Artificial Intelligence in the Life Sciences 3 (2023) 100052

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Editorial

Specific contributions of artificial intelligence to interdisciplinary life science research – exploring and communicating new opportunities

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

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Life sciences
Drug discovery
Interdisciplinary research

Artificial intelligence between hype and reality

Artificial intelligence (AI) is lauded as an auspicious problem solver in many areas. However, the understanding of AI methods is often limited. Hence, an aura of mystery –and also concern– might be generated around AI. For example, the expectation that machines would "think" independently and reach autonomous decisions beyond human reasoning is not factual. In science, the popularity and promise of AI mostly originate from notable advances in a few fields, but are also influenced by business-driven hype and unrealistic expectations.

As *Artificial Intelligence in the Life Sciences (AILSCI)* is completing its second year, this contribution aims to put developments in AI that are particularly relevant for the journal into scientific perspective. It is based upon –and further extends– two recent open access publications addressing a wide life science audience [1,2].

In computer science, various disciplines are covered under the term AI [3]. Among these, deep learning (DL) using deep neural networks (DNNs), a sub-discipline of machine learning (ML), has been responsible for recent progress in areas such as computer vision (image analysis) or natural language processing. These advances have greatly contributed to the popularity of AI in science. Robotics, another AI discipline, is a mainstay in industry and also plays an important role in laboratory automation. Furthermore, expert and recommender systems, which also belong to the AI spectrum, are explored in different scientific fields. Similar to AI-driven developments in physics, theoretical biology, or quantum chemistry, AI is beginning to impact the life sciences including early-phase drug discovery on a larger scale. Here, the term AI is for the most part synonymously used with DL when applied at interfaces between computation and experiment [2,4–6]. In medicine, DL is employed in different therapeutic areas [7] such as radiology or oncology [8,9]. In clinical practice, medical image analysis represents a prime growth area for DL [8,10].

For these developments in the greater life science arena, *AILSCI* represents a well-positioned publication venue. One of *AILSCI's* core values is ensuring high scientific standards of publications, including method development and AI applications.

Characteristics of deep machine learning

Generally, ML uses algorithms for the extraction of feature patterns from training data to classify test objects or address regression tasks. Hence, ML methods are statistical in nature and derive predictive models capturing linear or non-linear instance-feature relationships based on inference from data. DNNs are well suited for feature extraction from large volumes of unstructured data (such as pixels in images) and for learning new object representations. DL relies on systematic correlation of feature patterns and known class labels and derives models with decision functions that are not pre-programmed. Hence, there is nothing mysterious about this type of supervised "machine intelligence".

Shallow NNs were popular during the early stages of ML in biology, chemistry, and drug discovery, but were largely replaced by other approaches such as decision tree methods (random forest, gradient boosting), Bayesian modeling, or support vector machines. This was largely due to a general tendency of shallow NNs to overfit models to training data and their high sensitivity to varying parameter settings. The increasingly popular second-generation DNNs represent highly versatile computational architectures. In computer science, a great variety of DNNs and associated learning strategies have been introduced, some-

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

https://doi.org/10.1016/j.ailsci.2022.100052
Received 7 December 2022; Accepted 9 December 2022
Available online 11 December 2022

times described with terms like a network "jungle" or "zoo" [11]. This architectural variety has pros and cons. For many applications, alternative DNNs can be considered, but finding preferred solutions is not necessarily straightforward. Moreover, complex DNN models are often derived without demonstrating that their complexity is indeed required for the predictions tasks at hand. Compared to other ML approaches, DNNs are particularly rich in hyper-parameters and derivation of DNN models requires substantial knowledge, skills, and experience. Accordingly, although public domain software is available for constructing DNNs, DL is not an approach that is readily accessible to non-experts. There is a strong discrepancy between the mechanics of model building, which might be handled by less experienced users, and the evaluation of results and recognition of potential caveats or model errors, which requires much more expertise. Importantly, similar to other –but not all– ML methods, DNNs have notorious "black box" character [12], meaning that it is not transparent how these models reach their decisions. The black box of DNNs is a major issue in life science and drug discovery applications, as further discussed below.

Data heterogeneity

Life science and drug discovery data are highly heterogeneous in terms of volumes, composition, and complexity. Early-phase drug discovery concentrating on target validation, bioassays, compounds, and activity assessment is not a data-rich discipline compared to other areas where DL has made a strong impact. In early-phase drug discovery, data sets from medicinal chemistry are typically confined to test results for compound series and therefore limited in size. This also applies to data sets from, for example, probe investigations in chemical biology, time series experiments in biology, or confirmatory assays in biological screening. A consequence of data heterogeneity and sparseness is that sufficiently large data sets for "hungry" DNNs are often not available. Moreover, informatics approaches in the life sciences have traditionally employed pre-defined object (for example, target or compound) representations (descriptors) and not relied on representation learning. To further complicate matters, effective ML/DL models can also be generated on the basis of very small data sets [13]. Hence, there are many incentives to undertake expeditions into the DNN jungle, further analyze learning characteristics of different methods, and compare ML/DL models of different complexity.

Artificial intelligence in medicinal chemistry

As an exemplary field with many opportunities for practical applications, one may have a closer look at the state-of-the art of AI in medicinal chemistry, a core discipline of early-phase drug discovery.

ML already has a long history in medicinal chemistry, a traditionally conservative discipline. For more than two decades, ML methods have been used for compound property predictions and other applications. In medicinal chemistry, properties of interest for computational studies include, first and foremost, biological activities of small molecules, but also physiochemical (e.g. solubility) or *in vivo* properties (such as metabolic stability or toxicity). Predictions of such properties aim to support the key task in the practice of medicinal chemistry, that is, deciding which compound(s) to synthesize next. Over time, shallow NNs that were popular early on for property predictions were for the most part replaced by other ML methods, as discussed above. Importantly, in medicinal chemistry, chemical intuition, experience, and subjective decisions continue to play a major role. Accordingly, black box predictions that cannot be explained in chemical terms work against the acceptance of ML for practical applications. However, the popularity of DNNs and high expectations associated with DL are also changing computational medicinal chemistry.

Balancing high hopes

In a publication comparing black box models, explainable, and interpretable ML from a computer science perspective [14], Cynthia Rudin states the following:

> "*When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing.*"

This situation exactly applies to medicinal chemistry. Here, one typically does not work with very large sets of low-resolution data (such a pixels comprising images) for representation learning, a notable strength of DNNs. Instead, in medicinal chemistry projects, available data sets are mostly small and well-defined molecular representations are used (e.g. molecular graph-based descriptors). Such conditions explain observations that are frequently made: At least for property prediction in medicinal chemistry, there typically is little, if any detectable advantage of DNNs over other (simpler) ML methods.

In her article, Cynthia Rudin makes another key point that is equally applicable to medicinal chemistry and drug design:

> *"There is a widespread belief that more complex models are more accurate, meaning that a complicated black box is necessary for top predictive performance. However, this is often not true …"* [14].

One may rephrase this point and emphasize that methodological complexity does not necessarily scale with predictive performance. Clearly, no complex ML approach is "validated" until it is not conclusively demonstrated that simpler methods do not yield comparable results; a critical issue that is often not sufficiently considered in the literature.

New opportunities

While molecular property predictions in medicinal chemistry currently only benefit little from DL, DNNs open the door to novel types of applications, for example, in generative molecular design [15,16] or chemical reaction analysis and modeling [17–19], which would be difficult to tackle using other ML methods. Generative compound design aims to construct chemically novel compounds with desired properties. In these and other areas, recent DL applications illustrate the scientific heterogeneity of the field. For example, despite reports of individual success stories in molecular design [20], caveats remain along the way [21] and consistently successful applications of a given approach on different compound classes are rare. While various compound design strategies can be implemented using DNNs the jury is still out whether or not DL is capable of generating new chemical entities that are of higher quality than others. There are recurrent claims by AI-driven drug discovery enterprises of breakthroughs in generating novel compounds in record times. However, these claims can typically not be assessed scientifically because the data are kept proprietary. In general, scientifically sound reports of practical AI applications with a clear positive impact on medicinal chemistry are still rare [22].

Big data trends

Benefits of DNNs are often associated with learning from "big data". While early-phase drug discovery is not a data-rich discipline, as noted above, big data trends on a relative scale are also observed in medicinal chemistry [23]. Major public repositories of compound data have grown to contain millions of bioactive compounds with activity annotations for thousands of biological targets. In addition, more than 200 million commercial compounds are offered for medicinal chemistry. Of course, for investigators working in particle physics, genomics, or social media, such data volumes might be a far cry from what they perceive

as big data. However, for medicinal chemistry, which is traditionally not data-driven, these data volumes are challenging. On the other hand, data-driven approaches provide new opportunities for the further development of medicinal chemistry as a scientific discipline [24].

The situation is different for predictive modeling. In medicinal chemistry, ML is mostly applied at the level of individual target-directed projects using relatively small data. Each of these projects provides a specific context for modeling. In data science, the context dependence of data structuring and analysis is known to work against generalization of knowledge extraction, which requires abstraction from project-based data sets and project-specific analysis criteria [25]. By contrast, in medicinal chemistry, project focus takes center stage and confines the applicability of ML. Furthermore, the predominant small data framework in medicinal chemistry also suggests alternative strategies for ML. Rather than heavily investigating methodologies whose strengths depend upon large data volumes, approaches such as transfer learning [26] or active learning [27,28] can be applied that are capable of predicting molecular properties or new compounds on the basis of sparse data. Transfer learning makes it possible to use data from related prediction tasks (targets) for modeling; active learning derives predictive models from minimal sets of informative training instances. In medicinal chemistry, these approaches are particularly relevant for addressing novel targets with interesting disease biology for which only limited compound information is available. All in all, there is much room for further computational developments with practical utility for medicinal chemistry.

Model impact and acceptance

Returning to the greater life science arena, there are other areas where DNNs have achieved unprecedented advances such as in de novo protein structure prediction [29]. Regardless, DL will ultimately only become an integral part of interdisciplinary life science research if it measurably impacts experimental programs. Importantly, further establishing DL in interdisciplinary settings is only possible if life science investigators and drug discovery practitioners agree to rely on predictions for experimental design. This requires increasing model acceptance in interdisciplinary research. As is the case with any new technology, time will be required until DL can realize its potential in this area. However, there are specific requirements that must be met to further increase the confidence of experimentalists in predictive models.

Rationalizing predictions

Experimentalists are naturally reluctant to rely on predictions that are difficult or impossible to understand. Given the black box nature of DNNs, this presents a major obstacle for the acceptance of such models for experimental design. Therefore, increasing attention is being paid to approaches for “explainable AI” (XAI) that make it possible to rationalize the results of ML/DL models and interpret predictions in chemical or biological terms [30,31]. Among others, these include methods for the identification of features making largest contributions to individual predictions or the determination of feature sets that are minimally required to produce an accurate prediction. Closely related to XAI approaches are methods to quantify the uncertainty of predictions [32–34]. Obtaining uncertainty estimates also helps to build confidence in predictive modeling. Although there are ML approaches that yield prediction uncertainties, for example, probabilistic (Bayesian) modeling [34], most methods including DNNs produce endpoints without uncertainty estimates, which are subject to further analysis.

Prospective applications

The ultimate assessment of the potential of ML/DL for the life sciences depends on prospective applications, that is, predictions leading to experimental evaluation in real-life projects. In interdisciplinary research settings, prospective applications require confidence in predictive modeling. However, it must also be considered that predictions relying on project data are inevitably influenced by data generation processes and data integrity [35] and that predictions might succeed or fail for different reasons. Accordingly, an incorrect prediction might not necessarily be attributable to methodological failure. Hence, in prospective applications, predictions should be analyzed, explained, and carefully evaluated within a given project context and care should be taken not to prematurely generalize successes or failures.

Conclusion

The rise of AI in many scientific fields comes along with promises and new challenges. AI and DL are often synonymously used, although DL is only a part of the AI spectrum. In the life sciences and drug discovery, many DNN variants are being considered for method development and different applications. However, in interdisciplinary research, DL will ultimately only meet expectations if it positively impacts experimental programs. This will require time, making immediate breakthroughs unlikely, especially in drug discovery. For example, medicinal chemists currently witness methodological advances in synthesis design through DL. However, these new approaches are not easily accessible in the practice of medicinal chemistry. Thus, transforming expert-dependent computational models into robust and readily usable computational tools is one of the grand challenges in the field. More prospective applications will also be essential to further advance DL in interdisciplinary research and demonstrate true impact on high-profile projects. In turn, this will require narrowing the gap between DL and experiments and further increasing the confidence of practitioners in predictions.

For *AILSCI,* method development efforts, practical applications, and perspectives concerning the development of AI approaches at interfaces with the life sciences are equally relevant. High scientific quality is an essential criterion for publication in *AILSCI*, which also strongly encourages data sharing and open science initiatives. Furthermore, diversity of authors and research topics is highly desired and submissions of “off-the-beaten-path” contributions are strongly encouraged, for example, as conceptual analysis, methods, or viewpoint contributions. Notably, in line with its specific aims, *AILSCI* has launched several Themed Article Collections (TACs), which essentially correspond to special issues and focus on, for example, young investigators or women in AI. Going forward, these and other TACs are expected to play a major role for *AILSCI’s* further development.

Declaration of Competing Interest

The author declares that he has 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] Bajorath J. Artificial intelligence in interdisciplinary life science and drug discovery research. Future Sci OA 2022;8:FSO792.
[2] Bajorath J. State-of-the-art of artificial intelligence in medicinal chemistry. Future Sci OA 2021;7:FSO702.
[3] Rapaport WJ. What Is artificial intelligence? J Artif General Intell 2020;11:52–6.
[4] Webb S. Deep learning for biology. Nature 2018;554:555–8.
[5] Leite ML, de Loiola Costa LS, Cunha VA, Kreniski V, de Oliveira Braga Filho M, da Cunha NB, Costa FF. Artificial intelligence and the future of life sciences. Drug Discov Today 2021;26:2515–26.
[6] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23:1241–50.
[7] Wang F, Casalino LP, Khullar D. Deep learning in medicine – promise, progress, and challenges. JAMA Intern Med 2019;179:293–4.

[8] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10.
[9] Farina E, Nabhen JJ, Dacoregio MI, Batalini F, Moraes FY. An overview of artificial intelligence in oncology. Future Sci OA 2022;8:FSO787.
[10] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Ann Rev Biomed Eng 2017;19:221–48.
[11] van Venn, F. The neural network zoo (2016). https://www.asimovinstitute.org/neural-network-zoo/.
[12] Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20–3.
[13] Siemers FM, Feldmann C, Bajorath J. Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity. Cell Rep Phys Sci 2022;3:101113.
[14] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15.
[15] Meyers J, Fabian B, Brown N. De novo molecular design and generative models. Drug Discov Today 2021;26:2707–15.
[16] Tong X, Liu X, Tan X, Li X, Jiang J, Xiong Z, Xu T, Jiang H, Qiao N, Zheng M. Generative models for de novo drug design. J Med Chem 2021;64:14011–27.
[17] De Almeida AF, Moreira R, Rodrigues T. Synthetic organic chemistry driven by artificial intelligence. Nat Rev Chem 2019;3:589–604.
[18] Struble TJ, Alvarez JC, Brown SP, Chytil M, Cisar J, DesJarlais RL, Engkvist O, Frank SA, Greve DR, Griffin DJ, Hou X, Johannes JW, Kreatsoulas C, Lahue B, Mathea M, Mogk G, Nicolaou CA, Palmer AD, Price DJ, Robinson RI, Salentin S, Xing L, Jaakkola T, Green WH, Barzilay R, Coley CW, Jensen KF. Current and future roles of artificial intelligence in medicinal chemistry synthesis. J Med Chem 2020;63:8667–82.
[19] Bort W, Baskin II, Gimadiev T, Mukanov A, Nugmanov R, Sidorov P, Marcou G, Horvath D, Klimchuk O, Madzhidov T, Varnek A. Discovery of novel chemical reactions by deep generative recurrent neural network. Sci Rep 2021;11:1–15.
[20] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, McNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ. A deep learning approach to antibiotic discovery. Cell 2020;180:688–702.
[21] Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol 2020;38:143–5.
[22] Bajorath J, Kearnes S, Walters WP, Meanwell NA, Georg GI, Wang S. Artificial intelligence in drug discovery: into the great wide open. J Med Chem 2020;63:8651–2.
[23] Hu Y, Bajorath J. Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited. Future Sci OA 2017;3:FSO179.
[24] Bajorath J. Foundations of data-driven medicinal chemistry. Future Sci OA 2018;4:FSO320.
[25] Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. Big Data 2013;1:51–9.
[26] Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J. Transfer learning for drug discovery. J Med Chem 2020;63:8683–94.
[27] 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.
[28] Yu J, Li X, Zheng M. Current status of active learning for drug discovery. Artif Intell Life Sci 2021;1:100023.
[29] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[30] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy 2021;23:18.
[31] Rodríguez-Pérez R, Bajorath J. Explainable machine learning for property predictions in compound optimization. J Med Chem 2021;64:17744–52.
[32] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst (NIPS) 2017;30:6402–13.
[33] Hie B, Bryson BD, Berger B. Leveraging uncertainty in machine learning accelerates biological discovery and design. Cell Syst 2020;1:461–77.
[34] Lazic SE, Williams DP. Quantifying sources of uncertainty in drug discovery predictions with probabilistic models. Artif Intell Life Sci 2021;1:100004.
[35] Kearnes S. Pursuing a prospective perspective. Trends Chem 2021;3:77–9.