ROLE OF MACHINE LEARNING IN MODERN CHEMISTRY

Tsuber V.Y.
Ukrainian Medical Stomatological Academy, victoriya_tsuber@gmail.com

Machine learning (ML) is an application of artificial intelligence in which systems are able to learn by themselves from experience without being explicitly programmed. Machine learning creates algorithms that learn automatically without human assistance. Such algorithms are able to find new knowledge in massive quantities of data. Deep learning (DL) is a principal method of machine learning. In it, deep neural networks (DNN) are built of multiple layers. The signal that goes forth through the layers enables for extraction of higher-level features from the raw data. DNN have been used in multiple domains including drug design, bioinformatics, analysis of medical images and genetic data, prediction of material properties etc. In these fields, the DNN algorithms fared as well and sometimes even better than human experts [1]. There have been multiple domains where deep learning is applied to solve long-standing problems in chemistry. The arising big chemical databases have been successfully used to decipher data patterns with DL methods in quantum chemistry [2], molecular dynamics simulations [3], identification of new drug candidates [4], protein structure prediction, *in silico* discovery of novel molecules, to name but a few.

Ab initio quantum mechanical calculations have the potential to correctly predict structure and properties of chemical compounds, but their usage is limited by their high computational cost. For complex chemical compounds, ML is successfully used to decrease the high computational cost of *ab initio* methods such as DFT [5]. For instance, the accuracy of predictions of ground state and excited state properties of organic molecules made with DL methods surpassed that of *ab initio* calculations for small molecules [6]. In it, DNN models successfully calculated not only atomization energies of the organic molecules, but also their polarizabilities, ionization potentials and electron affinities [6]. In addition, DNNs can calculate energies from molecular orbitals [7] or represent the wavefunction [8].

Another excellent achievement of application of DL methods to solving chemical problems is their precise prediction of three-dimensional shape of proteins from their amino acid sequences by Google AI branch DeepMind. Proteins are vital to all metabolic functions and perform them due to their unique 3D shapes, as “structure is function” is an axiom of molecular biology. X-ray crystallography and cryo-electron microscopy are used to decipher molecular structures of proteins. However, the structures of about 170 thousand proteins have been investigated over the last sixty years, while there are over 200 million known proteins in total in all living species [9]. The ability to predict a 3D structure of a protein solely from its amino acid sequence has been the holy grail of protein chemistry. A host of diverse computational methods have been tried to solve the problem of protein structure prediction but they generally failed to predict accurately the structure of most proteins except for the smallest ones. Since 1994, researchers participate in the Critical Assessment of Structure Prediction contest (CASP) to compare their methods for protein structure predictions. The accuracy of the predictions is measured in global distance test (GDT), that measures the degree to which a predicted structure is similar to the structure data obtained through the use of experimental methods. By 2016, GDT scores of not more than 40 out of 100 (which is an absolute match) only could be reached for complex proteins [9]. Google DeepMind’s program, called AlphaFold first participated in the 2018 CASP using DL methods. AlphaFold fared better than about a hundred other teams in 2020, with a score of more than 90 in GDT for the majority of the proteins used in the contest [10]. DeepMind’ algorithm has learned on over 170 thousand protein sequences and structures from a huge public database. The algorithm belongs to attention networks, a DL method that breaks a larger problem into its constituent parts, then merges the discrete solutions to obtain the global outcome [10].

The progress of the data-driven discoveries greatly depends on the quality and size of the growing number of public chemical databases. PubChem is one of the largest public repositories [11]. PubChem is a database of chemical species and their activities in biological assays. PubChem consists of three interconnected domains, Substance, Compound and BioAssay. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH). Structure and compound datasets can be downloaded freely from PubChem via FTP. PubChem contains scores of data on small molecules with fewer than one hundred atoms. Another important public database is BindingDB [12]. It contains data on measured protein-ligand affinity associated with specific drug targets. BindingDB contains binding data for millions of small molecules and thousands of proteins. It also contains multiple tools with which the information can be searched, analyzed and integrated. ChEMBL [13] is a public database containing information on binding modes and functions for scores of bioactive compounds. The information is regularly extracted from the published literature, then sorted and organized so as to meet needs of researchers in drug discovery and chemical biology. Currently, the database contains millions of bioactivity measurements for more than one million compounds and thousands of protein targets.

Machine learning methods are set to attain an important place in chemistry. However, machine learning paradigm has its limitations. First and foremost, machine learning methods are known as “black box” methods, which means that performance of algorithms used in ML analyses do not have means of being tracked and explained. Secondly, high demand for computation power makes many ML methods out of reach for smaller institutions. Thirdly, ML algorithms need enormous amount of data in order to produce accurate results and therefore cannot be applied in domains where data procurement is complicated. Fourthly, results obtained in ML analysis of massive databases may reveal hidden patterns in the data but seldom bring about completely new knowledge.

1. Bengio Y., LeCun Y., Hinton G. Deep learning // Nature.–2015.–521(7553).–P.436–444.
2. Dral P. Quantum chemistry in the age of machine learning // J. Phys. Chem. Lett. – 2020. – 11, №6. – 2336–2347.
3. Lin S., Weitao Y. Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks // J. Chem. Theory Comput. – 2018. – 14, №3. – P.1442–1455.
4. Freedman D. Hunting for new drugs with AI // Nature. – 2019. – 576, №7787. – P.49–53.
5. Pereira F., Xiao K., Latino D., Wu C., Zhang Q., Aires-de-Sousa J. Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals // J. Chem. Inf. Model. – 2017. – 57, №1, P.11–21.
6. Montavon G., Rupp M., Gobre V., Vazquez-Mayagoitia A., Hansen K., Tkatchenko A., Müller K.-R., Anatole von Lilienfeld O. Machine learning of molecular electronic properties in chemical compound space // New J. Phys. – 2013. – 15, №9, P.95003.
7. Welborn M., Cheng L., Miller T. F. Transferability in machine learning for electronic structure via the molecular orbital basis // J. Chem. Theory Comput. – 2018. – 14. – P.4772–4779.
8. Sugawara M. Numerical solution of the Schrödinger equation by neural network and genetic algorithm // Comput. Phys. Commun. – 2001. – 140. – P.366–380.
9. Service R.F. “The game has changed.” AI triumphs at solving protein structures // Science. – 2020. – 370, №6521. – P.1144–1145.
10. “DeepMind’s protein-folding AI has solved a 50-year-old grand challenge of biology”. MIT Technology Review. Retrieved 2021-03-03.
11. Kim S., Thiessen P. A., Bolton E. E., Chen J., Fu G., Gindulyte A., Han L., He J., He S., Shoemaker B.A., Wang J., Yu B., Zhang J., Bryant S.H. PubChem Substance and Compound databases // Nucleic Acids Res. – 2016. – 44. – P.1202–1213.
12. Gilson M. K., Liu T., Baitaluk M., Nicola G., Hwang L., Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology // Nucleic Acids Res. – 2016. – 44. – P.1045–1053.
13. Papadatos G., Gaulton A., Hersey A., Overington J.P. Activity, assay and target data curation and quality in the ChEMBL database // Comput. Aided. Mol. Des. – 2015. – 29. – P.885–896.