The machine learning life cycle and the cloud: implications for drug discovery

Ola Spjuth, Jens Frid and Andreas Hellander

ABSTRACT

Introduction: Artificial intelligence (AI) and machine learning (ML) are increasingly used in many aspects of drug discovery. Larger data sizes and methods such as Deep Neural Networks contribute to challenges in data management, the required software stack, and computational infrastructure. There is an increasing need in drug discovery to continuously re-train models and make them available in production environments.

Areas covered: This article describes how cloud computing can aid the ML life cycle in drug discovery. The authors discuss opportunities with containerization and scientific workflows and introduce the concept of MLOps and describe how it can facilitate reproducible and robust ML modeling in drug discovery organizations. They also discuss ML on private, sensitive and regulated data.

Expert opinion: Cloud computing offers a compelling suite of building blocks to sustain the ML life cycle integrated in iterative drug discovery. Containerization and platforms such as Kubernetes together with scientific workflows can enable reproducible and resilient analysis pipelines, and the elasticity and flexibility of cloud infrastructures enables scalable and efficient access to compute resources. Drug discovery commonly involves working with sensitive or private data, and cloud computing and federated learning can contribute toward enabling collaborative drug discovery within and between organizations.

Abbreviations: AI = Artificial Intelligence; DL = Deep Learning; GPU = Graphics Processing Unit; IaaS = Infrastructure as a Service; K8S = Kubernetes; ML = Machine Learning; MLOps = Machine Learning and Operations; Paas = Platform as a Service; QC = Quality Control; SaaS = Software as a Service

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have high potential to revolutionize drug discovery, and is increasingly being used in various aspects of the discovery process [1–4]. Examples of applications using AI/ML include drug screening and repurposing [5,6], structure-based modeling [7], image-based analysis of cells using high-content microscopy [8], drug safety assessment [9,10], and de-novo molecular generation [11]. It is envisioned that AI will increase productivity and reduce attrition in drug discovery [12], a hypothesis that has been strengthened during the COVID-19 pandemic [13–15]. Examples of success stories include the identification of Halicin as an antibacterial molecule from a drug repurposing screen [16], and a phase I clinical study of DSP-1181 developed by Sumitomo Dainippon Pharma and Exscentia for the treatment of obsessive-compulsive disorder [https://www.bbc.com/news/technology-51315462]. Several new companies are emerging where AI is at the core of their business, such as Recursion Pharmaceuticals with several ongoing phase 1 and phase 2 trials [https://www.recursionpharma.com], Benevolent AI [https://www.benevolent.com/], and Insilico Medicine [https://insilico.com/].

Training useful ML models primarily requires access to sufficiently sized dataset of observations with labels; an example would be a large set of compounds run in an assay where each individual experiment constitutes an observation and the label is the result of the assay. It is also important that the data is of high quality. No ML methodology can rescue a dataset with poor labeling [17], such as assays with high bias or high variance in the results. Collecting, harmonizing, storing and preprocessing data can be a very challenging task, even if the data is generated within the same organization. In some cases data might come from experiments having different setups, and sometimes data is distributed, further contributing to challenges [18]. Having a well-planned informatics infrastructure with databases and data management pipelines greatly improves the throughput and reduces the errors when assembling datasets for ML modeling.

Many life science instruments continuously increase their throughput, and new types of experiments emerge, capable of producing detailed depictions of biological processes [19]. This leads to ever-increasing data sizes – both in the number of objects studied but also in the resolution and dimensionality of each observation. Growing data sets enable more accurate AI modeling to be performed, but also contribute to challenges in data storage and the necessary computational infrastructure required for training.
Article highlights

- Integrating machine learning (ML) in drug discovery necessitates managing the entire life cycle of data and models.
- The ML life cycle comprises steps for data collection, preprocessing, model training and validation, model serving as well as predictions and inference.
- MLOps is a set of practices and a category of software that aims to support all stages in the ML modeling lifecycle, and can facilitate reproducible and robust ML modeling in iterative drug discovery.
- Cloud computing provides scalable and resilient access to compute infrastructure that assist in MLOps and the machine learning life cycle.
- Containerization and platforms such as Kubernetes together with scientific workflows can facilitate reproducible and resilient analysis pipelines for continuous ML modeling in drug discovery projects.
- Emerging federated approaches holds great promise to enable multi-organization collaboration on ML. This will require a combination of private and cloud infrastructure.

Recently, Deep Learning, or Deep Neural Networks have emerged as a method that in many cases has been demonstrated capable of producing more accurate predictions in applications related to life science [20-22]. Convolutional Neural Networks have been especially useful for training AI models on images, such as captured by microscopy in cell-based experiments [8,23]. More recently, DeepMind’s AlphaFold was able to very accurately predict a protein’s three-dimensional shape (https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology). Deep Learning requires, apart from the necessary data management infrastructure, substantial computing power. For this reason, accelerators such as GPUs are today more or less a requirement to train even medium-sized AI models using DL methods.

The ML life cycle describes the entire workflow from data to inference and contains many more challenges than the actual model training, including evaluating, deploying and serving models, managing a resilient computational infrastructure, and supporting reproducible data pipelines to enable continuous re-training (see Figure 1). These steps come with different types of risks, and this can be summarized under the software engineering framework of ‘technical debt’ [24]. In drug discovery, this is a traditionally overlooked issue and it is important to address all parts of the machine learning life cycle.

In this manuscript, we discuss how cloud computing can assist and alleviate many of the challenges related to ML modeling when training and serving models aimed at drug discovery, with the required steps making up the ML life cycle. We conclude with discussing working with private and sensitive data in cloud environments, and its potential impact on drug discovery projects.

2. Cloud computing and containerization

Cloud computing refers to delivering and consuming configurable computing resources as networked services on demand [25]. At the heart of cloud computing is virtualization technology where a hypervisor software is used to abstract away, partition and share underlying hardware resources as virtual machines (VMs). A cloud-computing environment pools virtual resources, including networking and storage, and allows the user to rapidly scale resources up and down, with minimal interaction with the service provider. Due to its wide-spread adoption, cloud computing is today used in virtually all aspects of drug discovery, for an overview see [26]. In fact, drug discovery is a constantly changing process, and cloud-based solutions enable an organization or lab to continuously adapt and re-purpose underlying hardware resources to meet evolving needs. Three main service levels can be distinguished: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) with an increasing level of abstraction of underlying hardware (Figure 2). The most easily consumed service is IaaS, relieving organizations of the burden of maintaining an on-premises computational infrastructure, see also D’Agostino et al. for a review of practical and economical aspects for computational drug discovery [27].

Clouds come in four principal deployment models: Public cloud, private cloud, hybrid clouds and community clouds.
Public clouds are provisioned for use by a general public, often in large data centers spread over multiple geographic regions. Private clouds are often deployed on premise for use by a single organization, while community clouds are used by multiple organizations with e.g. similar regulatory concerns. Hybrid cloud refers to the combination of at least two of the other deployment models, often a public and private cloud. Hybrid cloud assumes a software layer that enables seamless application portability between the involved infrastructures [25]. OpenStack [28] is the major open-source software stack for operating cloud environments. Being modular in nature, it has seen a large adoption both in academia [29] and as a basis for smaller, national and local cloud providers.

Major public cloud providers such as Google Cloud Platform, Microsoft Azure, Amazon EC2 and Alibaba cloud, all host extensive and extendable catalogs of virtual appliances – virtual machine images (VMIs) with preinstalled software environments ready to be deployed and used. Such VMIs can greatly simplify distribution of tools in the drug discovery pipeline and help improve reproducibility [30]. While VMIs provide an effective abstraction layer for application portability, there is a significant overhead in emulating a complete operating system. Containerization provides a lightweight alternative that packages everything needed to run an application – the code, its runtime dependencies and settings, as a standard, portable unit of software. Multiple containers run on top of the host system’s operating system kernel, and consume far less resources than VMs. Containerization helps applications run in a standardized way across platforms ranging from public cloud to private cloud to laptops. To automate deployment and management of a large number of such containerized applications in production, container orchestration platforms are used. The currently most widely used solution is Kubernetes (K8s), originally developed and open-sourced by Google. A Kubernetes cluster on top of local hardware is an increasingly common alternative for a private cloud platform, providing scientist easy-to-use and scalable access to CPUs and GPUs. It is also common to run K8s on top of VMs in a cloud, and all major cloud providers offer on-demand container orchestration services based on K8s. Containerization is integral to the microservices architecture pattern, where complex software is broken down into small, loosely coupled services with clearly defined roles. Examples of PaaS based on Kubernetes in life science are the PhenoMeNal Virtual Research Environment [31,32] and OpenRiskNet (www.openrisknet.org), realized as microservices platforms on Kubernetes.

There are a great number of tools available as SaaS for drug discovery, a few examples includes SwissTargetPrediction for prediction of protein targets for small molecules [33], SwissADME for evaluating pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules [34], various APIs and GUIs for prediction of molecular properties such as logD [35]. Larger suites of integrated tools available as SaaS are commonly proprietary and commercial offerings.

Finally, we like to comment on the practical IT knowledge needed to work with cloud services. SaaS services typically require very little practical skills since services tend to be consumed directly via a browser-based UI. However, to work
efficiently with IaaS and PaaS, we recommend basic knowledge on the Linux operating system. It is also very useful to know the basics of Python programming.

3. Machine learning life cycle in drug discovery

The ML life cycle consists of multiple steps with varying infrastructure and software requirements (see Figure 3). In all these steps, cloud IaaS can help reduce the need to buy and maintain a large amount of on-prem compute infrastructure to get started, and for many of the steps higher level services are available as public cloud services. Private cloud and on-premise K8s clusters are good complements in that they let a scientist quickly repurpose the same hardware to meet the varying application needs in the different stages.

3.1. Data collection

Drug discovery comprises data from many different types of experiments, and with an increasing use of AI modeling it becomes crucial to have a data management solution in place that enables rapid access to data so that it can be efficiently used. The data collection step in the AI life cycle can make drug discovery comprise different tasks such as forming experiments to produce data, extracting existing data from databases, or making a selection of data to be used for modeling. Cloud computing here offers scalable capacity and on-demand infrastructure for storage, databases and middleware with no up-front costs. Further, hosting data in cloud enables easy and rapid access to data when modeling is done in cloud environments [36]. This also facilitates collaboration, for example, by giving others access to data volumes. There are many proprietary cloud-based data platforms used in drug discovery, but most are integrated with analytics solutions and are often not flexible enough to be integrated in larger multi-vendor process flows.

3.2. Data preprocessing

When data has been selected and assembled, it is very common to perform a set of data preprocessing steps. This includes dealing with artifacts such as duplicate data, missing values, normalization, augmentation, and quality control. Here, cloud computing offers many opportunities to facilitate the preprocessing, simplifying construction and execution of pipelines or workflows for transparency, reproducibility, and robustness. Many workflow systems enable declarative specifications of analysis pipelines with a built-in capability to execute workflows on public as well as private clouds. Nextflow [37] is a widely used workflow engine in life science research, and can execute workflows on IaaS resources directly but also Kubernetes clusters. Argo (https://argoproj.github.io/argo/) and Pachyderm [38] are two workflow systems that are Kubernetes native. Specifying preprocessing steps in a Notebook deployed on cloud resources is a common way to run preprocessing steps, and has the added advantage of supporting visual interpretations. A complete workflow of all preprocessing steps that can run in the cloud makes the preprocessing reproducible, portable and scalable. In addition, many PaaS services targeted at big data preprocessing and analysis are readily available as public cloud services, including Databricks based on Apache Spark.

3.3. Model training and validation

When the dataset has been constructed the next step is model development, comprising steps such as model training and validation. In most cases this includes an optimization process and hyperparameter tuning and could be a both time-consuming and resource-demanding process. Cloud computing holds many opportunities for this stage in the ML life cycle. A common solution for practitioners, such as data scientists involved in drug discovery, is to invest in on-site powerful workstations equipped with a GPU accelerator, but this solution is neither scalable nor flexible within a team of scientists with a bursty need for compute power, and it also requires considerable up-front investments and maintaining on-site infrastructure over time. Procuring IaaS resources from a cloud provider is a compelling alternative, with access to large resources when needed and a pay-per-minute policy the up-front and maintenance costs for hardware are completely eliminated. However, the running costs for advanced infrastructure on cloud providers, such as GPUs or large VMs with a lot of RAM, should not be underestimated, and teams with sustained large needs for GPU compute increasingly often invest in setting up private K8s infrastructure to support a GPU-powered containerized workflow. One example in drug discovery is Moghadam et al. who scaled AI modeling using VMs procured at public cloud providers [39].

The software stacks used for AI modeling in drug discovery are large and diverse, but some prominent tools include Tensorflow, PyTorch and Scikit-Learn where the two first are focused mainly on Deep Learning. These and many other stacks are readily available in cloud environments, and there are both VMs and container images with all dependencies enabling rapid setups for single-node infrastructure. For scaling over many nodes, Kubernetes is widely used, and specialized frameworks for AI on Kubernetes. In drug discovery, Recursion Pharma uses Kubernetes to scale AI modeling over CPU and GPUs on both Google Kubernetes Engine and on-prem clusters (https://www.recurissionpharma.com/careers?gh_jid=2271704). BenevolentAI has a Kubernetes-based...
3.4. Model serving and inference

An often overlooked issue is to make developed AI models available to end users in a production-grade environment with adequate governance. Such model management and serving requires not only validation and versioning of models but also features such as privacy, access control, auditability, logging and monitoring, and a resilient infrastructure that can recover from failures. This is arguably the most technically challenging step of the ML modeling life cycle, but fortunately there is much to learn from modern software engineering. Cloud computing with its solutions for infrastructure-as-code has enabled new sets of practices for combining development (Dev) and operations (Ops), so-called DevOps. Clouds and Kubernetes have become popular in DevOps teams as they offer a flexible and scalable environment, and also provide many of the features needed for production-grade solutions. Serving ML models have the same prerequisites in terms of e.g. resilience and scalability as many other services, but also more specific challenges related to e.g. monitoring and governance from an AI perspective. Most public clouds offer services for serving and publishing ML models. In addition, several cloud-native open-source frameworks that leverage Kubernetes have been developed, including Tensorflow Serving (https://github.com/tensorflow/serving), Seldon (https://www.seldon.io/), Kubeflow (https://www.kubeflow.org/), and STACKn (https://github.com/scaleoutsystems/stackn).

Making models available over an API is an important step so that models can be consumed over a network by other software components, and it is also a prerequisite for cloud-based inference with predictions as SaaS. Examples in drug discovery include the OpenRiskNet platform for chemical risk assessment (www.openrisknet.org), built on OpenShift (a Kubernetes distribution by RedHat) that allowed for publishing services, such as AI models, and also adding a layer of discoverability and interoperability in annotating APIs using JSON-LD (https://doi.org/10.5281/zenodo.2597061). Another example is DeepCell Kiosk for scaling deployment and inference for microscopy images using TensorFlow, also built on top of Kubernetes [41].

4. Iterative drug discovery and the need for MLOps

Small-molecule drug discovery is an iterative process, and after initial high-throughput screening or other activities for lead identification, the subsequent lead optimization phase commonly follows the Design-Make-Test-Analyze (DMTA) cycle (see Figure 4) [42]. The cycle starts in the Design phase with deciding on the next round of compounds that are to be synthesized in the Make phase. After being produced, the new compounds are in the Test phase evaluated using different types of assays, and the results are analyzed to guide the next round of experiments. ML models can aid the decision-making in the Analyze phase both by models trained on data from the Test phase (commonly for on-target activity) and by models trained on global data collected over multiple drug discovery projects (e.g. for safety endpoints and off-target effects). The project-specific models are usually smaller, and could, for example, be one or more SAR-models developed around one or more scaffolds [43]. These are commonly developed by the scientists themselves within the drug discovery project, but might still need to be consumed by several members in the team. Models should preferably be deployed and made available over a network. The global models are commonly larger, since they are based on data from multiple projects [44], and since they are applied in each cycle and in several projects.
there is a much higher demand for production-grade model serving and accessible APIs. The global data available within a drug discovery organization is however continuously updated, and global models need to be continuously retrained to include the most recent data. While traditional ML modeling and model serving has to a large extent been a manual process in drug discovery organizations, it can create delays in the DMTA cycle and if e.g. global models are not updated it might mean that predictions are made on models that are not trained on all available data. The emerging area of MLOps – Machine Learning and Operations – can here provide both a systematic way of working as well as cloud-based software to support the process.

MLOps describes the joint undertaking by data engineers, data scientists and operations professionals with the intention to cover the entire life cycle of ML modeling in production environments. In drug discovery, MLOps here tries to bridge the disconnect in many organizations between pharmaceutical professionals, AI modelers, and service providers for hosting production-grade ML-models and to enable collaborations. Figure 5 illustrates the continuous deployment nature often associated with MLOps: the aim is to automate the processes from data preparation and experimentation to ML model training and validation to delivery of production-grade models with standard API endpoints, with the purpose of ensuring that up-to-date models become available to the scientist in a timely manner as data is added.

All major cloud providers offer services that cater to the stages of MLOps, and there are many commercial and proprietary systems. Example of open-source frameworks that can be used both in public cloud and on private Kubernetes infrastructure include KubeFlow (https://www.kubeflow.org/), STACKn (https://github.com/scaleoutsystems/stackn), and H2O.ai (www.h2o.ai).

The vast majority of AI models in drug discovery are today trained on data in a batch-wise fashion, where datasets are generated and preprocessed to form a standalone training set that is then progressed to AI modeling. Datasets are either generated within the same organization, or assembled elsewhere and downloaded. A critical component is adequate quality control (QC), and many people currently involved in AI modeling operate on datasets where such QC has already been performed, in many cases without disclosing the details. When AI models are used in decision-making, there is a larger requirement on training on updated data, and it becomes critical to close the loop from data generation to a predictive model in production [42,45]. There are a few activities reported related to using Active Learning in drug discovery, where algorithms decide where the next data point should be generated [46,47]. There are also studies showing the advantages of integrating ML in iterative drug screening [48–50]. In order to harness the full power of such continuous ML modeling, all components including data extraction, preprocessing, QC, training, validation and deployment must be integrated into a reproducible pipeline with adequate versioning.

5. Machine learning using private, sensitive and regulated data
In the machine learning life cycle, data is the key asset and in drug discovery this data is often sensitive, private or regulated. In this situation, transferring data to a cloud outside of the administrative control of the data owner is not an option. A hybrid cloud setup – where some computational resources are located inside the organizations domain and some outside – can help a single organization overcome these challenges internally, by replicating standardized infrastructure on many sites in different countries. Seamlessly managing applications over such hybrid clouds is however not a trivial task.

![Figure 5. MLOps.](image-url)
In general, ing sets databases and sharing knowledge across organizations, employing machine learning, and applying trusted schemes to facilitate data sharing can lead to great promise to enable the joint construction of models across organization and data silos, without the need to ever move or disclose data. Figure 6 illustrates the main idea – multiple organizations can collectively train a ML model by each doing completely local model updates and employing an aggregation scheme in an iterative fashion. There are many possible variants of the scheme employed for aggregation, ranging from centralized versions supported by a trusted third party to completely decentralized trustless variants, each with their own pros and cons [52]. In either case, a practical challenge in all such systems will be to standardize and manage the code needed to be executed on each organization’s own infrastructure. Again, containerization and orchestration frameworks can greatly facilitate that process. In drug discovery, the appealing idea of precompetitive alliances in areas of mutual interest, such as in safety assessment of hits, leads, and candidates would be propelled if federated learning could be realized. The MELLODY consortium (https://www.mellody.eu/) is targeting such analysis over proprietary databases of pharmaceutical companies. There have also been efforts for aggregation of predictions on non-disclosed datasets applied to drug discovery problems [54]. Federated learning has attracted major interest recently, and open source general software frameworks include OpenMined (https://www.openmined.org/), FATE (https://github.com/FEDn) and FEDn [55] (https://github.com/scaleout systems/fedn).

6. **Expert opinion**

AI is establishing itself as a key component in modern drug discovery, and especially Deep Learning approaches have become popular in many different areas. This trend however presents many challenges for drug discovery organizations, both for providing sufficient computational resources and software to its scientists but also for managing and serving production-grade models to end users within the organization. Cloud computing offers many components that assist in the machine learning life cycle, such as easy and scalable access to computational resources and an ecosystem of tools developed to aid data scientists to work effectively with ML modeling. Many of these tools rely on containerization technology and together with platforms such as Kubernetes they can greatly aid and help streamline the ML life-cycle. But apart from technical resources and platforms, there is also a need for scientists and engineers to adopt MLOps best practices. For scenarios where continuous modeling is required, such as when data is updated in regular intervals or when engaged in human-in-the-loop or active learning, the demands on reproducibility and traceability of data and models become even more important to cater for all steps of the ML life cycle. Relying solely on public cloud providers has many advantages in terms of not having to maintain any infrastructure locally, but has implications on costs and data privacy. In most real-world scenarios, drug discovery organizations will need to
combine both local, private cloud infrastructure and public cloud. Distributed ML on such hybrid clouds can be an interesting solution; however, this is currently a technically challenging setup. Federated ML is another upcoming technology that can be used for privacy preserving and collaborative ML, such as in precompetitive alliances, and this has high potential for future drug discovery to unlock hidden data among organizations.

**Funding**

This project was financially supported by FORMAS (grant 2018-00924), Swedish Research Council (grant 2020-03731), the Swedish strategic research programme eSENCE, and the Swedish Foundation for Strategic Research (grant BD15-00085B16-0046).

**Declaration of interest**

O Spjuth, J Frid and A Hellander are co-founders of Scaleout Systems AB, a company involved in developing the open-source STACKn and FEbN software solutions. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

**Reviewer disclosures**

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

**ORCID**

Ola Spjuth [http://orcid.org/0000-0002-8083-2864](http://orcid.org/0000-0002-8083-2864)

Andreas Hellander [http://orcid.org/0000-0001-7273-7923](http://orcid.org/0000-0001-7273-7923)

**References**

Papers of special note have been highlighted as either of interest (∗) or of considerable interest (++) to readers.

1. Yang X, Wang Y, Byrne R, et al. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem Rev. 2019;119:10520–10594.
2. Chan HCS, Shan H, Dahoun T, et al. Advancing Drug Discovery via Artificial Intelligence. Trends Pharmacol Sci. 2019;40:801.
3. Kim H, Kim E, Lee I, et al. Artificial Intelligence in Drug Discovery: a Comprehensive Review of Data-driven and Machine Learning Approaches. Biotechnol Bioprocess Eng. 2020;25:895–930.
4. Nayariseri A, Khandelwal R, Tanwar P, et al. Artificial Intelligence, Big data and Machine Learning approaches in Precis-ion Medicine & Drug Discovery. Curr Drug Targets. 2021;22:631–655.
5. Lim H, Xie L. Machine learning strategies for identifying repurposed drugs for cancer therapy. Drug Repurposing Cancer Ther. 2020;pp 55–79. DOI:10.1016/j.protger.2019.808-12-819668-7.0003-8
6. Gardner S,Das S, Taylor K. AI enabled precision medicine: patient stratification, drug repurposing and combination therapies. Artif Intell Oncol Drug Discovery Dev. 2020. DOI:10.5772/inTechpen.92594
7. Batool M, Ahmad B, Choi S, et al. Discovery Paradigm. Int J Mol Sci. 2019;20:20.
8. Chandrasekaran SN, Ceulemans H, Boyd JD, et al. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov. 2020;20:145–159.
   - Highlighting the increasing importance and needs for AI for Image-based drug discovery.
9. Valerio LG Jr. Predictive computational toxicology to support drug safety assessment. Methods Mol Biol. 2013;930:341–354.
10. Basile AO, Yahi A, Tatonetti NP. Artificial Intelligence for Drug Toxicity and Safety. Trends Pharmacol Sci. 2019;40:624–635.
11. Kotsias P-C, Arús-Pous J, Chen H, et al. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. Nat Mach Intell. 2020;2:pp 254–265.
12. Mak -K-K, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today. 2019;24:773–780.
   - The role of AI in drug discovery and future prospects.
13. Keshavari Arshadi A, Webb J, Salem M, et al. Artificial intelligence for COVID-19 drug discovery and vaccine development. Front Artif Intell. 2020;3:65.
14. Russo G, Reche P, Pennisi M, et al. The combination of artificial intelligence and systems biology for intelligent vaccine design. Expert Opin Drug Discov. 2020;15:1267–1281.
15. Zhou Y, Wang F, Tang J, et al. Artificial intelligence in COVID-19 drug repurposing. Lancet Digit Health. 2020;2:e667–e676.
16. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. Cell. 2020;181:475–483.
17. Gudivada V, Apon A, Ding J. Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. Int J Adv Software. 2017;10:1–20.
18. Ma’ayan A, Rouillard AD, Clark NR, et al. Lean Big Data integration in systems biology and systems pharmacology. Trends Pharmacol Sci. 2014;35:450–460.
19. Biology MV. The big challenges of big data. Nature. 2013;498:255–260.
20. Angermueller C, Pärnrama T, Parts L, et al. Deep learning for computational biology. Mol Syst Biol. 2016;12:878.
21. Webb S. Deep learning for biology. Nature 2018;554:555–557.
22. Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects. Drug Discov Today. 2019;24:2017–2032.
   - The role and future aspects of using Deep Learning in drug discovery.
23. Kersent A, Harrison PJ, Spjuth O. Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes. SLAS Discov. 2019;24:466–475.
24. Scully D, Holt G, Golovin D, et al. Hidden technical debt in machine learning systems. Adv Neural Inf Process Syst. 2015;28:2503–2511.
   - Highlights the importance of a holistic view on putting ML in production.
25. Grance PNMT. NIST Definition of Cloud Computing. Special Publication (NIST SP) - 800-145; 2011.
26. Sobeslav V, Maresova P, Krejcar O, et al. Use of cloud computing in biomedicine. J Biomol Struct Dyn. 2016;34:2688–2697.
27. D’Agostino D, Clematis A, Quarati A, et al. Cloud infrastructures for in silico drug discovery: economic and practical aspects. Biomed Res Int. 2013;2013:138012.
28. Sefraouli O, Aissaoui M, OpenStack EM. Toward an open-source solution for cloud computing. Int J Comput Appl. 2012;55:38–42.
29. Toor S, Lindberg M, Falman I, et al. SNIC Cloud Science (SSC): a National-Scale Cloud Infrastructure for Swedish Academia. 2017 IEEE 13th International Conference on e-Science (e-Science). Auckland, New Zealand. 2017. pp. 219–227.
30. Schaudrangrat N, Lampa S, Simeon S, et al. Towards reproducible computational drug discovery. J Cheminf. 2020. DOI:10.1186/s13321-020-00408-x
31. Peters K, Bradbury J, Bergmann S, et al. PhenoMeNaL: processing and analysis of metabolomics data in the cloud. Gigascience 2019;8:8.
32. Emami Khoonsari P, Moreno P, Bergmann S, et al., Interoperable and scalable data analysis with microservices: applications in metabolomics. Bioinformatics 35(19): 3752–3760. 2019.
   - Introducing and demonstrating microservices in Kubernetes environment as a means for scalable, production grade services.
• The need for workflows to enable reproducible analysis pipelines.

38. Novella JA, Emami Khoonsari P, Herman S, et al. Container-based bioinformatics with Pachyderm. Bioinformatics 2019;35:839–846.

39. Moghadam BT, Alvarsson J, Holm M, et al. Scaling predictive modeling in drug development with cloud computing. J Chem Inf Model. 2015;55:19–25.

40. Lampä S, Dahlö M, Alvarsson J, et al. A workflow library for agile development of complex and dynamic bioinformatics pipelines. Gigascience 2019;8. DOI:10.1093/gigascience/giz044.

41. Bannon D, Moen E, Schwartz M, et al. DeepCell Kiosk: scaling deep learning–enabled cellular image analysis with Kubernetes. Nat Methods. 2021. DOI:10.1038/s41592-020-01023-0.

42. Schneider P, Walters WP, Plowright AT, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19:353–364.

• Discussing the needs for ML in drug discovery.

43. Mok NY, Brown N. Applications of systematic molecular scaffold enumeration to enrich structure–activity relationship information. J Chem Inf Model. 2017;57:27–35.

44. Lambrinidis G, Tsantili-Kakoulidou A. Challenges with multi-objective QSAR in drug discovery. Expert Opin Drug Discov. 2018;13:851–859.

45. Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17:97–113.

• Outlining the needs for ML modeling when increasing the level of automation in drug discovery.

46. Reker D, Schneider G. Active-learning strategies in computer-assisted drug discovery. Drug Discov Today. 2015;20:458–465.

47. Eisenstein M. Active machine learning helps drug hunters tackle biology. Nat Biotechnol. 2020;38:512–514.

48. Ahmed L, Georgiev V, Capuccini M, et al. Efficient iterative virtual screening with apache spark and conformal prediction. J Cheminform. 2018;10:8.

49. Dreiman GHS, Bictash M, Fish PV, et al. Changing the HTS paradigm: AI-driven iterative screening for hit finding. SLAS Discov. 2021;26:257–262.

50. Svensson F, Norinder U, Bender A. Improving screening efficiency through iterative screening using docking and conformal prediction, J Chem Inf Model. 2017;57:439–444.

51. Banegas-Luna AJ, Imbernón B, Llanes Castro A, et al. Advances in distributed computing with modern drug discovery. Expert Opin Drug Discov. 2019;14:9–22.

• Discussing the role of distributed computing infrastructure for drug discovery.

52. Kairouz P, Brendan Mcmahan H, Avent B, et al. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning: Vol. 14: No. 1. 2021. http://dx.doi.org/10.1561/2200000083

53. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10:12598.

• Introducing and explaining the concept of federated learning

54. Spath O, Carlsson L, Gauraha N Aggregating predictions on multiple non-disclosed datasets using conformal prediction. arXiv [stat. ML]. 2018. [cited 2021 Feb 01]. Available: http://arxiv.org/abs/1806.04000

55. Ekmeefjord M, Ait-MLouk A, Alawadi S, et al. Scalable federated machine learning with FEDN, arXiv [cs.LG]. 2021. [cited 2021 Feb 01]. Available: http://arxiv.org/abs/2103.00148