International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 2 ǁ February 2023 ǁ PP. 40-43

A Review of Research on Molecular Property Prediction Methods Based on Deep Learning

Yuanbing Song[1],Jinghua Chen[1],Zhichong Ma[1], Yingdawei Fang[1]
[1]*University of Shanghai for Science and Technology, Shanghai, China*
Corresponding Author:Jinghua Chen

Abstract
At present deep learning is widely applied in various fields because it can accurately show excellent performance. This research review comprehensively analyzes the current status of research on deep learning in molecular property prediction applications from three aspects: supervised learning, semi-supervised learning, and unsupervised learning, as well as compares the advantages and disadvantages of each. The current research on molecular property prediction using supervised learning is the most extensive. However, the research trend has moved toward semi-supervised learning and unsupervised learning.
Keywords:*Deep learning,molecular property prediction.*

Date of Submission: 24-01-2023 Date of acceptance: 07-02-2023

I. INTRODUCTION

Artificial intelligence is a strategic technology that will lead the future. Deep learning is an important part in the field of artificial intelligence. Deep learning has become one of the common techniques to solve problems in many fields [1]. It is also widely used in the field of molecular property prediction to provide experts with reliable prediction results and to assist them in further delving into the intrinsic molecular relationships.

The deep learning approach has significant advantages compared with traditional experiments. Deep learning can predict the properties of thousands of molecules without the involvement of a large number of people. It also saves a lot of consumption of raw materials and prevents the waste of resources. In contrast to machine learning, the deep learning approach is an end-to-end learning that does not require hand-designed molecular descriptors. It extracts molecular features from the data [2]. Thus it also reduces the impact of expertise limitations on the prediction results.

In the following, three deep learning methods, supervised learning, semi-supervised learning, and unsupervised learning, are used as entry points to discuss the current status of research on the application of these methods in molecular property prediction methods, and to analyze their advantages and disadvantages.

II. REVIEW OF RELEVANT LITERATURE

There have been many research efforts using deep learning techniques to accomplish molecular property prediction tasks. In the process of molecular property prediction, deep learning methods can be classified into three categories according to whether labeled molecular data is used to train the model or not: supervised learning, semi-supervised learning and unsupervised learning (as shown in Figure 1).The schematic diagram of their work is shown in Figure 2.

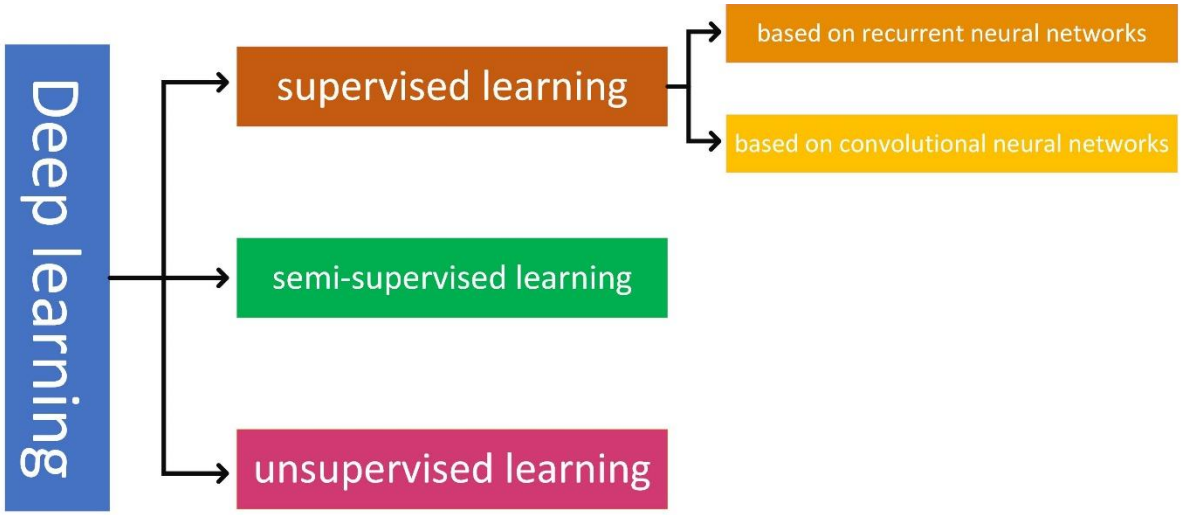

Figure1: Classification of deep learning methods

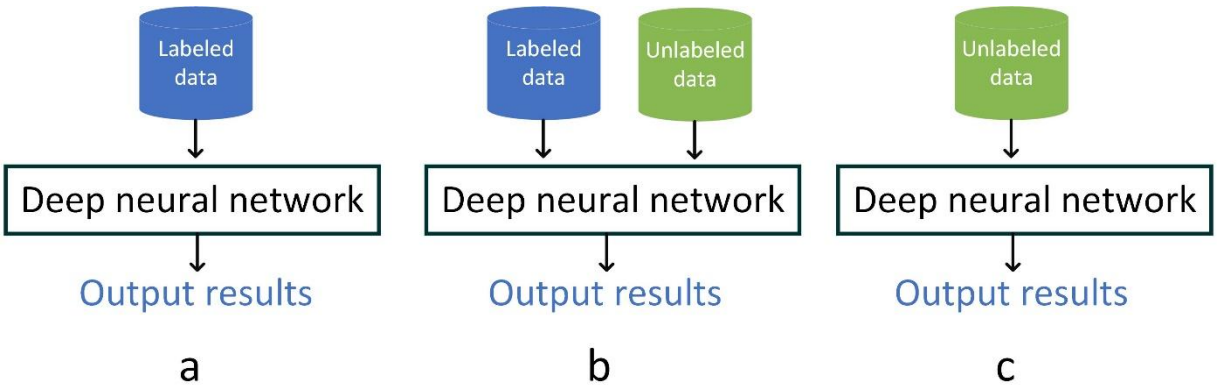

Figure2: Schematic diagram for the work of deep learning methods(a:supervised learning;b: semi-supervised learning;c: unsupervised learning)

2.1 SUPERVIESD LEARNING

Supervised learning is the use of labeled data throughout the process of training a model. Usually, for a particular domain, deep learning methods are studied starting from supervised learning. Therefore, the current research using supervised learning for molecular property prediction is the most popular. In the field of molecular property prediction, supervised learning methods are further subdivided into: supervised learning methods based on recurrent neural networks and supervised learning methods based on convolutional neural networks.

2.1.1 Supervised learning methods based on recurrent neural networks

The hidden layers of recurrent neural networks are cyclic, which results in the value of the hidden layer being influenced by both the current input value and the value of the hidden layer at the previous moment.ALTAE-TRAN H et al [3] proposed a model suitable for low data molecular property prediction using one-time learning, focusing on the identification of new categories, introducing iterative refinement of long and short-term memory, and using it in combination with graph neural networks to improve the learning of small molecules capability, allowing the learning of complex metrics.MANSIMOV E et al [4] developed a deep generative graph neural network (DGGNN) to generate molecular conformations by learning energy functions in the data in an end-to-end manner, generating conformations that are easily observed in experiments. The root-mean-square deviation (RMSD) between the generated conformation and the reference conformation of this model is lower compared to the conventional force field method, and the conformation generated by this model is closer to the reference conformation, and the generated molecular conformations are not similar, maintaining conformational diversity while being computationally fast. However, the best conformation of this method is farther away from the reference conformation than the traditional force field method, so the two methods can be

combined by first using a deep generative graph neural network to filter the generated conformations and later using the traditional force field method to obtain the best conformation.SHINDO H et al [5] proposed an accurate and effective gated graph recurrent neural network (GGRNet) for learning molecular representation and predicting molecular properties. The GGRNetmodels molecules as directed complete graphs, where each atom has a 3D spatial coordinate, updates hidden atomic vectors through inter-atomic distances, shares atomic vector parameters in all layers, and uses input embeddings as jump connections in order to speed up training. Allowing the inclusion of arbitrary features helps to learn molecular representations, and the model has great potential in molecular graph learning. However, there are fewer expression functions in the model and there is a clear deficiency in updating and readout functions.FEINBERG E N et al [6] studied intramolecular interactions and non-covalent interactions between different molecules and proposed the PotentialNet family of models, which are phased gated graph neural networks using a cross-validation strategy for protein-ligand binding affinity. Later, the multi-task PotentialNet model (MT-PotentialNet) was proposed to represent molecules as graphs and learn the most relevant features. Due to the limitations of traditional GNNs in capturing graph features, BOURITSAS G et al [7] designed a topology-aware messaging scheme based on substructure encoding "graph substructure network" (GSN), which exploits structural features extracted by subgraph isomorphism. Preserving the limitations of the standard GNN and the complexity of linear networks, structural information is introduced in the aggregation function to break the local symmetry.But the model has poor expressiveness and generalization ability, and the model cannot directly infer the important substructures from the data.

2.1.2 Supervised learning methods based on convolutional neural networks

The output of the relative position local neurons of the previous layer is used as the input of the later layer, sharing the weight information, and as the network level increases deeper features can be obtained to model more complex linear relationships.XIAOLIN P et al [8] introduced a model for predicting pKa using graph convolutional neural networks, called MolGpKa.It automatically learns the chemical features associated with pKa and uses the learned features to make reliable predictions.HENTABLI H et al [9] modified 2D fingerprint descriptors to propose a new molecular matrix representation Mol2mat and developed a deep convolutional neural network to predict the biological activity of compounds.PENG G et al [10] proposed a fragment-based graph convolutional neural network (F-GCN) for predicting atomic and interatomic properties. It extracts atomic and interatomic features using molecular fragments centered on the target chemical bond. More advanced descriptors are used to further improve the model performance.HUSSEIN H-H et al [11] combined molecular energy and chemical descriptors to propose a 3D convolutional neural network for predicting the absolute binding affinity of protein-ligand complexes. Since the molecular energy can be voxelized, it can be used to improve the predictive power of the neural network. Currently the model has not been extended to molecular dynamics simulations.

2.2 SEMI-SUPERVIESD LEARNING

Semi-supervised learning is the use of both labeled and unlabeled data throughout the training process. In this paper, self-supervised learning is classified as a special case of semi-supervised learning. Self-supervised learning uses unlabeled data for pre-training and then fine-tunes the pre-trained model using labeled data to accomplish molecular property prediction.ZHONGKAI H et al [12] designed an active semi-supervised graph neural network (ASGN) that jointly utilizes information from molecular structure and molecular distribution, considering both graph-level and node-level information. An active learning strategy was used to select new molecules.Li J et al [13] developed the Mol-BERT model. The pre-trained BERT module can extract molecular substructure information using a large amount of unlabeled data, and then molecular properties can be predicted by fine-tuning. Better results were achieved for the classification task, but poorer performance was achieved for the regression task.YUYANG W et al [14] proposed MolCLR model with pre-training using large amount of unlabeled data for comparative learning of molecular representations through graph neural networks. In addition, three molecular graph enhancement strategies were proposed, i.e., atomic masking, bond deletion, and subgraph removal.Li H et al [15] pointed out two major problems at present: ill-defined pretraining tasks and limited model capacity. To address these two major problems, a knowledge-guided graph transformer pretraining (KPGT) model was proposed to represent molecular graphs as line graphs, while introducing path encoding and distance encoding.

2.3 UNSUPERVIESD LEARNING

Unsupervised learning is the use of unlabeled data throughout the training process. There is no research on unsupervised learning in the field of molecular property prediction. Many studies have used unsupervised learning for feature extraction of molecular representations in the pre-training part. For molecular property prediction, the pre-trained model is fine-tuned using labeled data, so for the molecular property prediction task, these studies are semi-supervised learning methods.

III. SUMMARY OF DEEP LEARNING: ADVANTAGES AND DISADVANTAGES

In this section, different deep learning methods are discussed and Table 1 is drawn based on their advantages and disadvantages as follows:

Table 1Advantages and disadvantages of deep learning methods

deep learning methods	Advantages	disadvantages
Supervised learning	1. It has a definite knowledge of the category of the sample data. 2. The model can achieve excellent performance	1. A large amount of labeled data is needed. 2. Label data is difficult and expensive to obtain.
Semi-supervised learning	1. It solves the problem that there is insufficient label data. 2. It can make full use of the large amount of unlabeled data to extract features.	1. The goodness of pre-training can only be evaluated by downstream tasks. 2. It has a high complexity of algorithm.
Unsupervised learning	1.The sample data is sufficient. 2. More complex models can be modeled.	1. There is no monitoring mechanism, and the goodness of the model cannot be judged. 2. The accuracy of the model is low.

IV. CONCLUSION

This paper composes and summarizes the molecular property prediction methods based on deep learning, and analyzes the current research status from three aspects: supervised learning, semi-supervised learning and unsupervised learning. Deep learning has achieved many research results in the field of molecular property prediction, especially in supervised learning. Currently, the focus of research is toward semi-supervised and unsupervised learning approaches. The semi-supervised and unsupervised learning methods can be a good solution to the challenge of insufficient labeled data.

REFERENCES

[1]. DU X, CAI Y, WANG S, et al. Overview of deep learning; proceedings of the 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2016.
[2]. ZHANG W, YANG G, LIN Y, et al. On definition of deep learning; proceedings of the 2018 World automation congress (WAC),2018.
[3]. ALTAE-TRAN H, RAMSUNDAR B, PAPPU A S, et al. Low data drug discovery with one-shot learning [J]. 2017, 3(4): 283-293.
[4]. MANSIMOV E, MAHMOOD O, KANG S, et al. Molecular geometry prediction using a deep generative graph neural network [J]. 2019, 9(1): 1-13.
[5]. SHINDO H, MATSUMOTO Y. Gated Graph Recursive Neural Networks for Molecular Property Prediction [J]. arXiv preprintarXiv:1909.00259, 2019.
[6]. FEINBERG E N, SUR D, WU Z, et al. PotentialNet for Molecular Property Prediction [J]. ACS Central Science, 2018, 4(11): 1520-1530.
[7]. BOURITSAS G, FRASCA F, ZAFEIRIOU S, et al. Improving Graph Neural Network Expressivity via Subgraph Isomorphism Counting [J]. arXiv preprint arXiv.2006.09252, 2021.
[8]. XIAOLIN P, HAO W, CUIYU L, et al. MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network [J]. Journal of Chemical Information and Modeling, 2021, 61(7):3159-3165.
[9]. HENTABLI H, BILLEL B, FAISAL S, et al. Convolutional Neural Network Model Based on 2D Fingerprint for Bioactivity Prediction [J]. International Journal of Molecular Sciences, 2022, 23(21):13230.
[10]. PENG G, JIE Z, JIE Z, et al. General QSPR Protocol for Atomic/Inter-atomic Properties Predictions: Fragments based Graph Convolutional Neural Network (F-GCN) [J]. chemrxiv.14094903, 2021.
[11]. HUSSEIN H-H, CE Z, THOMAS L, et al. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks[J]. Journal of chemical information and modeling, 2020, 60(6): 2791-2802.
[12]. ZHONGKAI H, CHENGQIANG L, ZHENYA H, et al.ASGN: An active semi-supervised graph neural network for molecular property prediction, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 731-752.
[13]. LI J, JIANG X J W C, COMPUTING M. Mol-bert: An effective molecular representation with bert for molecular property prediction [J]. 2021, 2021: 1-7.
[14]. YUYANG W, JIANREN W, ZHONGLIN C, et al. Molecular contrastive learning of representations via graph neural networks [J].Nature Machine Intelligence, 2022, 4(3): 279-287.
[15]. LI H, ZHAO D, ZENG J. KPGT: Knowledge-Guided Pre-training of Graph Transformer for Molecular Property Prediction [J].arXiv preprint arXiv:2206.03364, 2022.