Artificial Intelligence in the Life Sciences 2 (2022) 100051

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Revisiting active learning in drug discovery through open science

Jürgen Bajorath

Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, Bonn D-53115, Germany

Active learning (AL) is a machine learning (ML) approach designed to minimize the amount of training data for developing predictive models [1–3]. The underlying idea is the iterative selection of most informative training instances to gradually refine ML models and improve their predictive performance. For labeled and unlabeled training data, different selection strategies have been introduced [3], attempting to balance exploration (sampling of representative data) and exploitation (focusing on most desired prediction outcomes). There are related approaches at the interfaces between computation and experiment such as iterative biological screening [4]. In this case, ML models are built to prioritize small subsets of screening compound libraries for experimental evaluation and include newly identified hits to re-train the models for the next round of selection. Iterative subset selection and model refinement aim at identifying the majority of available hits while limiting the number of database compounds that are experimentally tested.

Despite the increasing popularity of data-hungry deep learning approaches in drug discovery, high-quality compound activity or *in vivo* property data are often limited. This is a major reason why ML approaches capable of operating in sparsely populated data spaces are attractive. Moreover, data generation might sometimes be rather expensive and time-consuming. In such situations, AL-driven predictions might greatly contribute to, for example, focusing experimental design on the most promising compounds. Hence, although AL has already been considered for about two decades in pharmaceutical research [1], it continues to be a topical –and intellectually stimulating– approach.

In a new contribution to AILSCI, Thompson et al. make use of AL to address an expensive modeling task, that is, the calculation of relative binding free energy (RBFE) of test compounds as a measure of potency alterations [5]. In lead optimization, RBFE computations, which date back to the 1980s [6,7], have become popular for predicting potent compounds as candidates for synthesis [8,9]. Recent progress in RBFE analysis has largely been due to increasing computational power, graphics processing unit (GPU) computing, and advances in conformational sampling procedures [8]. However, although RBFE calculations can now be carried out on a larger scale to guide candidate selection and reduce synthetic efforts [9], the calculations continue to be computationally demanding and represent a substantial cost factor for larger compound libraries, as pointed out by Thompson and colleagues [5]. In other words, RBFE analysis is a time-consuming computational exercise to limit even more expensive experimental efforts. For a given lead optimization series, or multiple series pursued in parallel, each compound would be subjected to RBFE calculations to ultimately select the most likely candidate(s) for further increased potency. As long as the pool of potential candidates remains small, iterative RBFE calculations followed by synthesis might be readily feasible; if the pool becomes large –or if sizeable computationally enumerated libraries are investigated– RBFE calculations become rather challenging.

How can one apply AL to reduce the magnitude of RBFE analysis? From a given compound library, a confined subset must be selected for which RBFE calculations are carried out. On the basis of these data, one then develops an ML model to predict RBFE values from chemical structure. The trained model is used to predict RBFE values for the remaining library compounds and select another subset of promising candidates for RBFE calculations, the results of which are then used to re-build the ML model and further refine the predictions. This process, reminiscent of iterative screening, is continued until a pre-defined number of compounds from subsets has been investigated, aiming to identify the most potent library compounds via ML. It is based on the premise, of course, that ML is computationally much less expensive than RBFE analysis.

This scheme was investigated by Thompson et al. Starting from a computational library of congeneric compounds, the authors were able to detect 75 of the 100 top-ranked RBFE compounds by sampling of only 6% of the library [5]; an impressive result. Three earlier studies (including two preprints) already applied AL in the context of RBFE analysis (using distinct system set-ups), as discussed by the authors. The first of these studies [10] reported a similar success rate in identifying top-scoring compounds on the basis of a comparably small library sample. However, the work by Thompson and colleagues reaches far beyond these earlier studies, and many others in the AL field, for several reasons. For benchmarking the AL approach, the authors required RBFE data. Therefore, they generated a library of 10,000 congeneric compounds and computed RBFE values for all of them. Notably, the authors made the entire library with RBFE data and custom code generated for their analysis publicly available to ensure full reproducibility and enable follow-up investigations; an outstanding contribution to open science. Moreover, going beyond earlier studies, Thompson et al. systematically explored five ML methods and different main parameter settings for AL including the *(i)* selection of the initial subset, *(ii)* number of compounds sampled per iteration (size of each subset), and *(iii)* acquisition function (for selecting training instances). Especially the acquisition function is often thought to play a critical role for AL. The authors found that their AL results were robust and surprisingly insensitive to the use of alter-

E-mail address: bajorath@bit.uni-bonn.de

https://doi.org/10.1016/j.ailsci.2022.100051
Received 2 December 2022; Accepted 2 December 2022
Available online 5 December 2022

native ML methods and different parameter settings (including random selection of the initial data subset and the use of a greedy (exploitation-based) acquisition function). The largest influence on the results was observed for the number of compounds sampled per iteration, with a preference for subsets comprising at least 60 compounds [5]. The observed insensitivity of the AL results to varying parameter settings is particularly interesting. While the insensitivity might partly depend on the characteristics of the compound system under investigation, it also indicates that the choice of acquisition functions is less critical for AL than often assumed and that small training sets can be sufficient for generating well-performing ML models.

The work of Thompson et al. provides new insights into AL and strongly supports open science.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

No data was used for the research described in the article.

References

[1] 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;43:667–73.

[2] Reker D. Practical considerations for active machine learning in drug discovery. Drug Discov Today Technol 2019;32:73–9.

[3] Yu J, Li X, Zheng M. Current status of active learning for drug discovery. Artif Intell Life Sci 2021;1:100023.

[4] Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Discov 2002;1:882–94.

[5] Thompson J, Walters WP, Fenga JA, Pabon NA, Xu H, Goldman BB, Moustakas D, Schmidt M, York F. Optimizing active learning for free energy calculations. Artif Intell Life Sci 2022;2:100050.

[6] Kollman PM. Molecular modeling. Ann Rev Phys Chem 1987;38:303–16.

[7] Beveridge DL, Dicapua FM. Free energy via molecular simulation: applications to chemical and biomolecular systems. Ann Rev Biophys Biophys Chem 1989;18:431–92.

[8] Abel R, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing drug discovery through enhanced free energy calculations. Acc Chem Res 2017;50:1625–32.

[9] Schindler CEM, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, Cappel D, Chekler E, Czodrowski P, Dorsch D, Eguida MKI, Follows B, Fuchs T, Grädler U, Gunera J, Johnson T, Jorand Lebrun C, Karra S, Klein M, Knehans T, Koetzner L, Krier M, Leiendecker M, Leuthner B, Li L, Mochalkin I, Musil D, Neagu C, Rippmann F, Schiemann K, Schulz R, Steinbrecher T, Tanzer EM, Unzue Lopez A, Viacava Follis A, Wegener A, Kuhn D. Large-scale assessment of binding free energy calculations in active drug discovery projects. J Chem Inf Model 2020;60:5457–74.

[10] Konze KD, Bos PH, Dahlgren MK, Leswing K, Tubert-Brohman I, Bortolato A, Robbason B, Abel R, Bhat S. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. J Chem Inf Model 2019;59:3782–93.