AI FOR NUCLEIC ACIDS (AI4NA)

1 WORKSHOP MOTIVATION AND DESCRIPTION

AI community has focused on proteins. Since the publication of AlphaFold2 Jumper et al. (2021) in 2021, there has been a huge wave of interest in AI-driven protein research. This breakthrough has had a profound impact on structural biology, drug discovery, and biotechnology, leading to new biological insights and advanced AI tools for the design and engineering of proteins. Similarly, machine learning conferences have seen a surge of papers applied to structural biology and drug design, but most work has focused on proteins and small molecules. Although AlphaFold2's success has also drawn attention to nucleic acids—RNA and DNA—plenty of opportunities for AI for nucleic acids research still remain to be explored. With this workshop, we aim to shift the spotlight to nucleic acids, hoping to spark collaborations and innovation at the intersection of machine learning and nucleic acids research. The workshop will discuss the unique challenges of modeling nucleic acids compared to proteins, towards driving real-world applications and the impact of AI research in diagnostics, therapeutics, and biotechnology.

Why AI for nucleic acids? Motivated by the success of AlphaFold, several deep learning methods for RNA tertiary structure prediction have been developed Shen et al. (2022); Wang et al. (2023); Li et al. (2023b); Baek et al. (2024); Pearce et al. (2022). Yet, RNA structure prediction has not reached its 'AlphaFold moment', trailing behind protein structure prediction in terms of both performance and data scale by a decade Schneider et al. (2023). This underscores the two biggest challenges of RNA structure modeling: the limited number and relatively poor documentation of RNA structure data. In contrast, the amount of unlabeled nucleic acid sequence data is huge, which has led to the emergence of RNA and DNA language models Chen et al. (2022); Zhou et al. (2023); Li et al. (2023a); Nguyen et al. (2024); Penić et al. (2024). Language models are used as foundation models for a range of downstream function or property prediction tasks with limited labeled sequences.

Several emerging research directions make AI for nucleic acids uniquely challenging: Beyond individual RNA structures, modeling intermolecular interactions Kang et al. (2020); Sun et al. (2021); Abramson et al. (2024) is critical for the design and discovery of nucleic acid medicine. Examples include designing mRNA therapeutics and predicting their stability and degradation proneness Zhang et al. (2023), as well as RNA-targeting small molecules and finding druggable RNA binding pockets Childs-Disney et al. (2022). Another exciting line of work explores generative AI techniques, inspired by their success in protein design, for designing bespoke RNAs with desired functional properties Joshi et al. (2024); Nori & Jin (2024); Anand et al. (2024). Deep learning models that improve the accuracy of DNA sequencing and assembly reliability are crucial for identifying genetic markers for diseases, enabling researchers to identify potential targets for new therapeutics and advance precision medicine Poplin et al. (2018); Baid et al. (2023); Stanojevic et al. (2024); Vrček et al. (2024).

Aims of workshop. Nucleic acids are at the forefront of biology today: the last two Nobel Prizes in Physiology or Medicine were awarded for RNA research. Nucleic acids are central to understanding how life works as well as the focus of recent pharmaceutic innovations. Our goals are to introduce the AI community to cutting-edge research challenges in modeling RNA and DNA, as well as catalyze the development of AI systems that advance our understanding of nucleic acids. Additionally, we aim to highlight the unique challenges associated with nucleic acid data, such as its complexity and the difficulty in applying familiar techniques like data augmentation, which differ significantly from more intuitive domains like images or text.

Hosting a workshop on nucleic acids at ICLR, one of the most impactful international machine learning conferences, will serve as a bridge between domain experts from drug discovery and biomedicine with top AI researchers who may not have worked on this type of biological data but are keen to make a positive real-world impact. Thus, we hope that the AI for Nucleic Acids workshop will bring the community together, foster new collaborations, and drive AI research towards direct benefits in nucleic acids biology and therapeutics.

1.1 SCOPE

This workshop will aim to popularize AI applications for nucleic acids and introduce nucleic acid research challenges to the broader AI community. Thus, the topics will focus on applications of AI and novel AI methods for RNA and DNA research including, but not limited to:

- **Nucleic Acid Structure and Function** (RNA secondary and tertiary structure prediction, RNA function analysis, NA interactions) - the sequence-structure-function paradigm refers to the fact that the sequence determines the structure, and structure determines the function of RNA molecules. For structure prediction, the main problem is that current models struggle with accuracy and generalization across different RNA molecules and the goal would be to achieve the accuracy obtained for proteins (e.g. AlphaFold). For function, we need AI models that predict RNA functions based on their sequence, structure, and/or interactions, consequently discovering functional elements of the molecule. Furthermore, developing models that accurately predict interactions between RNA/DNA and other RNA, DNA, protein, or other small molecules is especially difficult, but also important for understanding both cellular processes and developing new drugs.
- **Foundation and Generative Models for Nucleic Acids** (Multimodal NA foundation models, Generative models for NAs) - developing ML models specialized for nucleic acids. The primary problem with RNA compared to proteins is that there is far less comprehensive structural, functional, and experimental data available for RNA. Foundation models, pre-trained on massive amounts of (unlabeled) data, can be particularly useful in such areas where data might be scarce or fragmented. Additionally, combining different modality data may be a successful way to deal with the data scarcity problem, e.g. combining sequences with structural probing data. Aside from the foundation models, a growing number of generative models for NAs have emerged recently. Some of the attempts include generating sequences from given structures and generating optimal mRNA sequences for vaccine design. Advancements in such models could also lead to breakthroughs in designing new RNA/DNA molecules for therapeutic or biotechnological purposes.
- **Nucleic Acids in Therapeutics** (NA drug design and discovery, NA modification, NA mutations) - designing drugs that target nucleic acids is still in the early stages. The goal is to use AI to find and optimize drug molecules for better targeting and binding to specific nucleic acid structures. In addition, the roles of modifications and impacts of mutations are still not properly understood. Any advancements in these areas help with a better understanding of these changes, which can be key in understanding genetic diseases, developing precision medicine, interpreting gene expression, etc.
- **Genomic Data Analysis** (Genome reconstruction, Gene expression, Calling genetic variants, Pairwise and multiple NA sequence alignment, Single-cell transcriptomics and genomics) - understanding gene expression patterns and accurate variant calling is crucial for understanding disease predisposition, personalizing treatments. It helps in both basic biological research and applied fields like drug development. On the other hand, better alignment methods are key to many genomics applications, such as genome assembly, evolutionary studies, and variant calling. Current methods struggle with scalability and accuracy in aligning large numbers of sequences. Finally, the complexity and noise inherent in single-cell data make it difficult to extract meaningful insights. Dealing with noisy data and combining different data modalities may bring advancements in these fields.

2 LOGISTICS

2.1 FORMAT

The workshop will be held in person. For those unable to attend in person, we will provide recorded sessions, slides, posters, and the online proceedings of the accepted workshop papers. If an invited speaker cannot attend the workshop, we will make sure the speaker will attend virtually over a communication platform. If an author of a selected paper for the poster presentation cannot attend due to exceptional circumstances, we will allow them to publicize their work on the workshop website.

2.2 AUDIENCE

We anticipate around 150 attendees which amounts to roughly 3% of the total ICLR 2025 participants. We expect a diverse audience at the intersection of machine learning and biology. Additionally, we seek to engage ML practitioners looking for interesting problems in nucleic acids research, and biologists who are exploring new AI methods to apply to their problems. We also expect to have industry researchers specialized in this field of research.

2.3 REACH OUT

We plan to actively promote the workshop through a variety of channels, including our website, social media platforms, and partnerships with both academic and industry collaborators. With a diverse group of organizers from different institutions, we will be able to reach a wide audience by leveraging our collective networks and connections. Additionally, we will collaborate with our sponsors to spread information about this workshop, helping us reach researchers from the industry as well.

2.4 ACCESSIBILITY.

We will establish a web page for the workshop, which will be a central place for the calls for papers and reviewers, promoting the workshop, and providing the planned time schedule and talk titles.

2.5 SUBMISSIONS

We will seek high-quality original submissions that fit the AI4NA scope. We will consider full-paper submissions up to five pages in length (excluding references and appendix) and tiny-paper submissions up to two pages in length (excluding references, appendix and URM statement). The reviews will be double-blind, ensuring fairness in the process. The review process will be conducted through OpenReview. The reviewers will be instructed to ensure novelty and to mark any previously published work. The accepted papers will be available on the workshop website, but it will be communicated that the accepted papers are non-archival and can be published elsewhere in the future.

2.6 TENTATIVE TIMELINE

- Call for papers: 16 December 2024
- Submission deadline: 3 February 2025
- Reviewing period: 3 February - 26 February 2025
- Notification: 5 March 2025
- Workshop: 27 or 28 April 2025

3 TENTATIVE SCHEDULE

Invited talks. The workshop will feature six invited talks from industry and academic leaders. Each invited talk is structured with 25 minutes reserved for presentation and 5 minutes allocated for questions to encourage discussion. Each talk will cover one of the themes from the workshop scope. The speakers are well-known in their respective fields, with significant scientific achievements and great presentation skills.

Contributed talks. We plan to have three slots for two contributed talks in each. Contributed talks will have 10 minutes reserved for presentation. Contributed talks will feature the most interesting and outstanding peer-reviewed papers submitted to the workshop. At least one contributed talk slot will be dedicated to tiny-paper submissions. That way we will give visibility and opportunity for talented young researchers to present their ongoing or novel machine learning work and to support the under-representative minority.

Poster sessions. The workshop will have two 60-minute poster sessions. The poster sessions will feature both full and tiny papers. The poster sessions will give opportunities for discussion and bolster the interaction between the participants, allowing for more personal and detailed conversations.

Panel discussion. We will have a 40-minute panel discussion led by a moderator and 10 minutes reserved for Q&As. The idea is to discuss current challenges in drug design from the perspective of nucleic acids and identify where and how AI methods could help or improve the present state.

Coffee and lunch breaks. We plan to have two 10-minute coffee breaks and one 60-minute lunch break to further encourage the discussion between the participants.

Tentative schedule	
Time	Description
9:00 - 9:10	Opening remarks
9:10 - 9:40	Invited speaker 1
9:40 - 10:10	Invited speaker 2
10:15 - 10:25	Coffee Break
10:25 - 10:45	Contributed talks (2 talks)
10:45 - 11:15	Invited speaker 3
11:15 - 12:15	Poster session 1
12:15 - 13:15	Lunch break
13:15 - 13:45	Invited speaker 4
13:45 - 14:15	Invited speaker 5
14:15 - 14:25	Coffee break
14:25 - 14:45	Contributed talks (2 talks)
14:45 - 15:35	Panel discussion
15:35 - 16:35	Poster session 2
16:35 - 17:05	Invited speaker 6
17:05 - 17:25	Contributed talks (2 talks)
17:25 - 17:30	Closing remarks

4 Invited Speakers and Panelists

All of the invited speakers and panelists are confirmed unless explicitly denoted as tentative.

4.1 Invited Speakers

Yang Zhang (e-mail: zhang@nus.edu.sg) is a Professor at the School of Computing and the Yong Loo Lin School of Medicine at National University of Singapore (NUS), and the Cancer Science Institute of Singapore. Prior to joining NUS, Dr Yang Zhang worked as a Professor at the University of Michigan. His research interests are in AI and deep neural network learning, protein folding and structure prediction, and protein design and engineering. The honors that Dr Zhang received include the Alfred P Sloan Award, the US National Science Foundation Career Award, and the University of Michigan Basic Science Research Award. He has been recognized as a Thomson Reuters/Clarivate Analytics Highly Cited Researcher seven times since 2015.

Kishwar Shafin (e-mail: shafin@google.com) is a senior research scientist in Google genomics team where he leads the development of DeepVariant, DeepTrio, and DeepSomatic. Prior to this, he obtained PhD in Biomolecular Engineering and Bioinformatics from the University of California. His work was published in scientific journals such as Nature Biotechnology, Nature Methods, New England Journal of Medicine, Science, and Cell Genomics, among others. He's a member of the Telomere-to-Telomere (T2T) consortium, the Human pangenome reference consortium, and the Genome-In-A-Bottle (GIAB) consortium. In PrecisionFDA Truth Challenge V2, his work won awards for best performance in difficult-to-map and all benchmarking regions for Oxford Nanopore variant calling.

Wengong Jin (e-mail: w.jin@northeastern.edu) is an assistant professor at Khory College of Computer Sciences at Northeastern University. He is also a visiting research scientist in the Eric and Wendy Schmidt Center at Broad Institute. He obtained his PhD at MIT CSAIL, advised by Prof. Regina Barzilay and Prof. Tommi Jaakkola. His research focuses on geometric and generative AI

models for drug discovery and biology. His work has been published in journals including ICML, NeurIPS, ICLR, Nature, Science, Cell, and PNAS, and covered by such outlets as the Guardian, BBC News, CBS Boston, and the Financial Times. He is the recipient of the BroadIgnite Award, Dimitris N. Chorafas Prize, and MIT EECS Outstanding Thesis Award.

Charlotte Bunne (email: charlotte.bunne@epfl.ch) is an assistant professor at EPFL in the School of Computer and Communication Sciences (IC) and School of Life Sciences (SV). Before, she was a PostDoc at Genentech and Stanford and completed a PhD in Computer Science at ETH Zurich working with Andreas Krause and Marco Cuturi. During her graduate studies, she was a visiting researcher at the Broad Institute of MIT and Harvard hosted by Anne Carpenter and Shantanu Singh and worked with Stefanie Jegelka at MIT. Her research aims to advance personalized medicine by utilizing machine learning and large-scale biomedical data. Charlotte has been a Fellow of the German National Academic Foundation and is a recipient of the ETH Medal.

Quaid Morris (e-mail: mammene@mskcc.org) is a Full Member in the Computational and System Biology program at Sloan Kettering Institute and co-Director of the Graduate Program in Computational Biology and Medicine at Weill-Cornell Graduate School. Previously, he held a CCAI chair through the Vector Institute for Artificial Intelligence (AI); was a full professor at the University of Toronto, where he still holds courtesy appointments in Molecular Genetics and Computer Science; and an associate researcher at the Ontario Institute of Cancer Research. Quaid pursued graduate training and research in machine learning at the Gatsby Unit and obtained his PhD in Computational Neuroscience from Massachusetts Institute of Technology. Morris lab (http://www.morrislab.ai/) uses machine learning and artificial intelligence to do biomedical research, focusing on cancer genomics, gene regulation, and clinical informatics.

Stephan Eismann (e-mail: stephan@atomic.ai) leads the Machine Learning Team at Atomic AI, where he focuses on developing new technology to discover RNA-targeting medicines. Prior to joining Atomic, he did his PhD in the AI Laboratory at Stanford University where his research centered on novel machine learning algorithms to tackle complex problems in biomolecular structure prediction. Originally from Germany, he studied physics in Heidelberg and London before moving to the US. He is a co-founder of the Machine Learning for Structural Biology workshop at NeurIPS, and his research has been published in Science and Nature as well as at NeurIPS and ICML.

4.2 Invited Panelists

Patrick Schwab (e-mail: patrick.x.schwab@gsk.com) is a Senior Director of Machine Learning and Artificial Intelligence and Head of the Biomedical AI group at GSK.ai. His work aims to advance personalized medicine by utilizing machine learning, computational systems biology methods and large-scale health data, such as genetics, multi-omics, cell-based assays, and continuous measurements from smart devices and electronic health records, to better understand and treat complex diseases. Prior to joining GSK, he was a Principal Architect working on Machine Learning for Personalised Medicine at Roche in Basel, Switzerland and at Genentech in South San Francisco, US. Before joining Roche, he was a doctoral researcher working on Machine Learning for Healthcare at ETH Zurich. Prior to ETH Zurich, he spent 5 years building custom data-driven software solutions in industry. He holds a PhD in Machine Learning (2019) from ETH Zurich, Switzerland.

Edith M. Hessel (e-mail: edith@relationrx.com) is a Chief Science Officer at Relation Therapeutics. She is an immunologist with over 20 years' experience in drug discovery and development. She has successfully pioneered novel target discovery platforms and advanced multiple therapeutics from inception to clinical proof of concept, both in biotech and pharma. Edith previously worked at Dynavax Technologies and then GSK, where she was responsible for the creation of the Refractory Respiratory Inflammation Discovery Performance Unit, adding multiple programs to GSK's pipeline. Subsequently, Edith co-founded Mestag Therapeutics as Chief Scientific Officer, and then served as CSO at Eligo Bioscience in Paris. Edith currently serves on the Board of the Fraunhofer ITEM Institute and is a Trustee on the Board of the British Society for Immunology. She received her PhD from Utrecht University and trained as a postdoc at DNAX Research Institute in California. In recognition of her scientific accomplishments, she was awarded an Honorary Professorship at the University of Manchester.

Jennifer Listgarten (e-mail: jennl@berkeley.edu) [*Tentative*] is a Professor in the Department of Electrical Engineering and Computer Science, the Center for Computational Biology, and the Bioengineering program at the University of California, Berkeley where she holds the Jeffrey Huber and Angel Vossough Chancellor's Chair in Computational Biomedicine. She is also a member of the steering committee for the Berkeley AI Research (BAIR) Lab. From 2007 to 2017 she was at Microsoft Research, through Cambridge, MA (2014-2017), Los Angeles (2008-2014), and Redmond, WA (2007-2008). She completed her Ph.D. in the machine learning group in the Department of Computer Science at the University of Toronto, located in her hometown. She has two undergraduate degrees, one in Physics and one in Computer Science, from Queen's University in Canada. Jennifer's research interests are broadly at the intersection of machine learning, applied statistics, molecular biology and science. Her current research is primarily in understanding how machine learning can be used to advance protein engineering.

5 Organizers

5.1 Junior Organizers

Ivona Martinović (e-mail: ivona_martinovic@gis.a-star.edu.sg) is a PhD student at the National University of Singapore (NUS), affiliated with the Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), where she was awarded the Singapore International Graduate Award (SINGA). She obtained her Bachelor's and Master's in Computer Science at the University of Zagreb, Faculty of Electrical Engineering and Computing. Her research focuses on deep learning for RNA structure and function-related tasks. She was a co-organizer of Job Fair Meetup 2021, a five-day event designed to connect technology experts with curious students through talks, panel discussions, and workshops, which brought together over 50 companies and attracted more than 400 attendees.

Lovro Vrček (e-mail: lovro_vrcek@gis.a-star.edu.sg) is a PhD student at the University of Zagreb, Faculty of Electrical Engineering and Computing, and is also affiliated with Genome Institute of Singapore (GIS), Agency of Science, Technology, and Research (A*STAR). Prior to this, he obtained a Master's degree in theoretical physics at the University of Zagreb, Faculty of Science. His research focuses on employing graph neural networks to *de novo* genome assembly.

Chaitanya Joshi (e-mail: ckj24@cam.ac.uk) is a 3rd year PhD student at the Department of Computer Science, University of Cambridge, supervised by Prof. Pietro Liò. His research develops geometric deep learning methods for structural biology with applications to RNA design. His work has been awarded the Qualcomm Innovation Fellowship and A*STAR National Science Scholarship. Previously organized conferences: Learning on Graphs Conference (LoG) 2022, 2023.

Tin Vlašić (e-mail: tin_vlasic@gis.a-star.edu.sg) is a postdoctoral research fellow at the Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR). He obtained his PhD at the University of Zagreb, Faculty of Electrical Engineering and Computing, where he was awarded the Outstanding PhD thesis award. In the academic year 2021-2022, he was a Swiss Government Excellence Scholar with the Department of Mathematics and Computer Science, University of Basel. His research focuses on RNA language models and employing deep learning for RNA structure prediction.

Agata M. Kilar (e-mail: akilar@fas.harvard.edu) is a postdoctoral researcher at Harvard University, working with Elena Rivas. She earned her PhD at Masaryk University in the Czech Republic, where her research focused on discovering telomerase RNA genes across the Viridiplantae group. During her PhD studies, she was awarded the Barrande Fellowship and received the Visegrad Scholarship three times for her work on tracking telomerase RNA using computational methods. Currently, Agata is developing eHMMER, a method that enhances sequence homology sensitivity by integrating time-dependent evolutionary models into the profile HMM framework. She is also passionate about science outreach; during her undergraduate studies, she organized a science festival and has participated in numerous activities to share science with high school students.

5.2 Senior Organizers

Bruno Trentini (e-mail: brunod@nvidia.com) is a PhD candidate at the University of Oxford, and he holds an MSc in Data Science and Machine Learning from University College London. He was part of the organizing team of the ML4LMS workshop at ICML with over 200 attendees. His research focuses on AI applied to structural biology with an emphasis on geometric deep learning methods. Previously, he led Life Sciences Research Alliances at NVIDIA, where he now works as an Applied Research Scientist.

Max Ward (e-mail: max.ward@uwa.edu.au) is currently a Lecturer (equiv. Assistant Prof) at The University of Western Australia. He completed a PhD (2019) at The University of Western Australia, then spent several years at Google, and then did postdocs at the University of Adelaide and finally at Harvard University. Dr Ward was a finalist for the ACS 1962 Medal for the best PhD thesis in Western Australia and received a commendation from UWAs 2024 Early-Career Research awards in Physics Mathematics and Computing. He was on the program committee of ISMB/ECCB in 2023 and 2024, program committee for AJCAI 2022, and organised the RNA Design workshop at Computational Approaches to RNA Structure and Function in Benasque 2024.

Maria Brbić (e-mail: maria.brbic@epfl.ch) is an Assistant Professor of Computer Science and of Life Sciences at EPFL. She develops new ML methods and applies her methods to advance biology and biomedicine. Prior to joining the EPFL, she was a postdoctoral researcher at Stanford University. Her work has been recognized by many awards and recognitions such as MIT Rising Star, Outstanding PhD thesis award and Early Career Award by SIB. Previously organized workshops: ICML 2023 Workshop on Computational Biology, NeurIPS 2023 Workshop on AI4Science, ICLR 2024 Machine Learning for Genomics Explorations, NeurIPS 2024 AI for New Drug Modalities.

Bryan Hooi (e-mail: bhooi@comp.nus.edu.sg) is an Assistant Professor at the School of Computer Science, National University of Singapore (NUS). His research interests include machine learning, graphs, trustworthy machine learning, and biomedical applications. Before joining NUS, he got his PhD in ML from Carnegie Mellon University and his MS and BS from Standford University. His work has been recognized by awards such as ECML-PKDD 2018 Runner-Up Best Student Data Mining Paper Award and KDD 2016 Best Paper Award (Research Track). Previously organized workshops: KDD 2021 Outlier Detection and Description Workshop, WebConf 2024 Graph Foundation Model Workshop.

Fran Supek (e-mail: fran.supek@bric.ku.dk) is a Professor at the Biotech Research & Innovation Centre (BRIC), University of Copenhagen, and holds a part-time position as Group Leader at the Institute for Research in Biomedicine (IRB Barcelona). Dr. Supek's research focuses on computational genomics, cancer genetics, and the application of artificial intelligence techniques to biological data analysis. His work has been recognized with several prestigious awards, including the EMBO Young Investigator Programme membership (2020-2024). Dr. Supek has also been awarded highly competitive grants, such as the ERC Starting Grant (2018-2023) and ERC Consolidator Grant (2024-2029), and participates in two Horizon EU consortia, DECIDER and LUCIA. Dr. Supek has mentored early-career researchers through the EMBO YIP program and the Max Delbrück Centre "Aspire" program.

Pietro Liò (e-mail: pl219@cam.ac.uk) is a Full Professor at the Department of Computer Science, University of Cambridge. He is a member of the Artificial Intelligence Group, Cambridge Centre for AI in Medicine and Academia Europeae. His research interest focuses on developing computational methods to understand diseases complexity and address personalized and precision medicine, with a current focus on Graph Neural Networks for computational biology.

Elena Rivas (e-mail: elenarivas@fas.harvard.edu) is a Senior Research Fellow and Lecturer at Harvard University. Her major expertise is in noncoding RNAs, developing algorithmic approaches for RNA structure determination and identification, and developing mathematical models of RNA structure for the identification and characterization of evolutionarily conserved RNA structures. Moving from the field of theoretical physics, in which she obtained her PhD from the University of Zaragoza (Spain), she started her career in computational biology as a Research Assistant Professor

at Washington University (St. Louis). She organized two-week long workshops on "Computational Analysis of RNA Structure and Function", held in Benasque, Spain, eight times from 2003 to 2024.

Mile Šikić (e-mail: mile_sikic@gis.a-star.edu.sg) is a group leader at the Genome Institute of Singapore (A*STAR), where he heads the AI in Genomics lab. His research focuses on applying machine learning to develop novel models for *de novo* genome assembly, detecting modified DNA and RNA nucleotides, and predicting RNA structure. Additionally, his team works on the development of foundational models for RNA and DNA. In 2017, he initiated the International Summer School on Data Science in Split, Croatia, which continues to be held annually.

6 Diversity, Equity, and Inclusion

Our proposed workshop is dedicated to diversity, equality, and inclusion. This can be seen in our organizing team, invited speakers, and panelists.

- Our organizing team brings together 13 members from eleven institutions across academia and industry, representing four continents (Asia, Australia, Europe, North America) and eight countries (Australia, Croatia, Denmark, Singapore, Spain, Switzerland, UK, USA). The team is diverse in gender, ethnicity, and career stages, ranging from PhD students and postdoctoral fellows to assistant professors, full professors, senior researchers, and group leaders. Collectively, they cover a broad spectrum of expertise, including RNA and DNA research, as well as more theoretical machine learning. Some organizers have experience organizing conferences and workshops, but we also included early-career researchers without previous experience.
- The speakers and panelists are chosen from different fields of expertise (RNA-related, DNA-related, ML background), coming from both academia and industry from various institutions (nine different institutions for nine people), from different countries and continents (Asia, Europe, North America), different seniority levels (research scientists, senior directors, assistant professors, professors), ethnicity and gender.
- The program committee includes representation from different affiliations (industry and academia), seniority (master students, PhD students, professors, senior researchers), geographic locations (Asia, Europe, North America), gender and ethnicity. This way we expect to reduce the risk of biased evaluation of the submissions.
- Tiny-paper track is part of diversity, equity, and inclusion (DEI) initiative, and by having this track, we are offering opportunities for underrepresented, underprivileged, and first-time submitters to share their work at our workshop.

7 Previous Related Workshops

Over the past several years, there has been a growing number of biology-related workshops at top-tier ML conferences. While most of the workshops do not explicitly exclude ML research on nucleic acids, it is often overshadowed by ML research on proteins. In contrast, the proposed AI4NA workshop will emphasize NAs – DNA and RNA, by exploring how ML models can unveil the code and improve our understanding of NAs. Some of the previous biology-related workshops on ML conferences:

- **Workshops on Drug Design:** ICLR 2024 Workshop on Machine Learning for Genomics Explorations (MLGenX), ICLR 2023 Workshop on Machine Learning for Drug Discovery (MLDD), NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3), NeurIPS 2023 Workshop on Machine Learning in Structural Biology (MLSB 2023).
  These workshops focus on optimizing and discovering therapeutic candidates. MLDD, MLSB, and AI4D3 primarily focus on molecular design in drug discovery, and MLGenX emphasizes target identification. The proposed AI4NA workshop's scope includes drug design, but unlike the other workshops, the emphasis is specifically on NAs. We are specifically interested in NA drug design and NA interactions with other biomolecules that could give us a better understanding and create more opportunities for discovering therapeutics.

- **Workshops on Foundation Models for Biology:** ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery (AccMLBio).
  There is a growing gap between ML research on biology-related problems and the actual usage of ML in the lab or the clinic. This is obvious especially in the context of foundation models, where accessibility and efficiency concerns limit the adoption of these models by biologists and clinicians. AccMLBio aimed to bridge the accessibility and efficiency gap between ML research and lab use. In contrast, the proposed AI4NA workshop is not focused solely on the foundation models, nor do we specifically aim to close the gap between ML research and lab use. However, we seek for (multimodal) foundation models that can offer enriched representations for several specific or a broad set of NA-related downstream tasks.
- **Workshops on Generative ML for Biomolecular Design:** ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design (GEM), NeurIPS 2023 Workshop on Generative AI and Biology (GenBio).
  Generative AI models have led to tremendous discoveries, from image and text generation to protein folding and design. We are now able to predict protein structure from sequence alone, to characterize the function and interactions of biomolecules, to design such molecules never-before-seen in nature, and more. The goal of GenBio was to gain critical insights into the future of generative-AI-driven biology. Additionally, the GEM workshop tried to bring computationalists and experimentalists together. They explored the strengths and challenges of generative ML in biology and experimental integration of generative ML. The scope of the proposed AI4NA workshop includes generative ML for biology, but there is a highlight on nucleic acids and their interaction with other biomolecules. We especially look to bridge the gap between generative ML and RNA design, which can potentially offer many advantages in drug discovery research.
- **Workshops on Computational Biology:** ICML 2024 Machine Learning for Life and Material Science: From Theory to Industry Applications (ML4LMS), ICML 2023 Workshop on Computational Biology (CompBio 2023).
  The past workshops in computational biology have explored the broad spectrum of computational techniques applied to various parts of biology. The scopes of ML4LMS and CompBio were much broader than the scope of the proposed AI4NA workshop which focuses specifically on nucleic acids research and bringing together AI researchers and experts in nucleic acids.

8 Sponsors

Oxford Nanopore Technologies and NVIDIA have confirmed their sponsorship of the proposed AI4NA workshop, while we are still waiting for a few companies to confirm and are actively in search of other industrial labs for sponsorship. We expect around $10,000 to be collected from the sponsors, and the money will be spent in the following ways, prioritized from top to bottom:

- Registration and travel grants for minority groups and students
- Beverages and food during the breaks
- Best paper award (The committees will highlight the best submission accompanied with the following prospective prize: NVIDIA RTX A6000 – 48GB)
- Post-workshop gathering at a nearby restaurant

9 Program Committee

We have gathered a diverse group for our program committee. They come from various fields of expertise, levels of seniority, gender, and ethnicity to represent different views in the reviewing process. We will ensure the review load will not exceed more than three papers per reviewer. Having organizers from different labs and countries, we expect to have a diverse set of reviewers, mitigating the chance of a conflict of interest. We will use OpenReview as the submission and reviewing system. Here are the tentative program committee members (reviewers), and we are expecting to add more reviewers as we continue our outreach (additionally, we will add a call for reviewers on the official workshop web page):

Miquel Anglada-Girotto (CRG Barcelona), Rishabh Anand (Yale), Maciej Antczak (UT Poznan), Sara Bakić (NUS), Eugene Baulin (IIMB Warsaw), Nigel Chou (A*STAR Singapore), Valentin Debarnot (INSA Lyon), Krešimir Friganović (A*STAR Singapore), Jonathan Goeke (A*STAR Singapore), Amanda Guo (A*STAR Singapore), Arian Jamasb (Prescient Design / Roche), Vinith Kishore (University of Basel), Kiran Krishnamachari (A*STAR Singapore), Kelly Lindsay (University of Minnesota), Josipa Lipovac (University of Zagreb), Liang Li (A*STAR Singapore), Min Hao Ling (NUS), Simon Mathis (Cambridge), Alex Morehead (University of Missouri), Rafael Josip Penić (University of Zagreb), Anton Petrov (Riboscope Ltd), Anna Poetsch (TU Dresden), Bo Qiang (University of Washington), Martin Schmitz (NUS), Dominik Stanojević (University of Zagreb), Tim Stuart (A*STAR Singapore), Marcell Veiner (IRB Barcelona), Shreyas V (BITS Pilani), Sukwon Yun (UNC Chapel Hill), Yinghua Yao (A*STAR Singapore).

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pp. 1–3, 2024.

Rishabh Anand, Chaitanya K Joshi, Alex Morehead, Arian R Jamasb, Charles Harris, Simon V Mathis, Kieran Didi, Bryan Hooi, and Pietro Liò. RNA-FrameFlow: Flow matching for de novo 3D RNA backbone design. *arXiv preprint arXiv:2406.13839*, 2024.

M. Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature methods*, 21(1): 117–121, 2024.

Gunjan Baid, Daniel E Cook, Kishwar Shafin, Taedong Yun, Felipe Llinares-López, Quentin Berthet, Anastasiya Belyaeva, Armin Töpfer, Aaron M Wenger, William J Rowell, et al. Deepconsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nature Biotechnology*, 41(2):232–238, 2023.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

Jessica L Childs-Disney, Xueyi Yang, Quentin MR Gibaut, Yuquan Tong, Robert T Batey, and Matthew D Disney. Targeting RNA structures with small molecules. *Nature Reviews Drug Discovery*, 21(10):736–762, 2022.

Chaitanya K Joshi, Arian R Jamasb, Ramon Viñas, Charles Harris, Simon V Mathis, Alex Morehead, Rishabh Anand, and Pietro Liò. gRNAde: Geometric deep learning for 3D RNA inverse design. *bioRxiv*, 2024.

J. 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.

Qiang Kang, Jun Meng, Jun Cui, Yushi Luan, and Ming Chen. Pmlipred: a method based on hybrid model and fuzzy decision for plant mirna–lncrna interaction prediction. *Bioinformatics*, 36(10): 2986–2992, 2020.

Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, et al. CodonBERT: Large language models for mRNA design and optimization. *bioRxiv*, pp. 2023–09, 2023a.

Y. Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P Lydia Freddolino, and Yang Zhang. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14(1):5745, 2023b.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.

Divya Nori and Wengong Jin. RNAFlow: RNA structure & sequence design via inverse folding-based flow matching. *arXiv preprint arXiv:2405.18768*, 2024.

Robin Pearce, Gilbert S Omenn, and Yang Zhang. De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *BioRxiv*, pp. 2022–05, 2022.

Rafael Josip Penić, Tin Vlašić, Roland G Huber, Yue Wan, and Mile Šikić. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*, 2024.

Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018.

Bohdan Schneider, Blake Alexander Sweeney, Alex Bateman, Jiri Cerny, Tomasz Zok, and Marta Szachniuk. When will RNA get its AlphaFold moment? *Nucleic Acids Research*, 51(18):9522–9532, 2023.

T. Shen, Zhihang Hu, Zhangzhi Peng, Jiayang Chen, Peng Xiong, Liang Hong, Liangzhen Zheng, Yixuan Wang, Irwin King, Sheng Wang, et al. E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. *arXiv preprint arXiv:2207.01586*, 2022.

Dominik Stanojevic, Dehui Lin, Paola Florez De Sessions, and Mile Sikic. Telomere-to-telomere phased genome assembly using error-corrected simplex nanopore reads. *bioRxiv*, pp. 2024–05, 2024.

Lei Sun, Kui Xu, Wenze Huang, Yucheng T Yang, Pan Li, Lei Tang, Tuanlin Xiong, and Qiangfeng Cliff Zhang. Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research*, 31(5):495–516, 2021.

Lovro Vrček, Xavier Bresson, Thomas Laurent, Martin Schmitz, Kenji Kawaguchi, and Mile Šikić. Geometric deep learning framework for de novo genome assembly. *bioRxiv*, pp. 2024–03, 2024.

W. Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei, Fa Zhang, Zhenling Peng, and Jianyi Yang. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nature Communications*, 14(1):7266, 2023.

He Zhang, Liang Zhang, Ang Lin, Congcong Xu, Ziyu Li, Kaibo Liu, Boxiang Liu, Xiaopin Ma, Fanfan Zhao, Huiling Jiang, et al. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature*, 621(7978):396–403, 2023.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.