Suggestions from Pietro Lio's Research group

Proposals

1. Predicting the flow of information from DNA to RNA and RNA to Protein

Idea: Experimental techniques to measure multiple modalities within the same single cell are increasingly becoming available. The demand for these measurements is driven by the promise to provide a deeper insight into the state of a cell. Yet, the modalities are also intrinsically linked. We know that DNA must be accessible (ATAC data) to produce mRNA (expression data), and mRNA in turn is used as a template to produce protein (protein abundance). These processes are regulated often by the same molecules that they produce: for example, a protein may bind DNA to prevent the production of more mRNA. Understanding these regulatory processes would be transformative for synthetic biology and drug target discovery. Any method that can predict a modality from another must have accounted for these regulatory processes, but the demand for multi-modal data shows that this is not trivial. This project is based on Task 1 of the NeurIPS 2021 competition "Open problems in Single-Cell Analysis" (https://openproblems.bio/neurips_docs/about_tasks/task1_modality_prediction/).

Expected outcome: During this project you will learn about: bioinformatics, generative models, and deep learning.

Ideal target student: Interested in computational biology

People interested in (co)supervising: Arian, Ramon

2. Matching profiles of each cell from different omics modalities.

Idea: While joint profiling of two modalities in the same single cell is now possible, most single-cell datasets that exist measure only a single modality. These modalities complement each other in their description of cellular state. Yet, it is challenging to analyse uni-modal datasets together when they do not share observations (cells) or a common feature space (genes, proteins, or open chromatin peaks). If we could map observations to one another across modalities, it would be possible to treat separately profiled datasets in the same manner as new multi-modal sequencing data. Mapping these modalities to one another opens up the vast amount of uni-modal single-cell datasets generated in the past years to multi-modal data analysis methods. This project is based on Task 2 of the NeurIPS 2021 competition "Open problems in Single-Cell Analysis" (https://openproblems.bio/neurips_docs/about_tasks/task2_modality_matching/).

Expected outcome: During this project you will learn about: bioinformatics, generative models, and deep learning.

Ideal target student: Interested in computational biology

People interested in (co)supervising: Arian, Ramon

3. Learn a joint embedding from multiple omics modalities.

Idea: The functioning of organs, tissues, and whole organisms is determined by the interplay of cells. Cells are characterised into broad types, which in turn can take on different states. Here, a cell state is made up of the sum of all processes that are occurring within the cell. We can gain insight into the state of a cell by different types of measurements: e.g., RNA expression, protein abundance, or chromatin conformation. Combining this information to describe cellular heterogeneity requires the formation of joint embeddings generated from this multimodal data. These embeddings must account for and remove possible batch effects between different measurement batches. The reward for methods that can achieve this is great: a highly resolved description of the underlying biological state of a cell that determines its function, how it interacts with other cells, and thus the cell's role in the functioning of the whole tissue. This project is based on Task 3 of the NeurIPS 2021 competition "Open problems in Single-Cell Analysis" (https://openproblems.bio/neurips_docs/about_tasks/task3_joint_embedding/).

Expected outcome: During this project you will learn about: bioinformatics, generative models, and deep learning.

Ideal target student: Interested in computational biology

People interested in (co)supervising: Arian, Ramon, Pietro

4. Multi-tissue imputation of transcriptomics data

Idea: Gene expression (e.g. RNA-seq) is difficult to acquire for certain human tissues because they are inaccessible. The goal of this project is to develop a computational model to infer the expression measurements of difficult-to-acquire tissue types (e.g. heart, kidney, or brain tissues) given the transcriptomes of a variable number of accessible tissues (e.g. whole blood or skin).

Suggested readings (more material can be provided):

- Predicting tissue-specific gene expression from whole blood transcriptome. https://www.science.org/doi/10.1126/sciadv.abd6991
- Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833292/
- Gene Expression Imputation with Generative Adversarial Imputation Nets. https://www.biorxiv.org/content/10.1101/2020.06.09.141689v1
- Adversarial generation of gene expression data. https://www.biorxiv.org/content/10.1101/836254v1
- Neural Processes. https://arxiv.org/abs/1807.01622

Expected outcome: During this project you will learn about: bioinformatics, generative models, and deep learning.

Ideal target student: Interested in computational biology

People interested in (co)supervising: Ramon (+ Simon)

5. Reinforcement Learning for structure-based drug design

Idea: Reinforcement Learning (RL) has previously been proposed as a method for designing protein therapeutics with high binding affinity to a drug target (e.g. the SARS-CoV-2 Spike protein) https://arxiv.org/abs/2012.01736 . This is performed via an iterative process where an agent learns to mutate amino acids in a binding protein sequence to optimise a given score (in this case the molecular docking score as calculated by a physics based docking program).

Expected outcome: This project will try to design novel small molecules (~10-30 atoms) with a high binding affinity to a drug target using a combination of conventional molecular docking and RL.

Ideal target student: Interested in computational biology and reinforcement learning

People interested in (co)supervising: Charlie (+ Simon)

6. Self-supervised learning of ligand docking.

Idea: Attention-based protein language models (i.e. transformers) have been shown to be self-supervised learners of protein structure when only trained on protein sequences (https://openreview.net/pdf?id=fylclEqgvgd). This is desirable as there is a substantial gap between the number of proteins sequences and protein structures available. Similarly, in drug design, there are many small molecule compounds which we know to bind to a protein target but we do not know the 3D conformation of the binding pose.

Expected outcome: This project will try to use attention-based models to perform self-supervised learning of ligand docking.

Ideal target student: Interested in self-supervised learning and computational biology

People interested in (co)supervising: Charlie

7. Assessment of Protein-Protein/Protein-Ligand/Antibody Docking

Idea: Protein-Protein docking typically produces a list of plausible structures of what a protein complex may look like. These are typically ranked by a scoring function. Here, the goal is to learn this scoring function from data in order to rank the outputs of widely-used docking programs

Expected outcome: The student will gain familiarity with working with proteins structural data and state-of-the-art representation learning techniques.

Ideal target student: Interested in computational biology

People interested in (co)supervising: (Charlie?)

8. Protein Graph Fingerprinting

Idea: Proteins can naturally be represented as graphs of interacting amino acid residues. The goal of this project is to produce learnable fingerprints of protein graphs (there are numerous constructions possible) and to

Expected outcome: The student will gain familiarity with working with proteins structural data and state-of-the-art representation learning techniques. The student will have the flexibility to apply and benchmark the fingerprints in range of downstream tasks that appeal to their interest

Ideal target student: Interested in computational biology

People interested in (co)supervising: Arian, Pietro

9. Assessment of Protein-Protein/Protein-Ligand/Antibody Docking

Idea: Protein Graph informatics. Recently, the AlphaFold2 database has made available a huge number of protein structures (98%+ of the human proteome). The goal of this project is to apply ideas of graph theory and topological data analysis to mine these data for biological insight. The exact questions are open-ended and can be shaped in the context of the student's interest.

Expected outcome: The student will gain familiarity with protein structural data

Ideal target student: Interested in computational biology

People interested in (co)supervising:

10. Down Syndrome biomedical digital twin

Idea: Complex biomedical conditions like Down Syndrome cannot be studied from a unique point of view. The comorbidity landscape of such conditions requires having a complete overview on the human body at different levels, from genomics to metabolomics, from proteomics to physiology. This project aims at developing graph and generative neural models representing "biomedical twins", i.e. neural models considering the organism as a whole. The project will be based on fresh data made available by the European Consortium on Down Syndrome (https://go-ds21.eu/), encompassing a wide variety of omics and data modalities (from genomics to physiology to FitBit data).

Expected outcome: The student will gain familiarity with digital twins, graph and generative models, and Down Syndrome

Ideal target student: Interested in computational biology, multi-omics, graph and/or generative models

People interested in (co)supervising: Pietro, Ramon

11. Active learning for substructures in graphs.

Idea: Active learning is a machine learning paradigm where the focus is not on improving a model to perform better on a given dataset, but to devise algorithms that will help improve the dataset such that the model will perform better in subsequent train/test iterations. Active learning methods are commonly used in industry where they have the ability to improve their datasets (and thereby their models) in an economic manner. These methods use various approaches which utilise different sources of informativeness such as uncertainty to measure the usefulness of given unlabeled observations for labelling by an oracle.

This project explores the research and implementation of active learning techniques on graph structured data. Initial work on a general framework for active learning with PyTorch models has been completed, the student will be expected to research existing active learning techniques for graph structured data and extend them.

Expected outcome: The student will become familiar with the active learning paradigms, some of its methods, and GNNs. Will potentially contribute to PyTorch active learning library managed by co-supervisor. This will involve learning how to use PyTorch Geometric library.

Ideal target student: Interested in active learning, GNNs and graph methods, self-driven to research new methods for active learning on graph-structured data.

People interested in (co)supervising: Paul Scherer (pms69)

12. Multi-aspect node representation learning.

Idea: graph representation learning has to a large extent taken over machine learning research. The canonical task in GRL is node representation learning wherein we try to learn a vector representation of a node which contains information about its local/global neighbourhood. As an example of a citation network where nodes represent research papers and edges citations between papers. Each node is also associated with a feature vector describing the content of the

papers. The task is to classify the papers into different scientific categories such as “physics” or “cs” and so on. This is where graph representation learning methods made their mark as they are able to not only use the paper’s feature vector as part of the learning process but also the papers it is related to (and their feature vectors).

Now there is a plethora of node representation learning techniques. A bulk of these methods are centered around learning a single representation for a node. That is to say within a node classification scenario we seek to identity a single classification or role that characterises the node within the networks. In the context of clustering a social network this would equate to partitioning social network into non-overlapping sets. However, a much more realistic view is that people would belong to multiple communities to reflect that we have multiple personae, in biological contexts

This project aims to co-develop methods for learning node representations which jointly learn multiple representations of a node with respect to the various communities it is involved in.

Expected outcome: The student will become familiar with standard (single-aspect) and multi-aspect graph representation learning. Particularly good progress

Ideal target student: Interested in active learning, GNNs, self-driven to research new methods for active learning within GNN contexts.

People interested in (co)supervising: Paul Scherer (pms69)

13. Dynamic graph representation learning.

Idea: Dynamic graphs describe graphs that change over time, either in the nodes, edges, substructure features, structure, and sometimes all at once. This makes dynamic graphs particularly challenging to model. There are a number of open challenges to tackle in the modelling of dynamic graphs such as modelling arbitrary time stamps in continuous time dynamic graphs, modelling dynamic heterogeneous graphs, tackling scaling, feature oversmoothing in distant predictions and more. Each of these challenges can be tackled as part of a project and we can discuss which aspects to take forward.

Several students can apply for this project to tackle different aspects of dynamic graph modelling.

Expected outcome: The student will research dynamic graphs, prior experience with GNNs will be beneficial to get the most out of the project. Methods may be included into the Pytorch Geometric Temporal library.

Ideal target student: Interested in dynamic graph representation learning, self-driven to research new methods for discrete and continuous time dynamic graphs. Open ended project where many aspects can be explored and support will be given to research

People interested in (co)supervising: Paul Scherer (pms69)

14. Neural algorithmic reasoning for pseudotime trajectory inference

Idea: Generating trajectory inference (a.k.a. pseudotime) has been listed [1] as one (of the 11) of the key challenges in single-cell data science. Generating trajectory inference is the generation of a potential path a cell can undergo in its lifetime (from cell of type A to cell of type B -> C -> D…). Some proposed models [2] use minimum spanning tree algorithms (MST), but in order to build the graph on which MST is applied, several hand crafted feature transformation algorithms need to be applied.

Neural algorithmic reasoning provides an alternative to execute an algorithm, when the input data is not fully specified [3] (a concrete instantiation of the idea was recently accepted at NeurIPS [4]). The goal of this project is to apply ideas of neural algorithmic reasoning for calculation of pseudotime trajectories. Achieving this will circumvent the need of handcrafting features for the MST algorithm and may naturally allow for speedups by parallelising the algorithms on the GPU (neural algorithmic reasoning uses GNNs).

[1] https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-020-1926-6.pdf
[2] https://academic.oup.com/nar/article/44/13/e117/2457590
[3] https://arxiv.org/abs/2105.02761
[4] https://arxiv.org/abs/2110.05442

Expected outcome: The student will gain familiarity with single cell biology, neural algorithmic reasoning (main focus) and GNNs

Ideal target student: Interested in computational biology, computer science algorithms, GNNs

People interested in (co)supervising: Dobrik, Pietro (Pietro?)

15. DietVAEs for representation learning from limited data

International initiatives such as the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and The Cancer Genome Atlas (TCGA) are collecting multiple data sets at different genome-scales with the aim to identify novel cancer bio-markers and predict patient survival. To analyse such data, several machine learning, bioinformatics and statistical methods have been applied, among them neural networks such as autoencoders.

Although these models provide a good statistical learning framework to analyse multi-omic and/or clinical data, there is a distinct lack of work on how to integrate diverse patient data and identify the optimal design best suited to the available data.

Recent work [1] has focused on investigating various variational inference approaches [2], namely variational autoencoders, for integrating a variety of cancer patient data types (e.g., multi-omics and clinical data). The results from an extensive empirical study show that such integrative approaches yield relevant/good-quality data representations which, in turn, lead to accurate and stable diagnosis. These approaches, however, while relevant, stipulate that a certain (high) quantity of data is available at input. Although, in practice, for most clinical trials this is not the case. Namely, typically such trials produce high-dimensional heterogeneous data from only a few patients, thus challenging the currently developed approaches.

The project will focus on these challenges - designing, developing and evaluating approaches for training integrative variational autoencoders from limited data. This might include: (1) developing novel architectures and mechanisms for training them under such constraints, such as DietNets [3] and FSNets [4] (2) transfer learning approaches employing available data from public repositories such as The Cancer Genome Atlas (TCGA) During this project, you will learn: deep learning (PyTorch and/or Keras framework), variational/Bayesian inference, and make use of as well as integrating multiple data types - which is a strong preparation for a Phd or industry role in machine learning. The ideal candidate for this project will have a strong background in mathematics and statistics as well as good Python programming skills. Familiarity with bayesian methods and bioinformatics is a plus.

- [1] https://www.frontiersin.org/articles/10.3389/fgene.2019.01205/full
- [2] https://arxiv.org/abs/1312.6114
- [3] https://arxiv.org/abs/1611.09340
- [4] https://arxiv.org/abs/2001.08322v1

People : Nikola Simidjievski, Andrei Margeloiu, Mateja Jamnik, Pietro Lio