DECODE
Decoding the Language of Single-Cell Gene Regulation with Transformers

CUP J53C25002180001 - FIS-2024-02465

OVERVIEW of the Program Research interests and Faculty	The PhD program in Statistics and Computer Science at Bocconi University is a 4-year program taught in English that provides students a strong statistical, computational and mathematical background. The program features two curricula: Statistics (STAT) and Computer Science (CS). DECODE will be carried out within the Computer Science curriculum. In the first year, the program includes both shared and curriculum-specific compulsory courses for all enrolled PhD students. These provide a wide range of theoretical, methodological and computational skills that are essential for research in the Computer Science curriculum. The second year features reading groups and courses taught by international Visiting Professors. Students receive dedicated mentorship throughout their time at Bocconi. The faculty includes internationally recognized leading researchers in Statistics, Computer Science, Machine Learning, Probability, and Statistical Physics. The program also benefits from contributions of distinguished visiting professors who deliver short monographic courses. The PhD in Statistics and Computer Science is designed for highly motivated students aiming to pursue first-rate research careers in academia, research institutes and industry. Career opportunities also extend to central banks, financial institutions, government agencies, international organizations, and public health institutions. For more information about the program, please visit the website: https://www.unibocconi.it/en/programs/phd/phd-statistics-and-computer-science
Opening date	17th April 2026
Closing date	18th May 2026
Enrollment date	Refer to Art.9 “Enrolment deadlines” of the Call for details
Available places WITH FELLOWSHIP funded by Fondo Italiano per la Scienza (FIS)	**1**
Fellowship	€ 27,000 for year 1 and € 23,000 for years 2, 3 and 4.
Starting date	1st September 2026
Duration	4-year, full-time program
Principal Investigator	Prof. Andrea TANGHERLONI
Project overview	Single-cell multi-omics technologies, such as scRNA-seq (single-cell RNA

	sequencing, which profiles gene expression in individual cells) and scATAC-seq (single-cell Assay for Transposase-Accessible Chromatin using sequencing, which profiles chromatin accessibility), have dramatically advanced our understanding of cellular diversity and gene regulation. Public repositories now host tens of millions of single-cell profiles, yet effectively integrating these modalities to reconstruct Gene Regulatory Networks (GRNs) remains a major open challenge. Current methods suffer from oversimplified assumptions about linear relationships, separate analysis of each omics layer, loss of critical cross-modal insights, reliance on correlation rather than causation, and static modelling that fails to capture regulatory dynamics. A new generation of computational frameworks is needed to exploit the full scale and complexity of these data. The DECODE project addresses this challenge by developing a novel transformer-based computational framework that integrates scRNA-seq and scATAC-seq data to construct comprehensive models of gene regulation. The framework consists of three components: two specialised single-omics models—one for gene expression and one for chromatin accessibility—trained independently on large-scale atlases, and a cross-modal model that integrates the learned features from both modalities to identify regulatory interactions between genes and regulatory elements. All three models are built on an innovative mechanism that enables each gene or peak to attend to the full vocabulary while controlling computational complexity, overcoming a key limitation of the existing models. The biological applications of the framework focus on two critical areas. The first is the human immune system, investigating hematopoietic development and T cell differentiation to identify regulatory mechanisms underlying immune responses and inflammation. The second is Triple-Negative Breast Cancer (TNBC), where the framework will model cancer-specific regulatory networks and mechanisms of therapeutic resistance to identify novel drug targets. These efforts are expected to advance biomedical research by uncovering actionable mechanisms and accelerating the development of new therapies. Experimental validation through CRISPR perturbation experiments and a comprehensive benchmarking suite will generate robust tools and reference standards for the broader research community. The PhD candidate will join a research group at Bocconi University dedicated to AI applied to biology and medicine, working alongside postdoctoral researchers and collaborating with experimental biologists. The project offers a unique opportunity to contribute to a large-scale, well-funded research program spanning the design of deep learning architectures, multi-omics data integration, gene regulatory network inference, and translational applications in immunology and oncology.
Research objectives	The first objective is to help design, implement, and pre-train the single-omics models for scRNA-seq and scATAC-seq data. This includes building efficient training pipelines to stream data from public atlases such as the CELLxGENE Census. It also involves implementing and evaluating multitask curriculum

	learning strategies (e.g., masked language modelling, expression regression, and denoising. The candidate will systematically study the effect of model capacity, data volume, and attention mechanisms on representation quality. These models must scale to millions of cells, ensuring computational efficiency through mixed-precision training and multi-GPU parallelism. The second objective is to help develop and evaluate the cross-modal model that integrates gene expression and chromatin accessibility representations. The candidate will investigate training strategies for this integration. The candidate will also evaluate the quality of the inferred gene regulatory networks against experimental regulatory databases such as ENCODE (Encyclopedia of DNA Elements, a resource for functional elements in the genome), ChIP-Atlas (database for chromatin immunoprecipitation data), and RegNetwork (database of regulatory relationships), as well as perturbation screen data (experimental data showing effects of systematic gene modifications). The third objective focuses on biological applications of the framework. The candidate will use the trained models to investigate gene regulation in the human immune system and Triple-Negative Breast Cancer. They will perform comparative analyses between healthy and diseased conditions to find dysregulated gene programs and potential therapeutic targets. This work involves close collaboration with experimental partners. These partners will conduct CRISPR-based validation of key predicted regulatory interactions. The candidate will have a unique chance to bridge computational predictions and experimental biology. The fourth objective is to help develop a comprehensive benchmarking suite for single-cell foundation models. This includes defining standardized evaluation tasks such as imputation, clustering, cell-type annotation, batch correction, trajectory inference, and GRN inference. It also involves implementing computational and biological evaluation metrics. The candidate will conduct rigorous statistical comparisons with existing methods. Results will be published in top venues in both machine learning (NeurIPS, ICML, ICLR) and computational biology (Nature Methods, Genome Biology, Bioinformatics).
Profile of the candidate	The ideal candidate holds a Master's degree (or equivalent) in Computer Science, Data Science, Bioinformatics, Biotechnology, Mathematics, Physics, or Artificial Intelligence, with a strong quantitative background and interest in interdisciplinary research. Candidates from computational backgrounds should show curiosity toward biology. Candidates from life sciences should demonstrate solid programming and machine learning knowledge. A background in deep learning is highly desirable. Proficiency in Python and in scientific computing tools (PyTorch, NumPy, pandas) is essential. Experience with GPU-accelerated training, distributed computing, and HPC environments is a plus. Familiarity with single-cell analysis tools or genomic data formats is welcome but not required. Training in

	computational biology will be provided within the project. Comfort with large-scale datasets is expected. The candidate is expected to demonstrate strong analytical and problem-solving skills, the ability to work independently while contributing to a collaborative research environment, and excellent written and oral communication skills in English. Prior research experience—evidenced by a Master's thesis in a related area, publications, or contributions to open-source projects—will be a plus. The position is based at Bocconi University, in a research group focused on AI for biology and medicine. The candidate will have access to dedicated GPU infrastructure, cloud computing, large-scale single-cell data, and an interdisciplinary setting spanning computer science, statistics, and the life sciences. The PhD program enables international collaborations, conference attendance, and research visits. The selected candidate is expected to begin their doctoral studies in accordance with the university calendar.
Selection Criteria	The assessment is based on a candidate's: • Curriculum Vitae (mandatory); • Academic Records (mandatory); • International Graduate Admission Tests Scores (GMAT or GRE, not mandatory); • Statement of purpose (mandatory); • Up to two reference letters (not mandatory). The **assessment criteria** are as follows: • a maximum of **50 points** for the applicant's curriculum vitae and academic records; • a maximum of **10 points** for GMAT/GRE score; • a maximum of **40 points** for statement of purpose and reference letters. The Admission Board may decide to conduct a brief online interview to clarify aspects of a student's qualifications. No additional points are given to the interview. **A minimum total score of 60 points is required to be eligible for admission.**
Results	No later than 8th June 2026
Contact and Technical Support	phdadmission@unibocconi.it http://www.unibocconi.eu/admissionphd