Rapporteurs: Natalie Kucher and Sean Garin

NHGRI Machine Learning In Genomics Workshop: Tools, Resources, Clinical Applications and Ethics

Executive Summary

OVERVIEW

The NHGRI Machine Learning in Genomics Workshop was hosted by the NHGRI Genomic Data Science Working Group of Council to stimulate discussion around the opportunities and obstacles underlying the application of machine learning (ML) methods to basic genome sciences and genomic medicine, to define the key scientific topic areas in genomics that could benefit from ML analyses, and to identify NHGRI's unique role at the convergence of genomic and ML research.

The workshop was held virtually on April 13 and 14, 2021. Over 3,800 participants registered for the workshop from 73 countries, with over 1,000 attending on each day of the workshop. The workshop comprised two keynote talks and four scientific sessions highlighting cutting edge genomics applications of machine learning methods from 14 speakers. Members of the Genomic Data Science Working Group moderated the workshop sessions and created a set of recommendations for NHGRI from the workshop discussions. Recordings of the workshop are available for viewing on the genome.gov webpage.

Day 1 of the workshop started with a set of keynote talks. These talks highlighted how machine learning applied to the field of genomics is experiencing a boom and enabling analysis and prediction over many datatypes. Challenges in the field include generating explainable findings, enabling affordable training of deep learning models, increasing population representation in training datasets, and navigating privacy considerations of data sharing.

The sessions over Day 1 and Day 2 featured speakers described their work in the field, ranging from basic science and algorithms development to Ethical, Legal and Social Implications (ELSI), to clinical applications of machine learning. The speakers employed a broad range of methods, such as those adapted from image and language processing, specifically convolutional neural networks, graph neural networks, and transformers. The speakers provided many examples of how large, NIH-funded datasets have contributed to machine learning science (i.e., GTEx, Roadmap, ENCODE). Multiple speakers shared common problems, such as moving from a genetic variant to a robust prediction of its function; deep learning and associated technologies can integrate associated datatypes to address these problems.

In particular, some common themes that arose repeatedly across the two-day workshop are as follows. First, the constraints of limited training data can be ameliorated with meta-learning or transfer learning, or by leveraging data from one domain to build systems that can be transferred to domains with fewer data. Another means of dealing with limited data is to employ generative models and synthetic data, while remaining cognizant of the limitations of such approaches. Second, explainability and interpretability of machine learning methods are critical to engender trust in clinical applications. Such approaches for interpretation include causal-inference methods for understanding mechanistic relationships and evaluation of performance metrics for capturing the relevant insights. Related to this issue of trust, ethical, legal, and social issues must be front and center for machine learning models and datasets, as many populations are underrepresented in genomics datasets. Third, technical solutions are needed to enable use of genomic data while preserving privacy.

Fourth, speakers highlighted common characteristics of data structure and organization that underlie a successful ML modeling effort: 1) data must be FAIR and readily available to researchers, 2) there is high value in access to both raw and processed datasets, as well as annotations and metadata for use in future projects, and 3) transparency about the possible weakness from data generators and/or model creators can be extremely useful.

WORKSHOP STRUCTURE AND RECOMMENDATIONS

Day 1 of the workshop included a set of Keynote speakers and two sessions: Algorithm development and machine learning approaches in genomics; and Ethical, Legal and Social implications of machine learning in genomics. Day 2 of the workshop included two sessions: Data and resource needs for machine learning in genomics; and Machine learning in clinical genomics. Details of the agenda, including speakers and session moderators are included in Appendix 1.

The NHGRI Genomic Data Science Working Group of Council (see Appendix 2) provided the following recommendations for NHGRI based on their discussions of the scientific topics, opportunities, and barriers discussed in each session of the workshop.

RECOMMENDATION THEME 1: ALGORITHM DEVELOPMENT AND MACHINE LEARNING APPROACHES IN GENOMICS

The history and details of the field of algorithm development and recent cutting-edge innovations in the machine learning space were heavily discussed. The speakers delved into a variety of genomics applications, including protein structure and function prediction, generation and use of simulated datasets for population genetics, and modeling gene regulation with 3D interaction data. Although most recommendations were derived from the first topical session, other algorithm-related discussions arose in the data and resource needs and clincial genomics sessions.

Box 1. Algorithm Development and Machine Learning Approaches Recommendations:

- Increase emphasis on ML models that are not only predictive, but also can be used to infer causality and be interpretable.
- Spur development of machine-learning approaches that connect various types of data sets (e.g., population-based vs. functional genomics).
- Support research to address small sample size and diversity (e.g., less data hungry methods for ML with rich mechanistic and causal models).
- Support developing and applying methods to evaluate algorithms for unintended consequences (e.g., amplifying biases).
- Support research into existence and causes of algorithmic bias in machine learning approaches for genomics
- Delineate the research circumstances in which simulated datasets are useful and, if so, how such data can be generated in such a way as to be representative of actual data and not misleading. (e.g., ancestral population studies).
    - Generative Adversarial Networks (GANs) and adversarial training for creating “realistic” simulated data to train ML methods.

RECOMMENDATION THEME 2: ETHICAL, LEGAL AND SOCIAL IMPLICATIONS (ELSI) OF MACHINE LEARNING IN GENOMICS

Prospective ethical considerations are necessary for machine learning application to genomics to be equitable and safe. Discussions explored the limits of machine learning as a general tool due to the complexities and health disparities apparent in healthcare systems, the need to adopt technological solutions for preserving privacy, and how engaging healthcare providers and patients in clinical implementation of machine learning can help build trust. Recommendations reflect the need for a regulatory body to provide a framework for standards, the challenge of solving bias-related hurdles when using machine learning and artificial intelligence in genomics, and the need for teams with ethical machine learning in genomics at the foundation.

Box 2. ELSI Recommendations:

- Support partnerships with ELSI researchers, AI/ML tool developers, and genomic researchers to build a culture of ethical ML for both basic and clinical genomics.
- Set minimum requirements to define “explainability”, transparency, interpretability, reproducibility, and trustworthiness, and support efforts to develop machine learning models to reach these standards.
- Promote awareness of potential for and mitigation of models exacerbating biases.
- Consider implications for making clinical decisions based on data collected for ML/AI based methods.

RECOMMENDATION THEME 3: DATA AND RESOURCE NEEDS FOR MACHINE LEARNING IN GENOMICS

The third theme focuses on key opportunities for enabling machine learning in genomics, including making large datasets of diverse datatypes available for future analysis as well as decentralized and scalable compute resources. The workshop also highlighted how challenges of interpretability, reproducibility, and transparency of machine learning models for their accurate use remain to be addressed. Recommendations in this field can, and should, leverage existing efforts in machine learning and genomics. These recommendations include the support for sequencing of a diversity of genetic ancestries, the generation of genomics datasets with rich metadata, and the incorporation of best practices for data and metadata standardization.

Box 3. Data Generation and Resource Recommendations:

- Support more sequencing across the evolutionary tree for ML models that use evolution and information theory.
- Support data generation to address underrepresentation of different genetic ancestries in genomics datasets.
- Enable creative ways to augment observational data sets with rationally selected, model-driven experiments, including perturbations.
- Establish and support best practices for robust dataset generation with extensive, standardized metadata, including perturbation datasets.
- Facilitate early and easy access to both raw and processed data sets.

RECOMMENDATION THEME 4: MACHINE LEARNING IN CLINICAL GENOMICS

Some opportunities for clinical genomics included new and explainable AI techniques for cancer precision medicine, use of machine learning to understand the genetic architecture of complex phenotypes, and opportunities to predict clinically useful pharmacogenomics phenotypes for novel variations in important genes. Recommendations focused on supporting knowledgebanks and associated analyses, building infrascturure for supporting portable models, and leveraging genomics and clincial data for machine learning applications.

Box 4. Clinical Application Recommendations:

- Support work involving biobank-scale analyses (e.g., contribution of non-linear effects, contribution of environmental interactions, genetic effects shared across traits).
- Provide infrastructure to allow for training and deploying portable models that use data from many locations.
- Support knowledge resources, including the use of ML in knowledge resources to use predictions to fill gaps.
- Support research leveraging genomics and longitudinal clinical data in ML.
- Provide guidance on FDA approval considerations for ML models.

RECOMMENDATION THEME 5: TRAINING AND OUTREACH FOR MACHINE LEARNING IN GENOMICS

A theme that spanned topics throughout the workshop was growth in the field of machine learning in genomics. As the next generations of scientists will be increasingly familiar with enormous and complex datasets, genomics is a field of great potential for the incorporation of new technologies. Therefore, there should be extensive investment in training and outreach now, to set a foundation for future innovations in machine learning. Recommendations reflect the opportunity for increased community outreach in translation genomics, privacy, and machine learning, as well as cross-training future researchers in genomics, informatics, and ethical practices.

Box 5. Training and Outreach Recommendations:

- Engage the community via workshops on translational genomics, genomics and privacy, training the next generation of genomic data scientists (e.g. train the trainers), and other topics.
- Train researchers to develop comprehensive expertise in genomics, machine learning, and ELSI.
- Facilitate education on best practices for ML in genomics: e.g., comparison with baseline linear models; clarity on model explainability vs interpretability vs accuracy.

RECOMMENDATION THEME 6: COLLABORATION WITH INDUSTRY

A final theme that was discussed at the workshop was the need for academic collaboration with industry. Not only is this need in data sharing practices and standards, but also in ethics and transparency when using genomic data in machine learning. With all parties following strict ethical guidelines in data use, both collaboration and public trust will flourish.

Box 7. Partner with Industry Recommendations:

- Consider ways to attract the technology industry to genomics (e.g., supporting senior staff scientist positions).
- Promote transition and cross-fertilization of ML scientists between industry and academia.
- Engage industry in generation and analysis of large datasets needed to promote the advancement of biomedical AI/ML.

Appendix 1

DAY 1 AGENDA
April 13, 2021

11:00 a.m. – Welcome
Co-chairs: Trey Ideker, Ph.D. and Mark Craven, Ph.D.
Speaker: Eric Green, M.D., Ph.D., Director, National Human Genome Research Institute

KEYNOTE SESSION: WHAT ARE THE OPPORTUNITIES AND CHALLENGES FOR ML IN GENOMICS RESEARCH?
Moderator: Shannon McWeeney, Ph.D.
11:10 a.m. – Eric Topol, M.D., Scripps Research
Genomics in the Machine Learning Space
11:40 a.m. – Brad Malin, Ph.D., Vanderbilt University Medical Center
Challenges and Opportunities for Machine Learning in Genomics
12:10 p.m. – Q&A Session

12:40 p.m. – Break

SESSION 1: ALGORITHM DEVELOPMENT AND MACHINE LEARNING APPROACHES IN GENOMICS
Moderators: Trey Ideker, Ph.D. and Anthony Philippakis, M.D., Ph.D.
1:00 p.m. – Jian Peng, Ph.D., University of Illinois at Urbana-Champaign
Machine learning algorithms for structural and functional genomics
1:25 p.m. – Sara Mathieson, Ph.D., Haverford College
Automatic evolutionary inference using Generative Adversarial Networks
1:50 p.m. – Christina Leslie Ph.D., Memorial Sloan-Kettering Cancer Center
The 3D genome and predictive gene regulatory models
2:15 p.m. – Q&A Session

2:45 p.m. – Break

SESSION 2: ETHICAL, LEGAL AND SOCIAL IMPLICATIONS (ELSI) OF MACHINE LEARNING IN GENOMICS
Moderators: Dave Kaufman, Ph.D. and Eimear Kenny, Ph.D.
3:10 p.m. – Pamela Sankar, Ph.D., University of Pennsylvania
Machine learning: broadening the scope of ethical questions
3:35 p.m. – Varoon Mathur, AI Now Institute
Considerations for building ethical and socially responsible AI systems in Health Care
4:00 p.m. – Danton Char, M.D., Stanford University
Identifying and Anticipating Ethical Challenges with Machine Learning for Genomics
4:25 p.m. – Q&A Session

4:55 p.m. – Day 1 Wrap-up
5:00 p.m. – Adjourn

DAY 2 AGENDA
April 14, 2021

11:00 a.m. – Day 2 Opening

SESSION 3: DATA AND RESOURCE NEEDS FOR MACHINE LEARNING IN GENOMICS
Moderators: **Christina Leslie, Ph.D.** and **Mark Craven, Ph.D.**
11:10 a.m. – Alexis Battle, Ph.D., Johns Hopkins University
Integrative machine learning for regulatory genomics
11:35 a.m. – Anshul Kundaje, Ph.D., Stanford University
Machine learning for genomic discovery
12:00 p.m. – Gregory Cooper, M.D., Ph.D., University of Pittsburgh
Personalized Causal Machine Learning Using Genomic Data
12:25 p.m. – Q&A Session

12:55 p.m. – Break

SESSION 4: MACHINE LEARNING IN CLINICAL GENOMICS
Moderators: **Casey Overby Taylor, Ph.D.** and **Eric Boerwinkle, Ph.D.**
2:00 p.m. – Su-In Lee, Ph.D., University of Washington
Explainable AI for cancer precision medicine
2:25 p.m. – Sriram Sankararaman, Ph.D., University of California Los Angeles
Machine Learning for large-scale genomics
2:50 p.m. – Russ Altman, M.D., Ph.D., Stanford University
Deep learning to predict the impact of rare variation in drug metabolism genes
3:15 p.m. – Q&A Session

3:45 p.m. – Day 2 Wrap-up

4:00 p.m. – Adjourn

Appendix 2

NHGRI Genomic Data Science Working Group

Michael Boehnke, Ph.D., University of Michigan
Eric Boerwinkle, Ph.D., UTHealth School of Public Health and Baylor College of Medicine
Mark Craven, Ph.D., University of Wisconsin-Madison
Trey Ideker, Ph.D., UC San Diego
Gail Jarvik, M.D, Ph.D., University of Washington
Eimear Kenny, Ph.D., Icahn School of Medicine at Mount Sinai
Christina Leslie, Ph.D., Sloan Kettering Institute
Shannon McWeeney, Ph.D., Oregon Health & Sciences University
Casey Overby Taylor, Ph.D., Johns Hopkins University

Anthony Philippakis, M.D., Ph.D., The Broad Institute

NHGRI Organizing Committee

Eric Green, M.D., Ph.D.
Carolyn Hutter, Ph.D.
Valentina Di Francesco, M.S.
Shurjo Sen, Ph.D.
Kris Wetterstrand, Ph.D.
Natalie Kucher, B.S.
Sean Garin, B.S.