Artificial Intelligence in the Life Sciences 1 (2021) 100023

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Viewpoint

Current status of active learning for drug discovery

Jie Yu[a,b], Xutong Li[a,b], Mingyue Zheng[a,b,*]

[a] Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[b] University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China

ARTICLE INFO

Keywords:
Drug discovery
Active learning
Uncertainty estimation

ABSTRACT

Active learning has been widely used in drug discovery and design in recent years. In this viewpoint, we will briefly summarize applications of AL for drug discovery and propose two potential limitations of research in this field.

Main text

In most drug discovery projects, the scarcity of data sometimes hinders the application of deep learning models, which rely heavily on the quantity and quality of training data [1]. Moreover, generating training data by human selection requires a lot of resources and even leads to waste. For example, the newly "labeled" (i.e., designed and synthesized) compounds may not provide novel insights into the structure–activity relationship compared to the original training data. Therefore, a rational and objective sample selection method for deciding which samples should be labeled is essential. Active learning [2,3] (AL) has been proposed to play such a role.

AL is a subdiscipline in machine learning (ML) where the algorithm iteratively selects the most useful samples from the unlabeled dataset and queries the expert (or some other information source) to label them, so as to reduce the cost of labeling while improving model performance. To our knowledge, the application of AL algorithm to drug discovery starts with the pioneering work of Warmuth et al [4]. Their work highlighted the fitness of AL to process labeling tasks in pharmaceutical research and development. Recently, the topic has gained momentum, driven by the improved accuracy of ML prediction models [5]. Several promising studies have been reported for different drug development projects in the last two decades such as focused library design [6], rational *de novo* design [7], and drug combination [8]. Given the long-standing practical verification of its effectiveness, AL has shown potential to be an easily deployable technology to assist researchers in their molecular reasoning and experimental design.

In this viewpoint we summarize the promising applications of AL and review briefly some of the limitations to be addressed in the future.

A majority of the previously published studies focused on the ability of AL algorithm to identify desirable samples and enhance ML models through additional training data [9–11]. This ability makes it a well-suited choice for guiding labeling tasks. According to the needs of different applications, the preference of AL can be adjusted flexibly by defining different query strategies (selection functions). For example, some exploration-oriented query strategies quantify the uncertainty of models' predictions, and tend to select samples with novel structures to enlarge the applicability domain of models [10]. On the other hand, exploitation-oriented query strategies aim to select the samples with the highest property, such as compound binding affinities against a certain target, and the models based on such a query strategy often show enhanced hit rates [12].

In addition to the application of prospectively guiding labeling tasks, retrospective applications of AL on data sets with known labels have also been explored by researchers. First, deploying AL on labeled data sets can rapidly compare different model architectures and query strategies to obtain the best AL workflows before applying them in costly prospective studies [10,13]. Secondly, AL can serve as a data filtering tool [10] to remove redundant data that have already been understood by the ML model based on previous training data and thereby cannot offer any further knowledge. Finally, some studies reported that the subsample of training data chosen by AL are always balanced, even when the original dataset is highly imbalanced. It suggests that AL algorithm may also contribute as a data balancing technique [14].

The AL concept has been successfully applied to drug discovery for its promising applications mentioned above. However, some limitations are still present in AL algorithm. Some of the studies have reported that the improved hit rate triggered by the deploying of AL is closely linked

Abbreviations: AL, active learning; ML, machine learning.

* Corresponding author at: Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China.

E-mail address: myzheng@simm.ac.cn (M. Zheng).

https://doi.org/10.1016/j.ailsci.2021.100023
Received 25 November 2021; Accepted 25 November 2021
Available online 6 December 2021

to the performance of the ML model [5]. Moreover, the ability of models to pick balanced training data is tightly-bound to the complexity of the employed model architecture and its predictive capability [14]. Thus, the power of AL is heavily dependent on the model architecture. Recently, Transformer model and its derivatives have led to promising performance for property prediction and molecular representation [15]. However, they require a large number of parameters and expensive training process, which makes them unsuitable for AL that needs constant iteration. As a result, how to combine AL with powerful models like Transformer in a more economical way is a problem to be resolved in the future.

Uncertainty sampling is the most commonly implemented and best understood AL query strategy [5,10], and entropy sampling is one of the most widely used uncertainty-based query strategies due to its model-agnostic nature and ease of implementation. However, entropy sampling only captures the aleatoric uncertainty [16], and neglects the epistemic uncertainty [16]. It means that the entropy sampling, as an uncertainty quantification method, may be over-optimistic in some scenarios [17]. To address this problem, the epistemic uncertainty must be taken into consideration in the process of uncertainty quantification. Currently ensemble-based approaches [18] are still accepted as state of the art for epistemic uncertainty quantification [19]. But they usually require extensive computational costs and runtimes, which poses a major challenge to deploying them in iterative AL procedures. Therefore, to quantify the epistemic uncertainty in a fast, calibrated, and scalable way is also an urgent need. Although a few methods have been proposed to solve this problem, their effectiveness needs to be verified in the future [17].

Conclusion

In this viewpoint, we first introduce the promising and widely used applications of AL in drug design, including guiding experimental design and helping to remove of redundant information. It should be noted that currently there are multiple software applications, such as *Active Learning Glide* and *Active Learning FEP+* available in the Schrödinger Suite, which utilize AL to accelerate the drug discovery process. As claimed, an AL-based protocol successfully recovered more than 80% of the experimentally confirmed hits with a 14-fold reduction in compute cost [20]. It shows the high status of AL in alleviating the intractable computational cost of virtual screening of today's ultra-large chemical libraries. This viewpoint is also devoted to pointing out main limitations of AL algorithm. Because of its iterative nature, AL algorithm is not suitable for complex model architectures and ensemble-based epistemic uncertainty quantification methods, which may limit the applications of AL. Solving these problems is important for the development of AL and needs the joint participation and efforts of researchers of different fields.

Declaration of Competing Interest

Nothing declared.

Acknowledgments

This work was supported by National Natural Science Foundation of China (81773634).

References

[1] Saxe AM, Nelli S, Summerfield C. If deep learning is the answer, what is the question? Nat Rev Neurosci 2021;22(1):55–67.
[2] Smith JS, Ben N, Nicholas L, Olexandr I, Roitberg AE. Less is more: sampling chemical space with active learning. J Chem Phys 2018;148(24):241733–.
[3] Settles B. Active learning literature survey. University of Wisconsinmadison; 2010.
[4] Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C. Active learning with support vector machines in the drug discovery process. J Chem Inf Comput Sci 2003;43(2):667–73.
[5] Reker, D. (2020). Chapter 14: active learning for drug discovery and automated data curation. 2020.
[6] Jansen JM, De Pascale G, Fong S, Lindvall M, Moser HE, Pfister K, Warne B, Wartchow C. Biased complement diversity selection for effective exploration of chemical space in hit-finding campaigns. J Chem Inf Model 2019;59(5):1709–14.
[7] Schneider P, Schneider GJ. De novo design at the edge of chaos. J Med Chem 2016;59(9):4077–86.
[8] Park M, Nassar M, Vikalo H. Bayesian active learning for drug combinations. IEEE Trans Biomed Eng 2013;60(11):3248–55.
[9] Reker D, Schneider P, Schneider G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. Chem Sci 2016:3919–27.
[10] Ding X, Cui R, Yu J, Liu T, Zhu T, Wang D, Chang J, Fan Z, Liu X, Chen K, Jiang H, Li X, Luo X, Zheng M. Active learning for drug design: a case study on the plasma exposure of orally administered drugs. J Med Chem 2021.
[11] Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, Webster LA. Automated design of ligands to polypharmacological profiles. Nature 2012;492(7428):215–20.
[12] Grave KD, Ramon J, Raedt LD. Active learning for high throughput screening. international conference on discovery science. Berlin, Heidelberg: Springer; 2008.
[13] Ahmadi M, Vogt M, Iyer P, Bajorath J, Fr?Hlich H. Predicting potent compounds via model-based global optimization. J Chem Inform Model. 2013;53(3):553–9.
[14] Rakers C, Reker D, Brown JB. Small random forest models for effective chemogenomic active learning. J Comput Aided Chem 2017;18:124–42.
[15] Ying C, Cai T, Luo S, Zheng S, Ke G, He D, Liu TY. ArXiv Preprint; 2021.
[16] Tagasovska, N. & Lopez-Paz, D. (2018). Single-model uncertainties for deep learning. https://arxiv.org/abs/1811.00908
[17] Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model 2020;60(6):2697–717.
[18] Lakshminarayanan, B. Pritzel, A. & Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. https://arxiv.org/abs/1612.01474
[19] Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci 2021;7(8):1356–67.
[20] Yang Y, Yao K, Repasky MP, Leswing K, Abel R, Shoichet BK, Jerome SV. Efficient exploration of chemical space with docking and deep learning. J Chem Theory Comput 2021;17(11):7106–19.