ISSN 2959-409X

Application and difficulties of deep learning in drug target discovery

Xiaoqian Kuang

Jiaozhou No.1 Middle School
Email: xiaoqiankuang85@gmail.com

Abstract:

The integration of deep learning technology within the domain of drug target discovery represents a cutting-edge approach in the pharmaceutical sciences. This paper delves into the methods and prevailing issues associated with the application of deep learning in the identification of new drug targets. We initiate with an exposition of the fundamental concepts and methodologies underpinning deep learning, subsequently illuminating its utilization in the realm of drug target discovery. Specific attention is given to the deployment of architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) among others.

By interpreting experimental outcomes, it has been discerned that deep learning potentially enhances the precision of identifying drug targets. Nonetheless, this technological arena is not devoid of challenges. A significant hurdle lies in the interpretability of the models—understanding why and how these models arrive at their conclusions is often non-trivial. Moreover, the efficacy of deep learning is heavily reliant on the quality and volume of the datasets employed; insufficient or poor-quality data can severely impede the performance and reliability of predictive models.Through this inquiry, we aim to furnish novel insights into the utilization of deep learning for drug target discovery and underscore the challenges that must be addressed. The perspectives offered herein are vital to propel the progress of this burgeoning field, paving the way for more effective and efficient drug discovery processes.

Keywords:deep learning, drug target discovery, Neural network algorithm, convolutional neural networks, recurrent neura networks, multi-layer perceptron

1.Background and meaning

In the midst of rapid scientific and technological advancements, the urgency and demand for accelerated drug research and development have intensified. The pivotal challenge lies in the swift and precise identification of viable drug targets—a process essential for the creation of effective therapeutics. The advent of deep learning technology has ushered in a new era of potential solutions to this imperative issue. Building upon this context, this paper endeavors to further investigate the application of deep learning in the arena of drug target discovery. We aim to dissect the nuances of how deep learning algorithms can be harnessed to predict and validate new targets for drug development with greater accuracy and efficiency than traditional methods. This includes an examination of various deep learning models, such as CNNs, RNNs, and other advanced machine learning techniques that can analyze complex biological data. Concurrently, we recognize that the application of deep learning is not without its challenges. Critical issues such as the black-box nature of deep learning models—which affects their interpretability—and the dependency on large volumes of high-quality data will be scrutinized. Limitations pertaining to computational resources, algorithmic biases, and the need for robust validation methods to prevent overfitting are also discussed. This paper aims to contribute a comprehensive overview of the current state of deep learning in drug target discovery, while also highlighting the gaps and challenges that must be addressed. By doing so, it aspires to pave the way for future research and development, fostering advancements that could revolutionize the pharmaceutical industry and patient care.

The drug development process can be simplified into four main stages: target identification, discovery and optimization of lead compounds, preclinical research, and clinical research.

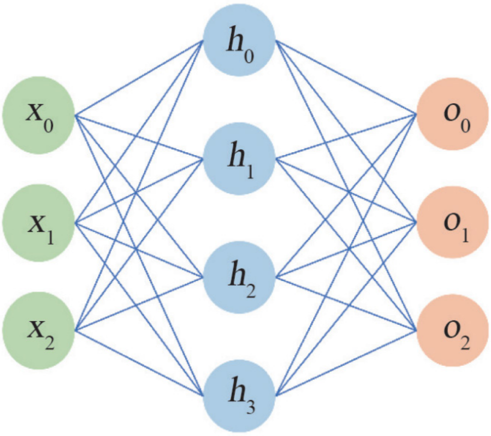
Among them, drug target identification is the first step in the modern drug development model, and also the key step to determine the success of new drug development. Current methods for drug target discovery are mainly analyzing genomic and proteomic data. These targets may be proteins, nucleic acids (DNA, RNA), or other biological macromolecules. Subsequently, researchers use cell biology, genetics and molecular biology methods to verify the feasibility of potential targets, including the functional mechanism of the target, the correlation of the target and disease, and the designability of drug molecules, and then to determine the drug target. The reported computational methods for drug target discovery have been divided into two categories: ① Strategies based on reverse molecular docking (e. g., IdTarget and TarFishDock), This class of method has large computational quantities, And is limited by the accuracy of the candidate target structures, For systems with unknown structure that cannot accurately predict ② on the premise that similar compounds have similar action targets, By comparing the structural similarities of the discovered active compounds to the active compounds of known targets, Establish an indirect association network between compounds and targets, Methods to reveal candidate drug targets (e. g., ChemMapper, PharmMapper, and SwissTargetPredictio n), This class of methods relies on data of small molecule target information, Therefore, it is less effective on small molecules with novel chemical structure.

2. Application of deep learning in drug-target identification

2.1Common neural network algorithms

2.1.1multi-layer perceptron

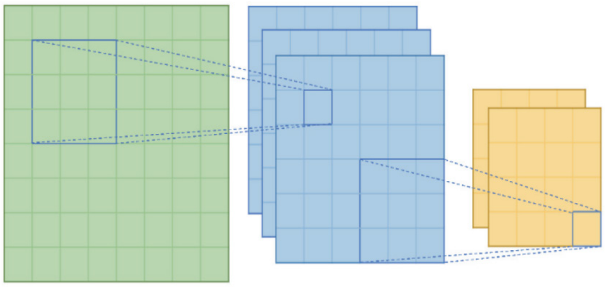
Multi-layer perceptron (MLP) is a fully connected network consisting of an input layer (input layer), one or more hidden layers (hidden layer), and an output layer (output layer) (Figure 1). Each neuron in the fully connected network is connected to the neuron in the previous layer, and the connections have weights. Therefore, each neuron can be calculated from all neurons in the previous layer, the formula is as follows: hi = wj * xj + bi, ok= hi * hi + bk, (1) where xj is the neuron in the input layer, hi is the neuron in the hidden layer, and ok is the neuron in the output layer. In order to avoid the limitation of linear dependence between network input and output, non-linear excitation functions (such as Sigmoid, tanh and Softplus) can also be introduced into the neurons of the perceptron, so that the input of neurons can be mapped to the output in a non-linear form



graph 1

2.1.2convolutional neural network

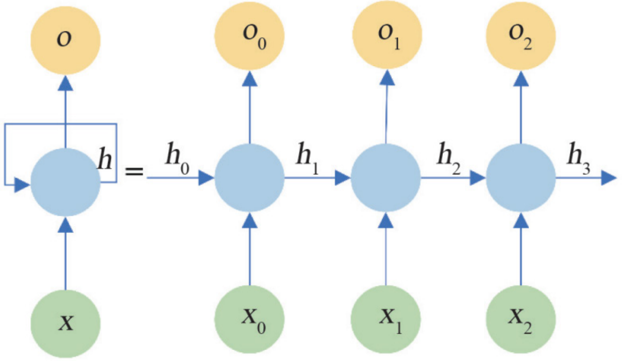
Convolutional neural network (convolution neural network, CNN) generally consists of convolutional layer and pooling layer. As the most important part of the CNN, the convolution computation can effectively reduce the parameters in the neural network. The network shown in Figure 2 consists of two convolutional layers, and each step in each layer is a convolution, expressed by the formula: (f * g) = f (τ) g (n- τ), (2) where f is the input data and g is the convolution kernel. The formula is a mathematically defined convolution, and is a one-dimensional form, while in deep learning, the form is often high-dimensional, and some modifications will be made in the implementation. The convolution operation of a layer is then a convolution layer. Increasing the number of layers of the convolution, using the residual network and pooling operations, can further optimize the convolutional neural network (such as VGG, Resnet, etc.).



graph 2

2.1.3recurrent neural network

The clic neural network (recurrent neural network, RNN) is a neural network that considers temporal feed forward. RNN takes into account the sequence of the input, that is, each input takes into account the previous output information, reflecting the “memory function”, and is the best choice for practical sequence analysis (Figure 3). The model is expressed as: oi, hi = f (xi-1, hi-1), (3)Where, xi is the neurons in the input layer, hi is the neurons in the hidden layer, and oi is the neurons in the output layer. Selective input of information before current learning can improve the effect of RNN, such as long-term and short-term memory (long short-term memory, LSTM), gated cycle unit (gated recurrent unit, GRU), etc.



graph 3

2.1.4Common deep learning tools

At present, there are various deep learning frameworks, such as PyTorch, TensorFlow, Paddle and Keras, which provide a platform for the construction of neural networks and bring great convenience to the development of application models based on deep learning. Up to now, there are many toolkits based on deep learning algorithms, such as DeepChem[1], DeepPurpose[2] and OpenChem[3]. On this basis, combined with the successful establishment of simple deep learning algorithms, it provides direct tools for users to use new data sets for training.

2.2apply

Protein-protein interactions are an important class of drug targets, and several drugs targeting protein-protein interactions have been successfully marketed. However, predicting protein – protein interactions remains challenging. In 2019, David Baker et al. [4] combined coevolutionary theory of protein amino acid sequence action sites and successfully predicted 1 618 E. coli protein pairs and 384 groups of unreported protein interactions of B. tuberculosis proteins by developing computational methods. However, for eukaryotes, including humans, coevolutionary analysis has less homologous sequence information. Therefore, the prediction accuracy of the coevolutionary analysis method based on the statistical algorithm is greatly limited. Deep learning method plays a certain role in promoting the development of the field (for example, DPPI [5] using convolution, random projection and full connection prediction three modules of neural network), through the protein protein function sequence amino acid sequence, protein interaction sequence, the accuracy regression curve auPR score of about 41% (human test set). In addition, such as MaSIF [6] used geometric neural network (geometric neural network), the protein surface geometry, chemical features and the interaction between biological macromolecules, established the protein protein-protein interaction and protein-small molecule interaction site prediction method, in the protein-protein interaction site to predict the median ROC AUC of each protein is 0.81.

3.Existing problems and challenges

In the process of drug target discovery, convolutional and recurrent neural networks are widely used. The convolutional neural network identifies the features of drug molecules by simulating the human visual system, and then realizes the identification of drug targets through the combination of features. The recursive neural network processes the data of sequential types by remembering the previous information and realizing the prediction of drug targets. The combination of the two can greatly improve the accuracy of drug target discovery. However, despite the strong potential in drug target discovery, there are several challenges. First, the model is not highly explanatory, and although deep learning models can provide high-precision prediction results, its internal computational process is complex, and this “black-box” nature makes the results of the model difficult to interpret. Secondly, the quality and quantity of data are also an important factor restricting the application of deep learning. High-quality data is the basis of deep learning model training, but in practice, it is often very difficult to obtain high-quality drug target data. In addition, training deep learning models requires a large amount of data, and the lack of data volume can also affect the performance of the model. To overcome these challenges, future research should focus on improving the interpretability of models and improving the quality and quantity of data. For example, the model can be improved by introducing attention mechanisms and increasing the quality and quantity of data by establishing public databases. Moreover, the performance of the model can be improved by improving the structure.

4.Exploration and innovation

To address the above challenges, several new algorithmic models were developed.

Cho et al [7] adopted a special GNN model, proposing the InteractionNet framework for predicting binding constants between drug and target. The InteractionNet model is an unconventional GNN model. In addition to the covalent bond, non-covalent effect is considered in modeling the drug-target system. Finally, the 20-fold crossover method is verified based on the PDBbind dataset. The root mean square error (root mean square error, RMSE) is 1.321, which is better than the PoteintialNet model (RMSE is 1.343).

deepDTnet, Model [8], this is in deepDR, the design of a new model, the model is optimized in the input and framework, enrich the information contained in the heterogeneous network, added more target related information, such as target-target similarity, target-disease information, while retaining the PCO, matrix and PPMI matrix representation, using multilayer DAE learning the implicit information of heterogeneous network. Compared to deepDR, deepDTnet has more predictive power with ROC-AUC of 0.963. Researchers have also tried to develop new network models by combining AE and other network models. For example, Peng et al. [9] proposed DTI-CNN model, which is characterized by Jaccard similarity coefficient combined with restart random walk algorithm (random walk with restart, RWR) to extract drug features and target features, and added CNN module to predict the final result after DAE layer. After training, ROC-AUC reached 0.9416, which is comparable to deepDTnet.

4.3Huang et al [10] proposed the SkipGNN model and suggested that the 2 directly connected nodes in the heterogeneous network do not necessarily have strong similarity, but that the similarity between indirect or jumping nodes may be more necessary. According to this idea, they to drug-drug, target-target, drug-target, gene-disease related information to build the heterogeneous network, extract the jump similarity information and build jump interaction diagram, at the same time combined with the original graph input to GNN model, finally through the decoder output drug and target interaction probability. The experimental results show that the SkipGNN model is better than other models, such as DeepWalk, graph convolutional neural network (graph convolutional neural network, GCN) and node2vec model.

5.epilogue

Deep learning shows great potential in drug target discovery, but it also faces some challenges. For example, the black-box nature of deep learning makes the interpretability of the results a puzzle. For drug target discovery, it is not sufficient to simply predict the possible targets, but it is more important to understand how these targets interact with the drug and how this interaction affects the physiological processes of the disease. In addition, deep learning models have a high demand for data, and in the field of drug development, high-quality data is often difficult to obtain. This is also an important bottleneck of deep learning in drug target discovery. To address these issues, in the future one could consider combining deep learning with other computational approaches, such as molecular dynamics simulations, to improve the interpretability of the model. At the same time, the problem of data scarcity can be addressed by transfer learning or semi-supervised learning. In addition, using the biological knowledge of drug targets, such as gene expression and protein structure information, can also further improve the performance of deep learning models in drug target discovery. Overall, although deep learning still faces some challenges in drug target discovery, through continuous attempts and optimization of various approaches, it is reasonable to believe that deep learning will play an increasingly important role in future drug target discovery.

Reference

[1] RAMSUNDAR B, EASTMAN P, WALTERS P, et al.Deep learning for the life sciences [M].Sevastopol: O' Reilly Media, 2019.

[2] HUANG K, FU T, GLASS L M, et al.DeepPurpose: a deep learning library for drug-target interaction prediction [J]. Bioinform, 2020, 36: 5545-5547.

[3] KORSHUNOVA M, GINSBURG B, TROPSHA A, et al. OpenChem: a deep learning toolkit for computational chemistry and

drug design [J].J Chem Inf Model, 2021, 61: 7-13.

[4] CONG Q, ANISHCHENKO I, OVCHINNIKOV S, et al.Protein interaction networks revealed by proteome coevolution [J].Science,2019, 365: 185-189.

[5] HASHEMIFAR S, NEYSHABUR B, KHAN A A, et al.Predicting protein-protein interactions through sequence-based deep learning [J].Bioinform, 2018, 34: 802-810.

[6] GAINZA P, SVERRISSON F, MONTI F, et al.Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning [J].Nat Methods, 2020, 17: 184-192.

[7] Cho H, Lee E K, Choi I S.Layer-wise relevance propagation of InteractionNet explains protein–ligand interactions at the atom level[J/ OL].Sci Rep , 2020, 10(1): 21155[2021-10-01]. https://doi/10.1038/s41598-020-78169-6.

[8] Zeng X X, Zhu S Y, Lu W Q, et al .Target identification among known drugs by deep learning from heterogeneous network[J].ChemSci , 2020, 11(7): 1775-1797.

[9] Peng J J, Li J Y, Shang X Q.A learning-based method for drug target interaction prediction based on feature representation learning and deep neural network[J/OL].BMC Bioinformatics , 2020, 21(13): 394[2021-10-01].https://doi/10.1186/s12859-020-03677-1.

[10] Huang K X, Xiao C, Glass L M, et al.SkipGNN: predicting molecular interactions with skip-graph networks[J/OL].Sci Rep , 2020, 10(1): 21092[2021-10-01].https://doi/10.1038/s41598-020-77766-9.