Annual Report

Giorgio Bini

September 2024

During my last year of Ph.D., my work has been intensely focused on deepening my understanding of key computational methods in biology, particularly through the application of Deep Learning (DL) and Machine Learning (ML) techniques to challenging problems in genomics and molecular biology. This report summarizes the progress made in several ongoing research projects, as well as the collaborative efforts and academic achievements that have marked the final phase of my PhD journey.

This year, I concentrated significantly on advancing my project on predicting RNA-RNA interactions directly from sequence data using a deep learning approach. The project is divided into two main aspects: a biology-oriented perspective and a prediction-focused analysis.

From the biological perspective, our research revealed the critical role of simple repeat sub-sequences in driving direct RNA-RNA interactions. In collaboration with Adriano Setti and Alessio Colantoni, we meticulously curated datasets of RNA-RNA interactions identified through transcriptome-wide approaches. These datasets were instrumental in training and evaluating our deep learning model, with a strong emphasis on biological relevance.

On the prediction side, we compared our method against traditional thermodynamic tools, achieving state-of-the-art performance across various datasets. This demonstrated our model's robust generalization capabilities, even when applied to datasets obtained from different experimental techniques used to capture RNA-RNA interactions. We are currently drafting a paper, where I am the first author, and the promising results suggest that this research could significantly enhance our understanding of key cellular mechanisms.

In the past year, I also submitted four papers. Two of these, in collaboration with Claudia Giambartolomei [3] and the Million Veteran Project Concosrtium [2], focused on evaluating the Enformer DL model's effectiveness in prioritizing Single Nucleotide Variants (SNVs). My contribution

involved a novel approach, where I compared the model's learning signals to traditional population-genetics metrics.

Another paper, co-authored with Magdalena Arnal Segura [5], expands on our previous work [1] by adapting our existing pipeline to study additional neurodegenerative diseases. This extension not only broadens the pipeline's application but also offers the potential for new discoveries in these complex conditions.

Another productive collaboration with Jonathan Fiorentino and Michele Monti has led to a forthcoming paper, currently under revision [4]. This research explores the use of machine learning to predict liquid-liquid phase separation (LLPS) proteins, which are crucial for cellular organization and function, with implications for diseases such as neurodegenerative disorders and cancer. Our study integrates physicochemical features with AlphaFold predictions, significantly enhancing our model's accuracy in predicting LLPS proteins and advancing our understanding of this critical biological phenomenon.

On the academic front, I successfully completed my final exam, "An Introduction to Optimization Over Time and Its Application to Online Machine Learning and Reinforcement Learning," taught by Prof. Gnecco. In this course, I deepened my knowledge of time-based optimization, focusing on discrete scenarios, and gained foundational insights into emerging areas such as Dynamic and Approximate Dynamic Programming, the Kalman Filter, and the use of Neural Networks for solving discrete-time N-stage optimization problems. These concepts will be invaluable as I continue to advance in my career.

This year, I made significant progress on my primary research project on predicting RNA-RNA interactions, laying a solid foundation for its completion. I am confident that I will finalize and submit this work for publication in the coming months. The experience has not only deepened my expertise in computational biology and machine learning but has also prepared me to make meaningful contributions at the intersection of these fields, driving further discoveries and innovations.

1. Your Name and Tutor(s):

    - Giorgio Bini (Tutor: Gian Gaetano Tartaglia)

2. List of Attended Courses and Exams:

    - An Introduction to Optimization Over Time and Its Application to Online Machine Learning and Reinforcement Learning

3. List of Publications:

   - Machine learning methods applied to classify complex diseases using genomic data [5] (paper submitted)
   - Enformer predictions applied to GTEx data (2 papers submitted)
   - Predicting RNA-RNA interactions with Deep Learning (in preparation)
   - Accurate Predictions of Liquid-Liquid Phase Separating Proteins at Single Amino Acid Resolution [4] (paper submitted)

References

[1] Magdalena Arnal Segura, Giorgio Bini, Dietmar Fernandez Orth, Eleftherios Samaras, Maya Kassis, Fotis Aisopos, Jordi Rambla De Argila, George Paliouras, Peter Garrard, Claudia Giambartolomei, et al. Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14(1):e12300, 2022.

[2] John Michael Gaziano, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, Jennifer Deen, Colleen Shannon, Donald Humphries, et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, 70:214–223, 2016.

[3] Liam Gaziano, Claudia Giambartolomei, Alexandre C Pereira, Anna Gaulton, Daniel C Posner, Sonja A Swanson, Yuk-Lam Ho, Sudha K Iyengar, Nicole M Kosik, Marijana Vujkovic, et al. Actionable druggable genome-wide mendelian randomization identifies repurposing opportunities for covid-19. *Nature medicine*, 27(4):668–676, 2021.

[4] Michele Monti, Jonathan Fiorentino, Dimitrios Miltiadis-Vrachnos, Giorgio Bini, Tiziana Cotrufo, Natalia Sanchez de Groot, Alexandros Armaos, and Gian Gaetano Tartaglia. Accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution. *bioRxiv*, 2024.

[5] Magdalena Arnal Segura, Giorgio Bini, Anastasia Krithara, George Paliouras, and Gian Gaetano Tartaglia. Machine learning methods applied to classify complex diseases using genomic data. *bioRxiv*, 2024.