Project suggestions from Prof. Pietro Lio 2024-25

Spatio-Temporal Graph Transformers for Time Series on Graphs.

Supervisor: Pietro Lio', pl219@cam.ac.uk

Time series data consists of a sequence of observations collected or recorded over a period of time. We can use a temporal graph structure to provide a correlation prior between multiple simultaneously evolving time series, represented as a (spatio-)temporal graph. Temporal graphs are useful for modeling complex dynamical systems and have applications in fields such as healthcare, biology, and finance, among others. Although this emerging area of research has recently garnered attention, it remains relatively underexplored, as most traditional graph representation learning literature has focused on static graphs. In this project, we aim to extend Graph Transformers to temporal graphs. Graph Transformers are an adaptation of the ubiquitous Transformer architecture, widely used in the large language model community for models such as GPTs and LLAMAs, to graph representation learning. Recent Graph Transformer architectures typically combine a message-passing layer, which leverages the locality inductive bias of the input graph, with a global attention mechanism to capture long-range dependencies. However, these architectures are generally tailored for static graphs. To generalize this approach for dynamic graphs with evolving node features and graph topology, our first step will be to introduce spatio-temporal attention. The student will be tasked with extending existing frameworks such as graph GPS with the aforementioned new form of attention, and to evaluate it in recently introduced benchmarks such as the Temporal Graph Benchmark. Since generally performing spatio-temporal attention can be computationally expensive there will also be the possibility of modeling time dynamics drawing inspiration from architectures such as Mamba (and state space models in general) and RWKV.

References:

1) "A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection" https://arxiv.org/pdf/2307.03759

2) "Elucidating Graph Neural Networks, Transformers, and Graph Transformers" https://www.researchgate.net/publication/378394991_Elucidating_Graph_Neural_Networks_Transformers_and_Graph_Transformers/stats

3) "Recipe for a General, Powerful, Scalable Graph Transformer" https://arxiv.org/abs/2205.12454, https://github.com/rampasek/GraphGPS4) "Temporal Graph Benchmark for Machine Learning on Temporal Graphs" https://arxiv.org/abs/2307.010265) "RWKV: Reinventing RNNs for the Transformer Era" https://arxiv.org/abs/2305.130486) "Mamba: Linear-Time Sequence Modeling with Selective State Spaces" https://arxiv.org/abs/2312.00752

Improving Graph Autoencoders for Latent 3D Graph Diffusion

Supervisor: Pietro Lio', pl219@cam.ac.uk

Typically, diffusion generative models are constructed directly in the product space of 3D graph topology, including nodes, edges, and coordinates. However, a recently proposed framework, Latent 3D Graph Diffusion, draws inspiration from latent diffusion models such as Stable Diffusion in the context of image generation. The authors primarily focused on developing a suitable low-dimensional latent space and an accompanying autoencoder for performing latent graph diffusion. In this work, the student will aim to enhance the original 3D graph autoencoder proposed by the authors by augmenting the training pipeline with additional loss functions beyond simple pointwise reconstruction of the input sample. Indeed, standard image autoencoders are not trained purely with pixelwise reconstruction metrics like MSE; they are often augmented with perceptual losses, such as LPIPS (based on the activations of a VGG network), and adversarial losses (based on GANs). Drawing inspiration from the image generation literature, we will adopt similar approaches to improve the latent 3D Graph Diffusion autoencoder and evaluate the impact of these enhancements on the downstream task of graph diffusion. This will involve introducing semantically meaningful perceptual losses in the context of graphs and implementing discriminators to add an adversarial signal during training, as well as retraining the graph diffusion model in the new latent space.

References:

1) "Latent 3D Graph Diffusion" https://openreview.net/pdf?id=cXbnGtO0NZ, https://github.com/Shen-Lab/LDM-3DG

2) "High-Resolution Image Synthesis with Latent Diffusion Models" https://arxiv.org/pdf/2112.10752

3) "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" https://arxiv.org/abs/1801.03924

4) "Generative Adversarial Networks" https://arxiv.org/abs/1406.2661

Improving gRNAde (Geometric Deep Learning for 3D RNA Inverse Design) revising GNN backbound and sequence model fine-tuning with LoRA

Supervisor: Pietro Lio', pl219@cam.ac.uk

Computational RNA design is generally approached as an inverse problem, where the goal is to create sequences that fold into a predefined secondary structure. This approach often neglects the 3D geometry and conformational flexibility of RNA molecules. To address these limitations, the gRNAde pipeline was recently introduced. gRNAde is a geometric RNA design framework that works directly with 3D RNA backbones, allowing for the design of sequences that account for both structural integrity and dynamic behavior. The core of gRNAde is a multi-state Graph Neural Network (GNN) that generates candidate RNA sequences based on one or more 3D backbone structures with unknown base identities. In this project, the student will focus on enhancing and benchmarking the original GNN-based architecture used in gRNAde. Additionally, we will explore the application of techniques such as Low-Rank Adapter (LoRA) fine-tuning, which is widely used in large language models, to refine the sequence decoder. While in the original paper the sequence decoder was employed as an evaluation metric, it would be intriguing to explore end-to-end training of the entire framework. LoRA allows to refine pre-trained Transformer models by updating only a small subset of the model parameters (particularly focusing on the attention layers), without requiring large computational resources.

References:

1) "gRNAde: Geometric Deep Learning for 3D RNA inverse design" https://arxiv.org/abs/2305.14749, https://github.com/chaitjo/geometric-rna-design

2) "LoRA: Low-Rank Adaptation of Large Language Models" https://arxiv.org/abs/2106.09685

3) "RNA secondary structure packages evaluated and improved by high-throughput experiments" https://www.nature.com/articles/s41592-022-01605-0

4) "RhoFold+: Accurate RNA 3D structure prediction using a language model-based deep learning approach" https://github.com/ml4bio/RhoFold

5) "Asymmetry in Low-Rank Adapters of Foundation Models" https://arxiv.org/abs/2402.16842

Complex diseases bioinformatics and machine learning

Supervisor: Pietro Lio', pl219@cam.ac.uk

This project's main focus will be on the integration of complex disease data taken from patients with Down syndrome or Multiple Sclerosis, with molecular RNA sequence data and clinical data. As an extension, this will then be used to train Graph Neural Networks (GNNs) used in forming a 'digital twin' of a patient. A digital twin is defined as a set of virtual information constructs that mimics the structure, context and behaviour of an individual or unique physical asset, that is dynamically updated with data from its physical twin throughout its life cycle, and that ultimately informs decisions that realize value. As modern medicine shifts to providing more precise and personalised treatment plans for patients, the importance of using digital twins to inform high-stakes decisions will only grow.

A combination of RNA sequence, methylation, metagenomic and clinical data will be used for this project and they all have arrived and are readily available. My supervisor has already obtained all of the ethical approvals needed to use this data. The aim of the project is to integrate together all of this data using a combination of deep learning and bioinformatics techniques. The project will involve analysing transcriptome and DNA methylation data of Down syndrome or Multiple Sclerosis patients to identify novel gene regulatory mechanisms associated with comorbidities in Down syndome/Multiple sclerosis patients. The student will implement higher order neural networks and dynamical causal networks.

Details and References from Pietro Lio', pl219@cam.ac.uk

Federated learning for multi-task classification in psychotic patients

Federated Learning (FL) is a privacy-focused machine learning paradigm that collaboratively trains models directly on edge devices [1-3]. Simulation plays an essential role in FL adoption, helping develop novel aggregation and client sampling strategies [4,5]. Federated learning operates by using the same prediction task across both client and server levels, ensuring privacy and security for the clients. An interesting question arises: *Can diverse classification tasks at the client level improve and generalize predictions at the server level?* In this project, we aim to explore this in the context of neuroscience, specifically focusing on psychotic patients. The goal is to investigate whether using clients to handle different neuroscience classification tasks (ranging from two to five) on the same dataset and modality can enhance the performance of a primary classification task at the server level. By integrating varied but related tasks, we aim to determine if the aggregated knowledge at the server level leads to better classification accuracy for the primary task: determining if a subject has suffered from auditory hallucinations within their lifetime. The student will learn the basics of federated learning, as well as some foundational concepts in neuroscience and psychotic disorders. The dataset is a pre-processed cohort of 600 patients (Top-OSLO [6]). Introductory material and code for the task will be provided. The student will need to familiarize themselves with the code structure and, as a first step, deliver a two-five tasks classification problem for two-five clients—each client focusing on one task (classifying psychotic vs. healthy patients, or identifying the presence or absence of a specific secondary sulcus region etc.).

References:

[1] McMahan B. H., et al., 2023. Communication-Efficient Learning of Deep Networks from Decentralized Data. https://doi.org/10.48550/arXiv.1602.05629

[2] Kairouz P., et al., 2019. Advances and Open Problems in Federated Learning. https://doi.org/10.48550/arXiv.1912.04977

[3] Wang J., et al., 2021. A Field Guide to Federated Optimization. https://doi.org/10.48550/arXiv.2107.06917

[4] Beutel D. J., et al. https://doi.org/10.48550/arXiv.2007.14390

[5] Sani L., et al., 2024. Pollen: High-throughput Federated Learning Simulation via Resource-Aware Client Placement. https://doi.org/10.48550/arXiv.2306.17453

[6] Mørch-Johnsen et al., https://doi.org/10.1093/schbul/sbw130

Contact: Michail Mamalakis, Lorenzo Sani, Pietro Lio {mm2703@cam.ac.uk, ls985@cam.ac.uk, pl219@cam.ac.uk }

Explainability of Federated framework for multi-task classification problem in psychotic patients to reveal new patterns.

In recent years, eXplainable Artificial Intelligence (XAI) has gained traction across a variety of AI tasks, driving the development of new methods and benchmarks aimed at enhancing fairness, complexity, and robustness in AI explanations. Previous studies have demonstrated XAI's potential to uncover novel patterns, as seen in our work [1,2], where we introduced a 3D XAI framework that improved the accuracy of explanations for deep learning models in neuroscience. This framework successfully revealed new biomarkers and sub-regions, illustrating how explanations can both expose novel insights and strengthen the reliability of AI models. Given this success, we now pose a new question: Can similar novel patterns be identified in complex and private settings like federated learning [3], particularly when dealing with multi-task classification problems? In this project, the goal is to use explanations to identify shared biomarkers across two neuroscience classification tasks: (1) distinguishing psychotic patients from healthy individuals and (2) detecting the presence or absence of a specific secondary sulcus region. This analysis will be performed both at the individual client level (for each task) and at the aggregated server level (classifying psychotic vs. healthy patients). By analyzing latent space representations and model explanations, we aim to uncover shared features across both tasks, potentially revealing common neural markers that are relevant to multiple psychiatric conditions. The student will gain foundational knowledge in federated learning, XAI, and core neuroscience concepts, particularly in relation to psychotic disorders. Working with a pre-processed dataset of 600 patients (Top-OSLO [4]), the student will be guided through the basics of the task. They will begin by implementing a two-task classification model across two clients, where each client focuses on one task: classifying psychotic vs. healthy patients or identifying a secondary sulcus region.

References:

[1] Mamalakis M., Macfarlane S. C., Notley S., et al., Deep multi-attention channels network to detect metastasizing cells in fluoresence microscopy images, https://doi.org/10.1016/j.compbiomed.2024.109052

[2] Mamalakis M., Vareilles H., Al-Manea A.,et al., An explainable three-dimension framework to uncover learning patterns: A unified look in variable sulci recognition., arxiv, arXiv:2309.00903, https://doi.org/10.48550/arXiv.2309.00903

[3] McMahan B. H., et al., 2023. Communication-Efficient Learning of Deep Networks from Decentralized Data. https://doi.org/10.48550/arXiv.1602.05629

[4] Mørch-Johnsen et al., https://doi.org/10.1093/schbul/sbw130

Contact: Michail Mamalakis, Lorenzo Sani, Pietro Lio {mm2703@cam.ac.uk, ls985@cam.ac.uk, pl219@cam.ac.uk }

Improving robustness of quantitative magnetic resonance imaging with XXX in subjects with Multiple Sclerosis

Prof Pietro Lio' (pl219@cam.ac.uk), Dr Fulvio Zaccagna (fz47@cam.ac.uk)

Multiple Sclerosis (MS) is an acquired autoimmune demyelinating disease affecting the brain and spinal cord, leading to cognitive decline and severe disability1. Subjects with MS are diagnosed and monitored using Magnetic Resonance Imaging, however, in daily clinical practice, images are only reviewed in a qualitative fashion.

Several quantitative MRI (qMRI) techniques, capable of in vivo quantification of imaging biomarkers, have been proposed to more accurately explore brain microstructure2 and metabolism3. In MS, these techniques have the potential to reveal pre-clinical inflammatory demyelination, affording a new therapeutic window.

However, to translate in clinical practice, a quantitative imaging biomarker should be reproducible, robust and accurate, and, at present, such a biomarker is still an unmet need.

This project will focus on Quantitative Susceptibility Mapping (QSM)4,5, a qMRI technique sensitive to differences between the magnetic responses of adjacent tissues that can estimate the concentration of several substances in vivo potentially revealing pre-clinical disease progression and enabling earlier treatment.
Preliminary results from a study lead by one of the co-supervisors of this project demonstrated higher variance along some clinically-relevant anatomical structures and in regions adjacent to areas prone to artifacts undermining the potential utility of the technique6. This project will use XXX to improve the robustness of qMRI data decreasing the variance due to noise and, in turn, allowing for a more reliable quantitative imaging biomarker.
The dataset is a pre-existing cohort of 100 subjects with MS and 50 healthy control subjects with already available imaging and clinical data6,7.

References

1. Reich, DS et al. Multiple Sclerosis. N. Engl. J. Med. 2018; 378(2), 169–180.
2. Sowa, P et al. Restriction spectrum imaging of white matter and its relation to neurological disability in multiple sclerosis. Mult. Scler. 2019; 25(5), 687–698.
3. Grist, J T et al. Imaging intralesional heterogeneity of sodium concentration in multiple sclerosis: Initial evidence from 23Na-MRI. J. Neurol. Sci. 2018; 387, 111–114.
4. Reichenbach, J. R. The future of susceptibility contrast for assessment of anatomy and function. NeuroImage 2012; 62(2), 1311–1315.
5. Deistung, A., Schweser, F. & Reichenbach, J. R. Overview of quantitative susceptibility mapping. NMR Biomed. 30: e3569. doi: 10.1002/nbm.3569.
6. Fiscone, C et al. Assessing robustness of quantitative susceptibility-based MRI radiomic features in patients with multiple sclerosis. Sci Rep 2023; 13: 1–16.
7. Fiscone, C et al. Multiparametric MRI dataset for susceptibility-based radiomic feature extraction and analysis. Sci data 2024; 11: 575.

Vision Language Model for Interpretable Whole Slide Image Description

Main supervisors: Professor Pietro Lio, Department of Computer Science
Professor Mireia Crispin, Department of Oncology

Co-supervisors: Dr. Zeyu Gao, Department of Oncology
Dr. Ines Machado, Department of Oncology
Dr. Shangqi Gao, Department of Oncology

Research Area: Multimodal Data Integration, Large Language Models, Computational Pathology

Project Proposal

State-of-the-art models in computational pathology have made strides in analysing whole-slide images (WSIs) using deep learning techniques. However, current models often lack interpretability, particularly in generating human-readable descriptions while finding the correspondence regions from the WSIs. There is a growing need for models that not only generate diagnosis reports but also provide the evidence regions within WSIs, linking the generated descriptions with detected regions to help clinicians better understand their diagnostic and prognostic significance.

Objectives

- To develop a vision-language model capable of generating descriptions of WSIs, leveraging large language models (LLMs) in combination with domain-specific vision encoders.
- To enhance model interpretability, providing explainable results (visualising the evidence regions) that clinicians can trust in diagnostic workflows.
- To improve upon state-of-the-art performance in both WSI analysis and interpretability using innovative LLM and fine-tuning methodologies.

Dataset

The student will be working with publicly available data from the following repositories:

- The TCGA dataset comprises 11.7k WSIs with clinical reports of 33 cancer types https://portal.gdc.cancer.gov/
- The GTEx dataset comprises 24.3k WSIs with descriptions, including 40 non-tumor tissue types https://gtexportal.org/home/histologyPage

References

1. Lu M Y, Chen B, Williamson D F K, et al. A visual-language foundation model for computational pathology[J]. Nature Medicine, 2024, 30(3): 863-874.
2. Lu M Y, Chen B, Williamson D F K, et al. A Multimodal Generative AI Copilot for Human Pathology[J]. Nature, 2024: 1-3.
3. Ahmed F, Sellergren A, Yang L, et al. PathAlign: A vision-language model for whole slide images in histopathology[J]. arXiv preprint arXiv:2406.19578, 2024.

Research environment : The student will engage in close collaboration with a diverse team comprising computer scientists, engineers and mathematicians drawn from various departments within the University. The student will work closely alongside fellow PhD students and postdoctoral researchers and collaborate with academic and industrial partners. The student will have access to office space, GPU hours for training the models (NVIDIA A100 GPUs) and will be expected to attend team meetings.

Graph representation learning for pathology-genomics integration

Main supervisors: Professor Pietro Lio, Department of Computer Science
Professor Mireia Crispin, Department of Oncology

Co-supervisors: Dr. Shangqi Gao, Department of Oncology
Dr. Ines Machado, Department of Oncology
Dr. Zeyu Gao, Department of Oncology

Research Area: Multimodal Data Integration, Graph Neural Networks, Computational Pathology, Genomics

Project Proposal:
Integrating pathology and genomics data enhances cancer diagnosis and treatment. However, most efforts have focused on the multimodal fusion of pathology and genomics in latent space, leaving the explainable interaction between these two domains largely unexplored. Graph neural networks have proven effective in data integration, particularly in connecting multi-scale data from the molecular to the physical level. Inspired by the success of graph-based methods in computational medicine, this project aims to:

- Construct genomic graphs to establish correlations between different genotypes.
- Explore pathogenetic graphs to integrate pathomics and genomics by linking histopathological phenotypes with molecular genotypes.
- Combine pathogenomics with graph neural networks for explainable survival analysis.

Dataset:
The student will be working with publicly available data from the following repositories:

- The TCGA dataset comprises 27 projects, 21 disease types, 131 primary diagnosis, 47 primary sites, and 133 tissues.
- Tissue slide from TCGA: 18,310 slide images
- Genomic profiles from TCGA: 97,610 RNA-seq
- Clinical data from TCGA: 10,966 cases with survival data

References:

1. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg J, Zou J. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng. 2020;4:827–34.
2. Cheerla A, Gevaert O. Deep learning with multimodal representation for pan-cancer prognosis prediction. Bioinformatics. 2019;35:i446–54.
3. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, Mahmood F. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans Med Imaging. 2022;41:757–70.

Research environment: The student will engage in close collaboration with a diverse team comprising computer scientists, engineers and mathematicians drawn from various departments within the University. The student will work closely alongside fellow PhD students and postdoctoral researchers and collaborate with academic and industrial partners. The student will have access to office space, GPU hours (NVIDIA A100 GPUs) for training the models and will be expected to attend weekly team meetings.

Titolo Using Graph Kernels in GNN and Knowledge graphs

Main supervisors: Professor Pietro Lio, Department of Computer Science (pl219@cam.ac.uk)
Professor Fabrizio Costa (f.costa@exeter.ac.uk)

Integrating data from several modalities is extremely relevant in computational biology, as it allows the emergence of system-level insights. Several approaches exist for integrating multi-modal data and learning from it. Graph representation methods stand out, as graphs are a natural way to represent biological systems, which are usually composed of complex networks of interacting elements. Graph kernels provide interesting ways of inputting domain knowledge into the predictions. However, graph neural networks are more flexible and can work with more complex inputs. A GNN based model able to learn to execute the high-level operations defined by a kernel, could therefore leverage domain knowledge to achieve even better performance in tasks such as recommendation networks, drug discovery and the detection of gene-disease correlations and operate on a more diverse set of data.

Tran, Van Dinh, Alessandro Sperduti, Rolf Backofen, and Fabrizio Costa. "Heterogeneous networks integration for disease–gene prioritization with node kernels." *Bioinformatics*36, no. 9 (2020): 2649-2656.

Originators: Pietro Lio (pl219@cam.ac.uk), Professor Fabrizio Costa (f.costa@exeter.ac.uk)