Ph.D. Thesis Project: Deep learning approaches as predictors of the cell's regulatory networks

Summary

Single-cell transcriptomics (scRNA-seq) has revolutionized biology and medicine by unrevealing the diversity of the cells constituting human tissues. However, the complex task of inferring gene regulatory networks has not yet profited from this revolution and tools have not been predictive enough to reach any usability. Knowing such a network for each cell type is expected to provide a more comprehensive view of the cell's states and explain some of the different cellular phenotypes and responses of the cells.

The aim of this Ph.D. project is to develop novel approaches using deep learning and specifically novel graph neural network approaches on large scRNAseq datasets, to assess their predictability in high-quality benchmarks and package them as an open-source Python library.

The methods developed during this project will be applied in collaboration with wet-lab biologists, in order to derive new biological knowledge from their in-house data.

This PhD project will impact both computational fields and biomedical fields, by developing rigorous methods that can reliably maximize the information extracted from complex multimodal datasets. In particular, applications of the tool will contribute to personalized medicine.

Background and objectives

Single-cell omics. Single-cell RNA sequencing (scRNA-seq) has revolutionized biology and medicine by unrevealing the diversity of the cells constituting human tissues. The possibility to assess cellular heterogeneity at a previously inaccessible resolution has profoundly impacted our understanding of development, of the immune system functioning, and of many diseases.

G*ene regulatory network prediction.* The task of inferring gene regulatory networks is complex and only informs part of the overall gene regulation. The regulation of a gene's expression is linked to promoters & enhancers, both distal and proximal, bound by transcription factors (TFs). But many other components can have an effect on the transcriptome: co-factors, histone & RNApol2 binders, as well as RNA binding proteins (RBPs), non-coding RNAs binding at the DNA and RNA level, gene localization, and genomic mutations changing the processing and degradation speed of RNA, can also have an impact.

Moreover, it is known that TFs have combinatorial, competitive, and compensating effects on gene expression. Whereby one can readily replace another and two TFs combined can have a different effect than the sum of their part.

Existing methods. Gene Regulatory Networks (GRN)s have often been constrained to TF-gene relationships. Typically from co-Expression networks, derived from the gene expression correlation across many RNAseq measurements. Other methods only use this as a base for TF-gene relationships, then epigenetic sequencing methods like ChIP-seq & ATAC-seq, help refine these coarse models by knowing the putative binding of TFs to proximal regulatory elements of such genes.

However, these methods are all giving measurements averaged over many different cells. They also do not inform on the direct relationship between a TF and a gene. for that, other methods have been developed and are low throughput and rarely used.

With scRNA-seq becoming so widespread, we have seen the rise of databases recouping tens of millions of measurements across a wide range of contexts. This led to the development of easy-to-use tools and pipelines to interrogate this data size. Foundation models are promising to learn from and correct batch effects across these millions of measurements. The knowledge within their weights could be used for many downstream tasks. However, they are still poorly trained. Have transformer layers that are not suited for the RNAseq modality and make -often false- predictions that are hard to interpret.

Graph Neural Networks (GNN)s: are a class of deep learning models designed to operate on graph-structured data. They are specifically tailored to handle and process data that can be represented as graphs, where the elements (nodes or vertices) in the graph are connected by edges.

Traditional neural networks are primarily designed for processing grid-like data such as images or sequential data such as text. However, GNNs extend this capability to graph-structured data by incorporating graph-related operations and architectures.

The key idea behind GNNs is to learn node representations by aggregating and combining information from their neighboring nodes in the graph. In other words, GNNs allow each node to gather information from its direct neighbors and use that information to update its own representation. This neighborhood aggregation process is typically repeated iteratively for multiple layers to capture increasingly complex patterns and dependencies in the graph.

Objectives. We wish to improve GRN predictions from scRNAseq data. Our approach is by:

- Using larger neural network models that scale linearly with the dataset size, taking advantage of the 10Ms of data points now becoming available.
- Using novel GNN layers that could reduce the “search space” of the model by constraining the set of possible topologies it is learning.
- Better pretraining and fine-tuning these models to the predictive task they have to perform and the constraints of the system they are predicting.

- Formulating better layers that correspond to the sparse interactions between genes and our current knowledge about their functions.
- Creating formal and rational benchmarks that best capture the ability of a GRN methodology.
- Assessing predictions and any usefulness or lack of it by having biologist test hypotheses using the model.

Impact of the project. This Ph.D. project will thereby contribute to methodological breakthroughs by providing new tools and methods to use neural networks on unstructured data, like scRNAseq. And to improve on the state of the art in GRNs prediction from scRNAseq.

The proposed methodologies will impact computational (bioinformatics, machine learning) and biomedical fields. Fields that would be facing similar challenges that could be solved by GNN layers, as well as other fields taking advantage of scRNAseq profiles, such as environmental research, industrial biotechnology, and biofuel studies.

The impact of our methods will be further enhanced by open-source distribution. Concerning the expected applications of our methods, GRN inference per cell type is expected to impact the clinic in the context of precision medicine.

Detailed description of the project

This project is organized into three Work Packages (WP)s, WP1 and WP2 will be devoted to the formulation, implementation, and testing of the methods while WP3 will be devoted to delivering usable tools and experimental results through their application in collaboration with wet-lab biologists.

WP1: Review of current tools and creation of a set of benchmarks

The first step is to review the fundamental question of what are GRNs. What are their limitations and what kind of inferences we can do? From that and the review of the existing benchmarking methods, we will define what exactly comprises the task of GRN inference and what we consider good benchmarks for it.

For example, a GRN should allow us to predict gene knockout effects from experiments such as perturb-seq[] on both high[] and low-time resolution measures. It should allow us to predict cell differentiation trajectories[]. overall, considering the likely complexity of the GRNs, the goal is also one of interpretability for the community.

A good benchmark should thus review all of these requirements from existing methods.

WP2: GNN model/layers to better predict TF-gene relationships

We posit that the best model is one that can scale linearly to the data size of scRNAseq samples. One that can learn in an unsupervised way, across datasets. Based on architectures such as the variational auto-encoder models for scRNAseq[] and the more recent transformer models[] seem well suited.

However, although great similarities exist between GNNs and transformers[], GNNs have proven abilities through their equivariances and transformers currently lack good, meaningful positional encoding for unstructured data like scRNAseq. Improvement in either to tailor them to the problem at hand would yield significant improvement that could extend to other applications of neural networks.

Moreover, the tool could be pre-trained or reuse models pre-trained on large scRNAseq databases like CellxGene[] to take advantage of all the data accumulated in the community. It might also help decrease the model's bias toward various batch effects and mislabeling.

We will then assess the abilities of this method of inference and of the addition of GNN layers jointly and separately to understand how each is improving performances.

Finally getting something such as a scaling law for large unsupervised models on scRNAseq would

WP3: Collaboration to test the model's prediction on novel data

Collaborative projects are the workhorse of interdisciplinary science and became even more impactful through the last decade. During this Ph.D. project, the goal is also to collaborate with biologists to assess the model prediction. Moreover, collaboration on specific biological questions would help assess the usefulness of our tool and possibly improve it or help define new avenues of research. The Cantini group and the Ph.D. students have contacts with experimentalists at Institut Pasteur, but also at Dana Farber Cancer Insitute, Broad Institute.

Additionally, if time allows, collaborations could be envisioned with computational labs. It is possible that, like other ML problems, the best performance on the task of GRN inference will be achieved in the future through large generalist pre-trained models also called "foundation" models. Such models are being designed in some labs and will need specific re-design, fine-tuning, and assessment. Taking advantage of previously gained expertise on this task. The Ph.D. end on a joint project over such a foundation model.

Provisional Schedule

1. **WP1**: Sept 2023 - Jan 2024 (4 months) Review of current tools and creation of a set of benchmarks
2. **WP2**: Jan 2024 - Dec 2024 (10 months) GNN model/layers to better predict TF-gene relationships.
3. **WP3**: Sept 2024 - Jul 2025 (6 months) Collaboration to test the model's prediction on novel data.
4. **Buffer Time** (16 months) writing papers (3mo), writing the thesis (2mo), posters, events, time off (5mo) additional collaborations (6mo).

References

[1] Badia-i-Mompel, P., Wessels, L., Müller-Dott, S. et al. Gene regulatory network inference in the era of single-cell multi-omics. Nat Rev Genet (2023). https://doi.org/10.1038/s41576-023-00618-5
[2] Rood, J.E., Maartens, A., Hupalowska, A. et al. Impact of the Human Cell Atlas on Medicine. Nat Med 28, 2486–2496 (2022). https://doi.org/10.1038/s41591-022-02104-7
[3] Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, Mascibroda L, Wagner EJ, Adelman K, Lithwick-Yanai G, Iremadze N, Oberstrass F, Lipson D, Bonnar JL, Jost M, Norman TM, Weissman JS. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell. 2022 Jul 7;185(14):2559-2575.e28. doi: 10.1016/j.cell.2022.05.013.
[4] Colin Megill, Bruce Martin, Charlotte Weaver, Sidney Bell, Lia Prins, Seve Badajoz, Brian McCandless, Angela Oliveira Pisco, Marcus Kinsella, Fiona Griffin, Justin Kiggins, Genevieve Haliburton, Arathi Mani, Matthew Weiden, Madison Dunitz, Maximilian Lombardo, Timmy Huang, Trent Smith, Signe Chambers, Jeremy Freeman, Jonah Cool, Ambrose Carr. "cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices". bioRxiv 2021.04.05.438318; doi: https://doi.org/10.1101/2021.04.05.438318
[5] Argelaguet, R., Arnol, D., Bredikhin, D. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 21, 111 (2020). https://doi.org/10.1186/s13059-020-02015-1
[6] Haotian Cui, Chloe Wang, Hassaan Maan, Bo Wang. "scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI".bioRxiv 2023.04.30.538439; doi: https://doi.org/10.1101/2023.04.30.538439
[7] Lingfei Wang, Nikolaos Trasanidis, Ting Wu, Guanlan Dong, Michael Hu, Daniel E. Bauer, Luca Pinello. "Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multi-omics".bioRxiv 2022.09.14.508036; doi: https://doi.org/10.1101/2022.09.14.508036
[8] Vinay K. Kartha, Fabiana M. Duarte, Yan Hu, Sai Ma, Jennifer G. Chew, Caleb A. Lareau, Andrew Earl, Zach D. Burkett, Andrew S. Kohlway, Ronald Lebofsky, Jason D. Buenrostro, "Functional inference of gene regulation using single-cell multi-omics". Cell Genomics, Volume 2, Issue 9, 2022, https://doi.org/10.1016/j.xgen.2022.100166
[9] Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, Stein Aerts. "SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks". bioRxiv 2022.08.19.504505; doi: https://doi.org/10.1101/2022.08.19.504505
[10] Chen, Y., Xu, L., Lin, R.YT. et al. Core transcriptional regulatory circuitries in cancer. Oncogene 39, 6633–6646 (2020). https://doi.org/10.1038/s41388-020-01459-w
[11] Beagan, J.A., Phillips-Cremins, J.E. On the existence and functionality of topologically associating domains. Nat Genet 52, 8–16 (2020). https://doi.org/10.1038/s41588-019-0561-1
[12] Mann R, Notani D. Transcription factor condensates and signaling driven transcription. Nucleus. 2023 Dec;14(1):2205758. doi: 10.1080/19491034.2023.2205758. PMID: 37129580; PMCID: PMC10155639.
[13] Bellot, P., Olsen, C., Salembier, P. et al. NetBenchmark: a Bioconductor package for reproducible benchmarks of gene regulatory network inference. BMC Bioinformatics 16, 312 (2015). https://doi.org/10.1186/s12859-015-0728-4
[14] Cannoodt, R., Saelens, W., Deconinck, L., & Saeys, Y. (2021). Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nature Communications, 7. doi: 10.1038/s41467-021-24152-2. Retrieved from https://doi.org/10.1038/s41467-021-24152-2
[15] Zappia L, Phipson B, Oshlack A. "Splatter: Simulation Of Single-Cell RNA Sequencing Data". Genome Biology. 2017; doi:10.1186/s13059-017-1305-0.

[16] Przytycki, P.F., Pollard, K.S. “CellWalkR: An R Package for integrating and visualizing single-cell and bulk data to resolve regulatory elements.” Bioinformatics (2022). https://doi.org/10.1093/bioinformatics/btac150
[17] Zhang, Ziqi, Chengkai Yang, and Xiuwei Zhang. "Learning latent embedding of multi-modal single cell data and cross-modality relationship simultaneously." bioRxiv (2021).
[18] Fischer, D.S., Dony, L., König, M. et al. Sfaira accelerates data and model reuse in single cell genomics. Genome Biol 22, 248 (2021). https://doi.org/10.1186/s13059-021-02452-6
[19] Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet (2023). https://doi.org/10.1038/s41576-023-00586-w
[20] Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimée Bastidas-Ponce, Marta Tarquis-Medina, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, Fabian J. Theis “Mapping cells through time and space with moscot” bioRxiv 2023.05.11.540374; doi: https://doi.org/10.1101/2023.05.11.540374
[21] Lotfollahi, M., Naghipourfar, M., Luecken, M.D. et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol 40, 121–130 (2022). https://doi.org/10.1038/s41587-021-01001-7
[22] Luecken, M.D., Büttner, M., Chaichoompu, K. et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods 19, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8
[23] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier & Nir Yosef. “A Python library for probabilistic analysis of single-cell omics data” Nature Biotechnology 2022 Feb 07. doi:10.1038/s41587-021-01206-w
[24] Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Scverse Community, Bonnie Berger, Dana Pe’er, Aviv Regev, Sarah A. Teichmann, Francesca Finotello, F. Alexander Wolf, Nir Yosef, Oliver Stegle & Fabian J. Theis “The scverse project provides a computational ecosystem for single-cell omics data analysis”. Nature Biotechnology 2022 Apr 10. doi:10.1038/s41587-023-01733-8.
[25] Huynh-Thu V. A., Irrthum A., Wehenkel L., and Geurts P.
Inferring regulatory networks from expression data using tree-based methods.
PLoS ONE, 5(9):e12776, 2019.
[26] A.-C. Haury, F. Mordelet, P. Vera-Licona and J.-P. Vert. TIGRESS: trustful inference of gene regulation using stability selection. BMC systems biology 6(1), 145-153, 2012
[27] Deniz Seçilmiş and others, GRNbenchmark - a web server for benchmarking directed gene regulatory network inference methods, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W398–W404, https://doi.org/10.1093/nar/gkac377
[28] Michael Saint-Antoine, Abhyudai Singh. “Benchmarking Gene Regulatory Network Inference Methods on Simulated and Experimental Data”. bioRxiv 2023.05.12.540581; doi: https://doi.org/10.1101/2023.05.12.540581
[29] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, *17*(2), 147-154. https://doi.org/10.1038/s41592-019-0690-6