Cell Genomics

CellPress
OPEN ACCESS

Perspective

Opportunities for basic, clinical, and bioethics research at the intersection of machine learning and genomics

Shurjo K. Sen,[1,*] Eric D. Green,[1] Carolyn M. Hutter,[1] Mark Craven,[2,3] Trey Ideker,[4] and Valentina Di Francesco[1,*]
[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53792, USA
[3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA
[4]Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
*Correspondence: sensh@mail.nih.gov (S.K.S.), valentina.difrancesco@nih.gov (V.D.F.)
https://doi.org/10.1016/j.xgen.2023.100466

SUMMARY

The data-intensive fields of genomics and machine learning (ML) are in an early stage of convergence. Genomics researchers increasingly seek to harness the power of ML methods to extract knowledge from their data; conversely, ML scientists recognize that genomics offers a wealth of large, complex, and well-annotated datasets that can be used as a substrate for developing biologically relevant algorithms and applications. The National Human Genome Research Institute (NHGRI) inquired with researchers working in these two fields to identify common challenges and receive recommendations to better support genomic research efforts using ML approaches. Those included increasing the amount and variety of training datasets by integrating genomic with multiomics, context-specific (e.g., by cell type), and social determinants of health datasets; reducing the inherent biases of training datasets; prioritizing transparency and interpretability of ML methods; and developing privacy-preserving technologies for research participants' data.

INTRODUCTION

Artificial intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs.[1] Within the broader field of artificial intelligence, machine learning is the study of computer algorithms that improve automatically through experience.[2] Genomics and machine learning have a shared history dating back nearly a quarter century, with the first applications of machine learning methods on DNA sequence data being reported soon after the beginning of the Human Genome Project. Nowadays, genomics is inherently a data-intensive field of research; in fact, since the advent of next-generation DNA-sequencing methods, truly massive volumes of exome, genome, and transcriptome sequencing data have been generated, often with rich and complex metadata annotations. This rich data landscape, which includes not just sequencing data but additional layers of information such as functional genomics and single-cell profiling, provides a natural resource for the use of machine learning to derive biologically and clinically meaningful insights.

Ever since early uses of machine learning in genomics (e.g., for defining protein-coding sequences in Sanger sequencing data) through to the current era of massively parallel DNA sequencing, machine learning has consistently been a versatile tool for annotating genomes and extracting knowledge from raw DNA sequence data.[3] Diverse applications of machine learning in genomics include genome sequence assembly, gene identification, annotation of gene function, genomic variant calling, modeling of sequence evolution, genome-wide association and genotype-phenotype predictions, inferring gene interactions, and many more.[4] This list of applications grows much larger when considering the other omics sciences, such as transcriptomics, proteomics, metabolomics, and metagenomics. Particularly in the last decade, through a combination of accelerated basic research in artificial intelligence coupled with advances in computational hardware, the application of machine learning to biomedical research questions has seen a sharp acceleration. This surge is not unique to genomics and has occurred with many biomedical research fields (e.g., those leveraging imaging data and electronic health records); however, the application of machine learning to each of these data types presents unique challenges that need to be addressed to empower the next phase of biomedical machine learning research. Such challenges include both domain-specific technical hurdles (e.g., developing standards for artificial intelligence "readiness" that are optimized for individual data types) as well as challenges that are common across all fields of biomedicine (e.g., the requirement for transparency and interpretability in machine learning algorithms, defined as the ability for humans to understand and be able to explain, in human terms, the decisions or predictions made by these algorithms).

In this context, the National Human Genome Research Institute (NHGRI), which spearheads genomics research at the US National Institutes of Health,[5] seeks to help define a path forward

for allowing machine learning to be used productively in genomics research. While the examples above represent some ways in which machine learning has already organically contributed to genomics research, the anticipated acceleration of developments and advances motivated the NHGRI to bring members of the genomics and machine learning communities together with bioethics researchers and social scientists to create a roadmap for convergence of these three fields in an ethical, transparent, and equitable manner. This convening role was approached through multiple routes, such as conversations with NHGRI awardees in these areas as part of the institute's strategic planning process (Table 1)[5] and hosting events such as the NHGRI's 2021 "Machine Learning in Genomics: Tools, Resources, Clinical Applications, and Ethics" workshop. Cumulatively, these community engagement exercises served to identify opportunities and obstacles underlying the application of machine learning methods to basic genome sciences and genomic medicine, to define the key scientific areas in genomics that could benefit from machine learning analyses, and to build a map for the NHGRI's unique role in pursuing those efforts. Several challenges facing the convergence of machine learning and genomics research and a set of recommendations were identified for the NHGRI to consider in developing its strategic priorities in this area (briefly described in the following section and summarized in Figure 1).

CENTRAL THEMES FOR MACHINE LEARNING IN GENOMICS

Data, algorithms, and other resources

Machine learning approaches in genomics and other biomedical research fields depend on comprehensive and systematic FAIRness of training datasets (Table 1)[6]; hence, early and easy access to both raw and processed datasets coming out of genomics research networks and consortia should be promoted. As an extension of this theme, more DNA sequencing data need to be generated from across different branches of the evolutionary tree to enable the development of models that use evolutionary and information theory principles. In addition, the development of machine learning approaches that leverage and integrate multiple data types (e.g., population genomics, functional genomics, and single-cell genome-wide imaging data) to generate biological insights is also a high priority. While data exist from hundreds of thousands of genome-wide association study (GWAS) samples, only a few thousand have expression quantitative trait locus (eQTL) data available,[7] and available gene expression data are very limited by the specific biological context from which they were collected (being mostly from adult tissue cell lines). Ideally, a variety of data types (epigenetic, expression, and genome sequencing) derived from different cell types, sample collection modalities, and sampled populations should be accessible to machine-learning methods developers. Training datasets should also be augmented with data derived from statistically designed, model-driven experiments, including perturbation assays.

Also of critical importance is the availability of experimental metadata annotation, with such shared genomics metadata optimized for machine learning approaches (including not only the sample descriptions in a structured standardized format but also quality control parameters). In this context, best practices for robust machine-learning-amenable dataset generation with extensive, standardized metadata should be developed, including ways to annotate perturbation datasets. Current strategies for releasing processed large-scale genomics data are geared toward formats designed for genome browsers or sequence data analysis pipelines rather than as input for machine learning models. This places an additional burden on users who would like to instead apply machine learning-based analytics to the datasets. As an example of good practices in this area, the NIH Common Fund Epigenomics Program (Table 1) was highlighted. Through the combined efforts of federal science administrators and awardees who were part of the program, a strong and consistent focus was maintained throughout the program life cycle on creating data that were made accessible from an early stage (almost 2 years before their initial paper was published) and were accompanied by consistently formatted and contextually deep metadata files derived from a consortium-wide "matrix of experiments."[8,9] Alongside the raw data and quality control metrics, the availability of such metadata allowed the final dataset to become a substrate for development of multiple machine-learning-based epigenome analysis tools.[10,11]

As they stand, the majority of machine learning algorithms used in genomics are typically developed for other research fields and retrospectively optimized for genomics research. An increased emphasis is needed on machine learning models that are built specifically with the challenges facing genomics being kept in mind that are not only predictive but can also be used to infer causality from genomic changes. To facilitate this, testing of functional biology insights derived from machine learning applications deployed on existing observational datasets would provide a greater experimental knowledge base for machine-learning-based causal modeling in genomics. This is particularly relevant, for example, when studying how genetic variation associates with phenotype and gene function, such as to understand the genomic architecture of complex disease phenotypes, and how variants impact gene expression. For complex disease phenotypes, while linear regression models may be used as a baseline, more complex non-linear methods like neural networks and random forests may yield new insights when there is evidence of non-linearity (especially when large training datasets are available). For example, large-scale biobank programs, such as the US-based All of Us Research Program, the UK Biobank, or national networks such as the Australian Genomics-supported Program in Advanced Genomic Investigation (PAGI) (Table 1), are fertile grounds for machine learning in genomic medicine, especially due to the comprehensive, high-quality, and multimodal data that are being collected in such projects. Applying machine learning in such contexts could elucidate the role of genotype-phenotype-environment interactions and the genetic effects shared across traits.

The accuracy and performance of most supervised machine learning models are inherently linked to the availability of suitably large training datasets; however, existing large and machine learning-amenable datasets are few and far between in genomics. Therefore, newer and less data-hungry methods that can still yield rich mechanistic and causal models in genomics are

Table 1. Summary of web resources for machine learning in genomics

Web resource	URL	Brief description
NHGRI strategic planning process	https://www.genome.gov/about-nhgri/strategic-plan/overview	NHGRI-led community engagement process to create a compendium of ideas and opportunities for human genomics research in the coming decade
FAIR data principles	https://www.go-fair.org/fair-principles/	guidelines to improve the findability, accessibility, interoperability, and reusability of digital assets with an emphasis on machine actionability
NIH Common Fund Epigenomics Program	https://commonfund.nih.gov/epigenomics	NIH-funded program to accelerate understanding of how genome-wide chemical modifications to DNA regulate gene activity without altering the DNA sequence itself
All of Us Research Program	https://allofus.nih.gov/	NIH program aiming to collect and study data from one million or more people living in the US to enable individualized disease prevention, treatment, and care
UK Biobank	https://www.ukbiobank.ac.uk/	a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants
Program in Advanced Genomic Investigation (PAGI)	https://www.australiangenomics.org.au/accelerating-precision-medicine-through-machine-learning/	Australian initiative to use machine l earning and artificial intelligence to uncover unprecedented insights from genomic and clinical information
Data Science for Health Discovery and Innovation in Africa (DS-I Africa)	https://dsi-africa.org/	Africa-centered program harnessing data science technologies to develop solutions to that continent's most pressing public health problems
NIH Common Fund Bridge2AI program	https://commonfund.nih.gov/bridge2ai	NIH program setting the stage for widespread adoption of artificial intelligence (AI) that tackles complex biomedical challenges
NIH AIM-AHEAD program	https://datascience.nih.gov/artificial-intelligence/aim-ahead	NIH program to increase the participation and representation of researchers and communities currently underrepresented in the development of AI/machine learning (ML) models
GenoMed4All	https://genomed4all.eu/about/	European Union project aiming to deploy "white box" AI models to diagnose, treat, and predict hematological diseases
Cell Maps for AI (CM4AI)	https://cm4ai.org/	a project seeking to map the spatiotemporal architecture of human cells and use these maps toward the grand challenge of interpretable genotype-phenotype learning
NHGRI Machine Learning in Genomics workshop	https://www.genome.gov/event-calendar/machine-learning-in-genomics-tools-resources-clinical-applications-and-ethics	NHGRI-hosted workshop to stimulate discussion around the opportunities and obstacles underlying the application of ML methods to basic genome sciences and genomic medicine

needed. Methods such as zero-shot learning, where the model is able to learn even in contexts not observed during training, may have applications in genomic medicine.[12] As an example, a model originally trained on cell culture data could be subsequently used on patient-derived xenograft data and eventually guide patient treatment plans. Generative adversarial networks (GANs) and adversarial training generate "realistic" simulated data to train machine learning methods by combining small amounts of biologically observed data with simulated data.[13] This creates a larger training dataset that is expected to have the same characteristics as the original dataset. However, the imbalance between effectively unlimited simulated data versus

Challenges facing the convergence of machine learning and genomics research:	Recommendations:
1. Limited availability of training datasets for machine learning models in genomics	1. Increase data generation across matched omics datasets with extensive, standardized metadata
2. Facilitating early and easy access to datasets coming out of consortium-based studies	2. Promote the creation of machine learning models that can integrate genomics and other omics data from multiple cell types, individuals, and time points
3. Need for context-specific data (e.g., from different cell types and longitudinal experiments)	3. Support functional testing of insights derived from machine learning applications
4. Difficulty in interpretation of outputs from machine learning models and causality inference	4. Develop machine learning methods for genomics that are less data-hungry but can yield rich biological, mechanistic, and causal insights
5. Limited ability of current machine learning models to include longitudinal data from clinical genomics studies	5. Consider and address social, environmental and health disparities of training datasets
6. Lack of diversity in populations currently represented in training datasets	6. Promote collaboration between researchers in bioethics, machine learning and genomics
7. Need for integrating multi-omic and social determinants of health data with genomics	7. Encourage machine learning algorithm development teams working in clinical settings to develop understanding of health disparities in training data
8. Lack of explainability for models applied in genomic medicine, leading to lack of trust	8. Promote transparency and interpretability of machine learning algorithms in genomics
9. Lack of partnerships between clinicians, ethics researchers, computer scientists, and geneticists	9. Develop a workforce of genomic researchers using machine learning approaches, attract machine learning practitioners to genomics, and perform outreach with communities who are currently underrepresented in genomic data
	10. Build and promote models for academic collaboration with industry

Figure 1. Summary of challenges and recommendations from machine learning in genomics

limited observed biological data could result in machine-learning models that more closely represent the idiosyncrasies of the simulator than the actual biology of the studied system. Given that for some studies, such as evolutionary genetics and ancestral population studies or forensic DNA analyses, high-quality biological samples for genomic studies are limited, it is critical to delineate the circumstances in which synthetic training datasets are useful and, when they are, how such data can be generated to be representative of actual biological or population data.

To maximize the availability of suitable training datasets, which may currently be compartmentalized in different data repositories under different institutional data governance and access policies, and national and international data access regulations (e.g., in the Data Science for Health Discovery and Innovation in Africa [DS-I Africa] program) (Table 1), federated data infrastructure for genomics is a crucial need. Specifically, federated data technology enables virtual unification of data from different sources under a uniform data model, while the underlying data stores operate autonomously and without data leaving their original locations. This would allow for a larger number of currently isolated genomics training datasets to become accessible to machine learning models, which could run federated queries as though the data were combined. Together with such federated data infrastructures, privacy-preserving technologies enabling safe and ethical data access should be pursued.[14]

Genomic medicine and ethical, legal, and social implications (ELSI) research

Machine learning models have several applications in the implementation of genomic medicine such as to recommend diagnostic tools and pharmacogenomic therapies based on the patient's genetic makeup.[15] However, before clinical implementation of such models can become a widespread reality, it is critical to address the underrepresentation of many ethnic groups and the social, environmental, and health disparities prevalent in clinical research and healthcare datasets.[16,17] Machine learning algorithms may exacerbate inherent biases in the

training data, leading to biased findings and contravening the fundamental biomedical ethical principle of "do no harm" (e.g., through falsely finding differences in the level of urgent care needed among equally sick patients from different ethnic groups).[18,19] Additionally, clinically underserved communities are unlikely to develop trust in machine-learning-guided genomic-based treatment plans unless health disparities research is incorporated from the start of the model-building process. To engender trust in the use of these approaches and to build a culture of ethical and transparent machine-learning applications in genomic medicine, partnerships should be promoted among the full spectrum of stakeholders. This includes clinical research participants, genomics and ELSI researchers, machine-learning scientists, and advocates for the clinical populations being studied (e.g., rare disease community organizers). As an example of this promoted partnership model, machine-learning model developers working in clinical settings could be required to develop an understanding of health disparities research as a prerequisite for applying their models to patient data. Such partnerships could also alleviate the concern that machine-learning algorithms used in genomic medicine may reduce the role of physicians, which may be attributed to clinicians and machine-learning scientists operating in siloed environments.[20,21]

Concurrently with inclusion of underrepresented minorities, other ELSI considerations include establishing standards and/or guiding principles for explainability, transparency, reproducibility, trustworthiness, and accountability regarding machine learning applications in genomic medicine.[22] At present, ELSI research at the interface of machine learning and genomic medicine reveals a multitude of scenarios in need of further research support or regulatory clarification. For example, when treatment plans for individual patients involve input from machine learning algorithms, how to assign accountability for adverse clinical outcomes among healthcare practitioners, regulatory bodies, and algorithm developers is currently unclear.[23] The development of tools such as feature attribution methods that measure the impact of each feature on a certain prediction should be promoted, as well as algorithmic impact assessment frameworks to promote transparency and accountability.[24,25] Ultimately, the application of machine learning approaches in the genomic medicine and healthcare delivery setting will require partnerships and collaborations with US domestic and international regulatory agencies, such as the US Food and Drug Administration and its worldwide counterparts.

Workforce development

The rapid expansion of genomics and its applications in precision medicine, together with the current surge of machine learning usage in biomedical research, should be put on a sustainable track by adequate investment in multidisciplinary workforce development, ideally targeting college-level students as a future resource in addition to doctoral students and postdoctoral fellows. Current training programs in genomics and machine learning are typically compartmentalized, as trainees usually have either experimental or computational exposure to genomics research, and, conversely, computer science and machine learning students usually have minimal genomics training. To be able to interpret massive and multidimensional datasets, genomics researchers should be introduced to the fundamentals of machine learning early in their career path. Vice versa, genomics offers machine learning practitioners opportunities to solve fundamental questions in biology and medicine, and hence corresponding efforts should be made to increase knowledge of genomics fundamentals among that community.[26] In addition, partnerships are encouraged between academic genomics research institutions and private industries that competitively recruit machine learning scientists with significant remuneration and benefits. A possible route to such synergies could be through establishing non-traditional scientist positions in academic research institutions that offer competitive salaries funded by industry for research projects that are of interest to both the academic and corporate stakeholders.

Conclusions

In this perspective, we identify key opportunities and challenges and set priorities for future activities in support of the adoption of machine learning approaches in genomics research. Not just at the NHGRI but also at NIH in general, promoting the convergence of machine learning and biomedicine is viewed as a high priority. In Figure 1, we have summarized the key challenge areas for this convergence as viewed through a genomics-focused lens and plan to address them in the near future, while also leveraging recent progress made by related activities at the NIH such as the Bridge2AI and AIM-AHEAD programs or the European Union's GenoMed4All project (Table 1). Indeed, the Bridge2AI program, which has been designed from the ground up for artificial and human intelligence to work in hand, has a substantial genomics component through the CM4AI data generation project, which seeks to map the spatiotemporal architecture of human cells and use these maps toward the grand challenge of interpretable genotype-phenotype learning (Table 1). However, at this time, the gap between the state of current datasets and the needs of the field remains massive. While not all ideas and gaps discussed in this perspective may get addressed by future NHGRI- and NIH-supported research programs, through a combination of community input solicited by the NHGRI and detailed analysis of existing NIH funding portfolios in relevant areas, we expect to develop an evidence-based strategy for the NHGRI to support the convergence of machine learning and genomics.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to members of the NHGRI Genomic Data Science Working Group (Michael Boehnke, Eric Boerwinkle, Gail Jarvik, Eimear Kenny, Christina Leslie, Shannon McWeeney, Casey Overby Taylor, and Anthony Philippakis) and three anonymous reviewers for helpful comments on the manuscript. Natalie Kucher provided important assistance with the engagement activities of the scientific community, and Darryl Leja and Julia Fekecs helped with the figures for this manuscript. The authors also wish to thank all the participants of the NHGRI Machine Learning in Genomics Workshop (Table 1).

AUTHOR CONTRIBUTIONS

S.K.S., E.D.G., C.M.H., M.C., T.I., and V.D.F. together conceived of, wrote, and performed analysis for the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. McCarthy, J., Minsky, M.L., Rochester, N., and Shannon, C.E. (2006). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Mag. *31*, 1955.
2. Mitchell, T.M. (1997). Machine learning (McGraw Hill).
3. Salzberg, S. (1995). Locating protein coding regions in human DNA using a decision tree algorithm. J. Comput. Biol. *2*, 473–485.
4. Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. *16*, 321–332.
5. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature *586*, 683–692.
6. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018.
7. Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. *53*, 1290–1299.
8. Satterlee, J.S., Chadwick, L.H., Tyson, F.L., McAllister, K., Beaver, J., Birnbaum, L., Volkow, N.D., Wilder, E.L., Anderson, J.M., and Roy, A.L. (2019). The NIH Common Fund/Roadmap Epigenomics Program: Successes of a comprehensive consortium. Sci. Adv. *5*, eaaw6507.
9. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
10. Pavlovic, M., Ray, P., Pavlovic, K., Kotamarti, A., Chen, M., and Zhang, M.Q. (2017). DIRECTION: a machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. Bioinformatics *33*, 2986–2994.
11. Huang, Y., Sun, X., Jiang, H., Yu, S., Robins, C., Armstrong, M.J., Li, R., Mei, Z., Shi, X., Gerasimov, E.S., et al. (2021). A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer's disease. Nat. Commun. *12*, 4472.
12. Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z., and Wu, Q.M.J. (2023). A review of generalized zero-shot learning methods. IEEE Trans. Pattern Anal. Mach. Intell. *45*, 4051–4070.
13. Viñas, R., Andrés-Terré, H., Liò, P., and Bryson, K. (2021). Adversarial generation of gene expression data. Bioinformatics *38*, 730–737.
14. Sarkar, E., Chielle, E., Gursoy, G., Chen, L., Gerstein, M., and Maniatakos, M. (2023). Privacy-preserving cancer type prediction with homomorphic encryption. Sci. Rep. *13*, 1661.
15. Goecks, J., Jalili, V., Heiser, L.M., and Gray, J.W. (2020). How machine learning will transform biomedicine. Cell *181*, 92–101.
16. Manolio, T.A. (2019). Using the data we have: Improving diversity in genomic research. Am. J. Hum. Genet. *105*, 233–236.
17. Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A., and Green, E.D. (2018). Prioritizing diversity in human genomics research. Nat. Rev. Genet. *19*, 175–185.
18. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science *366*, 447–453.
19. Kidd, C., and Birhane, A. (2023). How AI can distort human beliefs. Science *380*, 1222–1223.
20. Budd, K. (2019). Will artificial intelligence replace doctors? (AAMCNews).
21. Shen, L., Kann, B.H., Taylor, R.A., and Shung, D.L. (2021). The clinician's guide to the machine learning galaxy. Front. Physiol. *12*, 658583.
22. Novakovsky, G., Dexter, N., Libbrecht, M.W., Wasserman, W.W., and Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat. Rev. Genet. *24*, 125–137.
23. Char, D.S., Abràmoff, M.D., and Feudtner, C. (2020). Identifying ethical considerations for machine learning healthcare applications. Am. J. Bioeth. *20*, 7–17.
24. Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability (AI Now Institute).
25. Watson, D.S. (2022). Interpretable machine learning for genomics. Hum. Genet. *141*, 1499–1513.
26. Wilkinson, J., Arnold, K.F., Murray, E.J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., et al. (2020). Time to reality check the promises of machine learning-powered precision medicine. Lancet. Digit. Health *2*, e677–e680.