frontiers | Frontiers in Bioinformatics

TYPE Opinion
PUBLISHED 23 February 2023
DOI 10.3389/fbinf.2023.1121591

OPEN ACCESS

EDITED BY
Garrett M. Morris,
University of Oxford, United Kingdom

REVIEWED BY
Sabina Podlewska,
Maj Institute of Pharmacology, Polish Academy of Sciences, Poland

*CORRESPONDENCE
Luiz Miguel Camargo,
✉ luizmiguelc@googlemail.com

†PRESENT ADDRESS
John P. Santa Maria Jr
Computational Drug Discovery,
FogPharma, Cambridge, MA,
United States

SPECIALTY SECTION
This article was submitted to Drug Discovery in Bioinformatics, a section of the journal Frontiers in Bioinformatics

RECEIVED 12 December 2022
ACCEPTED 06 February 2023
PUBLISHED 23 February 2023

CITATION
Santa Maria JP Jr, Wang Y and Camargo LM (2023), Perspective on the challenges and opportunities of accelerating drug discovery with artificial intelligence.
Front. Bioinform. 3:1121591.
doi: 10.3389/fbinf.2023.1121591

Perspective on the challenges and opportunities of accelerating drug discovery with artificial intelligence

John P. Santa Maria Jr[1†], Yuan Wang[1] and Luiz Miguel Camargo[2*]

[1]Data and Translational Sciences, UCB Biosciences Inc., Cambridge, MA, United States, [2]17-09, LLC, Wellesley, MA, United States

KEYWORDS

artificial intelligence, machine learning, drug discovery and development, data, compute

1 Introduction

Technology has long been a driver of innovation and improvement in drug discovery (Gershell and Atkins, 2003; Pina, et al., 2009). Advancements within fields such as chemical engineering, high-throughput experimentation, and molecular biology have transformed the process of drug hunting from serendipitous discovery within historic dye collections to methodical science with empowered understanding of medicines' impacts on human health. Despite improvements in scope and cost-efficiency brought on by adopted technologies, our industry follows "Eroom's Law", an observed exponential decay in FDA new drug approvals per billion dollars of R&D investment (Scannell, et al., 2012), highlighting a key expectation-reward gap between technological investments and payout in therapeutics.

Drug discovery is, in practice, a chain of challenging decisions across diverse disciplines. In considering the journey of a therapeutic candidate from administration to *in vivo* target engagement, a discovery team must address numerous biological and chemical questions. No single technology or dataset has emerged that is powerful enough to tackle all of these questions in aggregate, and we must often drive decisions using surrogate or partial readouts in the absence of ethical and/or practical measurements (Scannell and Bosley, 2016).

Artificial Intelligence (AI), a technology that has gained a lot of publicity and investment, is unique in its promise to impact multiple challenges across the drug discovery pipeline (Lounkine, et al., 2012; Vamathevan, et al., 2019; Bender and Cortés-Ciriano, 2021; Gupta et al., 2021; Paul, et al., 2021; Renaud and Wang, 2021; Kumar et al., 2023). On one hand, when appropriately applied, AI can help us leverage advances in laboratory techniques, generated data, and computational algorithms to make the best decisions we can using the often incomplete information that we have. On the other hand, AI has already had some controversial and expensive failures in the industry, such as Watson AI for automated disease diagnosis and the clinical trial failure of Exscientia's DSP-1181, touted as the first AI-designed drug (O'Leary, 2022; Raleigh, 2022). History teaches us that technology integration happens in the context of the present-day and thus implementation is neither seamless nor immediate. Knowing the strengths and weaknesses of AI can help ensure its correct application and reduce the risks of both over- and underinvestment.

2 Machine learning (ML) and AI in drug discovery

The recent procurement of diverse biological and chemical data generated by 'omics technologies has fueled an AI revolution for biomedical sciences (Biswas and Chakrabarti,

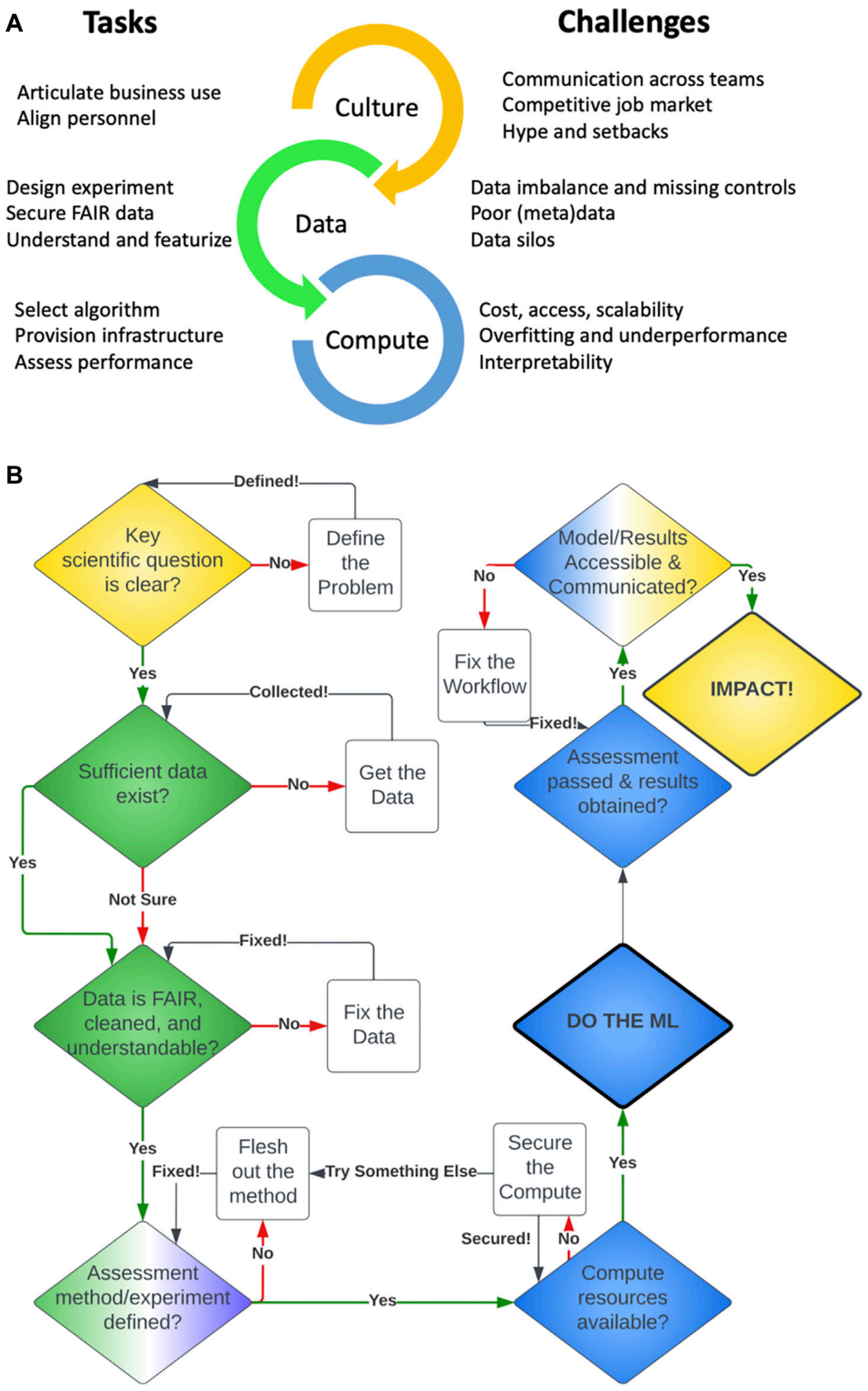

FIGURE 1
(A) Tasks required and challenges encountered in ML applications for drug discovery. **(B)** A decision flowchart for applying ML in drug discovery.

2020). The charge for data-hungry ML is to help us extract value from these experiments: identifying patterns beyond human recognition, distilling large and/or complex datasets, and generating predictions to inform future experiments. The hope is that resulting ML models built on biochemical data capture underlying principles describing molecules and the behaviors of living systems with implications for disease amelioration. Some of these principles have direct translation to the drug discovery process, such as enabling the engineering of antibodies with improved target affinity and reduced immunogenicity (Bachas, et al., 2022). Others have less immediate applicability—for example, functional understanding of AI-predicted RNA spliceoforms (Jaganathan, et al., 2019) toward new targets may be limited by our understanding of the nuanced contexts of disease, and ML-informed protein structures (Jumper, et al., 2021; Lin, et al., 2022) may not guarantee the identification of therapeutic binders, due to limitations such as the synthesizable chemical space of today's screening libraries. Nevertheless, each advance improves our ability to answer key scientific questions and make informed decisions, with industry impact proportional to the target and disease focus areas in which it can be applied.

The ability of AI to address diverse problems and data types is enabled in part by the modularity of ML architectures. In practice, observed success of an AI method for one task may lead to rapid trialing, tailoring, or even direct transfer of encoded modules and learnings for the next. However, domain-specific performance boosts are often achieved using customized scoring functions or connectivities and processing steps adapted to the input data. For example, advanced algorithms like AlphaFold2 and ESMFold illustrate how inputting amino acid sequence alignments together with procedures for facilitating information flow between modules and continuous refinement significantly improved prediction of protein structure (Jumper, et al., 2021; Lin, et al., 2022). Understanding input requirements of data and algorithms can help ensure correct application of these kinds of ML approaches to impact drug discovery.

3 Requirements and challenges for using AI

The existence of challenges in drug discovery does not guarantee AI as an immediate and practical solution. Required ingredients of data, compute, expertise, business utility, and a digital-savvy culture must first be assembled with conscious investment to first ensure readiness and implementability (Figure 1). Deficiencies in any of these elements limit the value we can generate with AI.

As the input for ML algorithms, the shape, quality, scope, and quantity of data matter. Data can be organized in many ways—for example, time series or multidimensional measurements can be stored within one wide or multiple narrow arrays—and conscious data shaping can mitigate misinterpretation and downstream reorganization. Meanwhile, quantity and type of input data can dictate AI method selection. Data such as protein structures or microscopy images often describe multi-dimensional interactions concentrated within localized contexts, and architectures best suited to learn from these data are often different from those that best learn from categorical or single point measurements. Augmenting limited data with new datasets or experiments can increase ML power and scope, though caution is required as diverse sources may introduce unique biases and noise. Alternatively, algorithms can work to mitigate deficiencies in data, such as inferring missing values, or over/under-sampling to improve balance in learning (Chawla, et al., 2002; Hessler and Baringhaus, 2021; Irwin, et al., 2021). Spending time understanding the data can be helpful in assessing when a problem is ready to be tackled with AI or when investment is needed in collecting data that would better inform the task at hand. Failure to assess and address data can decrease the accuracy, confidence, reproducibility, and applicability of downstream ML applications (United States Government Accountability Office, 2019), sometimes propagating even beyond to patient outcomes (Seyyed-Kalantari, et al., 2021).

Regardless of the format of input data, featurization, or the description of entities/samples based on measured or inherent properties, must be carefully performed to best complement the intended ML application. For example, small molecules can be featurized in many ways (David, et al., 2020): using SMILES, text-based descriptors of 2D structures; using atom coordinates describing 3D conformation; using calculated or measured physicochemical descriptors such as LogD and pKa; or even using methods such as extended-connectivity fingerprints (Rogers and Hahn, 2010) that iteratively capture atom connectivities. Choice of featurization affects result interpretability and actionability—ML to inform QSAR decisions for medicinal chemists might favor physicochemical descriptors that highlight relevant property changes for synthetic modifications, while ML for high-throughput screeners might benefit from topological descriptors to learn diverse scaffold hits within an experiment. Evaluating feature similarity is particularly important for generative chemistry endeavors, where AI-proposed molecules should be compared with training data to evaluate novelty (Walters and Murcko, 2020; Wills, 2021).

One often under-addressed component of data quality is metadata. Because biological systems are complex, drug discovery experiments often possess conditional features whose importance is realized only after subsequent measurements. Capturing information such as a cell line, associated genetic engineering, and time of measurements can enable downstream AI-based detection of systematic measurement biases, such as batch effects (Sprang et al., 2022), and meta-analyses across experiments. Metadata can also include interpretations of data—what were the hits in an experiment and what criteria were employed in their selection? This information is historically under-reported and difficult to assemble retrospectively. As a core enabler of findable, accessible, interoperable, and reusable (FAIR) data tenets (Wilkinson, et al., 2016), metadata should also include source information, especially when integrating external data. Because drug discovery data are diverse and acquired across disciplines and sources, (meta) data are rarely uniform. But, extra investment in data assembly and an infrastructure to enable data storage and sharing is often worth it, as failure to capture both data and metadata in the present may impair utility of data assets in future AI applications.

Compute, comprising algorithms and infrastructure, is the second requirement for AI. Compute facilitates both the execution of AI and interpretation of AI outputs. Computational algorithms should be selected with careful consideration of the data available and the problem at hand. Metrics, such as accuracy and area under the receiver-operator curve (AUROC) can facilitate quantitative evaluation of algorithm performance and comparison of new methods to the state of the art. Calculating these metrics requires definition of ground truth, which is supported by upfront investment in experimental design and inclusion of labeled control sample inputs. Published benchmarking datasets and tasks, such as GuacaMol (Brown, et al., 2019) and MoleculeNet (Wu, et al., 2018) for small molecules, and TAPE (Rao, et al., 2019) for protein structures, help standardize and contextualize these assessments, but may not exist for specialized applications.

Some ML methods, especially deep learning, are data and compute intensive, which may limit their implementability within organizations. Nevertheless, sharing centralized computational and data infrastructure across groups can help reduce cost. Computing environments must also be secure to protect proprietary data and sensitive patient data. Dedicated support from IT experts and/or cloud platform CROs can better ensure the timely and secure assembly of data, algorithm, and computational infrastructure on demand; otherwise time spent on this is time scientists spend away from solving the problems at hand. This is a significantly underreported and critical issue, as poor support delays insight delivery, and can decouple ML with expected time frames of decision making in drug discovery projects.

As a final key ingredient, organizational culture defines how AI can be successfully deployed, informing its business use and intended application. In most organizations, the generation of required input data, execution of AI, and validation of its impact can fall within separate groups. This requires machine learners to establish functional relationships with these key stakeholders to understand business needs and engineer AI solutions. Well-maintained relationships ensure continuous vision of actionability and mitigate both the risks of misapplication of AI and mis-/over-interpretation of data and results.

As data consumers, AI/data scientists are often the translators that convert data generated by laboratory scientists into viable inputs for machine learning models. In return, data scientists must ensure models and their results are continuously accessible and shared back with benchtop scientists. The two groups must work together to establish standards and practices for data/model capture and utilization, ensuring today's work informs smarter experiments tomorrow. AI-assisted Design-Make-Test and iterative screening cycles are great illustrations of how regular communication of data and ML results between groups or functions can accelerate hit identification and optimization (Pant, et al., 2018). Data and model siloing, especially in large organizations, impairs this communication and reduces accessibility and utility.

ML models are only useful if their outputs are actionable—in general, that they inform experiments or decision making toward the business impacts of developing a drug. As an example, for AI-generated compound hypotheses this can mean ensuring molecules are synthetically feasible and thus able to be tested. Business understanding also informs performance and time requirements for AI—a model must have higher specificity for hit identification if only a small number of generated compounds can be tested versus a large library. And, training a model for a week is infeasible if outputs are needed on shorter timescales. A breakdown in business understanding can lead to the problems of data scientists building hammers with no nails, or for securing screws.

Establishing a data-aware and AI-supportive culture can sometimes be a roadblock for AI utilization. This can manifest as organizational inertia or politics when AI automation obviates work that was previously performed with an established way of working, or when employees are asked to allot already limited time to new data initiatives. Hype and endorsements from executive leadership can be helpful in motivating diverse teams to support new AI endeavors, but can also lead to an expectation-reward error, as incidences of overpromising and under-delivering sever trust between computational and bench scientists.

To summarize, without good data, compute, people, and culture, including business utility, AI is not possible. Nor is AI the panacea for the absence of one of these components. Tapping into the true and sustainable impact of AI means maintaining these dependencies while tackling challenges that can arise.

4 Discussion—The future of AI in drug discovery

The generation of new quantities and types of biomedical data, together with continuous improvements in computation and lessons learned from ML applications, have driven evolution of the pharma AI landscape. There are key challenges ahead such as bringing ML to new data types and domains and in integrating diverse data to better inform current implementations. Some initial barriers to entry have fallen: democratization of data and algorithms with databases and code repositories have enabled sharing both within an organization and externally with the larger community. Even consortia, such as MELLODDY, facilitate federated learning, preserving confidentiality while pooling data across companies to improve ML models (for some; Blackburn-Owkin, 2022). Similarly, improvement of hardware and a growth in CROs providing infrastructure and workflow solutions make computation more accessible than ever. Nevertheless, the training of revolutionary transformer-based large language models like GPT-3 can cost tens of thousands or even millions in US$, requiring both big data and "big compute" (Sharir et al., 2020). Companies and institutions with the means to generate data at a scale beyond what is available to the public and those who can afford the compute requirements to train and deploy models at scale, will have a significant advantage over others. Though many large models are hosted openly for use, this also raises a reproducibility challenge as only those with access to big data and big compute will be able to validate them, compromising a key tenet of the scientific method.

The future of drug discovery will bring disruptive technologies and data that are unimaginable to us today. While it is possible that future AI will deliver a new generation of medicines, the belief that it will do so independently belies the complexity of the drug discovery process and undervalues the many scientific teams that contribute required inputs. Employing frameworks of cultural and data preparedness can ready us to tap into the data we have today with sustainable, thoughtful application of AI, and improve our probability of success in impacting human health through therapeutics.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

The authors would like to thank friends and family for help in editing the manuscript, especially Mark Kalinich and Massimo de Francesco.

Conflict of interest

JS, YW were employed by UCB Biosciences Inc. LC was employed by 17-09, LLC.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bachas, S., Rakocevic, G., Spencer, D., Sastry, A. V., Haile, R., and Sutton, J. M. (2022). Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv* 2022, 504181. doi:10.1101/2022.08.16.504181

Bender, A., and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Disc Today* 26 (2), 511–524. doi:10.1016/j.drudis.2020.12.009

Biswas, N., and Chakrabarti, S. (2020). Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer. *Front. Oncol.* 10, 588221. doi:10.3389/fonc.2020.588221

Blackburn-Owkin, A. (2022). Melloddy final results. Available at: https://www.melloddy.eu/y3announcement (Accessed October 17, 2022).

Brown, N., Fiscato, M., Segler, M. H. S., and Vaucher, A. C. (2019). GuacaMol: Benchmarking models for *de novo* molecular design. *J. Chem. Inf. Model.* 59 (3), 1096–1108. doi:10.1021/acs.jcim.8b00839

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Art. Intell. Res.* 16, 321–357. doi:10.1613/jair.953

David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: A review and practical guide. *J. Cheminform.* 12, 56. doi:10.1186/s13321-020-00460-5

Gershell, L. J., and Atkins, J. H. (2003). A brief history of novel drug discovery technologies. *Nat. Rev. Drug Disc.* 2, 321–327. doi:10.1038/nrd1064

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. (2021). Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. *Mol. Divers* 2021, 1315–1360. doi:10.1007/s11030-021-10217-3

Hessler, G., and Baringhaus, K. H. (2021). Artificial intelligence in drug design. *Molecules* 23, 2520. doi:10.3390/molecules23102520

Irwin, B. W. J., Whitehead, T. M., Rowland, S., Mahmoud, S. Y., Conduit, G. J., and Segall, M. D. (2021). Deep imputation on large scale drug discovery data. *Appl. AI Lett.* 2, e31. doi:10.1002/ail2.31

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell* 176 (3), 535–548.e24. doi:10.1016/j.cell.2018.12.015

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kumar, M., Nguyen, T. P. N., Kaur, J., Singh, T. G., Soni, D., Singh, R., et al. (2023). Opportunities and challenges in application of artificial intelligence in pharmacology. *Pharmacol. Rep.* 75, 3–18. doi:10.1007/s43440-022-00445-1

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* 2022, 500902. doi:10.1101/2022.07.20.500902

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486 (7403), 361–367. doi:10.1038/nature11159

O'Leary, L. (2022). How IBM's Watson went from the future of health care to sold off for parts. Available at: https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html (Accessed October 11, 2022).

Pant, S. M., Mukonoweshuro, A., Desai, B., Ramjee, M. K., Selway, C. N., Tarver, G. J., et al. (2018). Design, synthesis, and testing of potent, selective hepsin inhibitors via application of an automated closed-loop optimization platform. *J. Med. Chem.* 61 (10), 4335–4347. doi:10.1021/acs.jmedchem.7b01698

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., and Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discov. Today* 26, 80–93. doi:10.1016/j.drudis.2020.10.010

Pina, A. S., Hussain, A., and Roque, A. C. A. (2009). An historical overview of drug discovery ligand-macromolecular interactions in drug discovery: Methods and protocols. *Methods Mol. Biol.* 572, 3–12. doi:10.1007/978-1-60761-244-5_1

Raleigh, N. (2022). The future of AI drug discovery and development in immunology and GPCR research. Available at : https://pharmaphorum.com/digital/the-future-of-ai-drug-discovery-development-in-immunology-and-gpcr-research/(Accessed October 11, 2022).

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., et al. (2019). Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process Syst.* 32, 9689–9701.

Renaud, N., and Wang, Y. (2021). Artificial intelligence as an enabler for phenotypic drug discovery. *Phenotypic Drug Discov.* 77, 104–117. doi:10.1039/9781839160721-00104

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model* 50 (5), 742–754. doi:10.1021/ci100050t

Scannell, J., and Bosley, J. (2016). When quality beats quantity: Decision theory, drug discovery, and the reproducibility crisis. *PLoS ONE* 11 (2), e0147215. doi:10.1371/journal.pone.0147215

Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Disc.* 11, 191–200. doi:10.1038/nrd3681

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* 27, 2176–2182. doi:10.1038/s41591-021-01595-0

Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training NLP models: a concise overview. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/2004.08900 (Accessed February 5, 2023).

Sprang, M., Andrade-Navarro, M. A., and Fontain, J-F. (2022). Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. *BMC Bioinform* 23, 279. doi:10.1186/s12859-022-04775-y

United States Government Accountability Office (2019). *Artificial intelligence in health care: Benefits and challenges of machine learning in drug development.* Washington, DC: United States Government Accountability Office.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Disc* 18, 463–477. doi:10.1038/s41573-019-0024-5

Walters, W. P., and Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* 38, 143–145. doi:10.1038/s41587-020-0418-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Wills, T. (2021). Assessing structural novelty of the first AI-designed drug candidates to go into human clinical trials. Available at: https://www.cas.org/resources/blog/ai-drug-candidates (Accessed October 11, 2022).

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a