Using polygenic risk scores to address the heterogeneity of Autism Spectrum Disorder

Place of work/: Instituto Nacional de Saúde Doutor RIcardo Jorge

Supervisors: Hugo Martiniano (hugo.martiniano@insa.min-saude.pt); Astrid Vicente (astrid.vicente@insa.min-saude.pt);

Non-coding RNA ((ncRNA) are RNAs that do not encode proteins. These are the majority of human genes and several ncRNAs play a crucial role in regulating gene expression and other biological processes.

However, when compared with protein-coding genes, little is known of the function of most ncRNAs. The state-of-the-art approach to systematization of biological knowledge of genes and gene products is the Gene Ontology (GO). While annotation of proteins is well-developed (albeit not finished), this is not true for ncRNAs. Closing this knowledge gap is an essential step in understanding living systems. In particular when related to health and disease states.

In this project we propose the development of machine learning methods to predict GO annotations for ncRNA molecules. Using a dataset composed of a network of ncRNAs and their associations to each other and to protein-coding genes, the problem of predicting GO annotations can be framed as a link prediction task. For this purpose we propose the use of recently-developed graph neural networks, which have been shown to have excellent performance for link-prediction in complex networks.

The methods developed in the context of this project will be applied to the identification of pathogenic genetic variants in a cohort of children with Autism Spectrum DIsorder.

The candidate is expected to have knowledge of the Python programming language and an interest in machine learning methods.