Integrating Generative and Experimental Platforms for Biomolecular Design

Contents

1 Workshop Summary

Motivation. Fresh off a Nobel Prize in Chemistry, the field of biomolecular design and modeling aims to lower development costs and accelerate the discovery of solutions for medical, industrial, and environmental challenges. To do this, researchers in the field engineer proteins, ligands, and nucleic acids (e.g., DNA, RNA) to perform functions beyond those found in nature. Recently, generative machine learning (ML) has demonstrated remarkable potential to design novel and functional biomolecules [4]. Such extraordinary progress has been driven by the confluence of decades-worth of scientific experimentation data and the exponential growth in capabilities of both discriminative and generative ML. However, there remains a critical gap between biomolecular engineering and ML: many ML studies today strive for state-of-the-art outcomes on static benchmarks, often in isolation from their experimental counterparts. This disconnect may lead to misaligned evaluation metrics and impactful biological problems being overlooked. With generative modeling as a core focus of the current ML landscape, collaboration between biologists and ML experts is pivotal in ensuring research is geared toward addressing the most pressing challenges with experimental validations.

Overview. Our proposed workshop aims to align real-world biological problems with generative machine learning and foster interdisciplinary collaborations by appealing to experimental and computational attendees. The central themes of this workshop are:

1. Generative ML for biomolecular design.
    - *Inverse design.* While generative models are capable of designing diverse and novel biomolecules, an unsolved challenge is designing biomolecules for specific properties and constraints.
    - *Modeling biomolecular data.* The complexities of biological processes and the inherent experimental limitations are challenging for dataset generation and modeling. There is a need for domain-specific algorithms, capable of learning from both sparse and abundant data, that can uncover the governing mechanisms in these complex systems.
2. Integrating generative ML into the workflow of experimentalists.
    - *Adaptive experimental design.* After testing designs and conducting experiments, a unsolved challenge is determining how to best use the data to inform subsequent designs and experiments. Active learning, reinforcement learning, and Bayesian ML provide foundational approaches to develop relevant techniques.
3. Biological problems ripe for ML and development of useful *in-silico* oracles and benchmarks.
    - *Problem settings.* The workshop will foster dialogue on important biological problems that have abundant data or high-throughput data generation methods but lack satisfactory ML solutions.
    - *In-silico benchmarks.* The increasing utilization of *in-silico* oracles warrants investigation of how closely they align with real-world experiments. The workshop will encourage the development of robust and informative oracles.

Publication. The workshop is partnering with *Nature Biotechnology*, one of the most high-impact journals with a 5-year Journal Impact Factor of 56.9, where select papers will be invited for fast-track submission. This collaboration aims to elevate impact of the presented work and help attract top-tier submissions, especially from the biology sector, which is classically underrepresented at ML conferences. See Section 3 for details.

Last year's edition The GEM Workshop at ICLR 2024 in Vienna attracted 108 submissions from 142 institutions and 22 countries, where 56 papers were accepted. 7 papers were chosen for oral presentations and 7 papers were fast tracked to our partner journal, *Cell Systems*. 8 travel awards were given to students from underrepresented communities. We estimate that over 300 people attended the workshop, reflecting the growing interest and excitement around generative models in biology. The program featured a diverse set of topics, including cutting-edge research on protein, RNA, and biological circuit design, and applications of generative models for drug design. Attendees engaged in lively discussions during two dedicated poster sessions, a lively panel featuring young leaders in the field, and keynote talks from established, leading researchers, and more focused research presentations from selected trainees. The strong participation from both academia and industry underscored the relevance of generative AI to biological challenges and highlighted opportunities for future collaborations.

2 Logistics

2.1 Format

Our workshop will **be held in-person** with **virtual arrangements** for those unable to attend in-person. It will be a **large-attendance talk** format enriched with **two poster sessions**, a **panel discussion**, **contributed talks** for best papers at the workshop, and multiple **networking sessions**. We believe this format will strike a balance between delivering scientific advancements and building connections between the computationalists and experimentalists attending the workshop. We have planned multiple poster sessions and networking events to discourage attendees from breaking off into small, familiar groups and instead establish relationships with potential collaborators.

Virtual engagement. We will offer virtual engagement options for those unable to attend in person. Attendees will be able to access recorded sessions, posters, slides, and papers through our workshop website. We will require all accepted submissions to send us their posters and camera-ready papers as PDFs. In addition, we will provide links to papers published at our workshop and successfully fast-tracked in *Nature Biotechnology*.

2.2 Accessibility

Website. We will re-use our website[1] for the workshop. Accepted ML-track papers and biology-track abstracts will be released on our website before the workshop and maintained afterwards. We will feature talk titles and abstracts, as well as the final schedule for attendees. The sessions for each poster will also be made available as a public Google Sheets document. We will use OpenReview for the review process.

Resources. To promote mutual understanding between experimentalists and computationalists for productive and insightful discussions, we will provide tutorial videos covering key concepts in biomolecular design and generative ML. Furthermore, presenters will receive guidance on making their content more accessible and engaging for all attendees. We also will organize social mingles to promote networking between our diverse audience.

Travel awards. To foster diversity, equity, and inclusion (DEI), we will provide free workshop registration and fund travel for selected applicants using sponsorship funding, where the priority will be given to *students and minority groups based on DEI*.

2.3 Audience

Our audience would be diverse given the interdisciplinary nature of our workshop. Based on our previous year's experience and discussions with related workshop organizers, we anticipate around **300 attendees**. Our primary goal is to attract researchers working at the intersection of machine learning and biology. We also seek to engage pure ML researchers who are looking for applications for their work, as well as biologists who are exploring new machine learning techniques to address their problems. Additionally, we aim to draw in industry researchers actively involved in this field.

2.4 Seed Grants

This year, we have secured funding from Nvidia, Dimension Capital, and Dreamfold, and we are actively securing funding from Genetech, Microsoft Research, and other organizations—leaders in the intersection of generative ML and biology. This reflects their shared commitment to advancing innovations in biomolecular research. With these funds, we are launching small seed grants ($\sim$ \$1,000 each) to catalyze new collaborations between experimental and computational researchers attending the workshop. The goal is to provide initial support for promising projects that align with cutting-edge biological challenges and AI-driven solutions. To facilitate these collaborations, we will publicize a dedicated a "matchmaking" social event during the main conference, where participants with overlapping research interests can connect and brainstorm potential projects. Following this session, teams will be invited to submit concise, half-page proposals outlining their project idea, collaboration plan, and expected outcomes. Grants will be awarded based on the novelty of the idea, alignment with the workshop's themes, feasibility, and potential for impact. Funds will be disbursed after the workshop to the institutions to support project initiation. We envision these seed grants as a stepping stone toward meaningful, interdisciplinary partnerships. To further incentivize progress, we plan to invite the awardees back for short talks at the potential 2026 edition of the workshop, where they can share insights, showcase results, and inspire the next wave of AI and biology collaborations.

2.5 Broad Outreach

Our organizing and program committee members will leverage their extensive academic networks to increase workshop awareness in both generative ML and biology. This includes sharing event details within their institutions and with peers in the field. We will also collaborate with our sponsors and our connections in industry research labs, such as Valence, Recursion, Amgen, Genentech, Intel, and Microsoft, to disseminate information about this workshop. As we successfully did last year to extend our reach, we will employ various social media platforms, including Twitter, Facebook, LinkedIn, WeChat, and blog posts to foster interactions and engage with the general audience of the workshop.

2.6 Timeline

Main workshop deadlines:

- Workshop submission deadline: February 3rd, 2025.
- Workshop accept/reject notification date: March 5th, 2025.

We will then follow-up with selected papers for fast-tracking to *Nature Biotechnology*, where decisions are tentatively scheduled before May 2025.

3 Submissions

Submission tracks. The workshop submission is designed to attract high-quality original papers at the intersection between biomolecular design and generative AI. We will provide two separate submission tracks for topics described in Section 1:

[1] URL https://gembio.ai/

- **Machine learning track**. This track will feature generative machine learning advancements for biomolecular design where results are entirely *in silico*. The topics of the papers will include inverse design, biomolecular data modelling, and adaptive experimental design, *in silico* benchmarks.
- **Biology track**. This track will consist of papers which have *wet lab* experimental results. We will welcome hybrid works employing ML for experimental biomolecular design problems, as well as biological problem settings (e.g. high throughput techniques, single cell analysis) relevant to generative machine learning.

By providing a biology track and partnering with *Nature Biotechnology*, we hope to attract researchers from hybrid labs, as well as biologists who are pursuing state-of-the-art ML techniques for their research.

Submission guidelines. Both tracks will accept submissions up to 5 pages in length (excluding appendix). The biology track will also consider extended abstracts (up to 2 pages), similar to standard biological conferences. All submissions will be non-archival. We will be explicit in our Call for Papers that papers previously published at an archival venue will be rejected.

Review process. The review process will be double-blind, and they will be conducted through OpenReview. We anticipate around *100* submissions, based on our previous year's experience and discussions with organizers with related workshops. We aim to accept around *60* papers. We will recruit up to *100* reviewers from diverse backgrounds (see Section 10). Each submission will receive *3* reviews and each reviewer will review up to *3* submissions.

Conflict of interest. To prevent conflicts of interest, reviewers will not evaluate submissions from their department or from collaborators within the past 5 years.

Awards. Accepted papers with exceptional quality will be recognized through Best Paper and Distinguished Paper awards, as agreed upon by review scores and area chairs. Outstanding accepted papers will be selected for a series of contributed talks, offering a platform for further discussion among workshop attendees.

Partnership with *Nature Biotechnology* In collaboration with *Nature Biotechnology*, authors can opt for their paper to be considered for a fast-track review at *Nature Biotechnology*. Based on the workshop reviews, our organizing committee and editors at *Nature Biotechnology* will select high quality papers for an additional round of review at *Nature Biotechnology*. Accepted papers at *Nature Biotechnology* will form a special collection. This will be similar to our collaboration with *Cell Systems* in GEM-2024.

4 Tentative Schedule

We aim to create interdisciplinary discussions surrounding the formidable challenges in biology and how generative machine learning can tackle them. To create an engaging and inclusive workshop appealing to a diverse audience, our program features a variety of sessions. These sessions encompass distinguished keynote speeches (**invited talks**), selected contributed machine learning papers, experimental biology and *in-silico* modelling abstracts (**contributed talks**), engaging **poster sessions**, an insightful **panel discussion**, and **social mingles** in between and after sessions.

Invited talks. Each invited talk is structured with 25 minutes dedicated to the presentation and an additional 5 minutes reserved for questions. Each talk will focus on a theme of the workshop (see Section 1). We have invited emerging, young scientific leaders as speakers based on their well-known expertise, diverse scientific achievements, future research potentials, and exceptional presentation skills (see Section 5.1).

Contributed talks. For contributed talks, we will employ a rigorous peer-review selection process, guided by the diversity of topics and high reviewer scores, ensuring that we spotlight outstanding and impactful submissions. Our goal is that the talks strike a balance between biology and machine learning.

Poster sessions. Poster sessions follow contributed talks, offering a broader range of topics and a space for more personal and detailed conversations. Discussions around posters will foster connections and idea exchange amongst our participants. We purposely planned multiple poster sessions so presenters can also visit other interesting works.

Panel discussion. The panel discussion will spotlight senior leaders at the intersection of ML and biotechnology to discuss the past and the future of generative AI as applied to biomolecular design. We will identify current misalignment and challenges, unsolved biological problems where generative ML is set to disrupt the stage in the coming few years, as well as AI safety challenges. The panel discussion will be moderated by a workshop organizer, and will include a Q&A session with the audience (see Section 5.2).

Seed grant announcements Interested researchers can sign up to be grouped into pairs of experimentalists-computationalists at the beginning of the main conference. Each pair will brainstorm a concise research proposal to be shared in the workshop (only) with our panelists/speakers/sponsors during the second poster session. The rapid-fire session will enable researchers to pitch groundbreaking ideas to a panel of established researchers and venture capitalists, garnering immediate feedback and fostering collaboration. The selected proposals will be announced before closing remarks and will receive computational credits or cash prizes courtesy of our sponsors (see Section 2.4).

Social mingles. A key objective of the workshop is to bring experimentalists and computationalists together for collaborations. With support from our generous sponsors, we plan to provide lunch within the venue to foster continued interaction. Additional lunch activities, such as topic-specific, technique-based, or self-organized breakout sessions, can be arranged if the venue layout permits. We will also organize an after party to encourage further networking and idea exchange.

Tentative schedule	
8:50 - 9:00 AM	Open remarks
9:00 - 9:30 AM	Invited talk 1
9:30 - 10:00 AM	Invited talk 2
10:00 - 10:15 AM	Coffee break
10:15 - 11:00 AM	Contributed talks (4 talks)
11:00 - 12:00 PM	Poster session 1
12:00 - 1:00 PM	Lunch break
1:00 - 1:30 PM	Invited talk 3
1:30 - 2:30 PM	Panel discussion
2:30 - 2:45 PM	Coffee break
2:45 - 3:15 PM	Invited talk 4
3:15 - 3:45 PM	Contributed talks (3 talks)
3:45 - 4:40 PM	Poster session 2
4:40 - 4:50 PM	Seed grant awards
4:50 - 5:00 PM	Paper awards, closing remarks

5 Invited Speakers and Panelists

5.1 Invited Speakers

- Ava Amini (ava.amini@microsoft.com, confirmed) is a Senior Researcher at Microsoft Research. Her research focuses on developing new methods in machine learning and statistics, bioengineering, and nanotechnology and leveraged these approaches to to achieve new insights into cancer. In addition to research, she is a lead organizer and instructor for MIT Introduction to Deep Learning, MIT's official introductory course on deep learning.
- Florian Hollfelder (fh111@cam.ac.uk, confirmed) is a Professor in Chemical and Synthetic Biology at Cambridge University. He is an emergent leader in developing high-throughput experimental microfluidic devices for the discovery and understanding of functional proteins such as enzymes. More recently, he began to explore the intersection between microfluidics and generative machine learning for enzyme design.
- Brian Hie (brianhie@stanford.edu, confirmed) is an Assistant Professor of Chemical Engineering at Stanford University and an Innovation Investigator at Arc Institute. His research focuses on developing language-based generative machine learning for protein engineering and design. His work on structure prediction (ESMFold), genome-scale modelling (Evo), antibody design (ESM-IF1) have been recognized by journals such as *Science* and have become immensely useful tools in computational biology [2].
- Kotaro Tsuboyama (ktsubo@iis.u-tokyo.ac.jp, confirmed) is a Lecturer in Department of Chemistry and Biotechnology at University of Tokyo. His research focuses on experimental high-throughput analytical methods to study protein properties and functions, with an emphasis on de novo designed proteins. He employs biological and machine learning methods to study protein interaction prediction, protein stability, and cytosolic delivery of proteins.

5.2 Invited Panelists

- George Church (gchurch@genetics.med.harvard.edu, confirmed) is a Professor of Genetics at Harvard Medical School and Professor of Health Sciences and Technology at Harvard and MIT. He is Director of the U.S. Department of Energy Technology Center and Director of the National Institutes of Health Center of Excellence in Genomic Science. Renowned for his pioneering work in genome sequencing and synthetic biology, Church has made significant contributions to gene editing technologies, including CRISPR. His research also spans areas such as genome engineering, epigenetics, and personalized medicine. Church's work continues to shape the future of biotechnology, with potential applications ranging from disease treatment to bioengineering and environmental sustainability.
- Kyunghyun Cho (kyunghyun.cho@nyu.edu, confirmed) is a Professor of Computer Science and Data Science at New York University, CIFAR fellow, associate member of the National Academy of Engineering of Korea, and a senior director of frontier research at

the Prescient Design team within Genentech. His research interests span machine learning, natural language processing, and their applications in healthcare. He is widely recognized for his contributions to neural machine translation and recurrent neural networks, including the development of the GRU (Gated Recurrent Unit). He has served as a program chair of ICLR 2020, NeurIPS 2022 and ICML 2022, as well as serving as the founding co-Editor-in-Chief of the Transactions on Machine Learning Research.

- Barbara Cheifet (barbara.cheifet@us.nature.com, confirmed) is the Chief Editor of *Nature Biotechnology*. She holds a Ph.D. from Yale University, and spent 7 years at Genome Biology, including 4 years as Chief Editor, before joining *Nature Biotechnology* at the beginning of 2022 and becoming Chief Editor at the end of that year. *Nature Biotechnology*, with a 5-year Journal Impact Factor of 56.9, publishes new concepts in technology/methodology related to biological, biomedical, agricultural and environmental sciences as well as publishing commentary on the societal aspects of biotechnology research.
- Frank Noé (frank.noe@fu-berlin.de, confirmed) is a Professor at Freie Universität Berlin and a Microsoft Partner Research Manager in Microsoft Research (MSR) AI4Science. Frank has co-pioneered the Markov state modeling (MSM) approach for describing the long-time dynamics of proteins and other macromolecules, has developed several deep learning systems for molecular simulation, such as the Boltzmann Generator, and is an advocate of open research and software for the benefit of society.
- Sergey Ovchinikov (so3@mit.edu, tentatively confirmed) is a Professor of Biology at MIT. His lab focuses on using phylogenetic inference, protein structure prediction/determination, protein design, deep learning, energy-based models, and differentiable programming to tackle evolutionary questions at environmental, organismal, genomic, structural, and molecular scales, with the aim of developing a unified model of protein evolution.
- Charlotte Bunne (charlotte.bunne@epfl.ch, confirmed) is a tenure track assistant professor at EPFL. Her research aims to advance personalized medicine by utilizing machine learning and large-scale biomedical data. Charlotte Bunne's interdisciplinary research has won several (best paper) awards. Charlotte has been a Fellow of the German National Academic Foundation and is a recipient of the ETH Medal.
- Chaok Seok (chaok@snu.ac.kr, confirmed) is a Professor of Chemistry at Seoul National University and a CEO of Galux. Based on physical chemistry, polymer theory, and deep learning, her lab develops a biomolecular modeling program package GALAXY, which specializes in protein structure prediction, protein-ligand docking, and protein loop modeling. She aims to extend such technologies to design of new drugs and proteins.

6 Organizers and Biographies

Website page and email address in colored hyperlink. In bold we have highlighted previous organizing or related experience.

- Chenghao Liu (chenghao.liu@mail.mcgill.ca) is a PhD candidate at McGill University and Mila - Québec AI Institute, advised by Dima Perepichka and Yoshua Bengio. He is a co-founder of Dreamfold, a protein design start-up. He is a chemist by training, and his research is now focused on developing generative active learning methods for materials and drug discovery. **He was a co-organizer of the CQMF chemistry conference, and an organizer of the GEM-2024 workshop**
- Jarrid Rector-Brooks (jarrid.rector-brooks@mila.quebec) is a PhD candidate at the Université de Montréal and Mila - Québec AI Institute advised by Yoshua Bengio. His research aims to develop improved generative models specifically for the design of therapeutics with an eye towards high-throughput adaptive experimental design, along with a focus on designing performant methods for amortized sampling. **He was an organizer of the GEM-2024 workshop.**
- Soojung Yang (soojungy@mit.edu) is a PhD student at the Massachusetts Institute of Technology (MIT), advised by Rafael Gómez-Bombarelli. Her research focuses on exploring the protein conformational landscape by harnessing the power of machine learning and multiscale molecular simulations. Additionally, she collaborates with experimental scientists on therapeutic peptide discovery projects. **She was an organizer of the GEM-2024 workshop**
- Sidney Lisanza (lisanzas@gene.com) is a machine learning scientist at Prescient Design. He develops ML tools to expedite the protein design process both at the lead discovery stage and the subsequent optimization of candidates. He enjoys time with friends/family, being outside, listening to music, and preferably doing all simultaneously. **He was an organizer of the GEM-2024 workshop.**
- Francesca-Zhoufan Li (fzl@caltech.edu) is a bioengineering graduate student at the California Institute of Technology (Caltech), advised by Frances Arnold and Yisong Yue. Leveraging her experimental background, she focuses on predicting engineered protein functions with a particular interest in expediting wet-lab enzyme engineering. Her collaboration with Microsoft Research involved investigating the efficacy of pretrained protein language models for various protein engineering tasks. **She was an organizer of the GEM-2024 workshop.**
- Hannes Stärk (hstark@mit.edu) is a PhD student in EECS at MIT advised by Tommi Jaakkola and Regina Barzilay. His research focuses on generative models for biomolecules. He organizes and founded the annual Learning on Graphs Conference. He co-organized the ML on Graphs Workshop at WSDM 2022. He organizes the annual Molecular ML conference at MIT. **This is his first time co-organizing an ICLR workshop.**
- Jacob Gershon (jgershon@uw.edu) is a graduate student at the University of Washington within the Institute for Protein Design with David Baker. Jacob is working on developing deep generative models for de novo enzyme design, hoping to someday use these tools to design new materials that support sustainable fashion. **This is his first time co-organizing an ICLR workshop.**
- Lauren Hong (lauren.hong@duke.edu) is a PhD student in Biomedical Engineering at Duke University, advised by Pranam Chatterjee. As an experimentalist in a hybrid lab, Lauren focuses on leveraging generative language model-derived peptide binders to post-

translationally edit proteins for broad-scale therapeutic applications. She has co-organized the Nvidia Duke AI Day as well as the Quantitative Biodesign Seminar at Duke. **This is her first time co-organizing an ICLR workshop.**

- Pranam Chatterjee (pranam.chatterjee@duke.edu) is an Assistant Professor of Biomedical Engineering and Computer Science at Duke University. Research in his Programmable Biology Group exists at the interface of computational design and experimental engineering, specifically employing generative ML to design programmable proteins for genome, proteome, and cell engineering. He completed his SB, SM, and PhD from MIT and is the founder of two startups, Gameto, Inc. and UbiquiTx, Inc., that leverage AI to design the next generation of fertility solutions and therapeutic biologics, respectively. **He was an organizer of the GEM-2024 workshop.**
- Tommi Jaakkola (tommi@csail.mit.edu) is the Thomas Siebel Professor of Electric Engineering and Computer Science at MIT. His research covers theory, algorithms, and applications of machine learning, from statistical inference and estimation to natural language processing, computational biology, as well as recently machine learning for chemistry. His awards include Sloan research fellowship, AAAI Fellow, and many publication awards across the research areas. **He was an organizer of the GEM-2024 workshop.**
- Regina Barzilay (regina@csail.mit.edu) is a School of Engineering Distinguished Professor of AI & Health in the Department of Computer Science and the AI Faculty Lead at MIT Jameel Clinic. She develops machine learning methods for drug discovery and clinical AI. In the past, she worked on natural language processing. Her research has been recognized with the MacArthur Fellowship, an NSF Career Award, and the AAAI Squirrel AI Award for Artificial Intelligence for the Benefit of Humanity. **She was an organizer of the GEM-2024 workshop.**
- David Baker (dabaker@uw.edu) is the Henrietta and Aubrey Davis Endowed Professor in Biochemistry at the University of Washington. He serves as the director of the Rosetta Commons, a consortium of labs and researchers that develop biomolecular structure prediction and design software. He has co-founded many biotechnology companies such as Prospect Genomics, Icosavax, Charm Therapeutics, and Xiara Therapeutics. His work has pioneered computational and experimental methods for protein structure prediction and design, earning him the 2024 Nobel Prize in Chemistry. **He was an organizer of the GEM-2024 workshop.**
- Frances Arnold (frances@cheme.caltech.edu) is the Linus Pauling Professor of Chemical Engineering, Bioengineering, and Biochemistry and the Director of the Donna and Benjamin M. Rosen Bioengineering Center at Caltech. Renowned for pioneering directed evolution, she has been recognized by numerous awards, including the 2018 Nobel Prize in Chemistry. She has co-founded companies such as Gevo and Provivi and served on the boards for companies including Alphabet, Illumina, and Generate Biomedicines. Since January 2021, she has been an external co-chair of President Joe Biden's Council of Advisors on Science and Technology (PCAST). **She was an organizer of the GEM-2024 workshop.**
- Yoshua Bengio (yoshua.bengio@mila.quebec) is a Full Professor in the Department of Computer Science and Operations Research at Université de Montréal, as well as the Founder and Scientific Director of Mila and the Scientific Director of IVADO. Considered one of the world's leaders in artificial intelligence and deep learning, he is the recipient of the 2018 A.M. Turing Award. He is a Fellow of both the Royal Society of London and Canada, an Officer of the Order of Canada, and a Canada CIFAR AI Chair. **He was an organizer of the GEM-2024 workshop.**

7 Diversity, Equity, and Inclusion

We are dedicated to the cause of diversity, equity, and inclusion (DEI) in our proposed workshop. Our efforts span a wide spectrum of academic disciplines, cultural backgrounds, personal experiences, and identities, ensuring an inclusive environment for all. We are committed to creating a space where each individual feels valued and welcomed, regardless of ethnicity, gender, sexual orientation, affiliations, nationality, seniority, abilities, socioeconomic status, religion, backgrounds, experiences, viewpoints, perspectives, and beyond.

Our organizing committee members come from **8 institutions**[2] with expertise in more than **6 academic disciplines**[3]. Our organizers identify with more than **9 cultural backgrounds**[4] with **5 female organizers**[5] and include first-generation students[6].

We have **6 professors**, **7 PhD students**, and **1 industry scientist** organizers who work on a wide range of problems related to biomolecular design. Some of our student organizers are first time workshop organizers, while others bring prior experience from organizing conferences or events in various organizations (see Section 6). We provide many industry viewpoints through working or interning at (bio)tech companies such as Genetech and Microsoft Research on scientific applications with ML. Several of organizers are founders or scientific advisors to biotech start-ups and established pharmaceutical companies. The experience level of our student organizers span from 2nd to 5th year PhD students while our professors span from assistant to full professors.

We have thoughtfully invited speakers and panelists with DEI in mind. Each speaker and panelist represents a different academic institution or industry lab. Three out of seven of our panelists and one invited speaker are female, and they are all accomplished individuals in their field. Our panelists and speakers span three continents, Asia, Europe, and North America. We selected speakers and panelists who come from different backgrounds (e.g. computer science, experimental biology, chemical engineering, computational chemistry) and whose topics are orthogonal.

[2] Mila, MIT, Caltech, University of Washington, Duke University, Prescient Design, McGill University, and Université de Montréal.
[3] Including computer science, computational biology, biochemistry, physical chemistry, chemical engineering, and bioengineering
[4] Including Canada, China, Finland, France, India, Germany, Israel, Kenya, and South Korea
[5] Soojung, Francesca, Lauren, Regina, and Frances
[6] Jarrid, Soojung, and Francesca

In our commitment to fostering DEI, we are also dedicated to increasing global participation in our workshop. To facilitate international travel to the extent within our control, we will collaborate with the ICLR main organizers and provide assistance in obtaining invitation letters for those who may require them for visa and travel-related purposes. Our aim is to ensure that individuals from across the globe have the opportunity to join us, share their insights, and contribute to our collective learning and growth.

Finally, our workshop will promote a venue for safe, non-judgemental, and respectful discourse. We will set-up anonymous communication channels (phone, email, and form) to report any misbehaviour or DEI related issues.

8 Previous Related ICLR Workshops

An undoubtedly impactful application is combining ML with scientific applications. We list the most related workshops at previous ICLR conferences in order of relevance (first is most relevant).

- Machine Learning for Drug Discovery Workshop. ICLR 2022-2023. This workshop focuses on optimizing and discovering therapeutic candidates. Our workshop focuses on general purpose design of biomolecules – proteins, RNA/DNA – and more so on the generative perspective. The broader scope allows us to explore diverse applications; for example, enzyme engineering can significantly contribute to plastic degradation and gene editing with CRIPSR. Both workshops emphasize bringing ML closer to real-world evaluation and using ML as a core part of biological experimentation. An distinctive feature of our workshop is our aim to encourage the involvement of experimentalists, who have been underrepresented in prior ML conferences and workshops, despite their crucial role in generative biology. To achieve this goal, our workshop offers dedicated tracks and collaboration with a high impact biology journal, and our speakers and panelists come from a wide variety of backgrounds, ranging from experimentation to computational modeling.
- Deep Generative Models for Highly Structured Data. ICLR 2019 & 2022. This workshop focuses on incorporating structure into generative models for downstream use in real-world data modalities. While proteins and biological data is one application, it is not the focus whereas it is in our workshop. Furthermore, we focus on how to integrate generative models into real-world biological experiments.
- Machine Learning for Materials. ICLR 2023. The goals of this workshop are similar to ours but focused on materials. Here they emphasize identifying the unique challenges with designing useful materials and uncovering the meaningful tasks for novel ML techniques to be developed. We share the common aim of emphasizing the complexity of working with biomolecules and the need for directing ML research towards the most pressing problems in biomolecular design.
- Physics for Machine Learning. ICLR 2023. This workshop focuses on developing ML applications for physics. They also emphasize bringing ML closer to real-world applications in the physical sciences.

Why this workshop (and why now). It is an exciting time to work at the intersection of biology and ML. AlphaFold [1] undoubtedly shifted the direction of many ML researchers to work on structural biology and bioengineering applications. Experimental biology is reaching a inflection point of data generation becoming cheaper and faster than ever. Works such as RFdiffusion [4] published in *Nature* is a example of novel ML research combined with experimental validation to achieve unprecedented biomolecular design success [4]. In fact, our co-organizer David Baker was recently awarded the Chemistry Nobel Prize this year because of his contributions in computational protein design. The advent of ChatGPT and foundation models have also provided a possible road map to developing interactive systems between scientists and AI for scientific discovery [3]. However, the ChatGPT paradigm cannot be transferred to biology where data cannot be readily annotated or generated at scale. There will need to be important breakthroughs in generative AI tailored to biology and adaptive experimental design to reduce cost and time. We believe the time is ripe to bring together the different disciplines to chart a path towards rapid scientific discovery in the age of AI.

9 Sponsors

Nvidia, Dimension Capital, and Dreamfold have confirmed to be sponsors for the workshops, and Genentech and Microsoft has tentatively confirmed to be a sponsor. Their exact monetary contributions have yet to be confirmed but we expect at least $20,000. We are in active discussions with other industrial research labs for sponsorship. The money from sponsors will be used in the following ways ordered from highest to lowest priority:

1. Registration and travel grants for minority groups and students.
2. Lunch catering.
3. Best paper awards.
4. Seed funding for collaborative proposals.
5. Post conference gathering at a nearby restaurant.

10 Program Committee

The role of program committee members will be to help review the workshop submissions. Each organizer will recruit enough reviewers to provide 3 reviews for each submission. Since each organizer comes from a different lab with different personal connections, we expect

to have a rich pool or reviewers. We have prepared a list of reviewers who have tentatively confirmed to be reviewers. We have additional ways throughout our diverse connections to seek more reviewers if needed.

Patrick J. Almhjell, Ph.D., Simon Axelrod, Ph.D., Ava P. Amini, Ph.D., Emmanuel Bengio, Ph.D., Joey Bose, Ph.D., Tim-Henrik Buelles, Ph.D., Bianca Dumitrascu, Ph.D., Michael Galkin, Ph.D., Alex Hernandez-Garcia, Ph.D., Riashat Islam, Ph.D., Kadina Johnston, Ph.D., Michal Koziarski, Ph.D., Alex X. Lu, Ph.D., Pablos Lemos, Ph.D., Ge Liu, Ph.D., Sulin Liu, Ph.D., Santiago Miret, Ph.D., Ariane Mora, Ph.D., Nathan Frey, Ph.D., Andrei Nica, Ph.D., Ladislav Rampasek, Ph.D., Seongok Ryu, Ph.D., Lena Simine, Ph.D., Almer van der Sloot, Ph.D., Alexander Tong, Ph.D., Kevin K. Yang, Ph.D., Zichao Yan, Ph.D., Bruce Wittmann, Ph.D., Zach Wu, Ph.D., Tara Akhound-Sadegh, Lucas Arnoldt, Ron Boger, Shahar Bracha, Paul Bertin, Maria Carreira, Tianlai Chen, Itamar Chinn, MinGyu Choi, Felix Faltings, Jacob Gershon, Prashant Govindrajan, Sarah Gurev, Guillaume Huguet, Ian Humphreys, Moksh Jain, Andrew Kirjner, Maksym Korablyov, Daniel Levy, Seokhyun Moon, Sean Murphy, Peter Mikhael, Juno Nam, Lena Nehale Ezzine, Umesh Padia, Andrei Rekesh, Raman Samusevich, Wenxian Shi, Luca Thiede, Brian Trippe, Veronica Tarka, Tony Tu, Pascal Sturmfels, Akshay Subramanian, Allen Tao, Hannes Stark, Sophia Vincoff, Sasha Volokhova, Rachel Wu, Jason Yang, Dinghuai Zhang, Yinnuo Zhang, Wonho Zhung.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873): 583–589, 2021.

[2] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[3] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[4] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.