2.7999	0.6919	8.7390	0.6959

concentrated below the median and denser than those in the results of the other five models. In short, among these prediction models, CS-LightGBM had the smallest absolute error and better performance.

Figure 7 shows the learning curves of six models trained on the solubility datasets with the training sizes of 100, 400, 800, 1200, 1600, 2000, and 2446. As the amount of data increased, the performances of six models were gradually improved and eventually became stable. With the increase of training set size, the RMSE of GBDT finally converged to 1.0, SVR to 0.9, and the other four models finally converged to about 0.8. The CS-LightGBM model had the fastest convergence speed and the smallest convergence result. The MAE of CS-LightGBM had a more obvious decreasing trend and decreasing speed than that of the other models and finally converged to between 0.5 and 0.6. When the amount of data increased from 100 to 400, the R^2 of the six models increased significantly and gradually became stable. The R^2 of all models except for GBDT and SVR finally converged to 0.85.

Among the six models, the CS-LightGBM model showed significant advantage in accuracy and correlation. RF, XGBoost, SVR, and LightGBM had similar performances, and GBDT had the worst performance. The running time of each model is shown in Table 4. Although RF, XGBoost, SVR, and LightGBM had comparable performances in terms of accuracy and correlation, the difference in time cost was significant among the four models. LightGBM had the fastest training speed. However, under the same running time, the overall performance of CS-LightGBM was slightly better than that of LightGBM.

In Table 4, the error difference of each model is not significant enough. Therefore, a statistical test is used to determine whether there is a significant difference between models. As the experimental data presented in this paper showed non-normal distribution, the Wilcoxon signed rank method was used to test. The significance level of each model is shown in Figure 8. The smaller the P value, the greater the inconsistency between the actually observed data and the full hypothesis, and the more significant the test results are. As can be seen from Figure 8, the P value of RF, GBDT, XGBoost, SVR, LightGBM, and CS-LightGBM is less than 0.05, which proves that there are significant differences between them. Although there are differences between the remaining four models (XGBoost, SVR, LightGBM, CS-LightGBM), they are not significant. Among them, the P value of SVR and CS-LightGBM is 0.843.

https://doi.org/10.1021/acsomega.2c03885

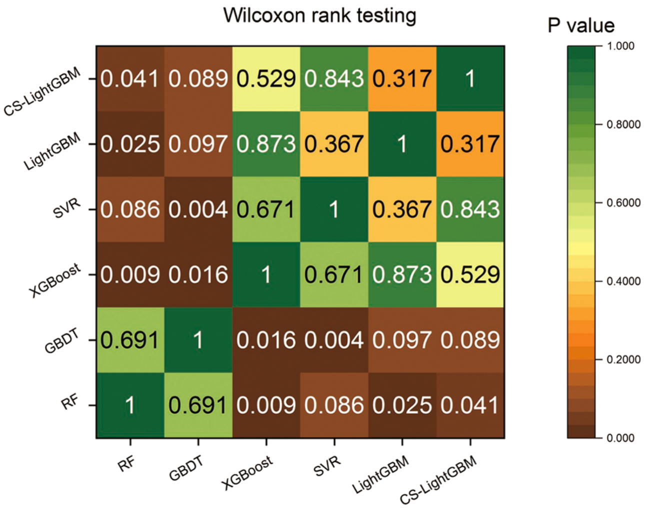


Figure 8. Significant difference test of each model.

However, it can be seen from Table 4 that CS-LighTGBM is superior to SVR in both running time and error.

In the comparison process, we found that the BO algorithm performed better than random search, grid search, and other search methods in parameter optimization. Therefore, we recorded the hyperparameter optimization results of LightGBM by BO and CS 10 times and evaluated the two prediction models with three performance metrics. Figure 9 shows the performance comparison of the two models. The fluctuation ranges of BO-LightGBM in the first, second, and third experiments were relatively larger. RMSE and MAE of BO-LightGBM basically showed an upward trend after the third experiment, whereas RMSE and MAE of CS-LightGBM fluctuated around 0.78 and 0.5, respectively. Similarly, R^2 of BO-LightGBM basically showed a downward trend, whereas R^2 of CS-LightGBM fluctuated around 0.856 in the third experiment. In addition, in nearly eight experimental results, CS-LightGBM had smaller errors and larger R^2 than BO-LightGBM, indicating that CS was more stable than BO. The CS-LightGBM model had a higher prediction accuracy, smaller errors, and more significant correlation compared to the LightGBM model.

Figure 10 shows the predicted and experimental values of each model in the test set. For a more direct comparison, two auxiliary lines $y = x + 1$ and $y = x - 1$ are introduced in Figure 10. Most of the data points of the CS-LightGBM model were well fitted to the line of $y = x$ and distributed between the two auxiliary lines. The data points of the CS-LightGBM model were more concentrated than those of the other models, although less data points were beyond the two auxiliary lines. The results indicated that the prediction error of the CS-LightGBM model was smaller than that of the other models and had a higher accuracy.

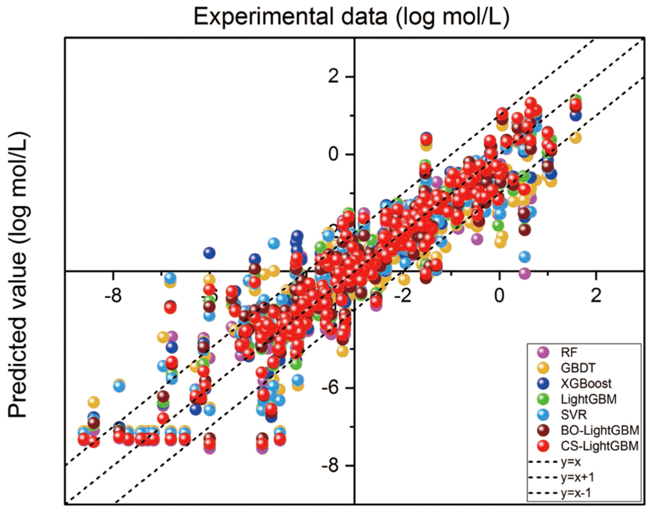


Figure 10. Experimental values and predicted values in models.

The comparison results of the seven models of RF, GBDT, XGBoost, LightGBM, SVR, BO-LightGBM, and CS-LightGBM indicated that the CS-LightGBM model had great advantages in prediction accuracy, correlation, and stability. CS-LightGBM is an excellent prediction model with high prediction accuracy, small deviation, high correlation, and strong stability.

4. CONCLUSIONS

In this paper, we optimized the hyperparameters of the LightGBM algorithm based on CS and established the CS-LightGBM model for predicting the aqueous solubility of compounds. The conclusions are drawn as follows.

(1) The CS-LightGBM model performed better in predicting the aqueous solubility of compounds and is superior to other comparison models in terms of prediction accuracy, correlation, and stability.

(2) Through the optimization with CS, the hyperparameter combination with minimum RMSE could be obtained for LightGBM within one time, thus providing a theoretical reference for the hyperparameter settings of various machine learning models.

(3) The CS-LightGBM model had good application prospects in chemistry, biology, medicine, materials, and other fields.

In the future, we will further explore the key technical and scientific issues of compound dissolution prediction and acquire

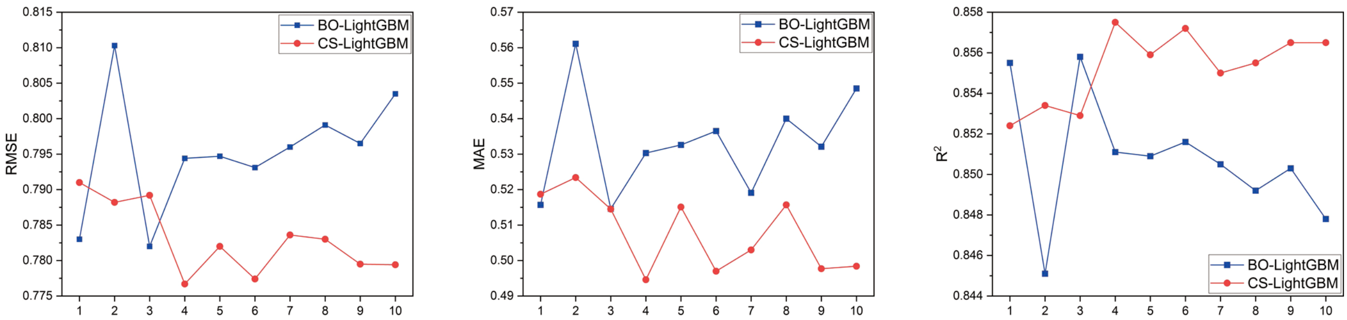


Figure 9. Performance comparison of BO-LightGBM and CS-LightGBM.

https://doi.org/10.1021/acsomega.2c03885

more data to establish a larger and more balanced database so as to reduce the impact of uneven data distribution on the model and better evaluate its prediction performance. Moreover, we will continue to investigate the solubility prediction problem using other state-of-the-art models and we will further study the relationship between the model and aqueous solubility to make the model more mechanized.

■ ASSOCIATED CONTENT

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

■ AUTHOR INFORMATION

Corresponding Author

Mengshan Li − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000 Jiangxi, China*; orcid.org/0000-0002-4093-6319; Email: limengshan@gnnu.edu.cn

Authors

Huijie Chen − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000 Jiangxi, China*

Hang Zhang − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000 Jiangxi, China*

Ming Zeng − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000 Jiangxi, China*

Bingsheng Chen − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000 Jiangxi, China*

Lixin Guan − *College of Physics and Electronic Information, Gannan Normal University, Ganzhou 341000 Jiangxi, China*; orcid.org/0000-0002-2554-5678

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c03885

Author Contributions

M.L. and H.C. designed the study; H.C., H.Z., and B.C. performed the research; M.L. and M.Z. conceived the idea; L.G. and H.C. provided and analyzed the data; H.C. and M.Z. helped perform the analysis with constructive discussions; and all authors contributed to writing and revision.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the support from the National Natural Science Foundation of China (Grant Numbers: 51663001, 52063002, and 42061067), the science and technology research project of the education department of Jiangxi province (Grant Number: 190745), and the Degree and postgraduate education and teaching reform research project of Jiangxi province(Multi thinking and "439" scientific research ability training mode for interdisciplinary academic degree postgraduate training).

■ REFERENCES

(1) Wang, J.; Hou, T.; Xiaojie, X. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **2009**, *49*, 571−581.

(2) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13*, 223−241.

(3) Das, T.; Mehta, C. H.; Nayak, U. Y. Multiple approaches for achieving drug solubility: an in silico perspective. *Drug Discovery Today* **2020**, *25*, 1206−1212.

(4) Marshall, G. R. Computer-Aided Drug Design. *Annu. Rev. Pharmacol. Toxicol.* **1987**, *27*, 193.

(5) Suzuki, T.; Ohtaguchi, K.; Koide, K. Computer-aided prediction of solubilities of organic compounds in water. *J. Chem. Eng. Jpn.* **1992**, *25*, 729−734.

(6) Corwin Hansch, J. E. Q.; Gary, L. Lawrence, Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347−350.

(7) Neera Jain, S. H. Y. Estimation of the Aqueous Solubility I Application to Organic Nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.

(8) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

(9) Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Bürger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* **2002**, *23*, 275−281.

(10) Vilas-Boas, S. M.; da Costa, M. C.; Coutinho, J. A. P.; Ferreira, O.; Pinho, S. P. Octanol−Water Partition Coefficients and Aqueous Solubility Data of Monoterpenoids: Experimental, Modeling, and Environmental Distribution. *Ind. Eng. Chem. Res.* **2022**, *61*, 3154−3167.

(11) Meftahi, N.; Walker, M. L.; Smith, B. J. Predicting aqueous solubility by QSPR modeling. *J. Mol. Graphics Modell.* **2021**, *106*, No. 107901.

(12) Bergström, C. A.; Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int. J. Pharm.* **2018**, *540*, 185−193.

(13) Klimenko, K.; Kuz'min, V.; Ognichenko, L.; Gorb, L.; Shukla, M.; Vinas, N.; Perkins, E.; Polishchuk, P.; Artemenko, A.; Leszczynski, J. Novel enhanced applications of QSPR models: Temperature dependence of aqueous solubility. *J. Comput. Chem.* **2016**, *37*, 2045−2051.

(14) Nakaoka, M.; Tran, K. V. B.; Yanase, K.; Machida, H.; Norinaga, K. Prediction of Phase Behavior of CO2 Absorbents Using Conductor-like Screening Model for Real Solvents (COSMO-RS): An Approach to Identify Phase Separation Solvents of Amine/Ether/Water Systems upon CO2 Absorption. *Ind. Eng. Chem. Res.* **2020**, *59*, 19020−19029.

(15) Song, F.; Xiao, Y.; An, S.; Wan, R.; Xu, Y.; Peng, C.; Liu, H. Prediction of Infinite Dilution Molar Conductivity for Unconventional Ions: A Quantitative Structure−Property Relationship Study. *Ind. Eng. Chem. Res.* **2021**, *60*, 14625−14634.

(16) Wang, J.; Hou, T. Recent Advances on Aqueous Solubility Prediction. *Comb. Chem. High Throughput Screening* **2011**, *14*, 328−338.

(17) Raevsky, O. A.; Polianczyk, D. E.; Grigorev, V. Y.; Raevskaja, O. E.; Dearden, J. C. In silico Prediction of Aqueous Solubility: a Comparative Study of Local and Global Predictive Models. *Mol. Inf.* **2015**, *34*, 417−430.

(18) Sluga, J.; Venko, K.; Drgan, V.; Novič, M. QSPR Models for Prediction of Aqueous Solubility: Exploring the Potency of Randić-type Indices. *Croat. Chem. Acta* **2020**, *93* (4), 311−319.

(19) Ge, K.; Ji, Y. Novel Computational Approach by Combining Machine Learning with Molecular Thermodynamics for Predicting Drug Solubility in Solvents. *Ind. Eng. Chem. Res.* **2021**, *60*, 9259−9268.

(20) Bahadori, B.; Atabati, M.; Zarei, K. Better prediction of aqueous solubility of chlorinated hydrocarbons using support vector machine modeling. *Environ. Chem. Lett.* **2016**, *14*, 541−548.

(21) Chevillard, F.; Lagorce, D.; Reynes, C.; Villoutreix, B. O.; Vayer, P.; Miteva, M. A. In silico prediction of aqueous solubility: a multimodel protocol based on chemical similarity. *Mol. Pharmaceutics* **2012**, *9*, 3127−3135.

(22) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855−1859.

(23) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150−158.

(24) Saraswathi K, S.; Bhosale, H.; Ovhal, P.; Parlikkad Rajan, N.; Valadi, J. K. Random Forest and Autoencoder Data-Driven Models for Prediction of Dispersed-Phase Holdup and Drop Size in Rotating Disc Contactors. *Ind. Eng. Chem. Res.* **2021**, *60*, 425−435.

(25) Zhao, Q.; Ye, Z.; Su, Y.; Ouyang, D. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. *Acta Pharm. Sin. B* **2019**, *9*, 1241−1252.

(26) Cui, Q.; Lu, S.; Ni, B.; Zeng, X.; Tan, Y.; Chen, Y. D.; Zhao, H. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Front. Oncol.* **2020**, *10*, No. 121.

(27) Ye, Z.; Ouyang, D. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *J. Cheminf.* **2021**, *13*, 98.

(28) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. In *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Proceedings of the 31st International Conference on Neural Information Processing Systems; Curran Associates Inc.: Long Beach, California, USA, 2017; pp 3149−3157.

(29) Rufo, D. D.; Debelee, T. G.; Ibenthal, A.; Negera, W. G. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics* **2021**, *11*, No. 1714.

(30) Yang, X. S.; Deb, S. In *Cuckoo Search via Levy Flights*, World Congress on Nature & Biologically Inspired Computing (NaBIC); IEEE, 2009.

(31) Gao, S.; Gao, Y.; Zhang, Y.; Li, T. Adaptive Cuckoo Algorithm with Multiple Search Strategies. *Appl. Soft Comput.* **2021**, *106*, No. 107181.

(32) Delaney, J. S. ESOL Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(33) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(34) O'Boyle, N. M. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminf.* **2012**, *4*, No. 22.

(35) Breiman, L. Random Forests. *Mach. Learn* **2001**, *45*, 5−32.

(36) Liang, W.; Luo, S.; Zhao, G.; Wu, H. Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics* **2020**, *8*, No. 765.

(37) Chen, T.; Guestrin, C. In *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785−794.

(38) Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A.; Tong, Z.; Lan, G.; Chen, Y. Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization. *Environ. Sci. Technol.* **2022**, *56*, 2572−2581.

(39) Baydaroğlu, Ö.; Koçak, K. SVR-based prediction of evaporation combined with chaotic approach. *J. Hydrol.* **2014**, *508*, 356−363.

https://doi.org/10.1021/acsomega.2c03885