# GENERATIVE AI IN GENOMICS (GEN$^2$): BARRIERS AND FRONTIERS


## ABSTRACT

Generative AI (GenAI) is transforming biology, with breakthrough applications like directed evolution in protein science. The parallel ambition to engineer cellular and tissue states in genomics is now a major frontier, yet progress is hampered by domain-specific roadblocks. Our workshop is designed to bridge this gap between GenAI's promise and its practical applications towards this goal. With recent large-scale data initiatives launched to support GenAI models creating an inflection point for the field, timing is ideal. Through a field-grounding keynote by a genomics expert, invited talks by GenAI practitioners, contributed presentations, and a moderated debate, we will bring together experts and early-career scientists from machine learning and experimental genomics to collaboratively define a roadmap for progress. Our program will target core, interconnected challenges across the development pipeline: from data generation priorities and model design for genomic hierarchies to biologically-grounded evaluation frameworks and interpretability. By defining promising research directions and critical evaluations, our ultimate goal is to catalyze a new generation of models for tangible biological impact.

**Tagline:** Advancing generative AI in genomics for tangible impact by identifying domain-specific challenges and promising directions.


## 1 WORKSHOP SUMMARY

Generative AI (GenAI) is transforming the modeling and synthesis of complex data across numerous domains, including biology. In protein science, GenAI models have demonstrated tremendous success in structure and mutational effect prediction [11; 1; 10], enabling powerful applications like directed evolution [7], where generative models propose protein variants that are iteratively screened and optimized for a targeted function. Similarly, in genomics, there is growing interest in using GenAI to design various perturbations that can shift cellular and tissue states towards desired targets in a controlled manner [4]. However, GenAI in genomics remains at a comparatively less mature stage to achieve this goal. While several areas have yielded compelling applications, such as regulatory sequence design [9], cellular trajectory modeling [8], and spatial tissue reconstruction [15], progress on some core challenges remains limited, with current methods struggling to deliver meaningful gains over simple baselines. For example, in perturbation-effect prediction, generative models have yet to demonstrate robust and consistent improvements over linear approaches, with recent studies suggesting they may fail to capture key biological effects [13; 2].

**Our workshop, *Gen$^2$: Advancing Generative AI in Genomics*, is designed to sharply focus on domain-specific roadblocks and open challenges that must be addressed to accelerate the impact of GenAI in genomics.** Through a field-grounding keynote, invited talks on systematic evaluations and data-specific architectural advances, and a moderated debate panel, we will discuss challenges and opportunities spanning the full development pipeline, such as:

- **Data generation priorities:** With the growing traction of large-scale pre-trained models that promise generalization across biological contexts and eventually enabling targeted perturbation design, multiple efforts are now underway to generate foundational datasets to support model training [5; 12; 16]. Yet, diverging views persist on which data modalities and levels of biological diversity should take priority to maximize downstream utility [3], **making this a timely and critical point of discussion;**
- **Defining downstream use cases and evaluation tasks for pre-trained models,** while also surfacing what they cannot achieve. As pre-trained models become more common in genomics, it is crucial to develop evaluation frameworks grounded in real-world biological relevance and clarify which downstream applications they can and cannot support;
- **Designing GenAI models for genomic inputs and hierarchies:** Many GenAI methods are developed for vision and language, where data typically has well-defined spatial or sequential structure. Genomics applications often adapt inputs to fit these architectures, e.g. by tokenizing expression data like text, potentially limiting expressivity. Moreover, genomic information is inherently hierarchical and interdependent, spanning

1D/3D regulatory signals determining cell states, and tissue-level signaling feeding back into regulation. Incorporating these into model design could improve expressivity and enable multiscale reasoning.

- **Mechanistic interpretability and counterfactual generation with GenAI:** Unlike applications in vision and language where generative performance is the primary goal, a model's utility in genomics depends heavily on providing mechanistic insight. A key application of this insight is counterfactual generation: predicting the outcomes of interventions on these mechanisms using limited observational data. Developing robust methods for both interpretability and counterfactual generation in emerging GenAI models is therefore essential for genomics applications;
- **Defining new genomics applications enabled by GenAI and new measurement technologies:** Much current work retrofits generative models to existing tasks. We aim to surface new biological applications made possible *because of* generative modeling and new genomic measurement technologies developed.

To foster productive discussion around these challenges, the workshop program is structured to build momentum from foundational context to focused debate (Section 2). We will begin with a 45-minute keynote by Dr. Sarah Teichmann, a recognized leader in genomics, who also frequently collaborates with the machine learning community. Her talk will give a primer to the GenAI practitioners on open questions in the field that are ripe for GenAI and also present a critical perspective on the current limitations and a forward-looking vision for how GenAI can transform genomics, followed by a 15-min interactive Q&A session. This will be followed by 30-min invited talks (with 10-min Q&A component) highlighting complementary perspectives across the development pipeline: systematic evaluations of the shortcomings in current approaches to perturbation modeling [13], architectural and modeling advances for representing biological data distributions, with applications on lineage tracing and spatial transcriptomic modeling [6; 15] and the recent advances and challenges in large-scale multimodal foundation models for regulatory genomic sequence design. The workshop will conclude with a moderated debate-style panel bringing together researchers who hold diversity of views on the state, trajectory and priorities of GenAI in genomics. By soliciting audience questions in advance through an open questionnaire on our workshop website and combining them with targeted prompts from the organizing committee to highlight the points listed above, we aim to encourage an evidence-driven discussion rather than repetition of consensus narratives. Finally, through contributed talks and interactive poster presentations on submitted works on areas listed among our scope, we will highlight the recent developments in the field across numerous applications.

To ensure the workshop's impact extends beyond the event, we will synthesize the key discussions and outcomes into a public report. This document will be disseminated through our website and social media to foster continued community engagement in guiding future research.

### 1.1 Scope for Submissions

We welcome submissions from a range of application areas and methodological topics:

**Core Genomics Applications of GenAI methods,** such as in

- **Single-cell -omics:** Multi-omic integration, conditional and cross-modal generation, single cell trajectory simulation and control, perturbation effect modeling, target identification, virtual cell models;
- **Spatial transcriptomics and tissue modeling:** Generative models for in silico tissue reconstruction, cell-cell interaction modeling, and spatial perturbation response;
- **Regulatory genomics and epigenomics:** DNA language models with multi-modality, design of regulatory elements for cell-type-specific activity, DNA & RNA sequence generation,
- **Population genomics:** Ancestry-aware generative models, rare variant simulation, generative modeling for variant effect prediction, genotype-phenotype modeling

**Methodological Themes,** such as

- **Adapting generative models to biological data:** Transformers, diffusion, flow-based generative models, and VAEs, tailored for biological data;
- **Pretrained foundation models for genomics:** Large-scale self-supervised learning across multiple scales and data modalities, genome-scale masked modeling;
- **Latent space interpretation:** Interpretation of large-scale foundation models, enforcing disentanglement of biological factors, linking latent variables to known biology;

- **Data-efficient generation:** Techniques to generate from few-shot data or under noisy, incomplete annotations common in genomics;
- **Evaluation and benchmarks:** Designing suitable benchmarking tasks and evaluation criteria, re-evaluating existing models with robust benchmarks.

## 2 Workshop Schedule & Invited Speakers

| Time | Event | Speaker | Affiliation | Topic |
|---|---|---|---|---|
| 9:00 | **Welcoming Remarks** | Organizers | | |
| 9:10 | **Invited Talk 1** | Sarah Teichmann | University of Cambridge | Keynote: Genomics grounding for GenAI |
| 10:10 | **Invited Talk 2** | Maria Brbic | EPFL | Challenges and benchmarks in GenAI for perturbation and tissue architecture modeling |
| 10:50 | | **Coffee Break** | | |
| 11:10 | **Invited Talk 3** | Nic Fishman | Harvard University | FM for perturbations |
| 11:50 | **Contributed talks** | ∼3 talks | | Selected work from submissions |
| 12:40 | | **Lunch** | | |
| 13:30 | **Poster Session 1** | | | Accepted submissions |
| 14:30 | **Invited Talk 4** | Gokcen Eraslan | Genentech | Multi-omic language models |
| 15:10 | **Panel Discussion** | Anshul Kundaje | Stanford University | |
| | | Sarah Teichmann | University of Cambridge | |
| | | Jean-Philippe Vert | Bioptimus | |
| | | Wengong Jin | Northestern University | |
| 16:10 | | **Coffee Break** | | |
| 16:30 | **Poster Session 2** | | | Accepted submissions |

Please note that the speakers have accepted our invitations.

**Sarah Teichmann, Ph.D.** is a Chair in Stem Cell Medicine at the University of Cambridge and Head of Cellular Genetics at the Wellcome Sanger Institute. She is a co-founder and co-leader of the Human Cell Atlas consortium, a landmark global initiative to map every cell type in the human body with genomics and single-cell technologies. Her research spans genomics, single-cell biology, and computational biology, with applications in understanding immune responses, tissue organization, and human development at cellular resolution. She earned her Ph.D. from the MRC Laboratory of Molecular Biology in Cambridge and has published over 250 papers in leading journals. Dr. Teichmann has received numerous prestigious honors, including the Francis Crick Medal and Lecture (2012), the Suffrage Science Award (2012), and the EMBO Gold Medal (2015). She was elected a Fellow of the Academy of Medical Sciences (2015), an ISCB Fellow (2016), and a Fellow of the Royal Society (2020). Her leadership has helped shape the field of computational genomics, particularly in advancing large-scale data integration, multimodal analysis, and machine learning–based cellular modeling.

As an expert in genomics and computational biology, she will deliver a keynote talk on genomic foundations for GenAI practitioners. Her talk will provide both a primer on emerging and open challenges in genomics that are ripe for generative AI applications, as well as a critical perspective on the limitations GenAI currently faces in real-world biological contexts. She will conclude with a forward-looking vision for how generative AI can transform the future of genomic science.

**Maria Brbic, Ph.D.** is an Assistant Professor of Computer Science and Life Sciences at EPFL in Switzerland. She launched her lab in 2022, where her group develops machine learning methods to advance biological and biomedical discovery. Prior to joining EPFL, she was a postdoctoral researcher at Stanford University and earned her Ph.D. in Computer Science from the University of Zagreb in Croatia. Her work has been recognized by several honors, including selection as a 2021 Rising Star in EECS and the Early Career Award from the Swiss Institute of Bioinformatics (SIB). Her recent research includes the development of diffusion models for spatial reconstruction of single-cell transcriptomic

data[15] and a systematic evaluation of the limitations of current deep learning and generative models for single-cell perturbation prediction [13]. Her work lies at the intersection of machine learning, computational biology, and precision medicine, and is helping to define new directions in data-driven approaches to cellular modeling.

In her talk, Dr. Brbic will discuss the systematically assessed challenges and frontiers in generative AI methods for understanding how cells respond to perturbations and organize into complex tissue structure.

**Nic Fishman** is a Ph.D. student in Statistics at Harvard University, where he works at the intersection of generative AI, multi-scale modeling, and causal inference. His research focuses on developing large-scale machine learning systems for scientific discovery, with applications in biology, healthcare, and social science. He was a 2021 Rhodes Scholar and received an NSF-GRFP (2023) and the James Mill Peirce Fellowship at Harvard (2023).

Fishman will deliver a talk on his recent work on Generative Distribution Embeddings [6], which introduces methodological advances in flow matching models for building pre-trained models that capture population-level structure, with applications in predicting cellular responses to perturbations and lineage tracing.

**Gokcen Eraslan, Ph.D.** is a Principal Scientist at Genentech, where he leads a research group focused on developing multi-modal genomic language models. He earned his Ph.D. in Computer Science from the Technical University of Munich in 2018. Dr. Eraslan's research bridges machine learning and regulatory genomics, with a focus on building large-scale foundation models that learn the complex syntax and semantics of the genome. His work includes the development of deep learning models that integrate multiple genomic modalities to understand gene regulation and enable regulatory element design for therapeutic applications.

Dr. Eraslan will present recent advances and open questions in multi-modal genomic language models for deciphering the regulatory and functional roles of non-coding DNA, and discuss how such models can support genomic sequence design for biological discovery and intervention.

**Anshul Kundaje, Ph.D.** is an Associate Professor of Genetics and Computer Science at Stanford University. His research focuses on deep learning approaches and interpretable AI methods to decipher gene regulatory mechanisms at scale, with the goal of characterizing noncoding genomic elements and understanding how genomic and epigenomic variation contributes to human disease. Dr. Kundaje has served as a computational lead for several major genomics consortia, including ENCODE and Roadmap Epigenomics, and has helped set standards for large-scale functional genomics analysis. His contributions have been recognized with numerous honors, including the NIH Director's New Innovator Award (2016) and an Alfred P. Sloan Research Fellowship (2014). His work bridges deep learning, regulatory genomics, and biomedical data science, and continues to influence both foundational research and translational efforts in genomics and precision medicine.

In addition to his scientific contributions, Dr. Kundaje brings a critical and thoughtful perspective on the current trajectory of generative AI in genomics, particularly regarding the potential, limitations and assumptions behind emerging large-scale pre-trained models. His perspective will contribute to a rigorous discussion during our panel.

**Jean-Philippe Vert, Ph.D.** is Co-Founder and CEO of Bioptimus, an AI-first biotechnology company building foundation models to transform biology and medicine. He is also a professor (on leave) at PSL University / Mines Paris, where he previously served as the founding director of the Centre for Computational Biology. Prior to founding Bioptimus, he was Chief R&D Officer at Owkin and a Research Lead at Google Brain, where he led efforts in applying large-scale machine learning to biological data. Dr. Vert's research lies at the intersection of AI for structured and high-dimensional data and computational biology, with applications spanning genomics, multi-omics integration, and predictive modeling of complex biological systems. He has authored over 190 peer-reviewed publications and is a member of the National Academy of Technologies of France. His work continues to shape the emerging landscape of foundation models in biology, bridging theoretical innovation with translational impact.

In our panel, Dr. Vert will offer perspective on the opportunities in applying large-scale pre-trained models to biological systems, drawing on his experience across both academic research and industrial AI development.

**Wengong Jin, Ph.D.** is an Assistant Professor at Northeastern University's Khoury College of Computer Sciences. He earned his Ph.D. in Computer Science from MIT, where his thesis focused on geometric deep learning models for representation learning, generation and screening of molecules. Prior to his faculty position, he was a Schmidt Center postdoctoral researcher at the Broad Institute. He received the BroadIgnite Award (2023), Dimitris N. Chorafas Prize (2022), and the MIT EECS Outstanding Thesis Award (2021). His current research explores generative modeling for genomics, with applications on integrating multi-modal data (e.g. spatial transcriptomics and pathology), and virtual

cell models to simulate cellular behaviors under perturbations. He bridges methodological developments in machine learning and AI with applications in computational biology, systems biology and drug development.

As an early-career researcher with strong technical foundations, Dr. Jin will help represent the rapidly growing community of GenAI practitioners, pivoting to genomics applications.

## 3 LOGISTICS

We expect around a total of 150 submissions and 250-300 participants. These estimates are based on the Whova app statistics of a relevant workshop, 2024 ICLR MLGenX, which reported a total of 470 attendees. We assume Whova statistics may be inflated and make a more conservative initial estimate.

All submissions will be handled on `openreview.net`. We will offer two submission tracks: **Regular Papers** (5–8 pages) and **Tiny Papers** (2–4 pages), the latter designed to encourage early-stage or exploratory contributions. All accepted papers will be **non-archival** and presented at least as posters during the workshop. Approximately the top 3-4 papers from the Regular Paper track will also be selected for short oral presentations. The organizing committee will go over the accepted submissions to ensure that previously published work is appropriately filtered out.

Accepted papers, talk abstracts, and presentation slides will be made available on the workshop website, which will be hosted on GitHub upon acceptance of this proposal. Panel discussion transcripts will also be published online to ensure virtual accessibility.

We will encourage presenters of accepted submissions to attend the workshop in person, while accommodating virtual attendance in cases of visa issues or other exceptional circumstances. In such cases, we will coordinate in advance to collect pre-recorded presentations to broadcast during the event.

To maximize visibility and community engagement, we will advertise the workshop through professional mailing lists, social media platforms, and relevant research communities such as computational biology, genomics, and machine learning. We will further promote the workshop through the professional networks of our organizers to enhance visibility across relevant research communities.

We do not anticipate any special technical or logistical requirements.

### 3.1 TENTATIVE TIMELINE & KEY DATES

- **Call for papers:** December 10, 2025
- **Submission deadline:** January 31, 2026 AoE
- **Review period:** February 5 - February 20, 2026, AoE
- **Author Notification:** February 27, 2026 AoE
- **Camera Ready Deadline:** March 10, 2026 AoE
- **Release of Accepted Submission:** March 11, 2026 AoE
- **Distribution of our Report on Outcomes:** May 29, 2026 *(as described in Section 1)*

## 4 DIVERSITY, EQUITY, INCLUSION & ACCESSIBILITY

We recognize that individual identities are not defined by any one feature and are instead an amalgamation of traits including but not limited to ethnicity, gender, age, able-bodiedness, culture, seniority, and work affiliation. As such, we are committed to fostering a diverse, equitable, and inclusive environment across all dimensions of identity throughout every aspect of our workshop, as detailed below. We are also open to any suggestions on how to further promote an inclusive environment.

- **Balanced Representation:** In our invitations to speakers, we strived for balanced gender representation and career stage. We similarly aimed for balanced in our organizing committee (OC), which is comprised of one Ph.D. student, three postdoctoral scientists, as well as industry scientists and faculty members across various levels of seniority.
- **Accessibility:** Accessibility is a priority for us. Towards this, we will ensure that:
  - All talks will include closed-captioning,

  - On-site volunteers will be available to assist anyone who may need support with mobility or navigation throughout the venue,
  - The front row of seats will be reserved for attendees who may require extra space or closer access, including individuals using wheelchairs or those with visual impairments,
  - All presenters are asked to design their materials to be colorblind-friendly, when possible,

  Additionally, we have reached out to the ICLR Diversity, Equity & Inclusion Chairs to explore the possibility of providing lactation rooms, prayer rooms and also quiet rooms for participants, who might need low stimulation environment during the workshop. After better understanding what additional accessibility accommodations may be available through the broader conference, we will communicate these resources clearly to our attendees ahead of the workshop and add this information on our website.

- **Financial Support:** Through sponsorships (Section 7), we are securing funds to offer financial assistance to people from underrepresented and underresourced groups and those whom the financial cost may be a barrier for attendance.

## 5 Previous & Related Workshops

Several recent NeurIPS, ICLR, and ICML workshops have explored intersections between machine learning (ML) and biology. Some focus on broad biological applications, while others target more mature GenAI domains such as protein structure or small molecule drug design. In contrast, **our workshop uniquely centers on the emerging intersection of GenAI and genomics**, a field marked by unique challenges and opportunities, as described in our motivation (Section 1).

Recently-launched large-scale genomic data initiatives designed for GenAI [5; 14; 12; 16], coupled with methodological advances driven by genomics applications [8; 6], have created an inflection point for progress. By addressing core genomics domains, such as regulatory genomics, spatial transcriptomics, and single cell modeling, we aim to highlight distinct opportunities at this frontier.

Unlike many other workshops on similar topics, we have invited not only experts working at the intersection of genomics and generative AI, but also researchers who **generate genomic data to answer fundamental biological questions**. By engaging those who best understand data generation and biological systems, we aim to uncover the most fundamental limitations of current approaches and identify the most feasible and promising directions for applying generative AI in genomics.

- **Broad ML in Computational Biology:** The ICML Workshop on Computational Biology (‘22, ‘23) and Learning Meaningful Representations of Life (LMRL) (NeurIPS ‘25, ICLR ‘25) broadly cover ML applications across diverse biological topics, without a specific focus on genomics or generative AI.
- **Generative AI and Foundation Models in Biology:** The Generative AI and Biology (GenBio), held at NeurIPS ‘23 and ICML ‘25, specifically target GenAI methodologies, but covers broad applications in biology. The Accessible and Efficient Foundation Models for Biological Discovery (AccMLBio), held at ICML ‘24, emphasized foundation models across biological domains. However, neither was centered specifically on challenges and opportunities in genomics.
- **ML/AI in Genomics and Nucleic Acids (DNA/RNA):** The Machine Learning Explorations in Genomics (MLGenX) and AI for Nucleic Acids (AI4NA) workshops, both held at ICLR ‘24 and ‘25, concentrate on ML/AI specifically for nucleic acids and genomics. However, MLGenX emphasizes genomic data analysis primarily for drug discovery and cell therapy target identification, while AI4NA covers RNA/DNA structure prediction and functional analyses.
- **ML for Structural Biology and Drug Discovery:** Several workshops emphasize structural biology and small molecule applications within biology rather than genomic, including Machine Learning in Structural Biology (MLSB), typically held at NeurIPS (‘23, ‘24), Generative and Experimental Perspectives for Biomolecular Design (GEMBio), held at ICLR ’25, Machine Learning for Drug Discovery (MLDD), held at ICLR ‘23, and AI for New Drug Modalities, held at NeurIPS ‘24.
- **Broad ML/AI for Science:** The AI for Science workshop, recently held at NeurIPS ‘23 and ICML ‘24, addresses broad ML/AI applications across multiple scientific disciplines, including physics, chemistry, materials science, and biology.

# 6 WORKSHOP ORGANIZERS

We have curated an organizing committee composed of experts in generative modeling and computational genomics, representing a range of career stages, seniority levels, and areas of expertise. Some members focus on theoretical and methodological aspects of generative AI, while others specialize in genomic applications. This collective breadth of knowledge has enabled us to invite a diverse and highly qualified group of speakers and panelists. The committee includes both senior researchers with prior experience organizing workshops at machine learning conferences and junior trainees contributing fresh perspectives. Below, we list the committee members in alphabetical order by last name.

## 6.1 ORGANIZING COMMITTEE

**Tamara Broderick, MIT** (Google Scholar) tbroderick@mit.edu
Tamara Broderick is an Associate Professor with tenure in the Department of Electrical Engineering and Computer Science at MIT. She completed her PhD at UC Berkeley. She completed master's degrees in physics and mathematics at the University of Cambridge and an AB in mathematics at Princeton University. She has been awarded the Presidential Early Career Award for Scientists and Engineers (2025), an ONR Early Career Grant (2020), an NSF CAREER Award (2018), a Sloan Research Fellowship (2018), and an Army Research Office Young Investigator Program award (2017). Her research focuses on uncertainty quantification, robustness quantification, and validation in machine learning.

**Tamara Broderick has organized 19 conferences and workshops. Of those, she has served as an organizer or on the advisory committee for 11 NeurIPS worskhops. Most recently she served on the advisory committee for the NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty.**

**Valentin De Bortoli, Google DeepMind** (Google Scholar) vdebortoli@google.com
Valentin De Bortoli is a research scientist at Google DeepMind and a chargé de recherche (equiv. to assistant professor) in the Center for Data Science in Ecole Normale Supérieure in Paris (on leave). He was previously a postdoctoral researcher at Oxford University and received his PhD from ENS Paris-Saclay. He has published papers in Nature, ICASSP, COLT, UAI, ICML, NeurIPS, TMLR and JMLR. His research lies at the intersection between applied probability, statistics and machine learning with a recent focus on the interplay between stochastic control, optimal transport and generative modeling.

**Valentin De Bortoli has organized the NeurIPS Workshop on Diffusion Models in 2022, 2023, the NeurIPS Workshop on Scientific Understanding of Deep Learning in 2024, the ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems in 2023 and co-organized the AISSAI Workshop on Machine Learning-Assisted Sampling for Scientific Computing in October 2022.**

**Pinar Demetci, Broad Institute** (Google Scholar) demetci@broadinstitute.org
Pinar Demetci is a Schmidt Center Postdoctoral Fellow at the Broad Institute of MIT and Harvard. She received her Ph.D. in Computer Science and Computational Biology from Brown University in 2023. Her research focuses on developing multi-modal AI models for genomic discovery, as well as probabilistic algorithms and machine learning models for integrating genomic data across measurement modalities and time points to study the genomic drivers of cell states. She has published papers in machine learning conferences and workshops, such as AISTATS, AAAI, NeurIPS Learning Meaningful Representations of Life, ICML Workshop in Computational Biology, as well as computational biology conferences and journals, such as MLCB, RECOMB, ISMB, Nature Neuroscience and PLOS Genetics. Her work has been recognized by 2022 Rising Stars in EECS by University of Texas at Austin. She previously served in program committees for NeurIPS Workshop on Optimal Transport and Machine Learning in 2023, MLCB conference in 2022-2023 and the ISMB conference in 2022-2024.

**Arnaud Doucet, Google DeepMind** (Google Scholar) arnauddoucet@google.com
Arnaud Doucet is a senior staff research scientist at Google DeepMind. He was previously a full professor in the Department of Statistics at Oxford University. He was an Institute of Mathematical Statistics Medallion Lecturer in 2016 and the Breiman Lecturer at NeurIPS 2024. He was also awarded the Guy Silver medal from the Royal Statistical Society in 2020. His research focuses on generative modeling, Monte Carlo sampling methods and uncertainty quantification.

**Arnaud Doucet has organized over 15 conferences and workshops over the years. He has most recently co-organized the ICLR Workshop Frontiers in Probabilistic Inference: Learning meets Sampling in 2025.**

**Dongshunyi "Dora" Li, Broad Institute** (Google Scholar) dongshun@broadinstitute.org
Dora Li is a Schmidt Center Postdoctoral Fellow at the Broad Institute of MIT and Harvard. She received her Ph.D. in

computational biology from Carnegie Mellon University in 2023. Her research focuses on developing machine learning methods for the study of spatial transcriptomics and tissue architecture. Her work has been published in computational biology conferences and journals, such as RECOMB, Nature Methods, Genome Research, Genome Biology, and Bioinformatics. Her work has been recognized by 2022 Rising Stars in EECS by University of Texas at Austin.

**Maria Skoularidou, Broad Institute** (Google Scholar) mskoular@broadinstitute.org
Maria Skoularidou is a Schmidt Center Postdoctoral Fellow at the Broad Institute of MIT. She received her Ph.D. in probabilistic machine learning with applications to genomics and healthcare from University of Cambridge in 2023. During her Ph.D., she was awarded the Trinity College Award in Mathematics by the University of Cambridge in 2020. Her research focuses on developing machine learning methods to unravel genetic factors that are associated with psychiatric disorders. Her work has been published at NeurIPS, ICML, JRSSB, IEEE Transactions on Information Theory, IEEE ISIT, IEEE SPAWC and IEEE ISI.

**Maria Skoularidou has been a member of the organizing committee of ICML (Accessibility and Diversity and Inclusion co-chair 2021-present), NeurIPS (DIA co-chair 2021; Social co-chair 2019) and ACM, FAccT (DIA Co-Chair 2025). She had also organized "Advances and challenges in machine learning programming languages" workshop in 2019 at the University of Cambridge.**

**Renzo Soatto, MIT** soatto@mit.edu
Renzo Soatto is a Ph.D. student in the Electrical Engineering & Computer Science department at the Massachusetts Institute of Technology (MIT). His research focuses on adapting modern machine learning methods for biomedical applications, including antibiotic discovery, genomics, and proteomics.

**Max Welling, CuspAI & U. Amsterdam** (Google scholar) m.welling@uva.nl
Prof. Dr. Max Welling is a full professor and research chair in machine learning at the University of Amsterdam and a Merkin distinguished visiting professor at Caltech. He is co-founder and CTO of the startup CuspAI in Materials Design. He is a member of the Royal Dutch Academy of Sciences, a fellow at the Canadian Institute for Advanced Research (CIFAR) and the European Lab for Learning and Intelligent Systems (ELLIS), where he served on the founding board. His previous appointments include Partner and VP at Microsoft Research, VP at Qualcomm Technologies, professor at UC Irvine. Max Welling holds a PhD in theoretical high energy physics, conducted under the supervision of Nobel laureate Prof. Gerard 't Hooft. He then switched fields to focus on machine learning, first as a postdoc at Caltech, under the supervision of Prof. Pietro Perona, and then as a postdoc under the supervision of Nobel laureate Prof. Geoffrey Hinton at UCL & U. Toronto. He has served as associate editor in chief of IEEE TPAMI from 2011-2015 and has served on the board of the European Lab for Learning and Intelligence Systems (ELLIS) until 2021, after co-founding it in 2018. Max Welling received the ECCV Koenderink Prize in 2010 and the 10-Year Test of Time Awards at ICML in 2021 and ICLR in 2024.

**Max Welling has served on the advisory board of the NeurIPS Foundation since 2015. He was the Program Chair and General Chair of the NeurIPS conferences in 2013 and 2014, respectively. He also served as Program Chair of AISTATS in 2009 and ECCV in 2016, and as General Chair and co-founder of MIDL in 2018. In addition, he has organized numerous workshops at NeurIPS, ICLR, and ICML, most recently "Frontiers in Probabilistic Inference: Learning Meets Sampling" at ICLR 2025 and "AI for Science" at NeurIPS 2025.**

### 6.2 Advisory Board

We would like to acknowledge the helpful feedback we received on our program from several individuals with expertise in genomics or generative AI and its scientific applications, who also had organizational experience:

**Arthur Gretton, UCL and Google DeepMind** (Google Scholar) gretton@google.com
Arthur Gretton is a Professor with the Gatsby Computational Neuroscience Unit, Director of the Centre for Computational Statistics and Machine Learning (CSML) at UCL, and Research Scientist at Google Deepmind. He holds degrees in Physics and Systems Engineering from the Australian National University, and a PhD with Microsoft Research and the Signal Processing and Communications Laboratory at the University of Cambridge. He previously worked at the MPI for Biological Cybernetics, and at the Machine Learning Department, Carnegie Mellon University. His recent research interests in machine learning include causal inference and representation learning, design and training of generative models, and nonparametric hypothesis testing.

Dr. Gretton previously organized 13 workshops at NeurIPS, most recently on Machine Learning meets Econometrics (MLECON) in 2021. He was Dali workshop co-chair (2019), Machine Learning Summer School London co-chair

(2019), tutorials co-chair for ICML (2018), workshops co-chair for ICML (2019), program co-chair, AISTATS (2016).

**Fei Chen, Broad Institute** (Google Scholar) chenf@broadinstitute.org
Fei Chen is a core institute member at the Broad Institute of MIT and Harvard and an associate professor in the Department of Stem Cell and Regenerative Biology at Harvard University. Dr. Chen's laboratory is building tools that bridge single-cell genomics with space and time, to enable discoveries of where cell types are localized within intact tissues, as well as when relevant transcriptional modules are active. Dr. Chen obtained his Ph.D. in biological engineering from the Massachusetts Institute of Technology and was a Schmidt Fellow at the Broad Institute. His awards include the National Institutes of Health Director's Early Independence Award, the Searle Scholars Award, the Burroughs Wellcome CASI Award, the Allen Distinguished Investigator Award, and a Merkin Institute Fellowship.

**Todd Golub, Broad Institute** (Google Scholar) golub@broadinstitute.org
Todd Golub is director and a founding core member of the Broad Institute of MIT and Harvard. He is recognized as a leader in cancer genomics and has pioneered the development of new cell-based approaches to drug discovery for cancer and other diseases. Dr. Golub is also the Charles A. Dana Investigator in Human Cancer Genetics at the Dana-Farber Cancer Institute and professor of pediatrics at Harvard Medical School. He has been recognized with multiple awards, including the Outstanding Achievement Award from the American Association for Cancer Research, the Paul Marks Prize for Cancer Research, and the Daland Prize from the American Philosophical Society. Dr. Golub is an elected member of the US National Academy of Medicine and the National Academy of Sciences.

## 7 Sponsors

We are actively fundraising through sponsorships to support underresourced researchers with travel awards and to recognize early-stage researchers with best student paper and poster awards. We are currently in conversation with GSK (GlaxoSmithKline), AstraZeneca, Bioptimus, Noetik, GenBio AI and Xaira.

## 8 Program Committee

The organizers will oversee the reviewing process and handle the potential conflict of interest in reviewing using OpenReview's reviewer matching support. We include below our initial list of 40 program committee members. If the workshop is accepted, we will recruit additional program committee members through our networks, as well as by posting PC sign-up forms on social networks and academic groups. We will additionally ask authors of each submission to nominate a reviewer among them. We aim to have 3 papers per reviewer and 3 reviews per paper.

Georgios Batzolis (University of Cambridge), Riccardo Conci (Harvard University), Kai Cao (Broad Insitutte), Hao Chen (University of Illinois Chicago), Uthsav Chitra (John Hopkins University), Monica Dayao (Chan Zuckerberg Biohub), Jun Ding (McGill University), Jiaqi Zhang (Brown University), David Fischer (Medical University of Vienna), Boyang Fu (Harvard University), Raluca Gordan (University of Massachusetts), Euxhen Hasanaj (GenBio AI), Hyeyeon Hwang (Brown University & Tinos Therapeutics &), Marcel Hirt (NTU Singapore), Marjan Hosseini (University of Connecticut), Geert Jan Huizing (GenBio AI), Sandeep Kambhampati (Harvard Univeristy), Agustinus Kristiadi (Vector Institute), Domna Ladopoulou (UCL), Da-Inn Erika Lee (University of Wisconsin-Madison), Derek Aguiar (University of Connecticut), Melih Yilmaz (Amazon - AWS Health AI Team), Jose Lugo-Martinez (Carnegie Mellon University), Cong Ma (University of Michigan), Weiguang Mao (Flatiron Institute), Ruochi Zhang (Broad Institute), Ghulam Murtaza (University of Washington), Chibuikem Nwizu (Brown University), Tuan Pham (Brown University), Surag Nair (Genentech), Aldo Pacchiano (Boston University), Easwaran Ramamurthy (GenBio AI), Jules Samaran (Institut Pasteur, Université Paris Cité, CNRS), Yihang Shen (Princeton University), Huangqingbo Sun (Standford University), Aarthi Venkat (Broad Institute), Yinjie Wu (MIT), Elizabeth Zhang (MIT) Liangyin Yin (Broad Institute), Sebastiano Cultrera di Montesano (Broad Institute).

## References

[1] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 630(8016):493–500.

[2] Ahlmann-Eltze, C., Huber, W., and Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods, pages 1–5.

[3] Baek, S., Song, K., and Lee, I. (2025). Single-cell foundation models: bringing artificial intelligence into cell biology. Experimental & Molecular Medicine, pages 1–13.

[4] Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., et al. (2024). How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell, 187(25):7045–7063.

[5] Chan Zuckerberg Initiative (2025). Czi launches billion cells project with 10x genomics and ultima genomics to advance ai in biology. https://chanzuckerberg.com/newsroom/billion-cells-project-launches-advance-ai-biology/. Press release.

[6] Fishman, N., Gowri, G., Yin, P., Gootenberg, J., and Abudayyeh, O. (2025). Generative distribution embeddings.

[7] Jiang, K., Yan, Z., Di Bernardo, M., Sgrizzi, S. R., Villiger, L., Kayabolen, A., Kim, B., Carscadden, J. K., Hiraizumi, M., Nishimasu, H., et al. (2024). Rapid in silico directed evolution by a protein language model with evolvepro. Science, 387(6732):eadr6006.

[8] Kapusniak, K., Potaptchik, P., Reu, T., Zhang, L., Tong, A., Bronstein, M., Bose, J., and Di Giovanni, F. (2024). Metric flow matching for smooth interpolations on the data manifold. Advances in Neural Information Processing Systems, 37:135011–135042.

[9] Lal, A., Garfield, D., Biancalani, T., and Eraslan, G. (2024). Designing realistic regulatory dna with autoregressive language models. Genome Research, 34(9):1411–1420.

[10] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130.

[11] Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. (2025). Boltz-2: Towards accurate and efficient binding affinity prediction. BioRxiv, pages 2025–06.

[12] Rood, J. E., Wynne, S., Robson, L., Hupalowska, A., Randell, J., Teichmann, S. A., and Regev, A. (2025). The human cell atlas from a cell census to a unified foundation model. Nature, 637(8048):1065–1071.

[13] Viñas Torné, R., Wiatrak, M., Piran, Z., Fan, S., Jiang, L., Teichmann, S. A., Nitzan, M., and Brbić, M. (2025). Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. Nature Biotechnology, pages 1–10.

[14] Wellcome Sanger Institute (2024). Generative and synthetic genomics programme. https://www.sanger.ac.uk/programme/generative-and-synthetic-genomics/. Accessed: 2025-05-27.

[15] Yu, T., Ekbote, C., Morozov, N., Fan, J., Frossard, P., d'Ascoli, S., and Brbić, M. (2025). Tissue reassembly with generative ai. bioRxiv, pages 2025–02.

[16] Zhang, J., Ubas, A. A., de Borja, R., Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan, U., Jones, M. G., Thompson, J. D., Tran, V., Pangallo, J., Papalexi, E., Sapre, A., Nguyen, H., Sanderson, O., Nigos, M., Kaplan, O., Schroeder, S., Hariadi, B., Marrujo, S., Salvino, C. C. A., Gallareta Olivares, G., Koehler, R., Geiss, G., Rosenberg, A., Roco, C., Merico, D., Alidoust, N., Goodarzi, H., and Yu, J. (2025). Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. bioRxiv.