JC2B

Junior Conference on Computational Biology 2025

November 13, 2025 | I2BC, Gif-sur-Yvette

Book of Abstracts

Junior Conference on Computational Biology 2025

Gif Sur Yvette
Bât. 21
Auditorium – I2BC
November 13, 2025

https://bioi2.i2bc.paris-saclay.fr/jc2b

JC2B 2025 is organized by all students of the **AMI2B** Master's program in Orsay, under the supervision of two faculty coordinators, with support from Université Paris-Saclay and I2BC.

Students of the **AMI2B** Master's (Orsay) : Full list available on request.
Faculty coordinators : **Anne Lopes** and **Daniel Gautheret**

Table of contents

The Paris-Saclay Junior Conference on Computational Biology (JC2B 2025 Edition) is an event organized by students of the AMI2B Master's program of Orsay's Sciences Faculty, with support from Université Paris-Saclay and I2BC.

This conference aims to bring together Master's students, PhD students, postdocs, and researchers around recent advances in bioinformatics. It provides a platform to :

- Present the work of young scientists;
- Strengthen connections between programs (BIBS, AMI2B Orsay/Évry);
- Foster intergenerational exchanges;
- Promote collaborations between research teams;
- Create networking opportunities for academic and industrial careers.

Acknowledgements

The JC2B 2025 Organizing Committee would like to express its sincere gratitude to Anne Lopes and Daniel Gautheret for their continuous guidance and support throughout the organization of this event.

We also thank Université Paris-Saclay and the Institute for Integrative Biology of the Cell (I2BC) for their institutional and logistical assistance.

Finally, our appreciation goes to all students of the AMI2B Master's program for their commitment, creativity, and teamwork, as well as to all speakers and participants who helped make the Junior Conference on Computational Biology 2025 a success.

Introduction
AI and Predictive Models in Bioinformatics

Artificial intelligence (AI) and predictive modeling have become essential tools in modern bioinformatics, enabling the analysis and interpretation of complex biological data. From genome sequencing to structural biology and health informatics, AI-driven approaches are reshaping how we identify patterns, infer biological mechanisms, and design new therapeutic strategies.

This session of the Junior Conference on Computational Biology 2025 highlights recent advances in machine learning, statistical modeling, and data-driven prediction across multiple levels of biological organization. The keynote lectures by Laurent Jacob (Sorbonne Université, CNRS) and Magali Richard (Université Grenoble Alpes, CNRS) will explore how AI can be leveraged to understand evolutionary structures and predict genomic and health-related outcomes.

Given the rapid advancement of AI in the biological sciences, this conference focuses on the following key areas:

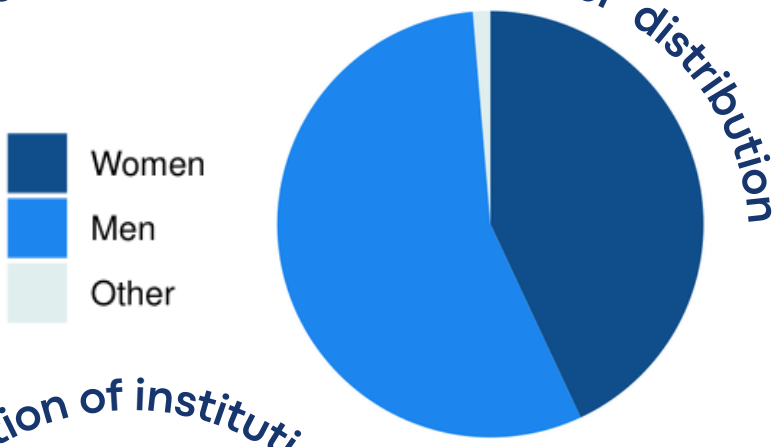
- AI applications in Structure and Evolution
- Predictive Modeling in Genomics and Health

By highlighting innovative AI-driven methods for data analysis, prediction, and biological interpretation, the JC2B conference seeks to inspire new research directions and promote the integration of computational and experimental approaches. Through keynote lectures, oral presentations, and interactive discussions, participants will gain deeper insights into how artificial intelligence is transforming our understanding of life at molecular, cellular, and systemic levels.

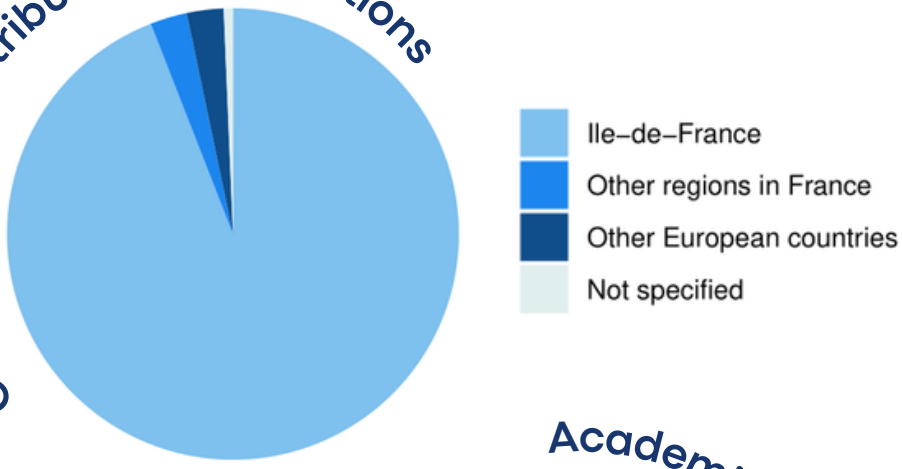
The JC2B in numbers !

151 participants
13 different universities
and schools
29 different laboratories
and companies

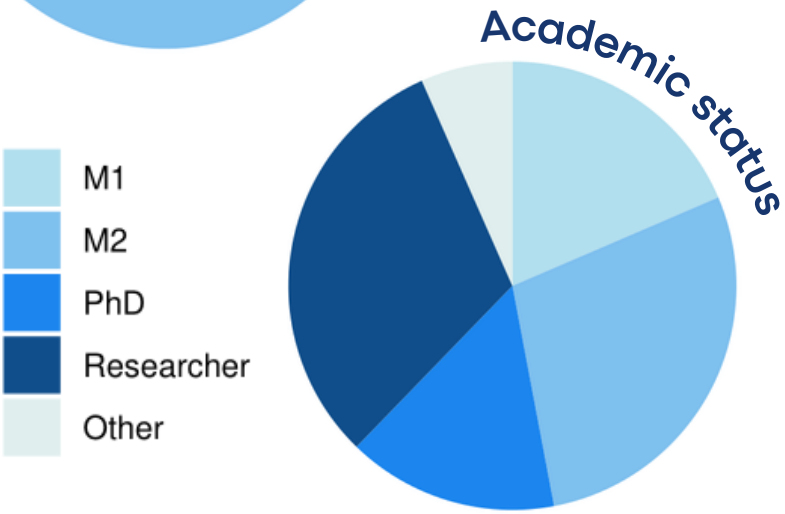
Gender distribution



Geographical distribution of institutions



Academic status



About JC2B and AMI2B Master's Students

This conference was organized by the students of the **Master 2 Analysis, Modeling and Engineering of Biological and Medical Information (AMI2B)** at *University Paris-Saclay.* The students coordinated all aspects of the event, from scientific program design and communication to logistics, graphic production, and publication of this abstract book, demonstrating the multidisciplinary and collaborative spirit at the core of the AMI2B program.

1. Management Team
In charge of creating and managing the registration system, handling participant submissions, and developing the website.

2. Communication Team
Responsible for conference promotion through mailing lists, social media, and poster design. They also handled contact with companies and laboratories for internship-related presentations.

3. Logistics Team
Managed administrative and logistical aspects such as travel arrangements for speakers, catering services, and coordination with external partners.

4. Organizational Team
Supervised the abstract collection and preparation of the abstract book. They were also responsible for creating and ordering personalized name badges and custom pins offered to participants, as well as coordinating session chairs and supporting speakers during the event.

Management Team
Registrations, Website Development

Communication Team
Emails, Social Media, Internship Fair

Logistics Team
Catering, Coffee Breaks, and Logistics

Organizational Team
Badges, Pins, Abstract Book, Event Setup, Session Chairs

Conference program

Morning session : Applications of AI in Structure and Evolution

- **9:00** : Opening speech by Anne Lopes and Daniel Gautheret
- **9:15** : **Keynote Speaker - *Laurent Jacob***, Computational and Quantitative Biology Laboratory, CNRS/Sorbonne University
- **10:00** : *Julie Fares* (Paris-Saclay University, ITODYS Paris Cité University), AGAPE (computAtional G-quadruplex Affinity PrEdiction): The first AI workflow for G4 binding affinity predictor

10:30 : Coffee break

- **11:00** : *Wissam Karroucha* (CBIO, Mines Paris, PSL University, Institut Curie) A comprehensive benchmark for RNA structure-function modeling
- **11:30** : *Simon Herman* (I2BC Paris-Saclay), Structural Space of Microproteins with Protein Language Models
- **12:00** : *Hélène Bret* (Lund University), Detecting rare codon in proteins sequences with a Transformer model

12:30 : Lunch buffet

13:30 : INTERNSHIP FAIR

Afternoon session : Predictive Modelling in Genomics and Health

- **14:00** : **Keynote Speaker - *Magali Richard,*** CNRS/Grenoble Alpes University
- **14:45** : *Chérif Seddik* (Paris-Saclay University, I2BC Paris-Saclay), Inferring latent chromatin folding signatures from bulk Hi-C data using NMF : application to the bistable expression of SPI-1
- **15:15** : *Jomar Sangalang* (Institut Gustave Roussy, Paris-Saclay), Integrative breast cancer multi-omics

15:45 : Coffee break

- **16:15** : *Shuhui Wang* (Dauphine University-PSL), Interpretable latent model predicts drug response and identifies biomarkers
- **16:45** : *Victor Reys* (BonvinLab, Utrecht University), Structural modelling and binding affinity prediction of the human PDZ-PBM interactome
- **17:15** : Closing note

Speakers and Abstracts

Neural inference in probabilistic models of evolution

Keynote speaker - Laurent Jacob

Researcher in Machine Learning and Statistics for Genomics
Computational and Quantitative Biology Laboratory, CNRS/Sorbonne University

Statistical inference in evolutionary genomics allows to estimate key parameters such as phylogenetic trees, diversification rates or reproductive numbers of pathogens. This inference relies on probabilistic models that represent how observed homologous sequences
evolved given these parameters of interest. Standard approaches such as maximum likelihood or Bayesian inference are well-suited to this problem but require to compute likelihoods, which can become very costly or impossible under realistic (and typically complex) probabilistic models. I will present an alternative likelihood-free approach exploiting samples rather than evaluations of the model. The methods I will present rely on neural network architecture that are specifically tailored to account for the symmetries of each inference problem.

AGAPE (computAtional G-quadruplex Affinitiy PrEdiction): The first AI workflow for G4 binding affinity predictor

Julie Fares

M2 AMI2B Paris-Saclay University,
ITODYS Paris Cité University

AGAPE (computational G-quadruplex Affinity Prediction) is a novel machine learning (ML)-based tool designed to predict the binding and stabilizing potential of small molecules targeting G-quadruplexes (G4s). G4s, prevalent in telomeres and oncogene promoters, are promising therapeutic targets, but designing selective binders remains challenging. Building upon a curated dataset of 1217 compounds annotated through Förster Resonance Energy Transfer (FRET) melting assays, AGAPE integrates 5666 molecular descriptors, both classical and quantum chemical. It captures features relevant to G4 recognition, driving researcher to predict the potential G4 stabilization of small molecules, including both organic ligands and metal complexes. Among the trained ML models, XGBoost achieved the best performance with an accuracy of nearly 91%, using 489 selected features. SHAP analysis highlighted descriptors related to molecular topology, polarizability, and electrostatic potential as key contributors to the classification. AGAPE is deployed through a user-friendly web interface supporting batch prediction and secure data handling and provides a robust and interpretable tool to accelerate the discovery of G4-stabilizing compounds, integrating quantum chemical information within an ML-driven cheminformatics framework.

A comprehensive benchmark for RNA structure-function modeling

Wissam Karroucha

PhD, CBIO, Mines Paris, PSL University, Institut Curie

The relationship between RNA structure and function has recently attracted interest within the deep learning community, a trend expected to intensify as nucleic acid structure models advance. Despite this momentum, the lack of standardized, accessible benchmarks for applying deep learning to RNA 3D structures hinders progress. To this end, we introduce a collection of seven benchmarking datasets specifically designed to support RNA structure–function prediction. Built on top of the established Python package RNAGlib, our library streamlines data distribution and encoding, provides tools for dataset splitting and evaluation, and offers a comprehensive, user-friendly environment for model comparison. The modular and reproducible design of our datasets encourages community contributions and enables rapid customization. To demonstrate the utility of our benchmarks, we report baseline results for all tasks using a relational graph neural network.

Structural Space of Microproteins with Protein Language Models

Simon Herman

PhD - Researcher
I2BC Paris Saclay

Microproteins—short proteins under 100 amino acids—are increasingly recognized as key regulators of diverse biological processes, yet their structural properties remain poorly understood. Traditional structure prediction methods often fail on such short, evolutionarily young sequences. Here, we present a deep learning–based computational framework leveraging protein language models (pLMs) to infer microprotein structural properties without relying on evolutionary data. Using embeddings from ProtT5-XL, we characterized thousands of annotated microproteins (UniProt) and potential intergenic ORF-encoded (iORF) microproteins across eukaryotic genomes spanning a wide range of GC content.

Detecting rare codon in proteins sequences with a Transformer model

Hélène Bret

Postdoc Researcher,
Lund University

The genetic code is highly redundant, with each amino acid encoded by one to six synonymous codons. However, codon usage is not random and influences RNA structure, regulatory signals, and ribosomal translation rates. Different codons are translated at varying speeds due to differences in tRNA availability: frequent codons, associated with abundant tRNAs, enable rapid elongation, while rare codons slow translation by relying on scarce tRNAs.

Rare codons, in particular, are non-randomly distributed. They play a fundamental role in regulating translation dynamics and have been maintained throughout evolution for functional reasons, such as co-translational folding or regulatory control. Despite their biological importance, rare codons are often neglected by existing optimization approaches (e.g., codon optimization or harmonization) or poorly predicted by current models due to their low representation in training datasets.

In this study, I trained a transformer-based model that takes an amino acid sequence as input and predicts the corresponding codon sequence, effectively reversing the natural translation process and explicitly accounting for synonymous codon choice.

This approach improves the prediction of rare codons compared to previous methods and shows strong agreement with a real-world case study on a mutational dataset of the TEM-1 β-lactamase enzyme. Analysis of internal model layers, attention heads, and gradient-based attribution revealed both short- and long-range sequence patterns as key determinants of codon choice learned by the transformer.

These results highlight the ability of deep learning models to capture subtle evolutionary and structural constraints underlying codon usage.

Computational approches to characterize cellular and spatial heterogeneity in cancer

Keynote Speaker - Magali Richard

Researcher in Computational Biology,
Laboratory of Translational Research and Innovation in Medicine and Complexity — CNRS, Université Grenoble Alpes

Cancer is a highly heterogeneous disease, with each tumor evolving as a multicellular, self-organized system. Tumors comprise diverse cell types of distinct origins, interacting dynamically to shape a complex ecosystem. This cellular heterogeneity is a major driver of cancer progression, yet remains challenging to observe, quantify, and interpret. Our limited ability to accurately estimate it continues to hamper a comprehensive understanding of oncogenesis. At the interface of bioinformatics, biostatistics, and oncology, our work develops computational strategies to analyze high-dimensional, multimodal molecular data, including single-cell and spatial transcriptomic datasets. These tools aim to disentangle tumor complexity and reveal its functional implications.

In this talk, I will present approaches we have designed to characterize tumor heterogeneity. I will illustrate them with two examples: (i) the construction of single-cell atlases to describe cellular heterogeneity in pancreatic adenocarcinoma, and (ii) the analysis of spatial organization to predict how heterogeneity shapes tumor evolution. Finally, I will highlight our efforts to foster collaborative benchmarking and evaluation of computational algorithms through data challenge frameworks. By enabling transparent comparison of methods, such initiatives foster reproducibility and accelerate the translation of methodological research into biological discovery.

Inferring latent chromatin folding signatures from bulk Hi-C data using NMF : application to the bistable expression of SPI-1

Chérif SEDDIK
M2 AMI2B Paris-Saclay University,
I2BC Paris-Saclay

Non-negative Matrix Factorization algorithms (NMF) have been shown to effectively extract distinct molecular signatures from various types of high-dimensional data, such as extracting cancer-related signatures from gene expression microarrays [Brunet et al., 2004]. These "signatures" can be viewed as latent components that can be inferred from bulk data, meaning that it is possible to estimate underlying meaningful biological patterns as state-specific signals that are not directly observed in mixed data.

Salmonella Pathogenicity Island 1 (SPI-1) is a genetic locus encoding a type III secretion system essential for host cell invasion. Its expression was characterized as "bistable" [Kortebi et al., 2025] : in a genotypically identical population of Salmonella cells under conditions that mimic early stages of host infection, only a small fraction of cells activate SPI-1 while the majority repress it. To study the 3D folding of active and silenced SPI-1 chromatin, Hi-C (High-throughput Chromosome Conformation Capture) was used. However, Hi-C averages contacts across the heterogeneous population, producing a “bulk” contact matrix that merges signals from all cells in the sample, thus masking cell-to-cell variability and complicating interpretation of chromatin organization in SPI-1ON versus SPI-1OFF subpopulations of cells.

By applying NMF to normalized Hi-C matrices of wild-type, mutant, and wild-type GFP-sorted samples of Salmonella cells, we demonstrate that it is possible to: (1) estimate the proportion of subpopulation-specific cells in each sample, (2) infer subpopulation-specific chromatin interaction profiles as contact matrices across all samples, and (3) reconstruct the observed bulk contact matrix as a weighted sum of (1) and (2). This method provides an alternative to cell-sorting-based Hi-C for analyzing heterogeneous bacterial populations where gene expression is bistable.

Integrative breast cancer multi-omics

Jomar Sangalang

PhD, Institut Gustave Roussy, Paris-Saclay

Endocrine therapy (ET) is the standard first-line treatment for hormone receptor-positive, human epidermal growth receptor 2-negative (HR+/HER2-) breast cancer (BC). Despite initial response, most patients develop ET resistance, driven by mechanisms that remain incompletely defined.

We performed ATAC-seq, RNA-seq, and whole-exome sequencing on 459 HR+/HER2- BC samples, including primary BC (pBC) and metastatic BC (mBC). pBC were obtained from patients treated with adjuvant ET who later relapsed or sustained remission, while mBC represented ET-resistant tumors.

PAM50 probability scoring identified five clusters (C1–C5) resembling intrinsic subtypes (Basal, Her2, LumA, LumB and Normal, respectively). In pBC, relapse rate was highest in clusters C1 and C2. Cluster C2 was enriched for TP53 alterations, while MAP3K1 alterations were more frequent in cluster C3. In mBC, ESR1 mutations characterized cluster C4, cluster C2 exhibited ERBB2 alterations, and cluster C3 was enriched with FOXA1 mutations. Gene set enrichment in clusters reflected their associated intrinsic subtypes, both in pBC and mBC. Cluster C4 was associated with higher estrogen response, while clusters C1 and C2 showed enrichment of Basal/Her2 gene sets, AP-1/TFAP2 TF motifs, and poorer survival.

In logistic regression models, the prediction of relapse in pBC showed the association of C2 cluster, ERBB2 alterations, high tumor grade, CHRNA9 expression and peaks linked to CLNS1A and KCNH2 with relapse. Predictive modeling of ET-resistant mBC revealed key prognostic markers, notably, C4 cluster classification, and alteration of EP300, UBR5, KMT2C and MDM2.

Our analysis of chromatin accessibility, transcriptomes, and exomes revealed molecular heterogeneity underlying relapse and ET resistance within HR+/HER2- BC. Distinct features across clusters highlight potential mechanisms of therapeutic escape and biomarkers for stratification of HR+/HER2- patients.

Interpretable latent model predicts drug response and identifies biomarkers

Shuhui Wang
PhD, Dauphine University-PSL

Predicting the effect of drugs on cell viability is a central challenge in drug discovery. Artificial intelligence holds the promise to considerably accelerate this process by leveraging rich cellular data such as transcriptomics. However, current models predict either transcriptomes or inhibitory concentrations using drug and cell descriptors, but they fall short on integrating these sources information. Here, we propose DORA (DOse-Response Autoencoder), a deep learning model that predicts changes in transcriptomes and viability in a dose-dependent manner, knowing the unperturbed cell state. By enforcing a latent space consistent with cumulative dose effects, DORA matches other methods approaches at predicting transcriptomes and substantially outperforms other latent representations at viability prediction. By exploiting the transcriptome-viability relationship, the model recovers known biomarkers involved in cell-specific viability and suggests novel candidates. Overall, DORA provides a unified framework delivering actionable biological insights for drug screening.

Structural modelling and binding affinity prediction of the human PDZ-PBM interactome

Victor Reys

Postdoc, BonvinLab, Utrecht University

PDZ domains are involved in major cellular functions, such as the localization of cellular elements and the regulation of pathways. They do this by interacting with partner proteins PDZ binding motifs (PBM), which are C-terminus linear motifs. With 266 PDZ domains and up-to 5000 potential PBMs that can be found in the human proteome, the corresponding interactome could involve more than 1 million putative interactions. Not only endogenous proteins are interacting with PDZ domain, but also pathogens are hijacking this system to better navigate through the cell, making the understanding of the underlaying interacting network of great interest.

While experimental data requires specialized setup and tedious work, in silico predictions together with the training of machine-learning models can bridge the gap between available binding affinities and the complete description of the interactome involved with the human PDZome. In this work, we investigate various machine-learning methods for the prediction of the PDZ-PBM interaction specificities and binding affinities; from sequence-based Bayesian models, to more complex deep-learning architectures using protein language models embeddings together with convolutional attention network for a global training on the human PDZome. Finally, structure-based information, first requiring to model tens of thousands of complexes, learned using graph convolutional neural network, are bringing additional contact-based features and therefore describing the system in its most complete form.

The accurate prediction of the interactions involving the human PDZ domains will not only complement our understanding of the complexity to maintain homeostasis in the human cell but also allow the deciphering of pathways used by pathogens to infect the cell and exploit its function, and therefore envision potential new therapeutic solutions.

Edited by the JC2B Editorial Commitee

JC2B 2025 Edition

Supported by Université Paris-Saclay and I2BC