Prediction of Drug-likeness of Central Nervous System Drug Candidates Using a Feed-Forward Neural Network Based on Chemical Structure

Yi-Gao Yuan, Xiao Wang*

School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, Jiangsu, China

Email: wangxiao@nju.edu.cn

Abstract: The modern medical science has been greatly advanced by the development of new drugs, despite the fact that the process of developing new drugs is costly and time-consuming. An accurate prediction method for the drug-likeness at the early stage of drug discovery is highly desirable, as it will facilitate the discovery process and reduce the overall cost, and eventually contribute to the well-being of human beings. Based on a central nervous system (CNS) drug dataset, we constructed an artificial neural network (NN) to predict the CNS drug-likeness of a given compound. Based on the published results, we first constructed a simple feed-forward neural network to learn and predict the possible correlations between twelve physiochemical properties and the CNS drug-likeness. The accuracy of prediction has reached 80%, which is higher than previous reports. The successful implementation of NN to predict the CNS drug-likeness indicated that NN could be a powerful tool for the prediction. Moreover, we further constructed a neural network based on the chemical structure, and the accuracy has reached 86%. We hope that these methods can serve as an applicable set of protocols for virtual drug screening.

KEYWORDS: Central Nervous System, Drug-likeness, Drug Screening, Neural Network

Introduction:

Modern medicine has advanced rapidly that many drugs are approved every year. However, the cost of developing a new drug is massive. Reducing the cost in the early stage of drug discovery such as lead compound screening is important for reducing the overall cost. Traditional lead compound screening methods include Lipinski's Rule of Five [1], and recently scientists have been using computational methods to screen lead compounds. In specific, when targeting the central nervous system (CNS) drugs, multiparameter optimization methods have been developed, which includes CNS multiparameter optimization (MPO)[2] and probabilistic MPO method [3]. In a broader sense where all drugs are considered, quantitative estimation of drug-likeness (QED) was proposed by Hopkins et al. [4] to give an estimation of whether a compound is likely to be druggable or not. The general form of these approaches is formatted as follows:

$$score = \Sigma w_k f(x_k);$$

where x_k is the physicochemical properties, f(x) is the function that maps x_k into a weightless score, w_k is the weight factor which measures the importance of each physicochemical property. All of these multiparameter optimization methods are relatively accurate, with an accuracy of approximately 70%. However, there is space for improvement, and in this paper, we demonstrate by implementing artificial neural network, the accuracy would be improved.

The artificial neural networks (ANN) are widely used in function approximation, classification and data processing. A typical ANN has three components: architecture, activation function, and learning rule. The common types of ANN include feed-forward networks, recurrent networks, and reinforcement networks. In this work, we will demonstrate the implementation of a feed-forward network (FFN) [5] to improve the accuracy of judging the drug-likeness of a chemical compound. An FFN consists of an input layer, a certain number of hidden layers and an output layer. The input layer receives values of the independent variables, and the hidden layers process input information based on an activation function and pass on the values to the next layer. The output layer receives values from the hidden layers and gives an output. The learning process of the FNN is achieved by supervised learning using a training sample set and backpropagation-gradient descendent algorithm. The learning process is monitored by a loss function, which represents the deviation from the true answer. A dataset from the work by Gunaydin [3] are also analyzed, which has 665 marketed orally available drugs and 14 corresponding physicochemical properties. Among them, 299 were reported to be capable of penetrating the brain-blood barrier which is denoted as CNS drugs. To make the analysis more operational, we herein redefine the CNS drug-likeness as the ability of crossing the brain-blood barrier.

Methods:

An FFN consists of an input layer with 128 neurons and a rectified linear unit (ReLU) activation function. The second layer has 32 neurons, and the third layer has 16 neurons. The output layer uses sigmoid as the activation function. Between each of the two layers, a dropout layer with a factor of 0.5 was added to prevent overfitting. The supervised learning uses Adam optimizer to reach minimal loss. The training dataset was adopted from Gunaydin's approach [3], tabulated and imported. Further data processing including cleaning null values, normalization was performed before the data is fed into the neuron network.

```python
def cns_nn_simple():
    model = keras.Sequential([
                keras.layers.Dense(128,activation=tf.nn.relu, input_shape=[len(normed_train_data.keys())]),
                keras.layers.Dropout(0.5),
                keras.layers.Dense(32,activation=tf.nn.relu),
                keras.layers.Dropout(0.5),
                keras.layers.Dense(16,activation=tf.nn.relu),
                keras.layers.Dropout(0.5),
                keras.layers.Dense(1,activation=tf.nn.sigmoid)
            ])


    model.compile(
                    optimizer='adam',
                    loss='binary_crossentropy',
                    metrics=['accuracy','binary_crossentropy']
                    )

    #model.summary()

    return model
```

Figure 1. Python code that constructs the simple feed-forward network using TensorFlow.

Results and Discussion:

The supervised learning process is monitored by the loss and cross-entropy, to verify that overfitting is not happening, we used a validation dataset which is completely different from the training dataset. Figure 2 shows how the loss and cross-entropy change as the supervised learning proceeds. The output layer of FFN uses a sigmoid activation function, which gives a value between 0 and 1. We take the output value and compare it with a cutoff value to define whether a compound is considered to be a CNS drug or not.

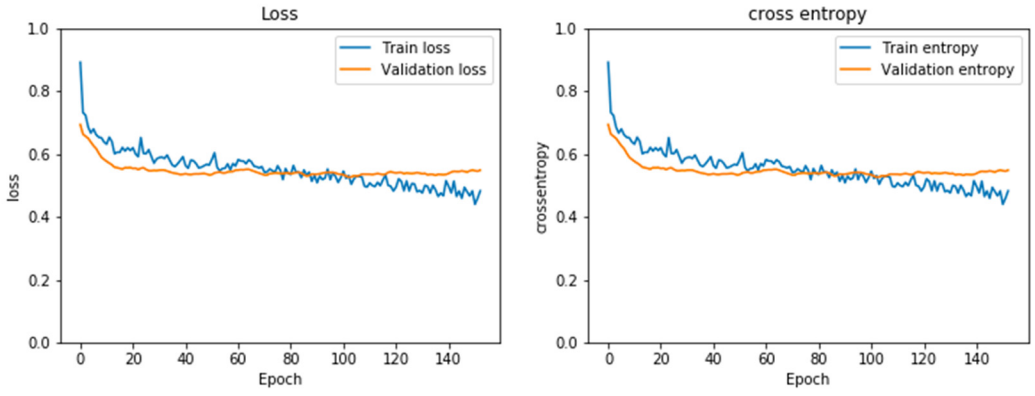

Figure 2. The loss of the training dataset and the validation dataset as a function of learning iteration. The cross-entropy is shown on the right. Both diagrams show that the neuron network is not overfitting because the training dataset has a similar loss compared to the validation dataset.

To visualize the capability of the FFN model, confusion matrixes and receiver operating curves (ROC) are utilized. Four quadrants, which represent the true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) predictions, are quantized in the confusion matrix. The ROC curve adopts the false-positive rate (FPR) as the x-axis and the true-positive rate (TPR) as the y-axis. The perfect model should be at (0,1), which means no false prediction since FPR is defined as FP/(FP+TN). Figure 3 shows the confusion matrix and the ROC for the simple FFN model. The true positive set, also known as the hit set has 217 drugs; and the true negative set, also known as the correct rejection set, has 317 drugs. This gives the accuracy of the simple FFN model, which is (TP+TN)/Total = 80.5%. Type I error, the false positive set has 49 drugs, and Type II error, the false negative set has 80 drugs. So the true-positive rate (TPR) = TP/(TP+FN) = 73.1%, and the false-positive rate (FPR) = FP/(FP+TN) = 13.4%. Based on the high accuracy, the high TPR, and the low FPR, the simple FFN model has a good performance, with better accuracy than former models. The ROC curve gives a clear illustration of where the cutoff value should be, which is marked with the green dot in Figure 3B.

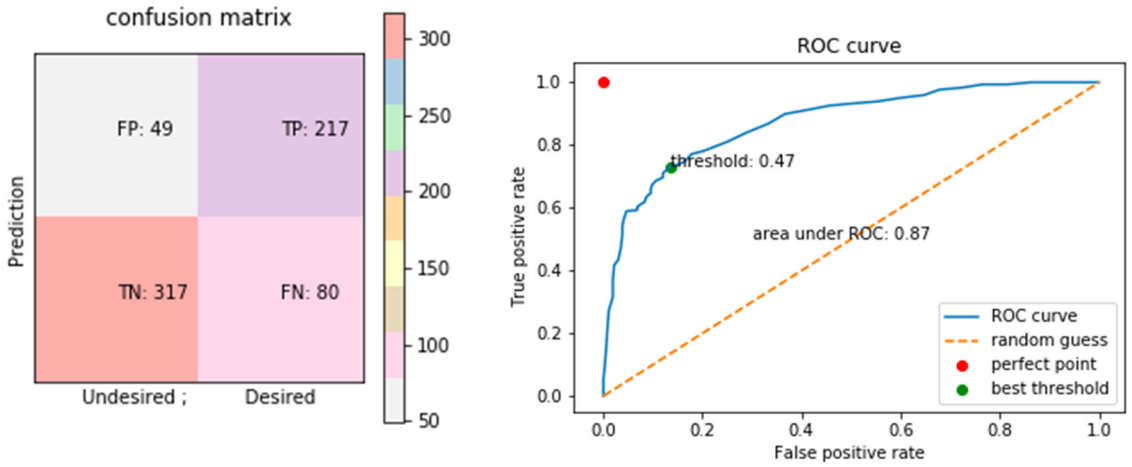

Figure 3. (A) Confusion matrix showing that the prediction accuracy is (TP+TN)/Total = 80.5%. (B) Receiver operating curve showing that the area under ROC is 0.87, and the best threshold is 0.47.

To further examine the applicability of the FFN model, we took advantage of an open-source chemical informatics python package called RDKit [6], which is able to generate the physicochemical properties of any molecule based on its chemical formula, the SMILES [7] only. Using the RDkit package, we generated 45 different physicochemical properties of each inputted SMILES, and then fed the values into the input layer of the FFN constructed before. The loss function and cross-entropy performance can be found in Figure 4. Both diagrams show that the neuron network is not overfitting because the training dataset has a

similar loss compared to the validation dataset. Under the condition that the FFN model does not overfit, the prediction based on the new FFN model is even better than the initial FFN model.

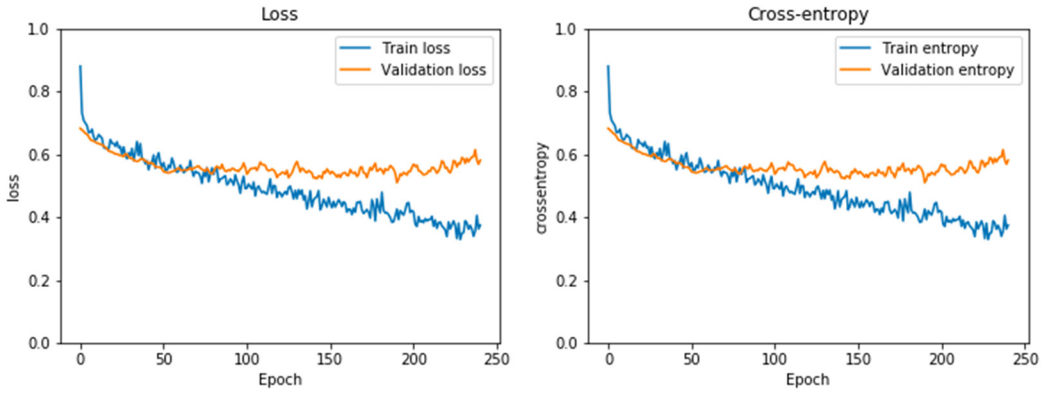

Figure 4. The loss of the training dataset and the validation dataset as a function of learning iteration. The cross-entropy is shown on the right.

Figure 5 shows the confusion matrix and ROC curve for the new FFN model. From the confusion matrix, TP has 260 drugs, and TN has 313 drugs. The FP and FN have 53 and 39 drugs, respectively. The accuracy reaches 86.4%, and the TPR reaches 87% while FPR is 14.5%. The ROC curve has an optimal cutoff value of 0.43, and the area under ROC reaches 0.91. This new FFN model shows better performance as compared to the initial FFN model.

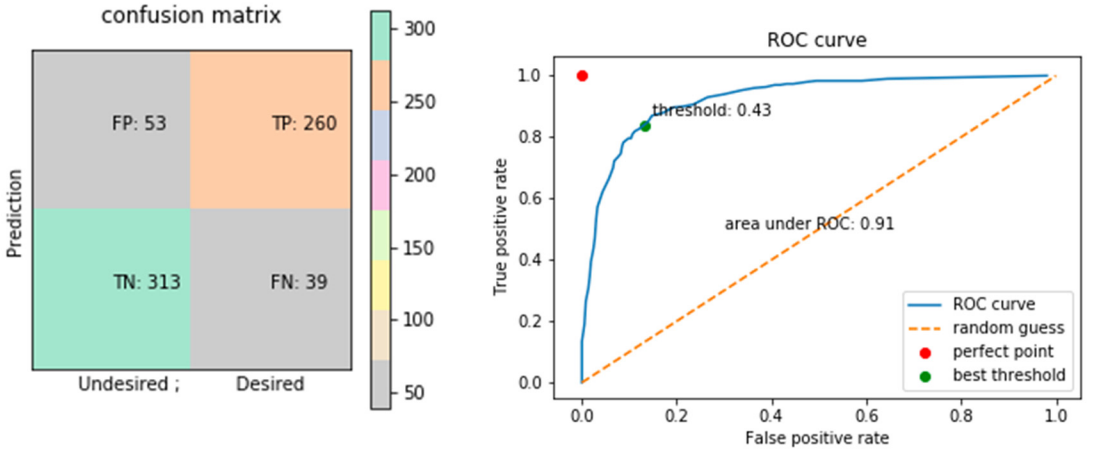

Figure 5. (A) Confusion matrix showing that the prediction accuracy is (TP+TN)/Total = 86.4%. (B) Receiver operating curve showing that the area under ROC is 0.91, and the best threshold is 0.43.

Nevertheless, achieving a high accuracy does not eliminate false prediction. Six drugs with false-positive classification are shown in Figure 6. Estradiol, for example, is a steroid hormone and proven to be able to penetrate the brain-blood barrier. However, the dataset put it in the wrong category as a false-positive. For other false predictions, we believe that the error of the method is the primary cause.

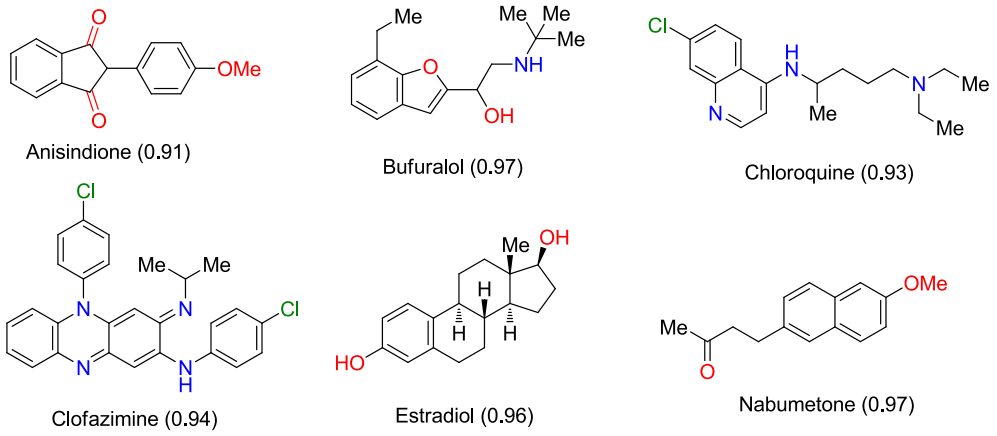

Figure 6. Some examples of drugs with false-positive predictions. The corresponding prediction score is in the parenthesis.

Conclusion

In this work, feed-forward networks are implemented to predict the CNS-drug-likeness, the accuracy has reached 80% using the simple FFN, and further improved to 86% by involving more physicochemical properties, higher than the previous reports. Although the current size of the dataset is limited to show the full potential of neural networks in processing large datasets, the primary objective has been achieved to verify the possibility to utilize the neural network in predicting drug-likeness of an organic small molecule. The investigation of larger datasets of CNS and other drugs using this FFN method is currently on-going in our lab to demonstrate a greater potential of utilizing neural networks in drug discovery.

References:

[1] LIPINSKI C A, LOMBARDO F, DOMINY B W, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1[J]. Advanced Drug Delivery Reviews, 2012, 64(1– 3): 4–17.

[2] WAGER T T, XINJUN H, VERHOEST P R, et al. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties[J]. ACS Chemical Neuroscience, 2010, 1(6) : 435.
[3] GUNAYDIN H. Probabilistic Approach to Generating MPOs and Its Application as a Scoring Function for CNS Drugs. [J]. ACS Medicinal Chemistry Letters, 2015, 7(1) : acsmedchemlett.5b00390.
[4] BICKERTON G R, PAOLINI G V, BESNARD J, et al. Quantifying the chemical beauty of drugs [J]. Nature Chemistry, 2012, 4(2) : 90 – 98.
[5] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural networks, 1989, 2(5) : 359 – 366.
[6] RDKit: Open-source cheminformatics; http://www.rdkit.org
[7] WEININGERD.SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [J]. Journal of chemical information and computer sciences, 1988, 28(1): 31 – 36.