

The Research Progress of AI-assisted New Drug Screening Technology

Ruilin Ma

College of Arts and Science, New York University, New York, 10003, USA

Abstract: Drug R&D is a systematic project with high investment, high risk and long period, and the traditional methods are faced with double pressures of cost and time. Computer-aided drug design (CADD) has introduced four steps: target discovery, confirmation, lead compound discovery and optimization, but the prediction accuracy is low and the consideration of drug formation is insufficient. With the development of artificial intelligence and big data, AI shows its potential in drug discovery, and improves the accuracy and efficiency of virtual screening through machine learning and deep learning. AI-driven virtual screening technology accurately captures molecular characteristics, enhances the enrichment of active molecules, reduces invalid tests and saves research and development costs. At present, AI-assisted virtual screening combines the advantages of traditional computing and AI learning, and combines ligand-based, structure-based and AI-driven technologies to provide innovative solutions for drug discovery. Technical breakthroughs such as deep generation model indicate that the field of drug discovery will face changes. It is of great value to study the application of AI-driven virtual screening technology to improve research and development efficiency and reduce costs.

Keywords: I Auxiliary; New Drug Screening; Virtual Screening Technology; Computer Aided Drug Design.

1. Introduction

Drug research and development is a complex system engineering with high investment, high risk and long period. The traditional development process includes the stages of target discovery and confirmation, lead compound optimization, preclinical and clinical research, etc. It often takes more than 10 years and costs more than 2-3 billion US dollars, which brings heavy cost burden to the pharmaceutical

industry. Traditional drug discovery relies on Qualcomm screening to obtain lead molecules, and it needs to synthesize and test tens of millions of molecules, which is time-consuming and costly and limits the technical popularization [1]. Computer-aided drug design (CADD) is introduced to alleviate this dilemma. Its research includes four steps: target discovery, confirmation, lead compound discovery and optimization (Figure 1). However, the traditional CADD has low accuracy in predicting binding affinity and lacks the consideration of drug formation.

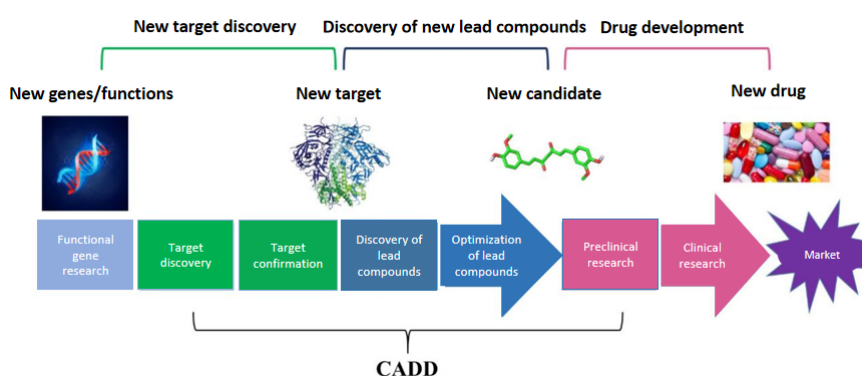


Figure 1. Flow chart of CADD new drug research and development

With the rapid development of artificial intelligence and big data, the potential of AI in the field of drug discovery has been effectively released, and machine learning and deep learning will also help improve the accuracy and efficiency of virtual screening. Virtual screening based on AI can grasp molecular information more effectively, enrich active molecules, reduce redundant testing, and reduce testing cost and testing time. At present, AI-assisted virtual screening combines the advantages of traditional computing and AI learning to provide a creative method. Technologies such as deep generation model break through the traditional boundaries, which indicates that the field of drug discovery will face changes, and it is of great value to study its

application.

2. Virtual Screening Technology Foundation

2.1. Ligand-based Virtual Screening (LBVS)

Ligand-based virtual screening (LBVS) is an important calculation method to identify potentially active compounds in drug discovery. It does not depend on the three-dimensional structure of the target protein, but predicts the activity of new compounds by analyzing the chemical characteristics of known active ligands. Its theoretical basis is the principle of ligand similarity, that is, molecules with

similar chemical structures often have similar biological activities [2]. The core technologies of LBVS are ligand similarity search and chemical fingerprint matching. The latter is an important method to encode molecular structural features, which can quickly compare and screen large-scale compound libraries [3,4]. There are typical models based on ligand similarity search, pharmacophore model and quantitative structure-activity relationship model (QSAR). These models explore the relationship between the structure and activity of compounds, which are expressed by some mathematical statistical models. In recent years, with the development of artificial intelligence technology and the use of deep learning convolutional neural networks, the prediction accuracy of LBVS has also been greatly improved. The common method is to evaluate the performance of LBVS model through internal data and external database, including ROC curve analysis, enrichment factor and so on. Although there are problems such as stereoisomerism of compounds and how to improve the success rate of novel structure identification, LBVS has been proved to be able to successfully identify novel drug lead molecules. Although more efforts are needed for LBVS to further develop the tasks of identifying active molecules in the external chemical space and understanding the interaction mechanism between complex biological systems.

2.2. Structure-based Virtual Screening (SBVS)

Structure-based virtual screening (SBVS) is a computational method to find compounds with novel structures and potential activities. By docking the compounds into the target protein binding pocket, the binding ability of

the compounds is evaluated by using a scoring function according to the matching degree of geometric shape and chemical characteristics. The biggest difference between SBVS and LBVS is that it depends on the target structure rather than the structural information of ligands, and it is easier to get new compounds with unique structures and functions. This method is mainly divided into molecular docking, molecular dynamics simulation and pharmacophore selection, among which virtual screening (DBVS) in molecular docking is the most commonly used method. In fact, there are also some successful cases that show that this research method is feasible and effective, such as the discovery of non-covalent forms of DprE1 inhibitors, new androgen receptor inhibitors, and new binding modes [5]. Nevertheless, SBVS still has the following shortcomings: for example, it depends on the three-dimensional structural data of the target protein, and the effect of SBVS will be limited if the target lacks relevant experimental reports. However, with the development of homologous structure construction technology and the development of AlphaFold protein structure prediction database, this limitation has been eased. However, the accurate calculation of free energy is not mature enough at present, and it is a pity that we can't accurately predict the binding capacity of compounds. Therefore, the contradiction between the calculation speed and the efficiency of virtual screening still needs to be further solved by researchers, hoping to solve these two problems simultaneously in future research.

2.3. Evaluation Index System

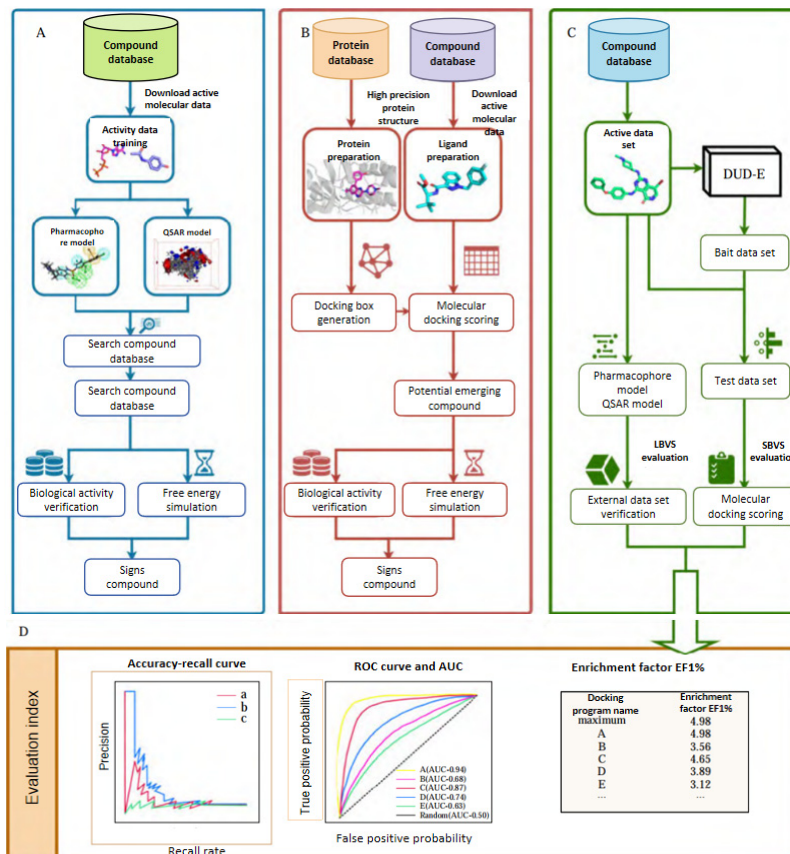


Figure 2. Process and evaluation index of virtual screening

Evaluating virtual screening is of great significance in the field of drug discovery, which can evaluate the efficiency and

reliability of screening strategies. Whether SBVS or LBVS, the evaluation system and core indicators are extremely

critical. At present, the evaluation of virtual screening efficiency is mainly done in the form of retrospective verification, that is, the historical data before the model is established to test the effectiveness of the model to predict the newly established model or strategy. The advantage of this evaluation method is that it can predict the credible evidence of the performance of the model in the "real world" environment. Commonly used criteria include accuracy, search efficiency, ROC curve and enrichment factor, among which accuracy refers to the proportion of effective drug candidates, while search efficiency reflects the ability of search tools to capture effective chemicals. ROC curve is used to visually represent the performance of the classification model, and the area under ROC (AUC) is used as a tool for quantitative evaluation, and the range of AUC is between 0 and 1 [6]. The final enrichment factor is used to determine the ability of virtual screening to find effective compounds. LBVS usually uses internal cross-validation and external data sets to verify the model performance, while SBVS uses active data sets and bait data sets generated by DUD-E to evaluate it.

3. Core Technology of AI in Drug Screening

3.1. Molecular Characterization Technology Innovation

As a key factor affecting the performance of AI model,

molecular representation has evolved from intuitive and readable form to special mathematical descriptors and fingerprints, and now it has developed into learning representation that can capture patterns and salient features in a vast chemical space, among which graph-based molecular representation is particularly popular. At present, AI model can express molecules in the following three ways: firstly, SMILES string can be used, which is a short and linear text coding form, and can describe molecular topological structure, connection mode between atoms, etc. AI model can be transformed into its internal judgment standard through these strings to learn and extract the characteristics of molecules; Secondly, molecular fingerprints are used to solve the problem. Commonly used molecular fingerprints include extended connectivity fingerprints (ECFP) and intermolecular system fingerprints (MACCS). Molecular fingerprints are mainly used to extract molecular structural features, query compound activity and compound similarity, etc. Finally, using a graph-based molecular expression, which is the most advanced graphical expression at present, more molecular structure information can be obtained, molecular atoms and chemical bonds can be simulated as nodes and lines in the graph, and the complete molecular three-dimensional structure information can be obtained by using graph neural network and other algorithms, thus effectively constructing a molecular attribute prediction model (see Figure 3).

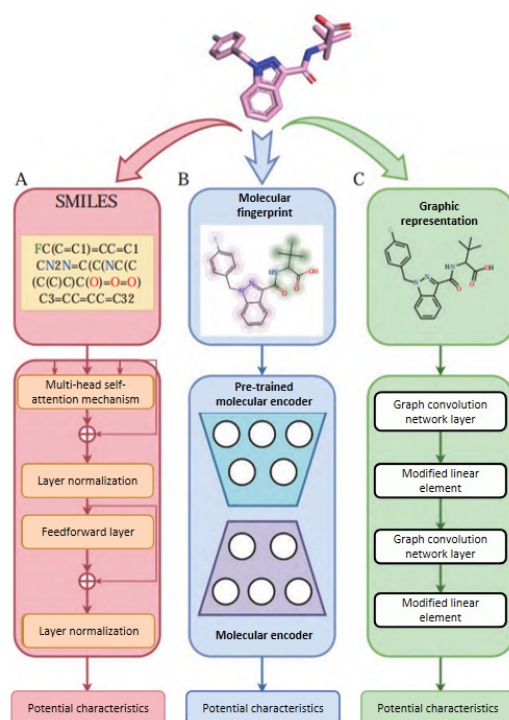


Figure 3. Common molecular characterization methods in AI model

3.2. AI-driven QSAR Modeling

With the progress of machine learning algorithm and the expansion of data scale, the traditional QSAR modeling method has been gradually supplemented and innovated by AI-driven method. Compared with the traditional linear regression model, the current QSAR modeling is more inclined to use machine learning methods such as support vector machine (SVM), artificial neural network and decision tree to generate nonlinear regression model, thus significantly

improving the prediction ability of the model. For deep learning, researchers have developed many new models, such as Transformer-CNN model based on Transformer technology, molecular model based on cloud 3D QSAR, Uni-Mol platform which can characterize 3D structure, and so on. These models have shown quite high efficiency in the aspects of molecular properties and conformation generation. When preprocessing the data, researchers used algorithms such as Alchemite to fill in the missing values of sparse drug data, and used methods such as RGIFE to delete redundant features

of the data set, which improved the training level of the model. In addition, the introduction of multi-task learning strategy enables the model to predict multiple related molecular properties at the same time, which overcomes the limitation of the traditional single-task QSAR model. Nevertheless, the QSAR model driven by AI still faces challenges such as insufficient interpretability and uneven data quality, which is also an important direction for future research.

3.3. Combined Affinity Intelligent Prediction

Traditional molecular docking software has obvious limitations in predicting drug-target binding affinity (DTA). The research shows that the average success rate of the best conformation obtained by mainstream commercial and academic docking software is 47%~68%, which is because the accuracy of entropy calculation by traditional scoring function cannot be guaranteed, and at the same time it relies too much on basic parameters such as molecular weight and distribution coefficient [7]. In order to overcome these shortcomings, researchers have developed a variety of prediction models constructed by deep learning methods. Models such as DeepDTA and WideDTA constructed by convolutional neural network (CNN) can adaptively learn sequence characteristics from protein sequences and drug SMILES strings. When the component of Graph Neural Network (GNN) is introduced, the situation has changed significantly. For example, PLANET can improve its generalization ability by training diversification. KarmaDock also uses GNN, which integrates self-attention mechanism, to continuously optimize molecular position information to get the best accuracy in PDBBind data set. Also in the application of DNN method, ArkDTA and DeepDock can significantly improve the prediction accuracy of binding sites and the scoring accuracy of affinity through attention mechanism and feature integration strategy [8]. These fully automatic learning models can eliminate the subjectivity of human experts in feature engineering processing to a certain extent, and can better grasp various complex molecular interaction

laws. However, the existing models still have the problems of generalization ability and interpretability, and the prediction results of novel skeleton compounds need further investigation.

3.4. Application of Pre-training Large Model

The application of pre-training large model in the field of drug discovery is becoming more and more mature, which has brought many breakthroughs. In the field of chemistry, ChemBERTa, a model based on BERT framework, has been able to learn useful properties of compounds from more than ten million unlabeled libraries, and achieved good performance, such as predicting molecular properties and drug-drug interactions. ChemBERTa has been proved to have the best performance in solving the drug-target association prediction problem through transfer learning and adaptation. AlphaFold has a certain impact on the field of protein structure prediction. By using multi-sequence alignment to obtain evolutionary information, more than 200 million kinds of protein have been accurately predicted through CNN and attention mechanism [9]. The latest version of AlphaFold 3 plays a very important role in the decoding of molecular behavior and biological computing. In the field of molecular generation, MolGPT, as a molecular generation model based on GPT structure, has great potential to create good, unique and novel molecules in terms of characteristics, efficiency and diversity through masked self-attention mechanism. It is superior to or consistent with the current machine learning framework. The effective use of these pre-trained large models not only improves the accuracy and efficiency of virtual screening, but also means that it will greatly expand the horizons of pharmaceutical chemistry exploration.

4. Practical Application of AI-assisted Drug Screening

4.1. Hierarchical Screening Strategy

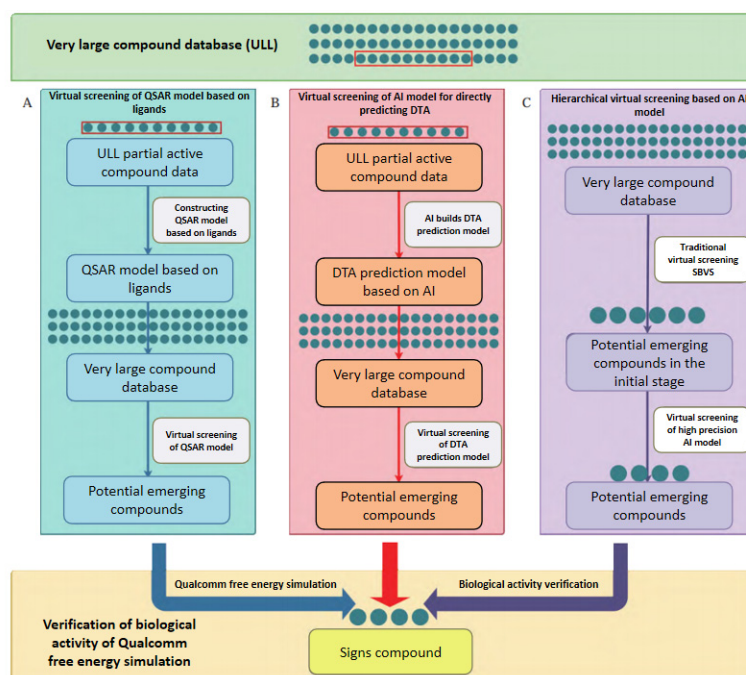


Figure 4. Flow chart of virtual screening based on artificial intelligence model

With the development of AI technology, researchers began to deeply integrate the AI model with the traditional

computer-aided drug design (CADD) method to construct a hierarchical screening strategy. This strategy overcomes the problems of poor generalization performance of traditional LBVS and high calculation cost of SBVS. From the perspective of model design, adding AI-based model re-scoring after traditional virtual screening undoubtedly improves the screening accuracy and accuracy. For example, the Deep Docking platform uses the scoring information of a subset of chemical libraries to train the depth model of QSAR, and efficiently screens large compound libraries by iteratively eliminating redundant molecules [10].

In the multi-stage screening process, the traditional virtual screening method is used for initial screening, then the high-precision AI model is combined for fine screening, and finally the candidate compounds are confirmed by molecular dynamics simulation or experimental verification (see Figure 4). This hierarchical strategy not only ensures the breadth of screening, but also ensures the accuracy of the results. In order to achieve the balance between accuracy and efficiency, researchers have developed ensemble learning methods such as ENS-VS, which use different kinds of machine learning classifiers to achieve the unification of the two goals without sacrificing high enrichment factors and greatly improving the screening rate [11]. At the same time, emerging technologies such as DeepDock and KarmaDock also use molecular and protein crystal structure data to train deep learning models to solve the problems of conformation search and scoring, so as to further improve the effect of enriching active molecules. This combination strategy of AI and CADD not only optimizes the virtual screening results, but also puts forward a more efficient and reliable drug development route. Nevertheless, the existing hierarchical screening methods still have some shortcomings, such as lengthy links and poor repeatability, which need to be further explored in standard process design and improving the generalization ability of the model.

4.2. Typical Application Case Analysis

In recent years, AI-assisted drug screening has achieved remarkable results in many important targets. In the study of PI5P4K- β target, Wei (2023) and others used XREINVENT reinforcement learning model to screen 15 million molecules and achieved an amazing success rate of activity detection of 52.94% [12]. Xie (2022) and others used the combined machine learning Mol2vec model to screen the AD targets in the database of 3738 natural small molecules, and obtained a high accuracy rate of 44.44% [13]. Zheng(2024) and others found metabolic dormant *Escherichia coli* inhibitors in the 800,000-molecule broadinstitute database by establishing a graph neural network model, and achieved an excellent detection rate of 46.67% [14]. In addition, Kang(2022) and others used HoTS model to search for inhibitors of P2X3 receptor in KCB library containing 150,000 molecules, and the test rate was only 3.2%, but they still found an efficient prodrug with $IC_{50} = 3.07 \mu\text{mol L}^{-1}$ [15]. As far as industry is concerned, pharmaceutical companies have generally adopted AI-assisted screening technology. For example, the PharmAIDiscoveryEngine system established by Gryniukova(2023) and others screened drugs targeting Sirtuin-1 from Enamine data set containing 2.6 million molecules, and even with a retrieval efficiency of 2.07%, some active molecules with IC_{50} value of $1.2 \mu\text{mol L}^{-1}$ were found, which has high commercial value [16]. The above cases fully demonstrate the effectiveness of artificial

intelligence-assisted screening tools in improving efficiency and reducing costs, which indicates that artificial intelligence-assisted screening technology will play an increasingly important role in the future drug research and development process.

5. Conclusion

Although AI has made great progress in the field of drug screening, there are still many challenging problems to be solved. For example, the problem of model interpretability, deep learning model is usually called "black box", and its decision-making process is difficult to understand and explain, which is very difficult for scenes that need to deeply analyze the principle of drug behavior; Data quality problems, at present, many training data have problems such as noise and nonstandard marking errors, which will reduce the model performance; The problem of false positive prediction is caused by the over-fitting of the model to a specific training data set or the lack of enough counter-example information. In order to solve the above problems, future research will focus on multi-dimensional information fusion methods, such as using GNN, multi-modal neural network, knowledge map and so on to synthesize a variety of information; Develop explanatory AI technology and use XAI methods such as SHAP and LIME to improve the transparency model; Strengthen the close connection between experiment and calculation, and improve the accuracy of data by mixing data. Finally, AI, especially the deep generative model, breaks the boundaries of traditional personalized drug design, which may lead to important subversive progress in the field of drugs, and it may give birth to new personalized therapeutic drugs. The breakthrough of these technologies will promote the development of AI-assisted drug screening in a more efficient and reliable direction.

References

- [1] Li Xiaohui, Du Guanhua. Introduction to new drug research and evaluation [M]. People's Health Publishing House: 2022 11:497.
- [2] Gao Zhigang, Li Yueqing, Song Qiling, et al. Innovative experimental design of medicinal chemistry based on virtual screening technology [J]. Laboratory Science, 2024,27(04): 133-137.
- [3] Gao Yu, Wang Fengxue, Liu Haibo. The application of virtual screening technology in the research and development of new natural products [J]. International Journal of Pharmaceutical Research, 2020,47(08):602-608.
- [4] Liu Yutian, Zhao Shiyu, Lu Shaowa. Application progress of computer virtual screening technology based on molecular docking in new drug discovery [J]. Chemical Engineer, 2020, 34 (02):59-63.
- [5] Cui Herong, Chen Kedian, Wang Cheng, et al. Screening lead compounds against novel coronavirus based on molecular docking technology [J]. Northwest Journal of Pharmacy, 2021, 36 (03):449-454.
- [6] Zheng Mingyue, Jiang Hualiang. High-value data mining and artificial intelligence technology to accelerate innovative drug research and development [J]. Pharmaceutical Progress, 2021, 45 (07):481-483.
- [7] J E Z, A J V, W I A, et al. Discovery of antibiotics that selectively kill metabolically dormant bacteria.[J].Cell chemical biology, 2023,31 (4) :712-728.e9.

- [8] Lin W ,Min X ,Zhiqiang L , et al.Hit Identification Driven by Combining Artificial Intelligence and Computational Chemistry Methods: A PI5P4K- β Case Study.[J].Journal of chemical information and modeling,2023,63 (16) .
- [9] Anastasiia G ,Florian K ,Iryna M , et al.AI-Powered Virtual Screening of Large Compound Libraries Leads to the Discovery of Novel Inhibitors of Sirtuin-1.[J].Journal of medicinal chemistry,2023,66 (15) :
- [10] Xiangying Z, Haotian G ,Haojie W , et al.PLANET: A Multi-objective Graph Neural Network Model for Protein-Ligand Binding Affinity Prediction.[J].Journal of chemical information and modeling,2023,
- [11] Mook K K, Ingo L, Hojung N , et al.AI-based prediction of new binding site and virtual screening for the discovery of novel P2X3 receptor antagonists[J].European Journal of Medicinal Chemistry,2022,240114556-114556.
- [12] Chenglong X, XuXu Z ,Zhangming N , et al.Amelioration of Alzheimer's disease pathology by mitophagy inducers identified via machine learning and a cross-species workflow. [J]. Nature biomedical engineering,2022,6 (1) :76-93.
- [13] John J ,Richard E ,Alexander P , et al.Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873) : 583-589.
- [14] Francesco G, Vibudh A ,Michael H , et al.Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery.[J].ACS central science,2020,6 (6) :939-949.
- [15] Li J ,Liu W ,Song Y , et al. Improved method of structure-based virtual screening based on ensemble learning[J].RSC Advances, 2020, 10 (13) :7609-7618.
- [16] Zhe W, Huiyong S, Xiaojun Y, et al.Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power.[J].Physical chemistry chemical physics : PCCP, 2016, 18 (18) :12964-75.