PhD thesis project
2026 Call for application

Causal discovery and Information Theory for high-dimensional biological data

DEADLINE: **May 15, 2026**, starting date **Sept-Dec. 2026**

General information

Call	2026
PhD Funding	PRAIRIE – Paris School of AI
Keyword(s)	Causal discovery; Information theory; Representation Learning; Single-cell multi-omic data; Live cell time lapse imaging data.

Director(s) and team

Thesis supervisor	Hervé Isambert (DR, CNRS)
Research team	Computational Biology Lab, CNRS-UMR168, Paris
Research Institute	Institut Curie and Prairie – Paris School of AI, Paris

Description of the PhD thesis project

Discovering causal effects is at the core of scientific investigation but remains challenging when mostly observational data is available. In essence, causal discovery infers cause-effect relations from specific correlation patterns involving at least three variables, which goes beyond the popular notion that pairwise correlation does not imply causation. In principle, causal insights may also help improve the explainability and generalizability of AI methods. Yet, in practice, causal representations have been difficult to learn and interpret, in particular, for high dimensional data such as state-of-the-art biological data.

Single-cell multi-omic technologies and light microscopy imaging, routinely used in cell biology labs, produce massive amounts of gene expression and cell imaging data at single cell resolution. However, this wealth of high-dimensional biological data remains largely under-explored due to the lack of unsupervised methods and tools to interpret them without preconceived hypothesis. This highlights the need to develop novel Machine Learning and Artificial Intelligence strategies to better exploit the richness and complexity of the information contained in such high-dimensional cell biology data.

The Isambert lab recently developed novel causal discovery methods and tools (*i.e.* MIIC, CausalXtract, MIIC-sdg, CausalCCC, MIIC seach&score) to learn cause-effect relationships in a variety of biological or clinical data, from single-cell transcriptomic data (Fuzilier 2026, Donada 2026, Manfroi 2026, Dupuis 2025, Desterke 2020, Sella 2018, Verny 2027) and live-cell imaging data (Simon 2025, Parent 2025) to clinical data from medical records of patients (Ribeiro Dantas 2024, Sella 2022 & 2025, Cabeli 2020). These Machine Learning methods combine multivariate information analysis with interpretable graphical models (Lagrange 2024 & 2025, Li 2019), while allowing for unobserved latent variables, that are ubiquituous in biomedical applications. They outperform other causal methods on a broad range of benchmarks, achieving better results with much fewer samples, notably in the presence of nonlinear couplings between variables

(Lagrange 2025). Yet, MIIC and its related tools remain restricted to low-dimensional variables (*ie* categorical or continuous) and somewhat limited in the number of variables (*ie* up to a few hundreds). This PhD project aims to overcome these limitations.

The **first objective** is to extend these causal inference methods, originally developed for categorical and continuous variables, to high-dimensional variables such as (gene) sequences and (cell) images. To this end, Variational Auto-Encoders will first be used to obtain lower dimensional embeddings of relevant high-dimensional biological features. Alternatively, Information theoretic principles to estimate multivariate information between categorical or continuous variables (Ribeiro Dantas 2024, Cabeli 2021, Cabeli 2020) will be extended to high-dimensional settings (Isambert 2025) to circumvent the need of embeddings or feature extraction steps.

The **second objective** is to learn dimensionally reduced causal representations from high dimensional datasets in terms of number of variables, such as single-cell transcriptomic data including several thousands of genes. To this end, we will adapt our information theoretic framework to perform information-based feature selection prior to causal discovery analysis, so as to highlight the relevant gene interactions for a specific biological question, thereby providing human intelligible and testable explanations highlighting the most informative gene players. Multiple biological questions might also be addressed form the same dataset.

The **third objective** will be to exploit the resulting high-dimensional information theoretic framework and tools to study modern Deep Learning methods themselves. In particular, we will investigate how various neural network architectures process information during training, notably architectures relying on attention mechanisms. This line of research, rooted in our ability to estimate multivariate information in high dimension, will allow us to revisit and extend early theoretical frameworks based on information processing in neural networks (*eg* “information bottleneck”).

Finally, applications of this PhD project on actual biological data will be done in close collaboration with our biologist colleagues at Institut Curie, Institut Necker and Sorbonne University, as well as with other teams from Prairie-PSAI Institute for the information-based analysis of neural network information processing.

Related publications

- Cabeli V, Verny L, Sella N, Uguzzoni G, Verny M*, Isambert H*, Learning clinical network from medical records based on information estimates in mixed-type data. **PLoS Comput. Biol.** 16(5):e1007866 (2020).

- Cabeli V, Li H, Ribeiro-Dantas M, Simon F, Isambert H, Reliable causal discovery based on mutual information supremum principle for finite dataset in Why21 at **Neural Information Processing Systems** (**NeurIPS’21**) (2021).

- Desterke C, Petit L, Sella N, Chevallier N, Cabeli V, … , Isambert H, Jaffredo T, Charbord P*: Inferring gene networks in bone marrow Hematopoietic Stem Cell-supporting stromal niche populations. **iScience** 23(6):101222 (2020).

- Donada A, Hermange G, *Tocci T*, … , Isambert H, Cournède PH, Laplane L, and Perié L*. Clonal memory of cell division in humans diverges between healthy haematopoiesis and acute myeloid leukaemia, **Nature Commun**, *in press* (2026).

- Dupuis L, Debeaupuis O, Simon F, Isambert H*. CausalCCC: a web server to explore intracellular causal pathways enabling cell-cell communication, **Nucleic Acids Res** 53: W125-W131 (2025).

- Fusilier Z, Clement A, Simon F, …, Isambert H, Goudot C, Pierobon P, Lennon-Dumenil A-M, Moreau HD*. Macrophages restrict tumor immune infiltration by controlling collagen topography, **Science Immunology**, 11(117):eadw8291 (2026).

- Isambert H, Efficient computation of Mutual Information and Conditional Mutual Information in High Dimension, **Technical Note** (2025). [available upon request]

- Lagrange N, Isambert H. An efficient search-and-score algorithm for ancestral graphs using multivariate information scores, **arXiv**, 2412.17508 (2024). [opted-in paper at **NeurIPS'24**]

- Lagrange N, Isambert H*. An efficient search-and-score algorithm for ancestral graphs using multivariate information scores for complex non-linear and categorical data, (**ICML'25**) PMLR 267:32164-32187 (2025).

- Li H, Cabeli V, Sella N & Isambert H*, Constraint-based causal structure learning with consistent separating sets. **Advances in Neural Information Processing Systems** (**NeurIPS'19**) 32, 14257 (2019).

- Manfroi B, Dang VD, Bui-Thi C, Jungmann A, Borzakian S, Tong Y, Beauvineau C, Dupuis L, Guffart E, Borst K, … , Isambert H, Prinz M, von Kries P, Specker E, Walter J, Mahuteau-Betzer F, Fillatreau S. Induced regulatory B cells stably expressing IL-10 cure CNS autoimmunity by targeting local myeloid cells. **Immunity**, *in press* (2026).

- Parent C, Honari H, Tocci T, Simon F…Isambert H, Wilhelm C*, Viovy JL*. Label-free machine learning prediction of chemotherapy on tumor spheroids using a microfluidics droplet platform, **Small Science**, 2500173 (2025).

- Sella N, Verny L, Uguzzoni G, Affeldt S & Isambert H*, MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. **Bioinformatics** 34 (13):2311-2313 (2018). [related server at https://miic.curie.fr]

- Sella N, Hamy AS, Cabeli V, Darrigues L, Laé M, Reyal F*, Isambert H*, Interactive exploration of a global clinical network from a large breast cancer cohort. **npj Digital Med.** 5, 113 (2022).

- Sella N, Guinot F, Lagrange N, Albou L-P, Desponds J, Isambert H. Preserving Information while Respecting Privacy: An Information Theoretic Framework for Synthetic Health Data Generation, **npj Digital Med.** 8, 49 (2025).

- Simon F, Comes MC, Tocci T, Dupuis L, Cabeli V, Lagrange N, Mencattini A, Parrini MC, Martinelli E*, Isambert H*. CausalXtract: a flexible pipeline to extract causal effects from live-cell time-lapse imaging data. **eLife**, 13:RP95485 (2025).

- Ribeiro-Dantas M, Li H, Cabeli V, Dupuis L, Simon F, Hettal L, Hamy AS, Isambert H*, Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients, **iScience**, 27(5):109736 (2024).

- Verny L, Sella N, Affeldt S, Singh PP & Isambert H*, Learning causal networks with latent variables from multivariate information in genomic data. **PLoS Comput. Biol.** 13(10):e1005662 (2017).

Expected profile of the candidate

Applicants should have a strong background in machine learning or computer science and a keen interest to analyze complex heterogeneous data of biological and medical interests. Applicants should be proficient in programming and willing to interact with scientists from different disciplines, from data scientists to biologists. Applicants are expected to show a clear capacity for independent and creative thinking. Experience on causal inference analysis is a plus but not required as long as the applicant has a strong motivation to learn.

Non-discrimination, openness and transparency. All PR[AI]RIE-PSAI partners are committed to supporting and promoting equality, diversity and inclusion within its communities. We encourage applications from a variety of backgrounds, which we will take care to select via an open and transparent recruitment process.

Application

Please send complete CV, Master's transcripts with marks and the name(s) of one or more references to herve.isambert@curie.fr. Informal inquiries are welcome.
Application deadline May 15th, 2026. Starting date: Fall 2026.