Prediction and Design of Cyclodextrin Inclusion Complexes formation with Machine Learning-based Strategies

Yiming Maa,b,c,§, Yue Niua,b,§, Huaiyu Yangc, Jiayu Daia,b, Jiawei Lina,b, Huiqi Wanga,b, Songgu Wua,b, Qiuxiang Yina,b, Ling Zhou a,b,\*, Junbo Gonga,b,\*

a School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, Tianjin 300072, People’s Republic of China

b The Co-Innovation Center of Chemistry and Chemical Engineering of Tianjin, Tianjin 300072, People’s Republic of China

c Department of Chemical Engineering, Loughborough University, Leicestershire, LE113TU, United Kingdom

§ These authors contributed equally to this work

AUTHOR INFORMATION

**Corresponding Author**

**Junbo Gong, Ling Zhou**

\*E-mail: [junbo\_gong@tju.edu.cn](mailto:junbo_gong@tju.edu.cn), [zhouling@tju.edu.cn](mailto:zhouling@tju.edu.cn)

Tel: 86-22-27405754, Fax: 86-22-27374971.

ABSTRACT:

This work aims to develop multi-purpose machine learning (ML)-based cyclodextrin inclusion complexes (ICs) formation predicting strategies in aqueous solution to replace traditional experimental approaches. A balanced dataset of drug relevant molecules was constructed with experimental verifications. Three ML models (artificial neural network, support vector machine, and logistic regression) were established and optimized for ICs formation prediction. In order to provide more reliable approaches for different prediction requirements, ML-based linear strategy, recall-first strategy, and precision-first strategy were further established based on the ML models to pursue the maximum recall or precision values. It has also been proved that the proposed recall-first strategy finds all positive samples as much as possible to avoid missing in prediction, and the precision-first strategy finds positive samples accurately to reduce the number of validation experiments. The ML-based prediction strategies for ICs formation were first established in this work and showed high accuracy and reliability.

*Keywords:* Inclusion complex, Machine learning, Artificial neural networks, Support vector machine, Logistic regression

1. **Introduction**

Inclusion complexes (ICs) are crystalline materials formed spontaneously due to the existence of van der Waals forces between guest molecules and cyclodextrins (CDs) in solutions1-3. This supramolecular strategy has been one of the most widely used methods in the field of environmental engineering, food, and drug delivery, as well as supramolecular chemistry and agriculture4,5. Especially, due to the characteristics of "rim hydrophilic and cavity hydrophobic", CDs were used to combine with drugs or small organic molecules in order to improve the water solubility and stability of hydrophobic small organic molecules6. In addition, CDs can affect the physicochemical properties of guest molecules, such as masking unpleasant smells like bitterness, controlling the release rate, decreasing volatility of compound, and realizing targeted therapy7-11. The preparation of ICs conforms to the concept of "improved new drugs" with superior efficacy and conforms to the global trend of new drug research and development in an environmentally friendly way12,13. Meanwhile, the wide application of ICs in industry, such as dissolving organic pollutants in water and soil and improving bioavailability and sustainability in food formulations, also proves that it is indeed a favorable tool with high efficiency, green fabrication and low energy consumption14-16.

CD-ICs have shown great potential in new drug and functional material development, however, there are no efficient screening approaches for researchers due to the poor understanding of the mechanism and the complex interactions in ICs 17,18. The trial-and-error experiments can be used as the screening approach, however, with poor efficiency and low success rates in ICs discoveries 19,20. Thus, it is essential to establish robust screening strategies for the prediction and directional design of IC formulations. Machine learning (ML) algorithms are powerful tools to identify patterns and make decisions with minimal background information for a particular problem. It has been successfully applied to multiple areas especially drug discovery and material sciences21-23. In recent years, researchers have attempted to apply a single ML model or statistical approaches to find the critical factors in CD-ICs formation24-27. However, lack of experimental verifications of the positive (CD-ICs can form) and negative (CD-ICs cannot form) datasets, and the imbalanced samples led to difficulties in model training and hyperparameter tuning. Meanwhile, there are limitations in using a single ML model to accurately find the most suitable CD-ICs. Therefore, developing prediction models based on multiple ML methods and using verified and balanced datasets can provide more accurate prediction results, reducing wasteful experimental screening works and improving the efficiency in the early-stage drug developments or expanding capacities for expensive compounds28,29.

Researchers are accustomed to selecting an ML model with the best performance in the testing set in drug discovery and material sciences. It is relatively easy to carry out the way of determining the best ML algorithm via a single or several evaluation metrics in the testing set. However, as we discussed earlier, this will reduce the comprehensibility of the model and lead to poor performance in generalization ability. Researchers are currently trying to optimize the ML models from several aspects to adapt to practical applications, including applying the new methodology for tackling limitations30, and selecting descriptors based on in-depth understanding of specific behaviors to improve prediction accuracy31-33. However, there are still challenges in efficiently selecting an ML model with high accuracy, which is widely applicable to multiple needs, because it is difficult to optimize all evaluation metrics including accuracy, precision, and recall in classification through a single ML model. Meanwhile, the lack of focused design for evaluation metrics will lead to unsatisfactory prediction performance, which still needs to be supplemented by a considerable number of experimental verifications to avoid prediction errors. Therefore, it is necessary to construct corresponding ML-based design strategies according to the different requirements of actual application scenarios.

In this work, widely-used β-CD was considered as the host molecule with the advantage of high stability and low price. The positive and negative datasets adopted in the models and strategies consist of 100 experimentally validated samples each, which were selected based on high commercial value and application potential. Three supervised ML algorithms, artificial neural network (ANN), support vector machine (SVM), and logistic regression (LR), were explored for the CD-ICs prediction of drug relevant molecules in aqueous solution. Three strategies using the combination of ANN, SVM, and LR models were developed and applied to meet various prediction requirements. In addition, five compounds, isonicotinamide, levulinic acid, prednisolone, 9-fluorenone and saccharin, were taken as a validation set to verify the accuracy and reliability of the ML-based screening strategies. According to the prediction results, three CD-ICs of prednisolone, 9-fluorenone and saccharin in the validation set were successfully prepared experimentally for the first time.

**2. Methodology**

Three successive steps including the construction, judgment and validation stage were constructed and shown in **Fig 1**. A large number of compounds were screened and corrected experimentally, and classified into positive IC group (recorded as lable 1) and negative non-IC group (recorded as lable 0). Experiments were conducted through a cooling crystallization process to identify new CD-ICs. 200 drug relevant molecules, with real or potential applications, were selected as the dataset with the same number of labels 1 and 0 to avoid class imbalance. Molecular descriptors were quantified by molecular structure analysis and molecular dynamic simulation, which were used as input after preprocessing to ensure or enhance performance and shorten the simulation time. The effective data matrix constructed by molecular descriptors and the corresponding labels were randomly divided into the training set and testing set with a fixed ratio as 70/30. In the judgment stage, three ML models, ANN, SVM, LR models, and three strategies (combinations of these ML models) were designed, optimized and validated on the same dataset. In the validation stage, the ML models and strategies were employed for assessing the possibility of ICs formation for five compounds to evaluate the feasibility and simplicity.



**Fig. 1**. Workflow of model structure used in this study. Details about the judgement stage are available in **Fig. 5**.

*2.1. Descriptor Definition*

The descriptors were defined based on molecular analyses for structure description, and molecular dynamic simulation for solute-solvent interaction quantification. The chemical structures of compounds were characterized and calculated by a set of standard computational Molecular Operating Environment (MOE) chemical descriptors (MOE molecular descriptors package34). Besides, solvation free energy was also obtained by the Materials Studio package (Accelrys Software Inc., USA) to quantify the interactions between molecules and β-CD32. In order to reduce the calculation time and retain the structure information and the interactions to the greatest extent, an unsupervised dimensionality reduction method principal component analysis (PCA) was used to condense the variables to 20 descriptors, which were available in **Table S1**. Moreover, the independence between the descriptors was analyzed shown in **Fig S1**, two descriptors showing a strong positive or negative correlation being selectively reserved. All ML models and strategies in this work were constructed based on these 20 descriptors.

*2.2. Machine learning Algorithms*

*2.2.1. Artificial neural networks*

ANN, one of the parallel computational ML models, exhibits powerful capabilities in data regression and pattern recognition in chemical engineering research35,36. Because the model parameters in the learning algorithm are tuned from known measured inputs and outputs, ANNs can describe arbitrary functions and the actual problems hidden beneath. An ANN with hidden densely connected layers, each of which has a different number of units, was evaluated and optimized by the 5-fold cross-validation test to mitigate overfitting.

*2.2.2. Support Vector Machine*

SVM is a supervised ML model which can efficiently perform a non-linear classification using the kernel trick and hyperparameter optimization to map the inputs into high-dimensional feature spaces37,38. The SVM model should establish and optimize to consider the balance among the recall, precision and accuracy.

*2.2.3. Logistic Regression*

LR analyzes the relationship between multiple independent variables and a categorical dependent variable, and it can estimate the probability of occurrence of an event by fitting data to a logistic curve39,40. The logistic function is helpful because it can take input with any values from negative to positive infinity, whereas the output always takes values between zero and one; hence, the model output is interpretable as a probability.

ML algorithms were implemented with custom-developed scripts using scikit-learn package version 0.24.2 and NumPy package 1.19.5 under Python 3.9.5. For each ML algorithm, hyperparameters were optimized by grid searching on a predefined hyperparameter space, and the best one was determined according to the evaluation metrics.

*2.3. Evaluation Metrics of Prediction Models.*

The following statistical parameters assessed the overall significance of the models to evaluate the classiﬁcation ability of the model. These metrics contain ACC (accuracy), precision, recall, F1 score, and MCC (Matthews correlation coeﬃcient), which were calculated based on the confusion matrix (TP, true positive; TN, true negative; FP, false positive; FN, false negative). The above five evaluation metrics were not only used to evaluate the predictive ability of the established models comprehensively, but also used to optimise three prediction strategies.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |
|  |  | (2) |
|  |  | (3) |
|  |  | (4) |
|  |  | (5) |

*2.4. Experimental Setup and Procedures*

*2.4.1 Materials*

β-CD was obtained from Shanghai Macklin Biochemical Co., Ltd and has the mass fraction purity equal or above 0.98. Prednisolone was purchased from Shandong Xinhua Pharmaceutical Co., Ltd has the mass fraction purity equal or above 0.98. Saccharin was purchased from HEOWNS Biochem Technologies LLC Tianjin with purity of 0.98, and 9-fluorenone was obtained from Meryer (Shanghai) Chemical Technology Co., Ltd with purity of 0.99. All were used without further purification. Other chemicals and reagents used in the screening experiments were of analytical reagent grade. All experiments were carried out using de-ionized water.

*2.4.2 Experimental Procedures*

A saturated aqueous solution of β-CD was obtained by dissolving 0.5g β-CD in 20 ml de-ionized water in a 40 ml brown glass with a cap at 40 °C, controlled by a water bath. The solution was continuously stirred at a constant temperature of 600rpm with a magnetic bar. The solid powder of a compound (one of 100 compounds) according to the molar ratio of 1:1 with β-CD, was slowly added to the vial in 20 minutes and the solution was kept stirring for 24 hours to reach thermodynamic equilibrium. Then the solution was slowly cooling down to 5 °C in 5 hours, and filtered to isolate the solid products (precipitation of the IC, β-CD or the compound). The products were washed with a small amount of de-ionized water and then dried in a vacuum drying oven at 40 °C for 12 h before analysis and characterization.

*2.4.3 Characterization*

Powder X-ray diffraction patterns of the β-CD, the individual compound and the products were determined by an X-ray diffractometer Rigaku (D/MAX 2500), with a copper target X-ray tube set to 40kV and 100mA with Cu Kα (1.54 Å) radiation. All samples were measured in the 2θ range of 2°–40° at a scan rate of 2° min−1 and the scanning step was 0.02°. Infrared spectra were collected by the KBr tablet method by a Nexus Fourier transform infrared spectrometer (Thermo Fisher, USA), focusing on a wavelength range of 4000 - 700 cm-1.

**3. Results and Discussion**

*3.1. Construction and Assessment of the ML Models: ANN, SVM, and LR*

A slight adjustment of the model structure and hyperparameters of the same machine learning algorithm may lead to significant differences in the classification performance of the ML models. The prediction ability for the ANN model was determined by factors including the number of nodes, network structure, and epochs. A comprehensive test on the influence of the network structures on the prediction ability was shown in **Table S2**. The optimization result indicated that the selected ANN structure, with 20 (inputs)−16 (nodes)−16 (nodes)−1 (output) containing two hidden layers, had the best performance in prediction accuracy. Although both relatively large and small structures can all get an acceptable prediction accuracy, the ANN with a large structure could not overfit and has a better generalization ability at the same time when the dataset size is not so large41. In addition, the set of epochs would contribute to the prediction results. The fluctuation of prediction accuracy and loss with the epochs were plotted in **Fig 2** and **Fig 3** (MSEin **Fig 3** represents the mean square error between the predicted and actual results). The training and validation accuracy raised significantly as the epochs within the range of 0-18. When the epoch exceeded 18, the prediction accuracy stabilized at around 0.9 without a significant improvement. Besides, the training and validation loss continued to decrease within 0-34 epochs, but the loss of the validation set increased after the 34th epoch, which can be considered as overfitting. Thus, applying the optimized ANN model (epochs = 38) to the testing set come to the result that yielded accuracy = 0.750, precision = 0.759, recall = 0.733, F1 score = 0.746, and MCC = 0.500 under the threshold value as 0.4 (**Table 1** and **Fig 4a**). The binary outputs of the SVM model were direct indicators of the ICs formation, which were significantly affected by the hyperparameter C and gamma value. According to **Fig S2** and **Fig S3**, it is found that the C value could dominate the values of accuracy, recall, and AUC (area under curve) in the range of 0-5. Therefore, it can be determined that the model has the best prediction ability when c=2.2 and gamma= 0.04. The optimal SVM model had accuracy = 0.783, precision = 0.774, recall = 0.800, F1 score = 0.787, and MCC = 0.567 (**Table 1** and **Fig 4b**). The prediction results yielded by the LR model could also indicate the probability of the ICs formation. The recall value decreased with the threshold value increased through the continuous changes within 0-1, and precision, accuracy, and F1 Score fluctuated around 0.7 (**Fig. S4**). Therefore, the optimal threshold was selected to be 0.4 considering all evaluation metrics, which can yield accuracy = 0.767, precision = 0.735, recall = 0.833, F1 score = 0.781, and MCC = 0.538 (**Table 1** and **Fig 4c**). Comparing the three ML models, the outcome from LR model had the highest recall rate, the SVM model yielded the highest F1 Score, and the ANN model had excellent prediction ability in some compounds (e.g., 3-Amino-4-picoline, alprostadil). However, it is noted that this inconsistency of multiple evaluation metrics hinders the choice of ML models in applications. It is possible to use three ML algorithms at first to get preliminary results and then carry out experiments based on the Ture Positive results obtained by the three models in this work. However, it is difficult but still possible to screen all compounds with the experimental verification, due to moderate size of the dataset. If there were very large size of the dataset or the limitations on the available experiments of the compounds (such as very high price or instability and toxicity of the compound), the use of the simple three algorithms could be further improved to avoid missing or errors. For this purpose, it is necessary to design the appropriate strategy for the ML models, ensuring the maximum accuracy of the prediction outcomes.

图表, 折线图

描述已自动生成

**Fig. 2.** Training accuracy−epochs and validation accuracy−epochs curves for the ANN model.

图表, 折线图

描述已自动生成

**Fig. 3.** Training loss−epochs and validation loss−epochs curves for the ANN model.



**Fig. 4.** Confusion matrix of (a) ANN model, (b) SVM model, (c) LR model, (d) ML-based linear strategy, (e) Recall-first strategy, and (f) Precision-first strategy on 60 compounds in the testing set.

**Table 1** Prediction results on 60 compounds in the testing set using three ML models and three strategies with optimal hyperparameters a

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Compounds | Experiment Results | ANN model | | LR model | | SVM model | ML-based Linear Strategy | Recall-First Strategy | Precision-First Strategy |
| Continuous Output | Binary Prediction | Continuous Output | Binary Prediction |
| 2,7-Dihydroxynaphthalene | + | 0.231 | - | 0.502 | + | + | + | + | - |
| 2-Naphthylacetic acid | + | 0.926 | + | 0.981 | + | + | + | + | + |
| Acetic acid | + | 0.038 | - | 0.100 | - | - | - | - | - |
| Benzamide | + | 0.953 | + | 0.998 | + | + | + | + | + |
| Chlorogenic acid | + | 0.972 | + | 0.973 | + | + | + | + | + |
| Hydroxytyrosol | + | 0.887 | + | 0.988 | + | + | + | + | + |
| N-acetyl-D-phenylalanine | + | 0.938 | + | 0.927 | + | + | + | + | + |
| Piperazine | + | 0.803 | + | 0.613 | + | + | + | + | + |
| Pefloxacin | + | 0.829 | + | 0.957 | + | + | + | + | + |
| Geraniol | + | 0.510 | + | 0.894 | + | + | + | + | + |
| Aspirin | + | 0.073 | - | 0.050 | - | - | - | - | - |
| 4,4-Dithiodipyridine | + | 0.932 | + | 0.822 | + | + | + | + | + |
| (R)-Fenoprofen | + | 0.534 | + | 0.786 | + | + | + | + | + |
| 3-Nitrophenylacetic acid | + | 0.062 | - | 0.074 | - | - | - | - | - |
| (+)-Borneol | + | 0.982 | + | 0.996 | + | + | + | + | + |
| (-)-Borneol | + | 0.978 | + | 0.997 | + | + | + | + | + |
| Hydrocortisone | + | 0.882 | + | 0.961 | + | + | + | + | + |
| Astaxanthin | + | 0.989 | + | 1.000 | + | + | + | + | + |
| Apigenin | + | 0.985 | + | 0.982 | + | + | + | + | + |
| 5-Methyl-2-isopropylphenol | + | 0.982 | + | 0.999 | + | + | + | + | + |
| Benzbromarone | + | 0.843 | + | 0.970 | + | + | + | + | + |
| Salicylic acid | + | 0.108 | - | 0.096 | - | - | - | - | - |
| 4-Phenyl-Pyridine-N-oxide | + | 0.628 | + | 0.941 | + | + | + | + | + |
| 4-Biphenylacetic acid | + | 0.443 | + | 0.743 | + | + | + | + | - |
| Cinnamic acid | + | 0.295 | - | 0.420 | + | - | - | + | - |
| Eugenol | + | 0.186 | - | 0.506 | + | - | - | + | - |
| Methylprednisolone | + | 0.887 | + | 0.955 | + | + | + | + | + |
| Formic acid | + | 0.386 | - | 0.091 | - | + | + | + | - |
| 1-Butanol | + | 0.966 | + | 0.832 | + | + | + | + | + |
| Isobutanol | + | 0.961 | + | 0.895 | + | + | + | + | + |
| L-Malic acid | - | 0.006 | - | 0.000 | - | - | - | - | - |
| Butyl hydroxyanisole | - | 0.663 | + | 0.869 | + | + | + | + | + |
| 3-Amino-4-picoline | - | 0.357 | - | 0.678 | + | + | + | + | - |
| 2,4,6-Triaminopyrimidine | - | 0.102 | - | 0.007 | - | - | - | - | - |
| Xylitol | - | 0.019 | - | 0.005 | - | - | - | - | - |
| 3,4-Dihydroxybenzoic acid | - | 0.085 | - | 0.020 | - | - | - | - | - |
| Azilsartan | - | 0.730 | + | 0.985 | + | + | + | + | + |
| Gallic acid | - | 0.085 | - | 0.011 | - | - | - | - | - |
| Myricetin | - | 0.887 | + | 0.697 | + | - | - | + | - |
| Catechol | - | 0.797 | + | 0.943 | + | + | + | + | + |
| 1,1-Binaphtyl | - | 0.700 | + | 0.909 | + | + | + | + | + |
| Alprostadil | - | 0.344 | - | 0.475 | + | - | - | + | - |
| Isovanillin | - | 0.173 | - | 0.229 | - | - | - | - | - |
| Dehydrocholic acid | - | 0.842 | + | 0.988 | + | + | + | + | + |
| β-Alanine | - | 0.043 | - | 0.010 | - | - | - | - | - |
| o-Vanillin | - | 0.152 | - | 0.216 | - | - | - | - | - |
| L-Pyroglutamic acid | - | 0.188 | - | 0.022 | - | - | - | - | - |
| 4-Amino-2-hydroxyprydine | - | 0.129 | - | 0.096 | - | - | - | - | - |
| Nicotinohydrazide | - | 0.294 | - | 0.190 | - | - | - | - | - |
| 2-Aminobenzamide | - | 0.195 | - | 0.188 | - | - | - | - | - |
| Ribavirin | - | 0.116 | - | 0.045 | - | - | - | - | - |
| 6-Chloronicotinic acid | - | 0.183 | - | 0.053 | - | - | - | - | - |
| DL-Methionine | - | 0.125 | - | 0.011 | - | - | - | - | - |
| D-valine | - | 0.070 | - | 0.017 | - | - | - | - | - |
| Phenyl salicylate | - | 0.767 | + | 0.721 | + | + | + | + | + |
| 2,3-Dimethoxybenzoic acid | - | 0.071 | - | 0.114 | - | - | - | - | - |
| Urea | - | 0.066 | - | 0.026 | - | - | - | - | - |
| Diethylenetriaminepentaacetic acid | - | 0.056 | - | 0.013 | - | - | - | - | - |
| 2-Hydroxy-3-methylbutyric acid | - | 0.100 | - | 0.019 | - | - | - | - | - |
| L-2-Hydroxy-3-methylbutyric acid | - | 0.087 | - | 0.015 | - | - | - | - | - |
| MCC Value | / | / | 0.500 | / | 0.538 | 0.567 | 0.567 | 0.575 | 0.503 |

a The shading in the table indicates the compounds forming ICs.

*3.2. Establishment and Verification of Strategies*

In the judgment step, three strategies for different design purposes were proposed: ML-based linear strategy, recall-first strategy, and precision-first strategy. The same training and testing sets were used as in section 3.1 for the individual ML model.

*3.2.1.* *ML-based linear strategy*

Ma et al.42 reported a strategy of combining different ML models for more precise prediction. The strategy was based on the least-squares method, which assigned different weights for three ML models to achieve the best overall prediction outcomes in the testing set. In this study, the ANN and LR model with continuous output and the SVM model with binary output were assigned different weights, and the size of the weight represented the contribution of the corresponding ML model (**Fig. 5**). The formula of ML-based linear strategy based on the least-squares method was as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

Where P represents the output probability matrix, ANN, LR and SVM represent the matrix of training set or testing sets, respectively, and a0, a1, and a2 represent the contribution of the corresponding matrix to the probability P, respectively. From the final result of a0=0.3453, a1=0.2032, a2=0.3274 (**Table S3**), it can be determined that each ML model has a positive contribution to the probability P of the strategy. Nevertheless, applying the testing set to the strategy did not get the expected prediction ability. The strategy does not have a significant improvement in F1 score. If only from the numerical point of view, the performance of the ML-based linear strategy on the testing set in this study was the same as the SVM model (**Fig. 4d** and **Fig. 6**). Because the output result of 0 or 1 of the binary output SVM model had a leading role, despite the similar weights of the three models. The prediction results of ML-based linear strategy were only numerical values, and there was no unique and in-depth understanding design based on logic to deal with prediction problems. Therefore, this strategy succussed in predicting spherical crystallization but was less successful in this study42. However, the strategy may perform better if we expand the data size in the testing set.

*3.2.2. Recall-First Strategy*

Recalling the CD-ICs from the testing set to the greatest accuracy is essential for the screening and prediction. To this end, a recall-first strategy based on three ML models was proposed (**Fig. 5**). The recall-first strategy aimed to find the compounds that can form ICs in the data set as much as possible, while, allowing errors of including more candidates unable to form CD-ICs in predicted. Therefore, the first layer is to screen through the ANN model, and the continuous values of the prediction results, which were greater than or equal to 0.1, entered to the second layer. In this layer, we defined that the ANN model output value below 0.1 cannot form CD-ICs, which was beneficial to reducing the running time while ensuring the recall accuracy. The second layer used the binary output SVM model, and the compounds with output label 1 were considered to be able to form ICs, and the others with label 0 entered the third prediction layer. The values above or equal to 0.35 of the LR model in the third layer were recorded as the formation of the ICs, and vice versa. Therefore, the IC groups had two sources: Y in the SVM model and Y in the LR model, and the non-ICs group also had two sources: N in the ANN model and N in the LR model. Applying the strategy to the testing set yielded the best recall value of 0.867. This strategy demonstrated accurate prediction ability such as cinnamic acid and eugenol (**Fig. 4e**, **Fig. 6** and **Table 1**).

*3.2.3. Precision-First Strategy*

Achieving precise CD-ICs prediction is another focus. A precision-first strategy (**Fig. 5**) based on three ML models was proposed. The purpose of the precision-first strategy was to give the most accurate evaluation for compounds to form ICs or not, and it allowed some ignorance of some molecules in the testing set with ability to form ICs. First of all, the ANN model was used for the first screening layer. The continuous output values above or equal to 0.5 entered the second layer, and the predicted values below 0.5 were directly recorded into the non-IC group. The other data set entered the second LR model layer for further evaluation, and the prediction values above or equal to 0.6 entered the third layer. Otherwise, it could be directly considered as non-IC group. The last step prediction used the SVM model with in the selected dataset through the first two layers. Therefore, the IC group had only one source, Y of the SVM model, while the non-IC group had three sources, N in the ANN model, N in the LR model and N in the SVM model. The prediction result precision was 0.778 for the testing set, which was the best prediction outcome. The strategy demonstrated accurate prediction ability in compounds such as 3-amino-4-picoline and myricetin (**Fig 4f**, **Fig 6** and **Table 1**). Despite the higher prediction accuracy, the sample number in the Ture Positive group was only 21, indicating the worst recall among the three strategies and three individual ML models (**Fig 4f**).

Diagram

Description automatically generated

**Fig. 5.** Three ML-based prediction strategies in the judgment section.



**Fig 6.** The comparation of the experimental values with the prediction results of ANN, SVM, LR models and ML-based linear strategy, recall-first strategy, precision-first strategy.

*3.3. The application of design and prediction on the validation ICs*

The further application should be executed using ML models and strategies to verify the accuracy and feasibility of the designed concepts. Isonicotinamide, levulinic acid, prednisolone, 9-fluorenone and saccharin(not used in the training and testing database), were chosen to validate the models and the strategies (**Table 2**, **Table S4** and **Fig 7**). For isonicotinamide, only the ANN model predicted that IC could be able to form ICs with the prediction value of 0.467 (higher than threshold 0.4). The SVM model, the recall-first strategy and the precision-first strategy all yielded 0, and the ML-based linear strategy (0.227) and the LR model (0.324) gave the negative results (**Table S4** and **Fig 7**). From the prediction results, it could be determined that there was high probability not to form IC. The experimental results also prove the prediction, that is, in addition to the ANN model, other ML models and strategies have successfully predicted that the isonicotinamide cannot form CD-IC. For levulinic acid, the LR model and recall-first model gave the positive results with the values of 0.420 and 1. Prediction results of the ML-based linear model (0.136), the ANN model (0.147), the SVM model (0) and the precision-first strategy (0) all indicated no possibility for the formation of IC in an aqueous solution (**Table S4** and **Fig 7**). This predicted result was a reasonable proof that, in the case of inconformity of the ML model prediction results, the recall-first strategy can maximize the recall value, while the precision-first strategy considered them as the negative samples. This was consistent with the intention of the original strategy design. For prednisolone and 9-fluorenone, all three ML models and three strategies yielded the same result that ICs could be formed (**Table 2**, **Table S4** and **Fig 7**), which was in agreement with the experimental results. For saccharin, the precision-first strategy (0), the SVM model (0), the ML-based linear model (0.260) and the LR model (0.382) all gave the negative results, but there was a positive result from the experiment that saccharin-CD IC could be formed (**Table 2**, **Table S4** and **Fig 7**). The successful prediction of the recall-first strategy (1) further proved the advantages of the design. It can be concluded that when the prediction results of the model and the strategy are consistent (both positive or both negative), there was no significant difference between the different prediction methods, such as compounds of isonicotinamide, prednisolone, 9-fluorenone. However, there were advantages of using the recall-first strategy for the compounds with contradictory prediction results, such as levulinic acid and saccharin, avoiding missing positive samples (which might be very valuable for applications). There were advantages for the precision-first strategy, reducing cost and the number of experiments by more accurately limiting the quantities of potential candidates.

Three new solid CD-ICs were obtained and characterized, which were confirmed by powder XRD **and** FTIR spectra shown in **Fig 8** to **Fig 10**. The FTIR spectra of prednisolone, β-CD, their physical mixture and inclusion complex were shown in **Fig. 8A**. β-CD gave a strong and broad absorption band at 3280 cm-1, which was attributed to O−H stretching vibrations of hydroxyl groups, and gave a characteristic peak at 2925 cm-1 that was assigned to −CH2 asymmetric stretching vibrations43. Meanwhile, it can be seen that prednisolone gave a characteristic peaks at 1706 cm-1 and 1656 cm-1 which corresponded to C=O stretching vibrations. The peak at 1615 cm-1 and 1601 cm-1 was attributed to C=C stretching vibrations44. By comparing the spectrum of the IC with the spectrum of the physical mixture composed of the compounds with same dosage ratio, there were obvious differences in the positions of characteristic peaks, indicating that we successfully prepared IC. The presence of C=O (1702 cm-1 and 1656 cm-1) and C=C (1594 cm-1) in the IC spectrum indicates no change in the basic skeleton of the prednisolone structural formula. Comparing of the IC and prednisolone infrared spectra, some peaks disappeared and some moved to a large extent, but there were no new peaks, indicating that there were no new chemical bonds. The −OH peak was broadened in the 3600 - 3100 cm-1 region in IC (the blue area), which was attributed to the presence of abundant hydroxyl groups in β-CD. Numerous typical bands of prednisolone were masked by the characteristic peaks of β-CD (the red area), indicating that prednisolone was encapsulated in the β-CD cavity. The absorption band of −OH moved from 3290 cm-1 to 3269 cm-1, which was a red-shift phenomenon, indicating that new hydrogen bonds were formed between prednisolone and β-CD45. The powder X-ray diffraction pattern of prednisolone, β-CD, physical mixture and inclusion complex are shown in **Fig 8B**. The diffraction peaks of the IC with that of the physical mixture confirmed that the physical mixture was just a simple superposition of the two components. The diffraction peaks of the ICs were normally different from those of the individual components. Some characteristic peaks of prednisolone (8.0°, 12.5° and 20.2° and 30.3°) disappeared from the diffraction patterns, while several new peaks (5.5°, 7.3°, 11.0° and 17.3°) appeared in the IC sample, confirming the crystallinity of the IC. At the same time, several peaks of the prednisolone (13.5°, 14.8°, 16.0°, 18.0° and 26.3°) still existed in the diffraction pattern of the IC, but the intensity decreased significantly.

The FTIR spectra and the powder X-ray diffraction pattern of saccharin, β-CD, their physical mixture and IC were shown in **Fig 9**. Saccharin gave a characteristic peak at 3089 cm-1, which was attributed to N−H stretching vibrations, and gave a characteristic peak at 1712 cm-1 that was assigned to C=O stretching vibrations. The peak at 1456 cm-1 corresponded to C=C stretching vibrations in the benzene ring. And the peak at 1173 cm-1 and 1151 cm-1 corresponded to S=O stretching vibrations. In spectra of IC, the intensity of the C=O infrared characteristic peaks at 1714 cm-1 dropped sharply, implying that the carbonyl group of saccharin might be encapsulated in the β-CD cavity, thus shielding the infrared signal. Comparing the broad -OH peak of IC from 3603 - 3000 cm-1 with that of β-CD, the peak shifted from 3290 cm-1 to 3272 cm-1, confirming that the saccharin molecule formed new intramolecular hydrogen bonds with β-CD. Comparing the powder XRD patten of each component, almost all characteristic peaks of the saccharin disappear from the diffraction pattern of the IC, and there were several new peaks (3.6°, 12.7°and 17.2°) confirming the crystallinity of the IC. Moreover, several crystalline peaks of β-CD still existed in the diffraction pattern of the IC. 9-Fluorenone also showed similar characterization results in **Fig 10**. There was an obvious red shift from 3309cm-1 to 3621 cm-1 of the -OH characteristic peak which was significantly stronger than the two cases of the compounds mentioned above. It indicated that stronger intermolecular hydrogen bonding were formed between 9-Fluorenone and β-CD. At the same time, although the characteristic peaks located at 1451 cm-1, 1595 cm-1, 1705 cm-1 and 2682 cm-1 in the 9-fluorenone spectrum are to a certain extent, they can still be found in the IC spectrum. The XRD patterns of pure 9-fluorenone contain one heightened characteristic peak in 2θ at 12.9°, which disappears in the spectrum of IC. Meanwhile, a new peak with higher intensity appeared at 4.9° in the spectrum of IC, and the intensity of the diffraction peak is greatly reduced in high degrees where 2θ is greater than 25°, confirming that 9-fluorenone-CD IC had been successfully prepared.

图形用户界面

描述已自动生成

**Fig. 7.** Experimental results and prediction results using three ML models and three strategies in (0a) Isonicotinamide, (0b) Levulinic acid, and (1a) Prednisolone, (1b) 9-Fluorenone and (1c) Saccharin. The experiment resluts are shown as 1 or 0 displayed in the bold black font. Prediction value = 1 (red) and Prediction value = 0 (blue).

**Table 2** Prediction results on 5 compounds using three ML models and three strategies with optimal hyperparameters

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Compounds | Experimental Results | ANN model | LR model | SVM model | ML-based Linear Strategy | Recall-First Strategy | Precision-First Strategy |
| Isonicotinamide | - | + | - | - | - | - | - |
| Levulinic acid | - | - | + | - | - | + | - |
| Prednisolone | + | + | + | + | + | + | + |
| 9-Fluorenone | + | + | + | + | + | + | + |
| Saccharin | + | + | - | - | - | + | - |

a The shading in the table indicates the compounds forming ICs.

图表, 直方图

描述已自动生成

**Fig. 8.** (A) FTIR spectra and (B) Powder X-ray diffraction pattern of prednisolone (PD), β-CD, physical mixture and IC of PD and β-CD.

图表, 直方图

描述已自动生成

**Fig. 9.** (A) FTIR spectra and (B) Powder X-ray diffraction pattern of saccharin (SAC), β-CD, physical mixture and IC of SAC and β-CD.

图表, 直方图

描述已自动生成

**Fig. 10.** (A) FTIR spectra and (B) Powder X-ray diffraction pattern of 9-fluorenone, β-CD, physical mixture and IC of 9-fluorenone and β-CD.

*3.4. Discussion on ML models and ML-based strategies*

For the three ML models, they were successfully predicted that prednisolone and 9-fluorenone could form ICs. The prediction results did not fully agree with the experimental results for the other three compounds in the new dataset, isonicotinamide, levulinic acid, and saccharin. If a single ML model was used to predict the formation of the ICs, it is highly possible to miss some compounds or to have errors in the negative samples which were predicted as positive samples. None of the three ML-based design strategies perfectly separated the label 0 and label 1 obtained in the experiments, but the design purposes all achieved. The recall-first strategy gave all samples of experimental label 1 in the new dataset despite predicting levulinic acid as label 1. Although the precision-first strategy missed saccharin in the positive samples, all the identified positive samples were able to form the ICs, which also demonstrated the effectiveness of the strategy. Overall, the strategy can be designed to emphasize a single evaluation metric, which is the best solution to improve the poor prediction ability of a single ML model under the limited training set.

For pharmaceutical industrials, it is important to include all the positive candidate compounds (able to form ICs), and it is acceptable to include small numbers of negative candidate, which is also applied to the screening of polymorphs, solvates, or hydrates of complexs in the drug development stage. For this purpose, the strategy should be designed to obtain a relatively large ‘positive’ data set (which has more polymorphs), even if some of them do not exist. Because researchers can determine whether the predicted crystal forms exist or not through only a few experiments. In addition, prediction precision is same important for predicting particular phenomena in the drug crystallization process because this will significantly reduce the amount of experiment, solvent usage, and labor cost42. For circumstances of very expensive, toxic, or unstable compound, the precision-first strategy would bring extra benefits for the screening, including making a more environmentally friendly pharmaceutical development.

**4. Conclusion and Outlook**

In this work, we first developed three ML models, ANN, SVM and LR models, for predicting the formation of the cyclodextrin inclusion complexes, and three ML-based strategies based on different prediction purposes demonstrated the high efficiency and high accuracy. Firstly, 200 compounds were screened in the cooling crystallization experiments to collect the balanced positive (forming the the CD-ICs) and negative dataset of the compound samples. The samples were quantitatively described according to the structure description and molecular dynamics simulation analyses, and the 200×20-matrix obtained after the PCA dimensionality reduction was obtained and used as the model input. ANN, SVM and LR models were constructed and optimized for searching for the best prediction performance. All three ML models showed moderate prediction success rate, but the inconsistency of evaluation metrics also limited further applications. Therefore, based on these evaluation metrics, we proposed three ML-based prediction strategies: ML-based linear strategy, recall-first strategy, and precision-first strategy. Furthermore, the applications of three models and three strategies to a new dataset have led to the discovery of three new CD-ICs: prednisolone-CD IC, 9-fluorenone-CD IC, and saccharin-CD IC.

Recently, more and more potential applications of ML models were reported in materials science, drug development, traditional chemical industry and other fields. One of the most significant advantages of our proposed strategies is to selectively meet high standard requirements such as maximum prediction precision or maximum recall value. The adjustments in predicting results undoubtedly provide more precious solutions to certain practical challenges. The strategy proposed in prediction ICs would be widely applied in drug and new functional material development areas, the recall-first strategy should be used to obtain a relatively large positive sample data set without missing some important candidates of the APIs or new materials, which will still highly reduce total screening costs and development period. In addition, precision-first strategies have more potential applications for the traditional chemical industry, such as industrial crystallization processes, since predictions such as particle size distribution or yield require accurate prediction results rather than lots of possibilities. Meanwhile, for the prediction of special processes or phenomena, such as the spherical crystallization process we reported, strategies similar to ML-based linear strategy can be adopted to balance the advantages and disadvantages of different ML models while also providing good prediction performance.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgements**

Y. Ma acknowledges the support of China Scholarship Council. The authors are grateful to the financial support of National Natural Science Foundation of China (No. NNSFC 22111530115) and the Tianjin Municipal Natural Science Foundation (No.21JCYBJC00600).

**References**

**1.** Zhang J, Ma PX. Cyclodextrin-based supramolecular systems for drug delivery: recent progress and future perspective. *Adv Drug Deliv Rev.* Aug 2013;65(9):1215-1233.

**2.** Park J, Park J, Lee J, Lim C, Lee DW. Size compatibility and concentration dependent supramolecular host-guest interactions at interfaces. *Nature communications.* Jan 10 2022;13(1):112.

**3.** Ji HB, Long QP, Chen HY, Zhou XT, Hu XF. β-Cyclodextrin inclusive interaction driven separation of organic compounds. *AIChE Journal.* 2011;57(9):2341-2352.

**4.** Morin-Crini N, Fourmentin S, Fenyvesi É, et al. 130 years of cyclodextrin discovery for health, food, agriculture, and the industry: a review. *Environmental Chemistry Letters.* 2021;19(3):2581-2617.

**5.** Tang P, Sun Q, Suo Z, et al. Rapid and efficient removal of estrogenic pollutants from water by using beta- and gamma-cyclodextrin polymers. *Chemical Engineering Journal.* 2018;344:514-523.

**6.** Brewster ME, Loftsson T. Cyclodextrins as pharmaceutical solubilizers. *Adv Drug Deliv Rev.* Jul 30 2007;59(7):645-666.

**7.** Topuz F, Uyar T. Electrospinning of Cyclodextrin Functional Nanofibers for Drug Delivery Applications. *Pharmaceutics.* Jan 2019;11(1).

**8.** Zhao JJ, Yang J, Xie Y. Improvement strategies for the oral bioavailability of poorly water-soluble flavonoids: An overview. *International Journal of Pharmaceutics.* Oct 2019;570:118642.

**9.** Szejtli J, Szente L. Elimination of bitter, disgusting tastes of drugs and foods by cyclodextrins. *European Journal of Pharmaceutics and Biopharmaceutics.* Oct 2005;61(3):115-125.

**10.** Assadpour E, Jafari SM. A systematic review on nanoencapsulation of food bioactive ingredients and nutraceuticals by various nanocarriers. *Critical Reviews in Food Science and Nutrition.* Oct 2019;59(19):3129-3151.

**11.** Tian BR, Liu YM, Liu JY. Smart stimuli-responsive drug delivery systems based on cyclodextrin: A review. *Carbohydrate Polymers.* Jan 2021;251:116871.

**12.** Hu QD, Tang GP, Chu PK. Cyclodextrin-Based Host-Guest Supramolecular Nanoparticles for Delivery: From Design to Applications. *Accounts of Chemical Research.* Jul 2014;47(7):2017-2025.

**13.** Topuz F, Holtzl T, Szekely G. Scavenging organic micropollutants from water with nanofibrous hypercrosslinked cyclodextrin membranes derived from green resources. *Chemical Engineering Journal.* 2021;419:129443-.

**14.** Rezaei A, Fathi M, Jafari SM. Nanoencapsulation of hydrophobic and low-soluble food bioactive compounds within different nanocarriers. *Food Hydrocolloids.* Mar 2019;88:146-162.

**15.** Del Valle EMM. Cyclodextrins and their uses: a review. *Process Biochemistry.* May 2004;39(9):1033-1046.

**16.** Landy D, Mallard I, Ponchel A, Monflier E, Fourmentin S. Remediation technologies using cyclodextrins: an overview. *Environmental Chemistry Letters.* Sep 2012;10(3):225-237.

**17.** Gao S, Liu YY, Jiang JY, et al. Encapsulation of thiabendazole in hydroxypropyl-beta-cyclodextrin nanofibers via polymer-free electrospinning and its characterization. *Pest Management Science.* Sep 2020;76(9):3264-3272.

**18.** Zhang MK, Lyu Y, Zhu X, Wang JP, Jin ZY, Narsimhan G. Enhanced solubility and antimicrobial activity of alamethicin in aqueous solution by complexation with gamma-cyclodextrin. *Journal of Functional Foods.* Jan 2018;40:700-706.

**19.** Ouyang D. Computational Pharmaceutics: Application of Molecular Modeling in Drug Delivery. 2015.

**20.** Dhoot AS, Fernandes GJ, Naha A, Rathnanand M, Kumar L. Design of Experiments in Pharmaceutical Development. *Pharmaceutical Chemistry Journal.* Nov 2019;53(8):730-735.

**21.** Zhao Y, Al-Fahdi M, Hu M, et al. High-Throughput Discovery of Novel Cubic Crystal Materials Using Deep Generative Neural Networks. *Adv Sci (Weinh).* Oct 2021;8(20):e2100566.

**22.** Song Y, Siriwardane EMD, Zhao Y, Hu J. Computational Discovery of New 2D Materials Using Deep Learning Generative Models. *ACS Appl Mater Interfaces.* Nov 17 2021;13(45):53303-53313.

**23.** Sousa T, Correia J, Pereira V, Rocha M. Generative Deep Learning for Targeted Compound Design. *Journal of chemical information and modeling.* Nov 22 2021;61(11):5343-5361.

**24.** Stephenson N, Shane E, Chase J, et al. Survey of Machine Learning Techniques in Drug Discovery. *Current Drug Metabolism.* 2019;20(3):185-193.

**25.** Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature.* Jul 2018;559(7715):547-555.

**26.** 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.* Nov 2019;9(6):1241-1252.

**27.** Di PW, Chen JH, Liu LP, Li WH, Tang Y, Liu GX. In silico prediction of binding capacity and interaction forces of organic compounds with alpha- and beta-cyclodextrins. *Journal of Molecular Liquids.* Mar 2020;302:112585.

**28.** Zhang SL, Wang JS, Lin ZH, Liang YY. Application of Machine Learning Techniques in Drug-target Interactions Prediction. *Current Pharmaceutical Design.* 2021;27(17):2076-2087.

**29.** Zhang L, Tan JJ, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today.* Nov 2017;22(11):1680-1685.

**30.** Vriza A, Canaj AB, Vismara R, et al. One class classification as a practical approach for accelerating pi-pi co-crystal discovery. *Chem Sci.* Dec 8 2020;12(5):1702-1719.

**31.** Devogelaer JJ, Meekes H, Tinnemans P, Vlieg E, de Gelder R. Co-crystal Prediction by Artificial Neural Networks. *Angew Chem Int Ed Engl.* Nov 23 2020;59(48):21711-21718.

**32.** Ma Y, Gao Z, Shi P, Chen M, et al. Machine learning-based solubility prediction and methodology evaluation of active pharmaceutical ingredients in industrial crystallization. *Frontiers of Chemical Science and Engineering*. 2021;16(4):523-535.

**33.** Wang W, Yang T, Harris WH, Gomez-Bombarelli R. Active learning and neural network potentials accelerate molecular screening of ether-based solvate ionic liquids. *Chem Commun (Camb).* Aug 14 2020;56(63):8920-8923.

**34.** Molecular Operating Environment (MOE), Version 2019.0102; Chemical Computing Group ULC: Montreal, 2019.

**35.** Szoplik J, Ciuksza M. Mixing time prediction with artificial neural network model. *Chemical Engineering Science.* 2021;246:116949.

**36.** Metta N, Ramachandran R, Ierapetritou M. A novel adaptive sampling based methodology for feasible region identification of compute intensive models using artificial neural network. *AIChE Journal.* 2020;67(2):e17095.

**37.** Jayaweera CD, Narayana M. Multi-objective dynamic optimization of seeded suspension polymerization process. *Chemical Engineering Journal.* 2021;426:130797.

**38.** Onel M, Kieslich CA, Pistikopoulos EN. A Nonlinear Support Vector Machine-Based Feature Selection Approach for Fault Detection and Diagnosis: Application to the Tennessee Eastman Process. *AIChE J.* Mar 2019;65(3):992-1005.

**39.** Lee J, He QP. Understanding the effect of specialization on hospital performance through knowledge-guided machine learning. *Computers & Chemical Engineering.* 2019;125:490-498.

**40.** Bian Y, Jing Y, Wang L, Ma S, Jun JJ, Xie XQ. Prediction of Orthosteric and Allosteric Regulations on Cannabinoid Receptors Using Supervised Machine Learning Classifiers. *Molecular pharmaceutics.* Jun 3 2019;16(6):2605-2615.

**41.** Nandy A, Duan C, Kulik HJ. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks. *Journal of the American Chemical Society.* Oct 27 2021;143(42):17535-17547.

**42.** Ma Y, Sun M, Liu Y, et al. Design of Spherical Crystallization of Active Pharmaceutical Ingredients via a Highly Efficient Strategy: From Screening to Preparation. *ACS Sustainable Chemistry & Engineering.* 2021;9(27):9018-9032.

**43.** Hu J, Shao DD, Chen CL, et al. Plasma-Induced Grafting of Cyclodextrin onto Multiwall Carbon Nanotube/Iron Oxides for Adsorbent Application. *Journal of Physical Chemistry B.* May 2010;114(20):6779-6785.

**44.** Aleem O, Kuchekar B, Pore Y, Late S. Effect of beta-cyclodextrin and hydroxypropyl beta-cyclodextrin complexation on physicochemical properties and antimicrobial activity of cefdinir. *J. Pharm. Biomed. Anal.* Jul 2008;47(3):535-540.

**45.** Zhao GJ, Han KL. Hydrogen Bonding in the Electronic Excited State. *Accounts of Chemical Research.* Mar 2012;45(3):404-413.