Annual Report

Giorgio Bini

September 2023

During my second year of Ph.D., I continued my exploration of the dynamic and fascinating intersection between Machine Learning (ML), Deep Learning (DL), and biology. Building on the foundation established in my first year, my focus remained on innovative applications of these techniques to various bioinformatics challenges, with a strong emphasis on collaboration and interdisciplinary learning.

A significant portion of my focus this year was dedicated to the advancement of my project on RNA-RNA interactions. Through rigorous research and development, our team achieved remarkable progress, attaining state-of-the-art performance on the dataset we assembled, composed of RNA-RNA interactions identified via transcriptome-wide approached. Our proposed model demonstrated its efficacy in predicting such interactions solely from sequence data. One of the standout achievements of our model is its ability to accurately predict the specific regions within the RNA sequences that are involved in the interaction. This insight provides valuable mechanistic understanding and highlights the depth of information encoded within the sequence data. We are currently in the process of validating our findings on external datasets, a crucial step towards establishing the robustness and generalizability of our approach. The promising results we have obtained indicate the potential impact of this work on understanding fundamental cellular processes. It's worth noting that I will serve as the primary author of this work.

Collaboration continued to be a driving force in my research journey. In partnership with Claudia Giambartolomei [4] and the Million Veteran Project Concosrtium [3], we submitted a paper that showcases the utility of the Enformer DL model [2] in prioritizing Single Nucleotide Variants (SNVs). This involved a unique perspective by comparing the Enformer signals with traditional population-genetics statistics, shedding light on the congruence between these approaches and highlighting the strengths of DL in this context. I've examined implications of using Enformer, in conjunction

with the Genotype-Tissue Expression (GTEx) project data, for multiple subfields of population genetics, including trans-ancestry analysis, for which dedicated papers are in the works.

Another fruitful collaboration with Magdalena Arnal Segura has resulted in the preparation of two upcoming papers. The first paper extends the pipeline we previously developed [1] to other neurodegenerative diseases, broadening the scope of its application and potentially uncovering novel insights into these conditions. The second work delves into the challenge of training a Machine Learning classifier to predict eQTLs (expression quantitative trait loci), sQTLs (splicing quantitative trait loci), and pQTLs (protein quantitative trait loci). This exploration is expected to contribute to a deeper understanding of the intricate interplay between genetic variations and molecular phenotypes.

Academic progress was marked by successful completion of two courses. *The Molecular Dynamics* course equipped me with a comprehensive understanding of this powerful technique, which holds great relevance in the fields of bioinformatics and biophysics. *Particle Swarm Optimization*, the focus of another course, enriched my toolkit with non-gradient-based optimization methods, essential for tackling the complexity inherent in computational biology problems.

Expanding my horizons, I participated in two advanced Machine Learning summer schools. The *Topics in Modern Machine Learning* school in Genova and the *Madrid UPM Machine Learning and Advanced Statistics Summer School* provided immersive experiences in cutting-edge topics such as Gaussian Processes, Bayesian Optimization, Support Vector Machines, and Regularized Learning. These experiences not only deepened my understanding but also nurtured connections with fellow researchers and students from diverse backgrounds.

As I reflect on the past year's accomplishments, it is evident that my dedication to interdisciplinary collaboration and advanced training has positioned me well for the challenges that lie ahead. The confluence of Machine Learning, biology, and computational techniques continues to yield promising results, propelling me towards a successful continuation of my Ph.D. journey.

1. Your Name and Tutor(s):
    - Giorgio Bini (Tutor: Gian Gaetano Tartaglia)
2. List of Attended Courses and Exams:

- Advanced Computational Physics: Molecular Dynamics (14/07/23)
- Advanced Computational Physics: Optimization (The exam will be taken in September)
- Training school (The exam will be taken in September)

3. List of Publications:

   - Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset alzheimer's disease [1]
   - Enformer predictions applied to GTEx data (3 papers in preparation)
   - Predicting RNA-RNA interactions with Deep Learning (in preparation)
   - Predicting eQTL-sQTL-pQTL with Deep Learning (in preparation)

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] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[3] 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.

[4] 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.