Assessing Drosophila neurons types based on topology and connectomics with GNNs

An increase in the speed and quality of modern neuroimaging techniques allows for large-scale connectomics studies of invertebrates, i.e. Drosophila. Connectomics is useful in generating biological insights about ensembles of neurons with interesting behavioural functions. New computational tools are needed to search and organise these data. For instance, one important question is how similar are neurons between two hemispheres of the fruit fly brain. To answer these questions a heuristic NBLAST algorithm is used. It considers both position and local geometry by decomposing neurons into short segments; matched segments are scored using a probabilistic scoring matrix. However, NBLAST ignores the topology of a neuron and its connectomics. On the other hand, neuronal connections in the brain can be represented as a graph. Graph Neural Networks have been shown to account both for local node features and adjacency matrix in the classification task. The aim of this project is to create classification model that would allow to determine the types of neuronal cells in order to improve neurons taxonomy.

Fixing neuronal morphology with graph generative models

An increase in the speed and quality of modern neuroimaging techniques allows for imaging individual neuronal cells in Drosophila. Previously, electron microscopy images were annotated manually in search of neuronal tracts. A skeleton of neuron can be represented as a directed graph with the beginning in the cell body. Today, machine learning algorithms (like flood-filling neural networks) vastly improve the tracing of neurons, but they leave tracers with a number of false positives and negatives. One artefact of the method (related also to the imperfect EM data collection process) is lack of continuity between certain neuronal branches.

Currently, as a remedy, a simplistic assumption is made to interpolate missing part of the neuron as a straight line. Generative graph models could help to render more biologically plausible replacement of the missing part of the neuron, which on the other hand could facilitate the process of visual inspection of annotations and neuron matching.

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 to have 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 Barbiero and pietro Lio’
Neural algorithmic reasoning for pseudotime trajectory inference

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 handcrafted 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).


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: Pietro Lio <pl219@cam.ac.uk>, Dobrik Georgiev <dgg30@cam.ac.uk>, Ramon Viñas Torné <rv340@cam.ac.uk>

Deep learning applied to immune repertoires for early cancer detection (computational)

The immune system is an exquisitely sensitive and highly specific in-built diagnostic system. It is now known that the immune system plays a central role in recognizing and responding to cancers - which has led to the remarkable success of immunotherapy. Immune cells rely on the diversity of B- and T-cell receptor sequences to recognize antigens including those caused by cancer. The entire "repertoire" of receptor sequences therefore represents a recorded history of the antigens to which we have been exposed, including those from a nascent cancer. Could these immune repertoires be used as a powerful early warning system for cancer? In this project we will explore this question by training a deep learning framework on a large database of T-cell receptor sequences (sequences that are responsible for the recognition of antigens) from 1000s of cancer cases and healthy controls. The aim is to "learn" which features of a T-cell receptor sequence best predict that it has successfully recognised a tumour and apply this to sequences from blinded data to assess to what extent one can predict cancer case / control status from T-cell receptor sequences alone.

Supervisor: Pietro Lio
Co-supervisors: Lilly Wollman, Jamie Blundel
Optimal encodings of T-cell receptor sequences for deep learning applications (computational)

T-cells rely on the diversity of their receptor sequences to recognize foreign antigens such as viruses. The part of the receptor that recognizes the antigen is a short (12-16 amino acids) but highly diverse region called the CDR3. A typical human will have between 10^8-10^10 different receptor sequences in their blood each of which is highly specific for a given antigen. A major goal in biomedical science is to learn which receptor sequences are associated with which diseases, and various deep learning frameworks have been used to attempt to learn these associations. However, this raises the question: how should one represent these short amino acid sequences? Do some encodings enable much better learning performance than others? What are the "right" kinds of encodings for this task? In this project the student will be given a large database of CDR3 sequences from cancer cases and healthy controls and will explore the classification performance of different types of amino acid encoding.

Supervisor: Pietro Lio
Co-supervisors: Lilly Wollman, Jamie Blundel

RCB & residual tumour cellularity predictor

Description: Digital Pathology images from Breast Cancer provide information at a single cell level of the tumour state and evolution. After being treated with neoadjuvant therapy, breast cancer patients show different levels of response to the treatment. An important measure of such response is the Residual Cancer Burden (RCB) score, based on the amount of cancerous cells remaining in the tumour bed after treatment. By using segmentation techniques, we are able to extract catalogues of cells from each patient's image, and perform further calculations to obtain for instance the distributions of cell densities and their heterogeneity.

In this project, the students will be developing an AI/ML driven tool to compute the Residual Cancer Burden for breast cancer patients after neoadjuvant therapy. There are multiple approaches that could be taken in order to compute the score, starting from basic computer vision regression to more complex approaches using Graph Neural Networks over the cell catalogues.

There are several challenges when working with medical data, such as data scarcity and interpretability, that the student will be tackling during this project. The methodology developed for RCB scoring will need to be scalable to low sample datasets (hundreds of patients). One solution is to develop techniques that are focused on individual patients (with hundreds of thousands of cells each) that are then able to generalise to different images. During their research, the student will also have to pay special attention to the interpretability of their models, they will need to be explainable and understood by the medical experts that will ultimately be using them.

In this project, the student can expect to gain experience on the implementation of AI/ML models on medical data. They will be learning about multiple network based approaches (CNNs for segmentation, DNNs for classification and potentially GraphNNs for mapping cellularity over the tumour cell distribution). The final goal will be to identify the optimal approach to extract RCB scores, so there will be flexibility in term of the model or methodology ultimately used. The student will be working closely with the AI - Integrative Cancer Medicine group in the Computer Lab, as well as collaborating with experts in pathology and Cancer Research UK.

Resources:
- https://www.nature.com/articles/s41598-019-50568-4

Supervisor: Pietro Lio
Co-supervisors: Helena Andres Terra