



PhD in DATA ANALYTICS AND DECISION SCIENCES - 40th cycle

THEMATIC Research Field: MACHINE LEARNING FOR POLYGENIC RISK PREDICTION OF COMPLEX NON-COMMUNICABLE DISEASES

Monthly net income of PhDscholarship (max 36 months)

€ 1400.0

In case of a change of the welfare rates during the three-year period, the amount could be modified.

Context of the research activity

Motivation and objectives of the research in this field

In genetics, complex diseases result from the interactions of multiple genetic variants and environmental factors. Population genomics employs polygenic risk scores (PRS) to estimate individual genomic risks for diseases by summing the effects of genetic variants. While PRSs hold promise for clinical applications, their effectiveness is hindered by several challenges: low accuracy due to simplistic additive modeling and ignoring gene-gene and gene-environment interactions, biased data from predominantly European ancestries affecting transferability to other ethnic groups, and restricted access to genotype databases limiting the creation of diverse and accurate predictive models. Identifying high-risk individuals is essential for targeted prevention and therapy, necessitating innovative approaches to overcome the aforementioned obstacles. Advanced machine learning (ML) methods offer a promising direction by capturing complex data patterns and integrating different data modalities (genetic, environmental, clinical) to enhance disease prediction and derive insight into its mechanisms.

Methods and techniques that will be developed and used to carry out the research

The research will focus on the development of novel representation learning and machine learning methods to enhance polygenic risk prediction. Specific methodologies will include:
- Representation learning of genotype (and potentially



	<ul style="list-style-type: none"> - Representation learning of genotype (and potentially also environmental and clinical) data; - Latent space arithmetic or explanation/attribution methods to identify the drivers of complex diseases; - (Deep) Transfer Learning and Generative Modeling to overcome data-access and privacy barriers and enhance PRS portability across ethnicities; - Advanced Deep Learning/Machine learning risk prediction models and multi-modal integration to account for gene-gene and gene-environment interactions
Educational objectives	To be able to critically analyse, model and integrate complex data sources to answer relevant biological and clinical questions. To develop skills in statistical data analysis, machine learning, deep representation learning, bioinformatics, and their application to human genetics research carried out on large-scale genomics biobanks.
Job opportunities	The profile of data scientist and the applications proposed in this project are of interest to a broad range of actors, including (but not limited to): public and private institutions dealing with healthcare, hospitals, clinical and pharmaceutical companies, as well as international institutions and research centres working in healthcare research, and policy makers in charge with healthcare governance.
Composition of the research group	1 Full Professors 1 Associated Professors 2 Assistant Professors 5 PhD Students
Name of the research directors	Matteucci, Di Angelantonio (HT), Massi (HT)

Contacts	
<p>Prof. Matteo Matteucci (DEIB) matteo.matteucci@polimi.it</p> <p>Emanuele Di Angelantonio (Fondazione HT) emanuele.diangelantonio@fht.org</p> <p>https://humantechnopole.it/en/people/emanuele-di-angelantonio</p>	



Additional support - Financial aid per PhD student per year (gross amount)	
Housing - Foreign Students	--
Housing - Out-of-town residents (more than 80Km out of Milano)	--

Scholarship Increase for a period abroad	
Amount monthly	700.0 €
By number of months	12

Additional information: educational activity, teaching assistantship, computer availability, desk availability, any other information
<p>List of Universities, Companies, Agencies and/or National or International Institutions that are cooperating in the research: Health Data Science Center (HT)</p> <p>Additional support</p> <p><i>Educational activities</i> (purchase of study books and material, funding for participation in courses, summer schools, workshops and conferences): financial aid per PhD student per year: 1st year: max 1.902,38 euro per student 2nd year: max 1.902,38 euro per student 3rd year: max 1.902,38 euro per student <i>Teaching and lab assistantship</i>: availability of funding in recognition of supporting teaching and lab activities by the PhD student. Further support is available for students who engage in activities of teaching or additional lab duties coherent with their academic mission and doctoral training. The PhD student is encouraged to take part in these activities, within the limits allowed by the regulations.</p> <p><i>Computer availability</i>: 1 st year: individual use 2 nd year: individual use 3 rd year: individual use <i>Desk availability</i>: 1 st year: individual use 2 nd year: individual use 3 rd year: individual use</p>