HIGH-PERFORMANCE COMPUTING MEETS HIGH-PERFORMANCE MEDICINE

Anurag Verma
Department of Medicine, Division of Translational Medicine and Human Genetics, University of Pennsylvania, Philadelphia, Pennsylvania , USA
Email: anurag.verma@pennmedicine.upenn.edu

Jennifer Huffman
VA Boston Healthcare System, Boston, Massachusetts, USA
*Email:*Jennifer.Huffman2@va.gov

Ali Torkamani
Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA
Email: atorkama@scripps.edu

Ravi Madduri
Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois, USA
*Email:*madduri@anl.gov

1. Introduction, Background, and Motivation

Artificial intelligence (AI) is making a big impact on patient experiences, clinician workflows, researchers, and the pharmaceutical industry work in the healthcare sector. In recent decades, technological advancements across scientific and medical disciplines have led to a torrent of diverse, large-scale biomedical datasets such as health, imaging data, clinical notes, lab test results, and other 'omics data. The dropping costs of genomic sequencing coupled with advances in computing allow unprecedented opportunities to understand the effects of genetics on human disease etiologies and has resulted in the creation of population-level biobanks like the Million Veteran Program[1], UKBioBank[2], PennBioBank[3]. As a consequence, the demand for novel computational methods, computational infrastructure, and algorithm improvements to efficiently process and derive insights from these datasets, particularly where it applies to clinical translational research, has dramatically increased. In addition to handling the sheer size and quantity of biomedical data, newly developed methods must also adapt and employ state-of-the-art AI algorithms that account for the unique complexities of biomedical datasets, such as sparseness, incompleteness, and noisiness of data, data multidimensionality such as clinical measurements from electronic health records, prescription drug data, environmental exposures. Additionally, these methods have to leverage the advances in high-performance computing like GPUs, faster inter-connects, and fast-access memory to help generate the needed insights at a faster rate.

The recent explosion of high-throughput experimental techniques for generating biological 'omics datasets (e.g., genomic, transcriptomic, or metabolomic) has led to a specific set of challenges related to the integration of biomedical with multi-omics data and second to the analysis of these integrated datasets. To begin to model complex phenotypic traits, modern statistical and machine learning methods must now draw from various datasets with diverse origins, such as from analogous data across multiple model organisms or from complementary data within the same species. It leads to challenges stemming from integrating biomedical and multi-omics data, including challenges related to the identification, visualization, and reproducibility of patterns elucidated from integrated datasets.

Data-intensive computing has firmly established itself as the fourth paradigm in scientific discovery. Advances in computing have propelled discovery in many physical sciences (cosmology, high energy physics, aerospace, to name a few). The data-intensive nature of computational problems in medicine and biomedical informatics warrants the use and development of advanced computing infrastructure and software methods. In recent years, advances in computational infrastructure, methods, and algorithms enabled storage and analysis of large-scale datasets (e.g., Exascale Computing Project, Cloud Computing, ESNet)[4]. These advances have created silos of excellence, and scientific discovery propelled by computation has been driven by computationally well-endowed groups. Though distributed computing in the cloud can dramatically improve the performance of complex computational analyses by reducing runtime and local storage requirements, it is still severely limited by the availability of cloud-compatible software packages. Gaps also exist for these packages to leverage supercomputing capabilities.

To address this, we invited experts leading the development and application of artificial intelligence and cutting-edge computing approaches to drive innovation in precision medicine. We discussed current breakthroughs in which our speakers are involved and the strengths and limitations of artificial intelligence in medicine. Our workshop session focused on four major domains of AI and computing 1) AI in Healthcare 2) Genomics in medicine 4) Exascale computing to advance precision medicine.

2. Workshop Presenters

The three-hour workshop will begin with an overview presentation of the workshop followed by four presentations. The workshop will conclude with a panel discussion session, which will be moderated by Drs. Torkamani and Verma.

2.1. Workshop Speakers

2.1.1. Rick Stevens, PhD - Rick Stevens is the Associate Laboratory Director of the Computing, Environment and Life Sciences Directorate at Argonne National Laboratory, and a Professor of Computer Science at the University of Chicago, with significant responsibility in delivering on the U.S. national initiative for Exascale computing and developing the DOE initiative in Artificial Intelligence (AI) for Science. At Argonne, Rick leads the Laboratory's AI for Science initiative and currently focusing on high-performance computing systems which includes leading a significant collaboration with Intel and Cray to launch Argonne's first exascale computer, Aurora 21, which will pursue some of the farthest-reaching science and engineering breakthroughs ever achieved with supercomputing, as well as a partnership with Cerebras Systems to bring hardware on site to advance the massive deep learning experiments being pursued at Argonne for basic and applied science and medicine with supercompute-scale AI. Stevens' research spans the computational and computer sciences from high-performance computing, to the

building of innovative tools and techniques for biological science and infectious disease research as well approaches to advance deep learning to accelerate cancer research. He also specializes in high-performance computing, collaborative visualization technology, and grid computing. Currently, he is the PI of the Bacterial / Viral Bioinformatics Resource Center (BV-BRC) which is developing comparative analysis tools for infectious disease research and serves a large user community; the Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer project through the Exascale Computing Project (ECP), which focuses on building a scalable deep neural network application called the CANcer Distributed Learning Environment (CANDLE); the Predictive Modeling for Pre-Clinical Screening (Pilot 1) of the DOE-NCI Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) project; and the Co-design of Advanced Artificial Intelligence (AI) Systems project focused on predicting behavior of complex systems using multimodal datasets. Rick has won numerous awards for his work, including two R&D 100 Awards and an HPCwire Readers' Choice Award. Rick was elected a Fellow of the American Association for the Advancement of Science (AAAS) in 2003 and since then is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in IEEE Computer Society, an ACM Fellow and a member of the Association for Automated Reasoning and the Association for Symbolic Logic

2.1.2. **Marylyn Ritchie, PhD** - Dr. Ritchie is a Professor of Genetics and Director of the Institute for Biomedical Informatics at the University of Pennsylvania School of Medicine. She is also Associate Director of the Penn Center for Precision Medicine, Director of the Center for Translational Bioinformatics, and Co-Director of the Penn Medicine BioBank. Dr. Ritchie is an expert in translational bioinformatics, with a focus on developing, applying, and disseminating algorithms, methods, and tools integrating electronic health records (EHR) with genomics. Dr. Ritchie has over 20 years of experience in translational bioinformatics and has authored over 375 publications. Dr. Ritchie was appointed a Fellow of the American College of Medical Informatics (ACMI) in 2020. Dr. Ritchie was elected as a member of the National Academy of Medicine in 2021; she is being recognized “for paradigm-changing research demonstrating the utility of electronic health records for identifying clinical diseases or phenotypes that can be integrated with genomic data from biobanks for genomic medicine discovery and implementation science.” Dr. Ritchie holds a Ph.D. from Vanderbilt University in Statistical Genetics, an M.S. from Vanderbilt University in Applied Statistics, and a B.S. in Biology from the University of Pittsburgh at Johnstown. Dr. Ritchie is also the host of two podcasts: she co-hosts The Biomedical Informatics Roundtable podcast with Dr. Jason Moore and the solo host of The CALM Podcast: Combining Academia and Life with Marylyn.

2.1.3. **Ravi Madduri -** Ravi is a computer scientist in the Data Science and Learning division at Argonne National Laboratory and is Senior Scientist at the Center of Research Computing at the University of Chicago. He is an innovation fellow at the Polsky Center of Entrepreneurship at University of Chicago. Ravi led several successful large projects in NSF, NIH and DOE. His research interests are in building sustainable, scalable services for science, reproducible research, large-scale data management and analysis. He co-leads the MVP-CHAMPION project, which is a collaboration between VA and DOE and developed methods to perform large-

scale genetic data analysis using DOE's high performance computing capabilities, including methods for generating PRS scores in Prostate Cancer, genome-wide PheWAS on Summit supercomputer. Additionally, Ravi is one of three key contributors to the National Institutes of Health $100M Cancer Biomedical Informatics Grid (caBIG), which linked 60 NIH-funded cancer centers and clinical sites engaged in cancer research. For his efforts in project management, tool development, and collaboration, Ravi received several Outstanding Achievement Awards from NIH. Ravi led the design and implementation of scientific and high-performance workflows under the caGrid toolkit. Ravi leads the Globus Genomics project (www.globusgenomics.org), which is used by thousands of researchers across the world for genomics, proteomics, and other biomedical computations on Amazon cloud and other platforms. He architected the Globus Galaxies platform that underpins Globus Genomics and several other cloud-based gateways realizing the vision of Science as a Service for creating, maintaining sustainable services for science. Ravi plays an important role in applying large-scale data analysis, deep learning to problems in biology. For his work on "Cancer Moonshot" project, he received the Department of Energy Secretary award in 2017.

2.1.4. **Jessilyn Dunn, PhD -** Dr. Dunn, is Assistant Professor in the Department of Biomedical Engineering at Duke University. She works on developing new tools and infrastructure for multi-modal biomedical data integration to drive precision/personalized methods for early detection, intervention, and prevention of disease. She leverages expertise in data science, engineering, informatics, medicine, biological sciences, and population health. Her works has direct implication by arming healthcare professionals with tools and information to detect illness and intervene early and to deliver the right treatment at the right time to the right person. Dr. Dunn received Ph.D. in Biomedical Engineering from Georgia Institute of Technology in 2015.

2.2.Panel Moderators

2.2.1. **Ali Torkmani PhD.** Dr. Torkamani is the Director of Genomics and Genome Informatics at the Scripps Research Translational Institute and Professor at The Scripps Research Institute. Dr. Torkamani's research centers on the use of genomic and informatics technologies to identify the genetic etiology and underlying mechanisms of human disease to define health risks and individualized interventions. Major focus areas include human genome interpretation, genomic discovery of novel rare diseases, comprehensive, genetically-informed machine- and deep-learning prediction of risk for common diseases, and digital communication of genetically-informed disease risk. He has authored over 100 peer-reviewed publications as well as numerous book chapters and Medscape references, and his research has been highlighted in the popular press. Dr. Torkamani's overall vision is to decipher that code in order to understand and predict interventions that restore diseased individuals to their personal health baseline.

2.2.2. **Anurag Verma PhD.** Dr. Verma is an Instructor in the Department of Medicine at the University of Pennsylvania and Associate Director of Clinical Informatics and Genomics for Penn Medicine BioBank. His research has focused on the study of the genetic basis of complex diseases using big data techniques with the main focus of studying the genetic architecture of multimorbidity, the phenotypic architecture of

common genetic risk, polygenic risk scores, and phenome-wide association studies to identify the complex phenotypic and genomic interactions that lead to complex disease. He has biomedical informatics expertise in the integration of genetic data with electronic health records (EHRs) from large biobanks, with extensive experience in analyzing large biobank datasets, including Penn Medicine BioBank, Million Veteran Program, Geisinger MyCode, and eMERGE network.

2.2.3. Jennifer Huffman PhD. Dr. Huffman is a member of the Faculty for the Department of Medicine at Harvard Medical School and the Scientific Director for Genomics Research within the Center for Population Genomics at the VA Boston Healthcare System. She is currently an investigator with the VA Million Veteran Program. She leads research investigations into the genetic contributions to cardiovascular risk factors and coordinates and implements several infrastructure programs for the program. This has also allowed her to actively participate in several collaborations with statisticians and computer scientists to improve analyzing "big data" methods.

References

1. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

2. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).

3. Kember, R. L. *et al. Polygenic Risk Scores for Cardio-renal-metabolic Diseases in the Penn Medicine Biobank*. http://biorxiv.org/lookup/doi/10.1101/759381 (2019) doi:10.1101/759381.

4. Tansley, Stewart, and Kristin Michele Tolle. The fourth paradigm: data-intensive scientific discovery. Ed. Anthony JG Hey. Vol. 1. Redmond, WA: Microsoft research, 2009.