Annual Report

Giorgio Bini

September 2022

My research activities during the first year of Ph.D. dealt with the application of Machine Learning (ML) and Deep Learning (DL) methods to relevant bioinformatics problems, ranging from genomics to molecular biology. My interest in these models, which are very suitable to analyze large amount of biomedical data, allowed me to collaborate with many colleagues of different backgrounds, such as molecular biologists, physicians and population geneticists. In particular, I took part in three different projects, under the supervision of Dr. G. G. Tartaglia [6].

The first one, published in the *Alzheimer's Dementia* journal [1], aims to highlight genomics profiles related to late onset Alzheimer's disease. ML models, together with the use of eXplainable Artificial Intelligence (XAI) methods, are particularly suited for this task, since we can prioritize a list of Single Nucleotide Variants (SNVs) without any prior assumptions about their genetic contribution to the traits. Our pipeline is generalizable and can also be used for the study of other diseases with genomic background predisposition, which will be of interest for future works.

My interest in Genomics have lead me to work together with Dr. Claudia Giambartolomei [4] who is specialized in identifying genetic variants that are related to various traits and diseases through Genome-Wide Association Studies (GWASs). We performed post-GWAS investigations with the use of a pre-trained DL model, called Enformer [2] in order to understand if DL can be used together with standard statistical approaches in order to prioritize variants affecting RNA expression levels (gene expression quantitative trait loci, eQTL) and protein quantitative trait loci, (pQTL). Specifically the results from this first analysis will be used to corroborate the evidence of a genetic effect of a SNV on the expression of a gene that has been prioritized from causal models of 12,000 phenotypes measured in 650,000 individuals as part of an effort from the Million Veteran Project Concosrtium [3].

The project to which I have mainly devoted my attention concerns the development of a DL-based model for RNA-RNA interactions (RRIs) pre-

diction. RRIs are important in many basic cellular activities including transcription, RNA processing, localization, and translation [5]. By combining embedded physico-chemical properties, such as secondary structure and hydrogen bonding our model aims to predict RRIs directly from the sequences. To the best of our knowledge, this is the first work performing RRIs prediction only on the basis of RNA sequence information alone.

The last two projects described above are still in progress; we expect to submit the papers during the next academic year and I will be first author in one of them.

The development of ML methods is revolutionizing computational biology because of impressive results in terms of model accuracy and predictive ability. However, most of these models are "black-box", while more interpretable solutions are claimed by the research community, specially in biological applications. XAI is a growing area of research that offers a solution for this problem. In August I took part in the XAISS summer school at TU Delft, an initiative that deals with interpretability and explainability of AI and machine learning models. It was an opportunity to update myself on the state-of-the-art methods, as well as an opportunity to build links with students and researchers who work in distinct but complementary fields to mine in an international environment.

In parallel to my projects, I attended courses and passed exams related to my study plan. The course *An introduction to optimization over time and its application to online machine learning and reinforcement learning*, by Prof. Gnecco, provided an introductory and comprehensive view of dynamic optimization problems. Some of the presented methods can be in the future used for analyzing biophysical systems (which will be also object of study in the course *Advanced Computational Physics*), such as molecular dynamics simulations.

The course *Theory and Practice of Learning from Data*, by Prof. Oneto, provided an overview of some ML models with a particular focus on statistics theorems. During the exam, which I successfully passed, we discussed how these notions relate to my projects.

Another course that I have followed with great interest is *Metodi di Simulazione Applicati alla Fisica*, by Prof. Ferrando and Prof. Parodi. The course gave me a comprehensive overview of Monte Carlo simulations, which are widely used for the analysis of physical systems. During the exam we focused on a specific case study, i.e. the reticular gas system.

In conclusion, as computational biology research requires different but complementary methodologies to be comprehensively studied, I strongly believe that the interdisciplinary training I have received this year will help

me in a successful continuation of the Ph.D. program.

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.

[5] Jing Gong, Yanyan Ju, Di Shao, and Qiangfeng Cliff Zhang. Advances and challenges towards the study of rna-rna interactions in a transcriptome-wide scale. *Quantitative Biology*, 6(3):239–252, 2018.

[6] Gian Gaetano Tartaglia, Amol P Pawar, Silvia Campioni, Christopher M Dobson, Fabrizio Chiti, and Michele Vendruscolo. Prediction of aggregation-prone regions in structured proteins. *Journal of molecular biology*, 380(2):425–436, 2008.