ICLR 2024 Workshop Proposal
Machine Learning for Genomics Explorations (MLGenX)

Tagline. Bring together communities at the intersection of machine learning and genomics to discuss areas of interaction and explore possibilities for future areas of research.

Abstract. The critical bottleneck in drug discovery is still our limited understanding of the biological mechanisms underlying diseases. Consequently, often we do not know why patients develop specific diseases, and many drug candidates fail in clinical trials. Recent advancements in new genomics platforms and the development of diverse omics datasets have ignited a growing interest in the study of this field. In addition, machine learning plays a pivotal role in improving success rates in language processing, image analysis, and molecular design. The boundaries between these two domains are becoming increasingly blurred, particularly with the emergence of modern foundation models that stand at the intersection of data-driven approaches, self-supervised techniques, and genomic explorations. This workshop aims to elucidate the intricate relationship between genomics, target identification, and fundamental machine learning methods. By strengthening the connection between machine learning and target identification via genomics, new possibilities for interdisciplinary research in these areas will emerge.

Format. In-Person[1].

1 Workshop Motivation and Description

The main objective of this workshop is to bridge the gap between machine learning (ML) and (functional) genomics, focusing on target identification—a pivotal yet lesser-known aspect of drug discovery for ML experts. Our goal is to explore this challenging aspect of modern drug development, where we aim to identify biological targets that play a critical role in modulating diseases. This exploration can help pave the way for impactful machine learning applications to accelerate the development of new therapeutics, including novel antibodies, cell therapy (Amini et al., 2022), gene therapy (Bulaklak & Gersbach, 2020), and RNA/DNA vaccines (Dailey et al., 2023).

In this workshop, we will delve into the intersection of ML and genomics, with a specific focus on areas where the availability of data has expanded due to emerging technologies (e.g., large-scale genomic screens, single cell, and spatial omics platforms). From a biological perspective, our discussions will encompass sequence design (Jain et al., 2022; Nguyen et al., 2023), genetic perturbations (Roohani et al., 2023; Tigas et al., 2022; Lotfollahi et al., 2023), single cell (Lähnemann et al., 2020; Dong & Kluger, 2022) and, spatial omics (Zhang et al., 2022), shedding light on key biological questions in target identification (Gupta et al., 2021). On the ML front, we aim to address topics such as interpretability (Lundberg et al., 2020), foundation models for genomics/biology (Ma & Wang, 2023; Oh et al., 2023), generalizability, and causal discovery (Ahuja et al., 2023; Lopez et al., 2022; Ke et al., 2023), emphasizing the significance of ML in advancing target identification.

Through this workshop, participants will gain valuable insights into the synergies between ML and genomics research, and help to refine the next-generation of applied and theoretical ML methods for target identification.

[1]The talks and panel discussion will be live on SlideLive.

1.1 Scopes

This workshop covers three topics in genomics, all essential for tackling the unique challenges presented in target identification:

- **Design of regulatory sequence elements:** Prediction and optimization of biological sequences, incorporating constraints and prior knowledge.
    a. Effectively miniaturize DNA/RNA sequences while preserving their key properties
    b. Multi-omics-based sequence design
    c. Modeling long-range genomic sequence interactions
    d. Tissue/cell-type specific sequence design
- **Inferring cellular communication via cell states and organization in tissues:** Causal representation learning to model cell states and cellular communities.
    a. Multi-omics data integration (single cell, spatial transcriptomics)
    b. Cell-cell interactions inference
    c. Mechanistic modeling of cells in their context to infer cellular function
    d. Modeling long-range interactions in single-cell and spatial omics
- **Perturbative biology:** Interpretable foundation models to understand cellular responses to perturbations.
    a. Translating genetic perturbations to understandable and actionable molecular changes
    b. Causal reasoning for learning gene regulatory networks
    c. Integrating multimodal perturbation readouts (transcriptomic and phenotypic) to better characterize the broader molecular effects
    d. Large-scale foundation models for predicting transcriptional outcomes of novel perturbations
    e. Generalizability of perturbation predictive models across cell lines and cellular contexts

1.1.1 From First Principles: AI for Genomics Exploration

The distinctive challenges posed by high-throughput omics data necessitate both the development of innovative deep learning architectures and present an intricate landscape for established machine learning domains, including:

- **Foundation models for genomics.** Transformer-based foundation models have gained significant strides in recent years manifesting emergent abilities (abilities that are not present in smaller-scale models but are present in large-scale models) in various NLP tasks. This is attributed to the models' capability to capture intricate relationships and dependencies among modalities of input data. By harnessing these emergent abilities, transformer-based foundation models can be leveraged to unearth meaningful insights from genomics data, facilitating the identification of disease-associated genes and advancing our understanding of complex genetic mechanisms.
- **Interpretability.** Expressive neural networks are often relied on to model high-dimensional and complex biological data. Extracting human-understandable and domain-specific biological hypotheses is a major challenge, because although the expressivity of these models allows for better predictive performance, it also limits the ability of scientists to derive high-level insights into the underlying biological system and its mechanisms. Broadly, the area of interpretability seeks to develop domain-agnostic or domain-informed methods that reliably recover scientific insight in ways other than relying solely on the target prediction.

- **Causality.** Current machine learning systems have improved significantly by using larger models and datasets, but they primarily rely on statistical correlations and struggle with tasks that require higher-level understanding. To address this issue, integrating ideas from causality into representation learning is a promising approach. Causal inference helps reason about the effects of interventions on a system and hypothetical scenarios. However, it often assumes causal variables are known, whereas real-world biological data is often high-dimensional and lacks meaningful causal structure. The goal is to learn low-dimensional, high-level causal variables and their relationships directly from unstructured data. This leads to representations that enable concepts like causal factors, intervention, reasoning, and planning.

1.2 Tentative Schedule

All invited speakers and panelists have been confirmed.

Time	Speaker	Affiliation	Areas of Expertise
Opening remarks			
9:00 - 9:10	Organizers		
Invited Talks I			
9:10 - 9:40	Su-In Lee	University of Washington	Interpretable AI principles
9:40 - 10:10	Bo Wang	University of Toronto	Foundation Models
Spotlight I			
10:10 - 10:50	Accepted spotlights		
10:50 - 11:00	Break		
Invited Talks II			
11:00 - 11:30	Silvia Chiappa	Google DeepMind	Intersection of Causality & ML, GRN Inference
11:30 - 12:00	Juan C. Caicedo	Broad Institute	ML for decoding cellular phenotypes using imaging and computer vision
Poster Session I			
12:00 - 12:45	Accepted papers		
12:45 - 13:30	Lunch Break		
Invited Talks III			
13:30 - 14:00	Marinka Zitnik	Harvard University	ML for Therapeutic Science
14:00 - 14:30	Max Jaderberg	Isomorphic Labs	Generative ML for Biology
Spotlight II			
14:30 - 15:10	Accepted spotlights		
15:10 - 15:15	Break		
Expert Panel Discussion			
15:15 - 16:35	Nicola Richmond	BenevolentAI	Knowledge Graph & LLMs
	Jure Leskovec	Stanford University	GNN & ML for Perturbation
	Kyunghyun Cho	NYU/Genentech	LLMs, Sequence Design
	Bianca Dumitrascu	Columbia University	ML for Single Cell, Multimodal data integration
	Lindsay Edwards	Relation Therapeutics	ML for functional genomics
	Jason Hartford	Recursion Pharma	Causal Discovery, Perturbation design
Poster Session II			
16:35 - 17:25	Accepted papers		
Closing remarks			
17:25 - 17:30	Organizers		

The workshop will feature six invited talks from industry and academic leaders, each of 30 minutes total (25 minutes presentation, 5 minutes Q&A). In order to foster interactivity during the workshop, we will have

two 45-minute poster sessions (one in the morning and one in the afternoon to accommodate time zone constraints) and one 80-minute-long **panel discussion** with domain experts. Two spotlight presentations of 40 minutes each will feature the most interesting works submitted to the workshop. Each session will feature 4 different papers (10 minutes per paper). *All invited speakers and panelists are confirmed.*

1.3 Attendees

We expect approximately 250 participants, which amounts to roughly 4% of the total ICLR 2024 attendees. This estimation is based on our prior experience in organizing similar workshops.

We expect the workshop to attract researchers both from machine learning and biology who are interested in diverse questions ranging from what ML can do for genomics as well as which datasets and questions in the next generation of therapeutics can contribute to advancing fundamental ML research.

1.4 Diversity

Our dedication to diversity, balance, equality, and inclusion is not only represented in the gender- balanced organizing team (including senior and junior researchers across several continents and institutions) but also in the invited speakers, panelists, reviewers, and the workshop's organizers:

- The organizing committee includes representation from different affiliations (academia and industry), seniority (graduate students, professors, and senior researchers), geographic location (Europe, USA), gender, and ethnicity.
- The speakers and panelists are chosen from different fields of expertise (both biology and ML), various institutions (both academia and industry), and different seniority levels, genders, and geographic locations.
- Three organizers will be involved for the first time in the organization of a workshop.
- The reviewers are also chosen with demographic, expertise, and seniority diversity in mind to minimize the risk of biased judgment of the submissions.

1.5 Accessibility, Advertisement, and Website

While the workshop will primarily be held **in person**, we will also utilize a range of digital tools to *engage with our online audience* and enhance interactivity throughout the event. These tools include Gathertown for poster sessions and break discussions, as well as virtual Q&As conducted on Zoom. Moreover, our website will serve as a central platform for disseminating the call for papers, promoting the workshop, and providing early access to the planned agenda and talk titles. This enables attendees to make choices about their attendance based on the content schedule in an accessible manner. We will promote the workshop in advance on our website, via our social media channels, and through collaborations with industry and academic partners to attract a diverse community of researchers interested in machine learning for target identification.

1.6 Related Workshops

While there are areas of overlap with recent workshops, MLGenX will focus on emphasizing specific open problems in the field of genomics. Unlike previous workshops that were closer to drug discovery and structural biology, our workshop focuses primarily on high-throughput omics techniques, e.g. single-cell and

perturbation data, which play a pivotal role in bridging genomics research and machine learning. More specifically, distinctions between our proposed workshop and related ones are as follows:

Drug Discovery: ICLR workshop on Machine Learning for Drug Discovery, NeurIPS Workshop on Machine Learning in Structural Biology (MLSB), and NeurIPS Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3).
The proposed MLGenX workshop distinguishes itself by delving into new and comprehensive topics. Unlike MLDD, MLSB, and AI4D3, which primarily focus on molecular design in drug discovery, MLGenX places a strong emphasis on *target identification*. This focus is pivotal not only as the initial step in molecule design but also for shaping the next generation of therapeutic domains, including RNA/DNA Vaccines, cell therapy, and gene therapy. From a machine learning perspective, MLGenX stands out by emphasizing foundation models applied to large-scale genetics and genomics data, causality, interpretability, and generalizability. This sets it apart from the MLDD, MLSB, and AI4D3 workshops, which primarily focus on active learning, graph neural networks, and generative models, respectively.

Computational Biology: NeurIPS Workshop on Generative AI and Biology (GenBio), ICML Workshop on Computational Biology (CompBio 2023), and NeurIPS Workshop on Learning Meaningful Representations of Life (LMRL).
The proposed MLGenX workshop on target identification stands apart from the previous workshops on computational biology because of its distinct focus and specialized objective. In fact, while past workshops in computational biology have explored the broader spectrum of computational techniques applied to various facets of biology, this workshop puts a strong focus on the three specific areas that connect deep learning to the field of target discovery which is so central in healthcare. From a machine learning perspective, MLGenX shares some similarities with representation learning in LMRL and generative models in GenBio. However, this workshop delves into the foundation models, generalizability, and effective methods for achieving *disentangled* representation learning and interpretability —challenging machine learning topics that we plan to thoroughly discuss during this event.

1.7 Tentative Timeline

- **Call for papers:** 15 December 2023
- **Submission deadline:** 3 February 2024
- **Reviewing period:** 3 February - 24 February 2024
- **Notification:** 1 March 2024

2 Workshop Organizers

Fabian Theis - Director of the Institute of Computational Biology and Professor at TUM Mathematics and Life Sciences, Germany (✉ fabian.theis@helmholtz-munich.de)
Fabian conducts research in the field of computational biology. The main focus of his work is the application of machine learning methods to biological questions, in particular as a means of modeling cell heterogeneities on the basis of single cell analyses and also of integrating "omics" data into systems medicine approaches. Since 2013 he has been a Full Professor of biomathematics at TUM, where he holds the Chair of Mathematical Modeling of Biological Systems, and director of the Institute of Computational Biology at the Helmholtz Zentrum München.
Previously organized workshops: Fabian organized multiple leading workshops on computational biology such as Workshop on Single Cell Genomics meets Data Science (2022) and Workshop on Computational Single Cell Genomics (2019).

Aviv Regev - Executive Vice President, Genentech, USA (✉ regev.aviv@gene.com)
Aviv is a computational biologist and systems biologist and Executive Vice President and Head of Genentech Research and Early Development in Genentech/Roche. She is a former core member of the Broad Institute of MIT and Harvard and former professor at the Department of Biology of the Massachusetts Institute of Technology. Regev is a pioneer of single cell genomics and of computational and systems biology of gene regulatory circuits. She co-founded and co-leads the Human Cell Atlas project.
Previously organized workshops: She has also played a role in organizing several conferences and workshops, including AI for Science: Progress and Promises (NeurIPS 2022), Human Cell Atlas General Meetings, and the yearly Single Cell Genomics Conference (SCG).

Arman Hassanzadeh - Software Engineer, Google, USA (✉ armanihm@google.com)
Arman is a software engineer at Google. His work is centered around developing multi-modal large language models. Prior to Google, Arman was a PhD student at Texas A&M University working with Nick Duffield and Mingyuan Zhou (UT Austin). His primary research interests are generative models, graph machine learning and Bayesian statistics.

Charlotte Bunne - PhD Student, ETH Zurich, Switzerland (✉ bunnech@ethz.ch)
Charlotte Bunne is an incoming assistant professor at EPFL and currently part of the Institute for Machine Learning and the ETH AI Center. She received her doctorate in Computer Science at ETH Zurich under the supervision of Andreas Krause and Marco Cuturi. Before, she visited the Broad Institute of MIT and Harvard as a research fellow and worked with Stefanie Jegelka at MIT. Her research aims to advance personalized medicine by utilizing machine learning and large-scale biomedical data.
Previously organized workshops: She has organized the Optimal Transport and Machine Learning Workshop at NeurIPS 2021 and 2023, the Diffusion Model Workshop (NeurIPS 2023), and the New Frontiers in Learning, Control, and Dynamical Systems Workshop at ICML 2023.

Eric Nguyen - PhD candidate, Stanford University, USA (✉ etnguyen@stanford.edu)
Eric is a PhD student at Stanford in the BioEngineering department, advised by Steve Baccus in Neurobiology and Chris Re in computer science. He focuses on novel and efficient neural network architectures for foundation models. Recently, he led the team that trained HyenaDNA, an ultralong-range genomic foundation model built with Hyena, a convolutional language model. He has a BS and MS in civil engineering from Berkeley and Stanford, respectively, and a MEng at Cornell in computer science.

Tommaso Biancalani - Director and Distinguished Scientist, Genentech Computational Sciences, USA (✉ biancalt@gene.com)
Tommaso is the head of the BRAID department (Biology Research — AI Development) which is part of the Computational Biology and Translation pillar within the Genentech Computational Science organization. The core mission of the BRAID team is to bridge foundational machine learning research with biology, with emphasis on target discovery. Prior to joining Genentech in 2021, Tommaso was at the Broad Institute of MIT and Harvard where he led a team working on the Human Cell Atlas project. Tommaso trained as a theoretical statistical physicist and completed his post-doctoral training at the Carl Woese Institute of Genomics and MIT.

Maria Brbic - Assistant Professor, EPFL, Switzerland (✉ maria.brbic@epfl.ch)
Maria is an Assistant Professor of Computer Science and, by courtesy, of Life Sciences at the Swiss Federal Institute of Technology, Lausanne (EPFL). She develops new machine learning methods and applies her methods to advance biology and biomedicine. Prior to joining the EPFL faculty in 2022, Maria was a postdoctoral fellow at Stanford University, Department of Computer Science, and was a member of the Chan Zuckerberg Biohub at Stanford.
Previously organized workshops: Maria organized different workshops in the past, including ICML 2023 Workshop on Computational Biology.

Ying Jin - PhD Candidate, Stanford Statistics, USA (✉ ying531@stanford.edu)
Ying is a PhD candidate at Department of Statistics, Stanford University, advised by Professors Emmanuel Candès and Dominik Rothenhäusler. Prior to Stanford, she obtained her Bachelor degree in Mathematics

from Tsinghua University. She works on conformal inference, multiple testing, causal inference, replicability, and data-driven decision making.

Ehsan Hajiramezanali - Principal AI Research Scientist, Genentech, USA (✉ hajiramezanali.ehsan@gene.com) Ehsan is a principal AI research scientist at Genentech in the DELTA team within the BRAID (Biology Research — AI Development) department. Before that, he was an AI research scientist at AstraZeneca. Ehsan received his PhD from Texas A&M University working with Xiaoning Qian and Mingyuan Zhou (UT Austin). His research lies at the intersection of machine learning and Bayesian statistics. He is interested in probabilistic methods, generative models, representation learning, relational learning, and multi-domain learning. *Previously organized workshops:* ICLR 2023 Workshop on Machine Learning for Drug Discovery, and MARBLE 2023 at ECML-PKDD.

3 Invited Speakers and Panelists

Silvia Chiappa. Silvia Chiappa is a Staff Research Scientist at Google DeepMind, where she leads the Causal Intelligence team, and she is an Honorary Professor at the Computer Science Department of University College London. Dr. Chiappa received a Diploma di Laurea in Mathematics from the University of Bologna and a PhD in Machine Learning from École Polytechnique Fédérale de Lausanne (IDIAP Research Institute). Before joining Google DeepMind, Silvia worked in the Empirical Inference Department at the Max-Planck Institute for Intelligent Systems (with Prof. Dr. Bernhard Schölkopf), in the Machine Intelligence and Perception Group at Microsoft Research Cambridge (Prof. Christopher Bishop), and in the Statistical Laboratory at the University of Cambridge (with Prof. Philip Dawid). Silvia Chiappa works at the intersection between statistical causality and machine learning. She is also interested in Bayesian reasoning, graphical models, variational inference, time-series models, deep learning, and ML fairness.

Bo Wang. Dr. Wang is the lead scientist of the artificial intelligence team for Peter Munk Cardiac Centre at University Health Network. He is also a Faculty Member of Vector Institute. His research areas include machine learning and computational biology. Dr. Wang obtained his PhD from the Department of Computer Science at Stanford University and has extensive industrial research experience at many leading companies such as Illumina. His PhD work covers machine learning algorithms for solving problems in computational biology with an emphasis on integrative cancer analysis and single-cell analysis. Bo Wang's long-term research goals aim to develop integrative and interpretable machine learning algorithms that can help clinicians with predictive models and decision support to tailor patients' care to their unique clinical and genomic traits.

Su-In Lee. Prof. Su-In Lee is a Paul G. Allen Professor of Computer Science at the University of Washington. She completed her PhD in 2009 at Stanford University with Prof. Daphne Koller. Before joining the University of Washington in 2010, Prof. Lee was a visiting Assistant Professor in the Computational Biology Department at Carnegie Mellon University School of Computer Science. She has received the National Science Foundation CAREER Award and has been named an American Cancer Society Research Scholar. Prof. Lee is a leading AI researcher and computational biologist, renowned for her pioneering work in developing AI/ML algorithms to advance biomedical sciences. She is widely recognized as a pioneer in explainable AI, significantly enhancing ML model interpretability.

Juan C. Caicedo. Juan Caicedo is a Schmidt Fellow and Principal Investigator at the Broad Institute of MIT and Harvard. Before becoming a Schmidt Fellow, Caicedo completed his PhD at Universidad Nacional de Colombia, where the community around him supported research, conferences, and internships in academic and industrial labs. As a result, his record enabled postdoctoral positions at the University of Illinois at Urbana–Champaign, and in Anne Carpenter's lab at the Broad. His main research interests are at the intersection of biology and machine learning, specifically for decoding cellular phenotypes and their interactions using imaging technologies and computer vision. His group aims to use machine learning to discover meaningful patterns and connections in diverse sources of biological data.

Max Jaderberg. Max Jaderberg is a research scientist and the Director of Machine Learning at Isomorphic Labs, an Alphabet company where he is reimagining the entire drug discovery process from first principles with an AI-first approach. Previously, Max Jaderberg was a research scientist at DeepMind and led the Open-Ended Learning research team. Prior to that, he co-founded Vision Factory, which was acquired by Google in 2014, and completed his PhD at the Visual Geometry Group, University of Oxford under the supervision of Prof. Andrew Zisserman and Prof. Andrea Vedaldi. His main research interests are in artificial intelligence: deep learning, reinforcement learning, and generative modeling. Please see his publications for some academic work.

Nicola Richmond. Nicola Richmond is the Vice President of AI at BenevolentAI. Nicola has a PhD in pure mathematics and has worked at the intersection of AI and drug discovery for 22 years. During her post-doc, she developed the algorithm that lies at the heart of a commercial product (GALAHAD) that has been used throughout the worldwide pharmaceutical industry to elucidate 3D pharmacophores for virtual screening. Nicola joined GlaxoSmithKline (GSK) in 2004 where she made several key contributions. Her statistical approaches for actioning high-throughput screening data are in continual use across GSK's early drug discovery portfolio, and her work on predicting the stable expression of therapeutic antibodies has reduced manufacturing cell line development timelines by 85%. Nicola also built and led the GSK.ai Fellowship Programme which has helped educate and inspire the next generation of brilliant minds who want to apply AI to drug discovery and impressively achieved a 50:50 ratio of women to men.

Jason Hartford. Jason Hartford is a Senior Machine Learning Research Scientist at Recursion Pharma. Prior to his current role, he completed a postdoc at Mila with Yoshua Bengio. Before his postdoctoral work, he obtained his PhD at University of British Columbia with Kevin Leyton-Brown. His research interests revolve around leveraging structural assumptions about data-generating processes to enhance the generalization capabilities of machine learning models beyond the observed distribution of training data. His work has encompassed the use of deep learning for causal inference and the design of deep network architectures tailored for permutation-invariant data. More recently, since starting at Mila, Jason has been dedicated to the pursuit of learning representations with identifiability guarantees.

Marinka Zitnik. Marinka Zitnik is an Assistant Professor at Harvard University in the Department of Biomedical Informatics. Dr. Zitnik is Associate Faculty at the Kempner Institute for the Study of Natural and Artificial Intelligence, Broad Institute of MIT and Harvard, and Harvard Data Science. Dr. Zitnik investigates the foundations of AI to enhance scientific discovery and facilitate individualized diagnosis and treatment in medicine. Before joining Harvard, she was a postdoctoral scholar in Computer Science at Stanford University. She was also a member of the Chan Zuckerberg Biohub at Stanford. She received her bachelor's degree, double majoring in computer science and mathematics, and then graduated with a Ph.D. in Computer Science from University of Ljubljana just three years later while also researching at Imperial College London, University of Toronto, Baylor College of Medicine, and Stanford University. Dr. Zitnik is an ELLIS Scholar in the European Laboratory for Learning and Intelligent Systems (ELLIS) Society. She is a member of the Science Working Group at NASA Space Biology. She co-founded Therapeutics Data Commons and is the faculty lead of the AI4Science initiative. She was named Kavli Fellow 2023 by the US National Academy of Sciences.

Lindsay Edwards. Lindsay Edwards CTO and Head of Platform at Relation. Previously, Lindsay was VP and Head of AI for Respiratory and Immunology at AstraZeneca. Originally a specialist in systems biology, he joined GlaxoSmithKline in 2014 from the Physiology faculty at King's College London. He started GSK's first Data Science group, was Head of Respiratory Data Sciences, Global Head of Respiratory Digital, Data and Analytics, then VP and Head of AI/ML for the UK and Europe before becoming VP of AI/ML Engineering under Kim Branson. He holds a DPhil in Physiology from Oxford, has published more than 40 articles in peer-reviewed journals and is an accomplished public speaker.

Bianca Dumitrascu. Bianca Dumitrascu is an affiliated lecturer in the Department of Computer Science and Technology (Computer Laboratory) at the University of Cambridge. She also holds the position of Departmental Early Career Fellow in the Accelerate Programme for Scientific Discovery. Her research is

situated at the crossroads of machine learning and genetics. Bianca's primary research focus revolves around comprehending how local molecular rules lead to emergent spatial patterns within biological dynamical systems. To achieve this, she employs techniques drawn from statistical optimization, statistical physics, and domain adaptation to identify contextual phenotypes in spatial transcriptomic data. Before her current position, Bianca was a Member in the School of Mathematics at the Institute for Advanced Study. She earned her Ph.D. in Computational Biology at Princeton University under the guidance of Barbara E. Engelhardt. Her doctoral research centered on the impact of experimental design in single-cell gene expression studies and the development of methods for structured, high-dimensional medical and genomic data. Bianca completed her undergraduate studies in Mathematics at MIT.

Kyunghyun Cho. Kyunghyun Cho is a professor of computer science and data science at New York University and a senior director of frontier research at the Prescient Design team within Genentech Research and Early Development (gRED). He is also a CIFAR Fellow of Learning in Machines and Brains and an Associate Member of the National Academy of Engineering of Korea. He served as a (co-)Program Chair of ICLR 2020, NeurIPS 2022 and ICML 2022. He is also a founding co-Editor-in-Chief of the Transactions on Machine Learning Research (TMLR). He was a research scientist at Facebook AI Research from June 2017 to May 2020 and a postdoctoral fellow at University of Montreal until Summer 2015 under the supervision of Prof. Yoshua Bengio, after receiving MSc and PhD degrees from Aalto University April 2011 and April 2014, respectively, under the supervision of Prof. Juha Karhunen, Dr. Tapani Raiko and Dr. Alexander Ilin. He received the Samsung Ho-Am Prize in Engineering in 2021. He tries his best to find a balance among machine learning, natural language processing, and life, but almost always fails to do so.

Jure Leskovec. Jure Leskovec is a Professor of Computer Science at Stanford University. His primary research domain lies in applied machine learning for large interconnected systems, with a specific focus on modeling complex, richly-labeled relational structures, graphs, and networks across systems of all scales. These systems range from the interactions of proteins within a cell to the interactions between individuals in a society. Jure's research has broad applications, encompassing commonsense reasoning, recommender systems, computational social science, and computational biology, with a special emphasis on drug discovery.

4 Advisory Committee

We received invaluable support and guidance from numerous domain experts from both academia and industry, in addition to our core organizers.

Aïcha BenTaieb - Principal Scientist, Genentech, USA (✉ bentaieb.aicha@gene.com)
Aïcha BenTaieb is a principal AI scientist at Genentech leading a research group focused on developing novel machine learning techniques applied to spatial biology in the BRAID department. Prior to joining Genentech, she was the director of the Pathology AI team at Tempus Labs and a scientist at Roche Tissue Diagnostics.

Gabriele Scalia - Principal Research Scientist, Genentech, USA (✉ scalia.gabriele@gene.com)
Gabriele is an AI research scientist at Genentech, where he leads the DELTA research group focused on developing new methods to support biology and biomedicine discovery. Before that, he was at the Broad Institute of MIT and Harvard, developing novel deep learning methods for genomics applications. He received his PhD in Computer Science and Engineering from the Polytechnic University of Milan, while conducting research at MIT as a visiting researcher. His research interests include graph learning, generative modeling, uncertainty estimation, and multi-modal learning.

Chandler Squires - PhD Candidate, MIT EECS, USA (✉ csquires@mit.edu)
Chandler is a PhD candidate at the Department of Electrical Engineering and Computer Science at MIT, advised by Caroline Uhler and David Sontag. His work is centered on developing methods to predict the effect of interventions in complex systems, with a particular focus on predicting the effect of genetic and molecular perturbations on cell state. His research interests include causal structure learning, representation

learning, and experimental design.

Sepideh Maleki - Postdoctoral Researcher, Genentech, USA(✉ malekis@gene.com)
Sepideh Maleki is a post-doctoral researcher at Genentech, specializing in the development of explainable models for predicting molecular perturbations. Before her current role at Genentech, Sepideh pursued her PhD in Computer Science at the University of Texas at Austin. Throughout her doctoral journey, she was involved in developing innovative machine learning and analytical methods tailored to adeptly process and analyze large-scale hypergraph and graph structured data. Her multifaceted expertise seamlessly bridges the domains of computer science and biology, promising significant advancements in both fields. *Previously organized workshops:* Graph Representations and Algorithms in Biomedicine at the Pacific Symposium on Biocomputing 2023 (PSB'23)

Moksh Jain - PhD Student, Université de Montréal and Mila, Canada (✉ moksh.jain@mila.quebec)
Moksh Jain is a PhD Student at Université de Montréal and Mila supervised by Yoshua Bengio. Prior to joining Mila, he was a Software Engineer at Microsoft. His research focuses on probabilistic inference and experimental design, along with applications in drug discovery.

Alex Tseng - Senior Research Scientist, Genentech, USA (✉ tseng.alex@gene.com)
Alex is an AI research scientist in the BRAID department at Genentech. He obtained his PhD at Stanford University, where his research was in improving and leveraging the interpretability of deep neural networks for genomics. His primary research interest is in figuring out how neural networks learn and store information so that they are more useful for scientific discovery. Broadly, his work focuses on methods that impose domain-specific inductive biases to neural networks, and methods to extract scientific knowledge that has been implicitly learned by these models.

Yashas Annadani - PhD Student, Helmholtz AI and TUM, Germany (✉ yashas.annadani@helmholtz-munich.de)
Yashas is a PhD student at Helmholtz AI and Technical University of Munich supervised by Stefan Bauer and co-supervised by Bernhard Schölkopf as part of the ELLIS doctoral program. Prior to his PhD, he obtained his masters degree from ETH Zurich. His research aims to design and predict large-scale genetic perturbations with machine learning. More broadly, his research interests include causality, representation learning and experimental design.

Nathaniel Diamant - Machine Learning Scientist, Genentech, USA (✉ diamant.nathaniel@gene.com)
Nathaniel is a research scientist at Genentech where he develops deep learning methods with applications in drug discovery and biology research. He has worked on graph-generative modeling, self-supervised learning, and uncertainty prediction. Prior to Genentech, Nathaniel worked on deep learning methods for cardiology and clinical applications at the Broad Institute and the MIT Computational Cardiovascular Research Group.

5 Program Committee

We have assembled a diverse program committee for the workshop from various levels of seniority, fields of expertise, gender, and ethnic background to represent different views in the reviewing process. We will ensure the review load will not exceed more than 3 papers per reviewer. This helps us to provide constructive and detailed feedback to the authors. If the workshop is accepted, we will announce the call for papers by *15 December 2023* with the submission deadline of 3 February 2024. The reviews will be available to the authors by 1 March 2024. It will be communicated that we will *not accept* submissions that have previously been published. Moreover, the reviewers will be instructed to ensure the novelty of the submissions in their scores and mark any previously published work as such. To prevent conflict of interests, we will ask both authors and reviewers to provide their affiliations and also existing conflict with institutions, or individuals. We will use Openreview as the submission and reviewing system. We will put the accepted papers on the workshop website and will communicate explicitly that the accepted papers are *non-archival* and can be published elsewhere in the future. Here are the tentative PC members (reviewers):

- Dino Oglic (AstraZeneca)
- Panagiotis Tigas (Isomoprhic Labs, Oxford)
- Mo Lotfollahi (Wellcome Sanger Institute)
- Samuel Stanton (Genentech)
- Julius Adebayo (Prescient Design)
- Gokcen Eraslan (Genentech)
- Alexandr Kalinin (Broad Institute)
- Kexin Huang (Stanford)
- Shantanu Singh (Broad Institute)
- Arash Mehrjou (ETH)
- Pascal Notin (Oxford)
- Talip Ucar (AstraZeneca)
- Yusuf Roohani (Stanford)
- Reza Abbasi Asl (UCSF)
- Emmanuel Bengio (Recursion)
- Tristan Deleu (Mila)
- Michael Poli (Stanford)
- Aliyah Hsu (UC Berkeley)
- Nan Rosemary Ke (Google)
- Randy Ardywibowo (Apple)
- Carlo De Donno (Helmholtz München)
- Yoyo Wu (City University of New York)
- Pouya M Ghari (UC Irvine)
- Mostafa Karimi (Amazon)
- Hannes Stärk (MIT)
- Ziqing Lu (Genentech)
- Emile Mathieu (University of Cambridge)
- Pietro Barbiero (University of Cambridge)
- Tal Ashuach (Berkeley)
- Shentao Yang (UT Austin)
- Wengong Jin (Broad Institute)
- Benedek Rozemberczki (Isomorphic Labs)
- Alex Tong (Mila)
- Niki Kilbertus (TUM)
- Seyednami Niyakan (TAMU)
- Yizhen Zhong (Freenome)
- Nick Pawlowski (Microsoft Research)
- Guadalupe Gonzalez (Genentech)
- Nikita Moshkov (BRC)
- Vignesh Ram Somnath (ETH Zurich)
- Lars Lorch (ETH Zurich)
- Stefan Stark (ETH Zurich)
- Zoe Piran (Hebrew University of Jerusalem)
- Philip Fradkin (University of Toronto)
- Reza Armandpour (Apple)
- Peiman Mohseni (TAMU)
- Andrew Jesson (University of Oxford)
- Emmanouil Angelis (TUM)
- Paul Bertin (Mila)
- Cecilia Casolo (TUM)

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

Leila Amini, Sara K Silbert, Shannon L Maude, Loretta J Nastoupil, Carlos A Ramos, Renier J Brentjens, Craig S Sauter, Nirali N Shah, and Mohamed Abou-el Enein. Preparing for car t cell therapy: patient selection, bridging therapies and lymphodepletion. *Nature Reviews Clinical Oncology*, 19(5):342–355, 2022.

Karen Bulaklak and Charles A Gersbach. The once and future gene therapy. *Nature communications*, 11 (1):5820, 2020.

Gabrielle P Dailey, Erika J Crosby, and Zachary C Hartman. Cancer vaccine strategies using self-replicating rna viral platforms. *Cancer Gene Therapy*, 30(6):794–802, 2023.

Mingze Dong and Yuval Kluger. Geass: Neural causal feature selection for high-dimensional biological data. In *The Eleventh International Conference on Learning Representations*, 2022.

Rohan Gupta, Devesh Srivastava, Mehar Sahu, Swati Tiwari, Rashmi K Ambasta, and Pravir Kumar. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity*, 25:1315–1360, 2021.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael Curtis Mozer, Christopher Pal, and Yoshua Bengio. Neural causal structure discovery from interventions. *Transactions on Machine Learning Research*, 2023.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35:19290–19303, 2022.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, pp. e11517, 2023.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Jun Ma and Bo Wang. Towards foundation models of biological image segmentation. *Nature Methods*, 20 (7):953–955, 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 2023.

Gyutaek Oh, Baekgyu Choi, Inkyung Jung, and Jong Chul Ye. schyena: Foundation model for full-length single-cell rna-seq analysis in brain. *arXiv preprint arXiv:2310.02713*, 2023.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, pp. 1–9, 2023.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. *Advances in Neural Information Processing Systems*, 35:24130–24143, 2022.

Linlin Zhang, Dongsheng Chen, Dongli Song, Xiaoxia Liu, Yanan Zhang, Xun Xu, and Xiangdong Wang. Clinical and translational values of spatial transcriptomics. *Signal Transduction and Targeted Therapy*, 2022.