Advances in Computer, Signals and Systems (2021) 5: 14-21
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2021.050204
ISSN 2371-8838

Polarized Surface Area and Oil-Water Partition Coefficient Predicted through the Machine Learning Based on Deepchem

Jin Li

Sichuan Normal University, Chengdu, Sichuan 610101, China

zblijin@stu.sicnu.edu.cn

Keywords: Machine learning, Polarized surface area, Oil-water partition coefficient, Deepchem, Drug-like

Abstract: Accurate prediction of the chemical information of compounds played a vital role in the discovery of drug-like properties. The Polarized surface area (PSA) and oil-water partition coefficient (AlogP) of drug-like compounds properties were predicted based on DeepChem. By comparing the four models (Random Forest, Deep Neural Network, Convolutional Neural Network, Graphic Convolution), it is shown that CNN has 94% accuracy in PSA, DNN has 81% accuracy in AlogP. DeepChem can easily build a platform for molecular machine learning to predict specific characteristic attributes by selecting data sets, which may provide the possibility for drug prediction.

1. Introduction

The discovery of drug-like compounds usually requires a large number of biochemical experiments, which measures the physicochemical properties and bioavailability of chemical compounds. It usually costs a lot to conduct these experiments, and therefore, accurate prediction of the key properties of drug-like properties has a guiding role in the improvement of existing drugs and the discovery of new drugs.

Machine learning is a new computational technique that has had a major impact in many fields and promoted many key progresses. Technics like speech-recognition all rely on multi-layer neural network learning to develop well. They also have a good effect on the prediction of physicochemical properties of drug-like compounds.

DeepChem is an open-source project aiming at promoting the democratization of deep learning in drug discovery, which contains data on the properties of over 70,000 compounds, molecular characterization models and common machine learning models [1]. For example, RF, DNN, CNN, etc. are all used to facilitate the discovery and development of drugs.

Machine learning frameworks seem well suited to predict the compound properties, since they allow multitask learning and automatically construct complex models [2]. This is especially important for the prediction of drug-like properties, because for most compounds, in terms of the few measurements available, it may not be possible to construct a valid representation in the prediction of traditional methods. In contrast, machine learning shows good performance.

The core challenge of drug molecular machine learning is to efficiently encode molecules into fixed-length vectors. Recently, studies have shown that it may be feasible to use SMILES strings for further learning tasks [3]. Modeling was performed using two wide range of molecular characterization methods and one of them is Extended-Connectivity Fingerprinting (ECFP) [4].

The present work firstly realized the use of random forest (RF) [5], the multi-task deep neural network (DNN) and the convolutional neural network (CNN) [6] and used the regression model; it also analyzed the key drug-like properties based on the graph convolution model (GC) [7], thus provided supports for drug prediction and compound.

This article carried out the modeling and quantitative prediction towards the two most common molecular properties in the literature on drug-like research: oil-water distribution coefficient (ALogP) and polarization surface area (PSA), and then provide a basis for drug prediction and synthesis. All calculations were done in the open-source machine learning package DeepChem [1].

2. Methods

The author divided the compounds into three sets to validate the predictive model, adopted attribute calculation to facilitate the experiments and finally adopted two forecast models to predict the drug-like properties. Besides, molecular featurization and evaluation standard also contributes to the discussion of the results.

2.1 Data Set

The current data set includes 1522 synthetic human BACE-1, which covers experimental data from at least 30 different laboratories, including research institutes, biopharmaceutical companies, etc. This data set can be loaded by the DeepChem built-in function [1].

2.2 Data Set Splitting

To validate the predictive model, the compounds were divided into a training set (80%), a validation set (10%), and a test set (10%). The training sets were used to train models, while validation sets were used to adjust hyperparameters of each model, and test sets were used to evaluate the model. Generally, data sets of machine learning use the method of random splitting data sets, but for molecular data, based on the label index as the benchmark model, it is convenient to compare and analyze the prediction performance of different machine models. In order to obtain more complete results, the data set is divided into 8 subsets, and the number of compounds is separated from 222 to 1522. And four kinds of neural network methods were mainly discussed to build the model on the training set, and then verify and test.

2.3 Attribute Calculation Methods

Two types of attribute calculation are used in the experiments, namely, AlogP (oil-water partition coefficient) and PSA (polarized surface area).

2.3.1 AlogP (Oil-Water Partition Coefficient)

The partition coefficient of oil-water (Log P) is the logarithm of the partition coefficient of the compound in the n-octanol/water two-phase system (the ratio of the concentration of the compound in n-octanol to water), characterizing the distribution of the substance in the two phases of oil and water. Log P measured by the ACD software is recorded as AlogP. Ghose et al [8] studied the physicochemical properties of 6304 drugs in the CMC database, indicating that 80% of the drugs

have AlogP between -0.4 and 5.6, an average of 2.52. Log P is the basic physical properties of drugs. Therefore, the study on the distribution of Log P is an important part of drug-like prediction, and it has a good application prospect in this field.

2.3.2 PSA (Polarized Surface Area)

Polarized surface area (PSA) is the sum of van der Waals surface areas of all polarized atoms [9]. It is also the main description indicator for evaluating the transportability of drugs in cells and can be used to predict the functions of human bodies. Veber et al [10] found that the flexible structure of the molecule has an important effect on molecular absorption through oral experiments in rats. When (1) PSA does not exceed 140 $Å^2$ (or the sum of the hydrogen bond donor and acceptor does not exceed 12) (2) the rotatable single bond does not exceed 10, the compound has a good bioavailability of 20-40%. The former PSA is related to the desolvation energy of the polarizing group permeating the cell membrane. In the experiments of Martin et al. [11], the PSA with negatively charged compounds have a decisive influence on bioavailability: when PSA$\leqslant$75 $Å^2$, the probability of utilization >10% is 85%, when PSA$\geqslant$150 $Å^2$, this probability dropped to only 11%, thus indicating that PSA is of great significance to drug-like prediction.

2.4 Forecast Model

Three forecast model, namely, Random Forest, Neural Network Model and Graphic Convolution are combined to better predict the key drug-like properties.

2.4.1 Random Forest

Random Forest (RF) [5], as developed in 2001 and first introduced by Breiman, is the principle of random forests into decision tree algorithms. In RF, multiple trees are generated with features randomly selected for each input quantity, and the predictions of the results are jointly determined by all trees and combined based on the majority vote rules. In this model, the number of trees was set to the default value of 100.

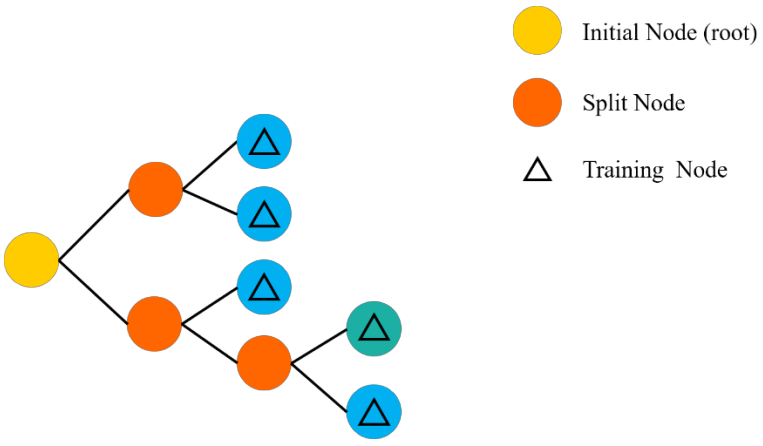

Fig.1 Structural Framework of Random Forest.

2.4.2 Neural Network Model

Neural Network (NN) is an important algorithm in machine learning and the core algorithm of deep learning. Two widely used neural network algorithms, Deep Neural Network (DNN) and Convolution Neural Network (CNN), are mainly discussed here. As is known to all, the multi-layer neural network consists of three parts: the input layer, the hidden layer, and the output layer, thus,

the neural network architecture seems to be very suitable for predicting the target.

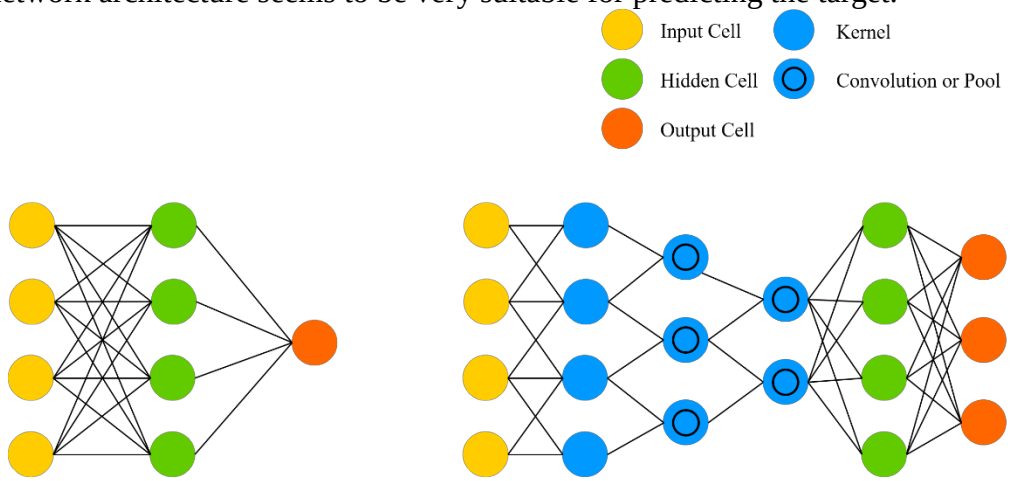

Fig.2 Structural Framework of Deep Learning Network (Left) and Convolutional Neural Network (Right).

Deep Neural Network (DNN) is composed of several hidden layers. For each assumed input (including output), it is independent of each other, and the layers are fully connected, that is, any neuron of the i-th layer must be connected to any of the neurons in the next layer i+1-th. In this model, the number of loops was set to 50.

Convolutional Neural Network (CNN) [6] uses the convolution kernel as an intermediary to avoid the direct connection between the upper and lower neurons. Generally, by the entire image as input, all the image information is shared in the same convolution kernel, and the image is subjected to convolution operation. After throwing and retaining the original positional relationship, the local structure can be mined, which is more suitable for images multitasking recognition processing.

2.4.3 Graphic Convolution

Graphic Convolution (GC) [7] is a new graph-based prediction model that extends the decomposition principles of circular fingerprints, which allows adaptive learning by using differentiable network layers. The useful information needed in the current task is extracted from it. Moreover, the graph convolution model treats molecules as undirected graphs, and the learning process can be applied to every atom and bonding atom of the molecular structure. This model uses RDKit to convert the SMILES string into a molecular graph based on Deepchem [1].

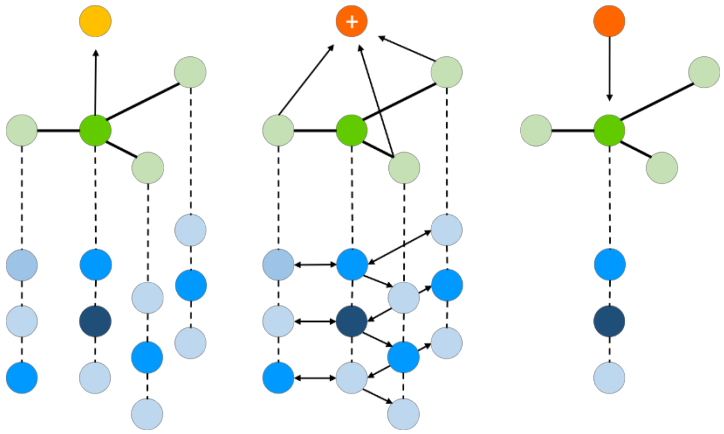

Fig.3 Structural Framework of Graphic Convolution.

2.5 Molecular Featurization

Models include (RF, DNN, and CNN) use ECFP features, and graph-based GC models use their circular convolution features.

Extended-Connectivity Fingerprinting (ECFP) [4] is widely-used molecular feature in chemical informatics. The molecules are decomposed into sub-modules from heavy atoms. Each sub-module is assigned a unique identifier and then extended with the identifier by binding.

Graph convolution (GC) [7] supports most graph-based models. It computes an initial feature vector and a neighbor list for each atom and further generates graph structures with the local chemical environment around the atoms and connectivity of the whole molecule.

2.6 Evaluation Standard

For this regression dataset MAE and R^2 as metrics for the model used to evaluate the effect of the model.

MAE refers to Mean Absolute Error, which represents the average of the sum of the absolute value of the predicted value and the true value, reflecting the accuracy of the model prediction.

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|(y_i - \hat{y}_i)| \qquad (1)$$

R^2, or R Squared, characterizes the statistic that reflects the degree of linear correlation between two variables.

$$R^2 = \frac{E^2[xy]}{\sigma_x^2 \sigma_y^2} \qquad (2)$$

3. Results and Discussion

The first step in the data analysis work is visualization. Therefore, the predictive properties AlogP and PSA are firstly visualized. The following figure shows the distribution of the two properties in the dataset. It can be seen that AlogP is mainly distributed between -2-6, while PSA is mainly distributed between 0-200 Å^2.

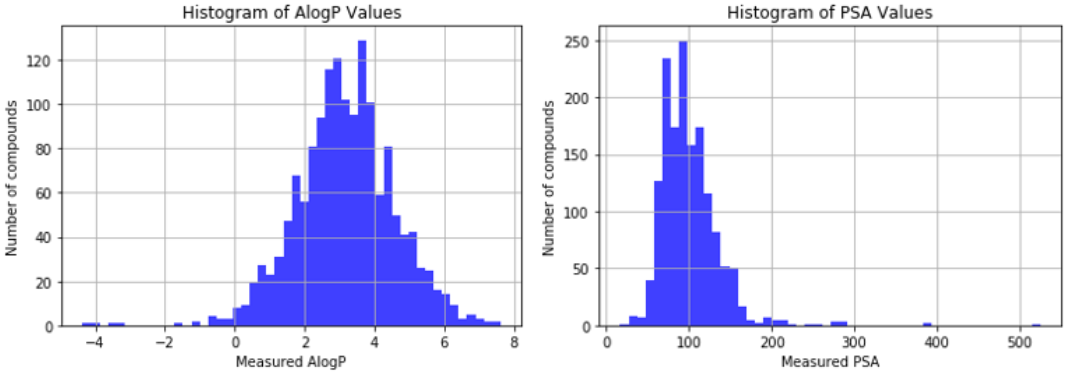

Fig.4 Distribution of Alogp and Psa Properties in the Complete Data Set 1522, ('Number of compounds', y-Axis) and ('Alogp or PSA' Measurements, X-Axis).

3.1 Model Comparison

The author compared the predictive performance of various machine learning models with various methods, especially the graph convolution method based on molecular graphs. Using molecular convolution will principally achieve better performance results.

Table .1. Performance comparison of target property prediction methods. The table gives the

mean standard deviation of R^2 and MAE values of the comparison algorithm on the test set. Overall, CNN (fourth column) performed best for PSA, while DNN (third column) performs better in the prediction of the AlogP.

PSA BEST	RF	DNN	CNN	GC
R^2	0.835±0.007	0.873±0.007	0.940±0.004	0.859±0.014
MAE	232.120±20.001	8.440±0.459	7.485±1.162	8.856±1.137
AlogP BEST				
R^2	0.722±0.025	0.813±0.003	0.799±0.004	0.765±0.037
MAE	0.751±0.079	0.507±0.021	0.518±0.062	0.525±0.052

The GC does not perform optimally and this could be related to the selection of the size and nature of the dataset. After expanding the size of the data set, it may have better performance.

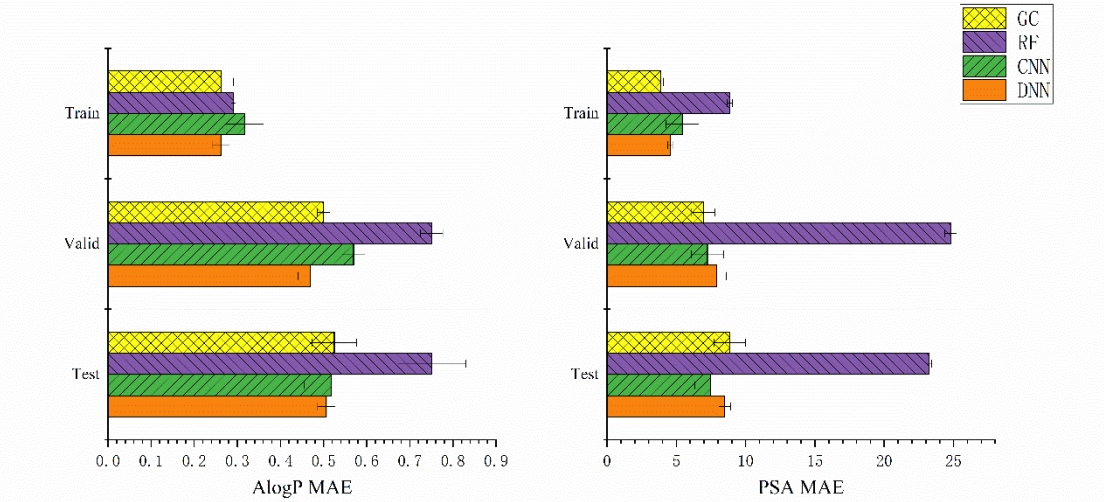

Fig. 5. The best performance of AlogP and PSA in the training set, validation set, and test set of MAE in different models, (‘data set type’, y-axis and ‘MAE best performance’, x-axis). Colors represent different predictive models, namely RF, DNN, CNN, GC. Each data point is the best performance of 5 independent operations, and the standard deviation is shown as an error bar. The MAE value of the RF in the PSA is one-tenth of its true value. It can be seen that for the PSA, the MAE of the RF is too large, and this model may not be suitable for PSA prediction.

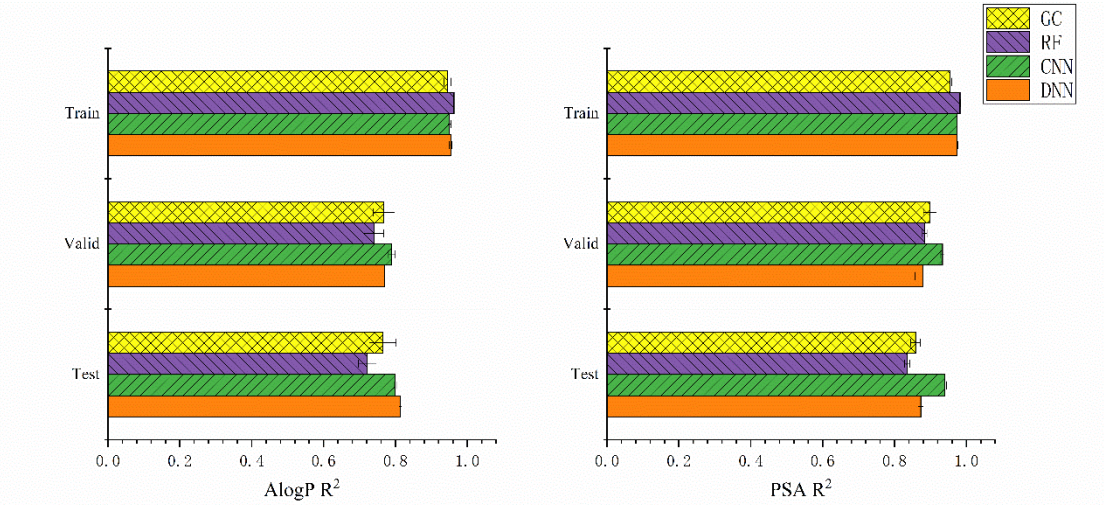

Fig. 6. The best performance of AlogP and PSA in the training set, validation set, and test set of R^2 in different models (‘data set type’, y-axis and ‘R^2 best performance’, x axis). Colors represent different predictive models, namely RF, DNN, CNN, GC. Each data point is the best performance of 5 independent operations, and the standard deviation is shown as an error bar. It can be seen that the

four models have a better prediction fit for PSA.

3.2 Data Set Size and Predictive Performance

By dividing the data set into subsets of different sizes, it can also get the relationship between the data set size and the predicted performance. By examining the correlation between MAE, R^2 values and the training set size (see Fig. 7). It can be observed that larger data sets in principle lead to better predictions. However, when the sample is small, the models also show good prediction correlation, and when the data set size reaches a certain value, the prediction result tends to be stable.

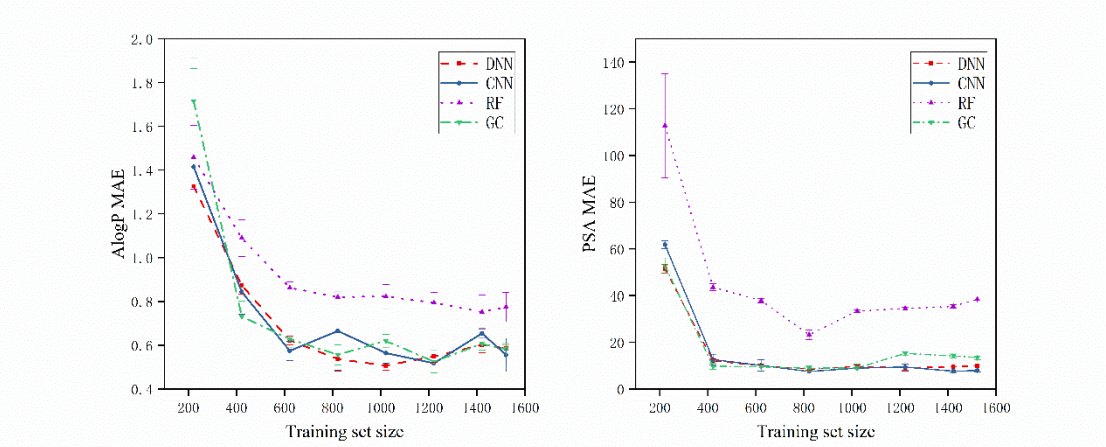

Fig.7. Trend graph of the predictive evaluation indicator MAE, ('MAE', y-axis and 'training set size', x-axis). Colors represent different predictive models, namely RF, DNN, CNN, GC. Each data point is the average of 5 independent operations, and the standard deviation is shown as an error bar. The MAE value of the RF in the PSA is one-tenth of its true value.

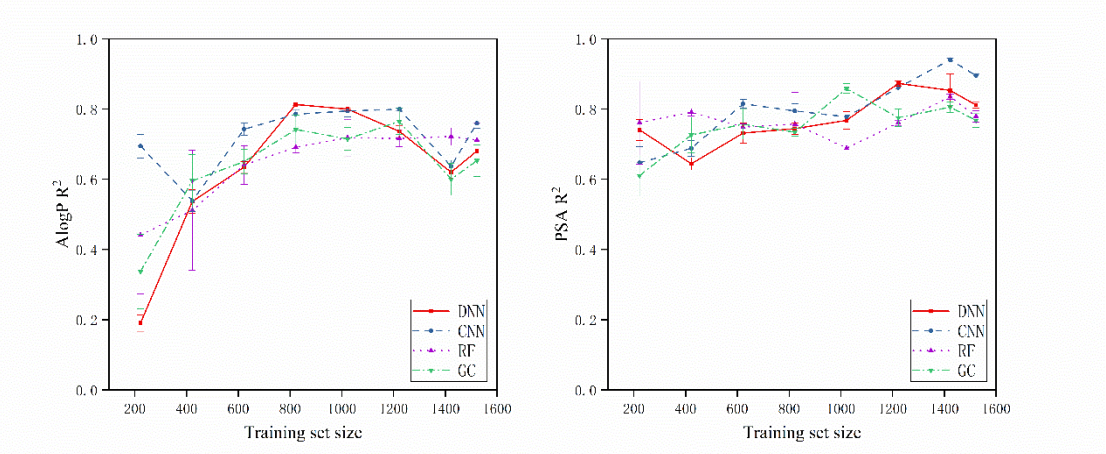

Fig.8 Trend Graph for Predictive Evaluation Index R^2, ('R^2', y-Axis and 'Training Set size', X-Axis).

Colors represent different predictive models, namely RF, DNN, CNN, GC. Each data point is the average of 5 independent operations, and the standard deviation is shown as an error bar. When the number of compounds in the data set exceeds a certain scale, it will lead to over-fitting, which will cause the prediction performance to decline.

4. Conclusion

Based on the DeepChem open-source software package, the key physicochemical properties PSA and AlogP in the drug-like properties were predicted by RF,DNN,CNN and GC, which reduces the common deviations in traditional methods. By predicting the various models of machine learning, it is observed that when the sample is small, GC based on graph convolution does not show obvious advantages, and the CNN is generally outperforming the other methods. Because it usually inputs the entire image and shares the parameters between neurons, the calculation of the convolution layer and pooling layer makes it better than other prediction methods in general. As the size of the data set increases, this performance will improve in principle. In general, when the prediction result of the machine learning model is not satisfactory, the changes of its hidden layer number, activation functions and molecular characteristics can be considered. The existing methods in DeepChem only allow separate experiments for different models and do not implement parallel processing of multiple data sets, which can be improved in future studies.

References

[1] Deep-learning models for Drug Discovery and Quantum Chemistry. https://github. com/deepchem/deepchem, 2018 (accessed July 25, 2018).

[2] Bengio, Y., A. Courville, and P. Vincent, Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 2013. 35(8): pp. 1798-1828.

[3] Gómez-Bombarelli, R., et al., Automatic chemical design using a data-driven continuous representation of molecules (2016). arXiv preprint arXiv:1610.02415.

[4] Rogers, D. and M. Hahn, Extended-connectivity fingerprints. Journal of chemical information and modeling, 2010. 50(5): pp. 742-754.

[5] Breiman, L., Random forests. Machine learning, 2001. 45(1): pp. 5-32.

[6] LeCun, Y., et al., Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998. 86(11): pp. 2278-2324.

[7] Duvenaud, D.K., et al. Convolutional networks on graphs for learning molecular fingerprints. in Advances in neural information processing systems. 2015.

[8] Ghose, A.K., V.N. Viswanadhan, and J.J. Wendoloski, A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. Journal of Combinatorial Chemistry, 1999. 1(1): pp. 55-68.

[9] Clark, D.E., Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. Journal of pharmaceutical sciences, 1999. 88(8):pp. 807-814.

[10] Veber, D.F., et al., Molecular properties that influence the oral bioavailability of drug candidates. Journal of medicinal chemistry, 2002. 45(12): pp. 2615-2623.

[11] Martin, Y.C., A bioavailability score. Journal of medicinal chemistry, 2005. 48(9): pp. 3164-3170.