ARTIFICIAL INTELLIGENCE IN CANCER RESEARCH

Group Leader
Ke Yuan

Postdoctoral Scientists
Chris Walsh[1]
Fran Young[2]
Kieran Lamb[2]
Seyed Mousavi[3]
Yilong Yang[4]

PhD Students
Dan Liu[5]
David Meltzer[6]
Farzaneh Seyedshahi[7]
Khanh Nguyen[1]
Kai Rakovic[8]
Lucas Farndale[9]
Rozeena Arif[10]
Robert Strange[11]
Tommy Stevens[12]

[1]UoG, Joint with David Chang, EU Horizon 2020 PANCAIM
[2]UoG, Joint with David Robertson, BBSRC
[3]UoG, Boehringer Ingelheim
[4]UoG, Prostate Cancer UK
[5]UoG, Joint with David Robertson, EU ITN
[6]UoG, Joint with David Chang, EPSRC
[7]CRUK SI, Joint with John Le Quesne, CRUK Early Detection of Cancer, IAMMED-Meso
[8]CRUK SI, Joint with John Le Quesne, Pathological Society
[9]CRUK SI, Joint with Robert Insall, MRC DTP
[10]UoG, Joint with Alfredo Castello & David Robertson
[11]UoG, Joint with David Robertson and Joe Marsh, MRC DTP
[12]UoG, Joint with Campbell Roxburgh and Joanne Edwards, CRUK Multidisciplinary Project

University of Glasgow

Modern-day cancer research is generating an unprecedented amount of data, from high-resolution medical images to large-scale genomic and transcriptomic datasets. Harnessing this data through advanced AI, machine learning, and statistical models opens new avenues for discovery and translation into clinical practice.

Our work focuses on developing state-of-the-art methods tailored to cancer research challenges, particularly in the analysis of imaging and sequencing data. By addressing critical questions, such as identifying predictive biomarkers and uncovering novel therapeutic targets, we aim to drive progress in precision oncology and improve patient outcomes.

Mapping histomorphological phenotype across cancer types

Pathology slides are pivotal for cancer diagnosis, capturing vast information about cell shapes, interactions, and tissue structures. However, current diagnostic practices underutilize this data due to reliance on broad annotations from pathologists. Manual labelling of cells or tissue patterns across slides is infeasible.

To address this, we developed Histomorphological Phenotype Learning (HPL), an AI system trained on unannotated pathology slides using self-supervised learning (Claudio Quiros *et al.*, 2024, *Nat Commun*). HPL extracts generalisable features from millions of image tiles, clustering them into Histomorphological Phenotype Clusters (HPCs), which represent distinct tissue and cellular patterns. HPCs can predict cancer types, patient outcomes, and correlate strongly with molecular signatures.

HPL has proven effective in mapping cancer phenotypes. In collaboration with the Le Quesne lab, we demonstrated its utility in lung adenocarcinoma, capturing known growth patterns and uncovering subclasses, such as immune-activity-dependent solid growth patterns. Impressively, HPCs outperform IASLC (International Association for the Study of Lung Cancer) recommended grading when predicting recurrence-free survival on an external validation cohort.

Beyond lung cancer, in collaboration with Le Quesne and Tsirigos (NYU) labs, we have demonstrated HPL's utility in mesothelioma (Seyedshahi *et al.*, 2024, *bioRxiv*), colorectal cancer (B Liu *et al.*, 2024, *bioRxiv*) and osteosarcoma (Coudray *et al.*, 2024, *Clin Cancer Res*). We are also working on pancreatic cancer with David Chang (School of Cancer Sciences, UoG), prostate cancer lymph node metastasis with Hing Leung, and radiotherapy response in rectal cancer with Campbell Roxburgh and Joanne Edwards (Cancer Sciences, UoG).

Leverage multiplexed images to improve AI models for routine pathology slides

Multiplexed imaging techniques like CODEX provide detailed molecular insights but are cost-prohibitive for large-scale use. We developed TriDeNT, an AI model that transfers features learned from multiplexed images to H&E images, improving the performance of H&E-based AI models (Farndale *et al.*, 2023, *arXiv*). TriDeNT has shown up to a 101% improvement in downstream tasks.

To address the scarcity of matched multiplexed and H&E data, we used generative AI to digitally restrain H&E images as IHC images. This approach enabled training on synthetic matched datasets, achieving up to a 5.6x error reduction compared to real multi-modal data (Farndale *et al.*, 2024, *arXiv*).

Protein language model based in silico deep mutational scan

Protein and RNA sequences, like human language, can be represented as sequences of words—amino acids or nucleotides. Transformer-based protein language models (PLMs) have demonstrated remarkable potential, with public protein databases now comparable in size to datasets used for training language models like GPT-3. PLMs, such as ESMFold, have achieved breakthroughs in predicting protein folding structures.

We explored the pretrained ESM model's capabilities in predicting variant effects using ~20 million SARS-CoV-2 sequences curated during the pandemic (Lamb *et al.*, 2024, *PLoS Comput Biol*). Remarkably, ESM-2, a version that had not encountered SARS-CoV-2 during training, accurately predicted variant effects from a single sequence. Using the Wuhan-1 spike protein as a backbone, we conducted an *in silico* deep mutational scan, quantifying the impact of amino acid changes at every position. The predictions revealed conserved and highly mutable regions, aligning well with traditional statistical metrics like entropy, which rely on observing mutations across multiple sequences. ESM-2's ability to derive such insights without prior exposure to SARS-CoV-2 highlights its generalizability and utility for variant effect prediction.

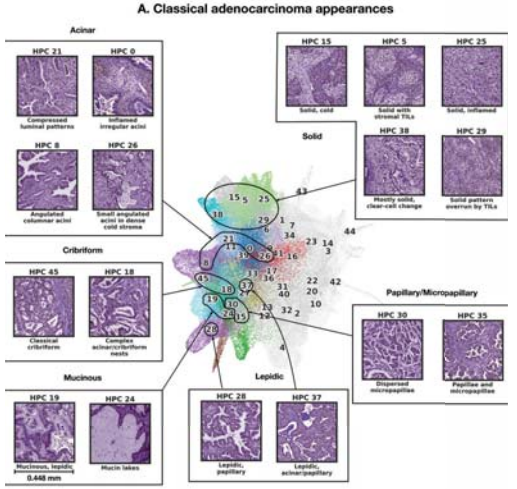

Figure 1. Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction of lung adenocarcinoma tile vector representations labelled by HPC membership). HPCs of interest are coloured, while other HPCs remain grey. The consensus was obtained after independent annotations of HPCs by 3 pathologists.

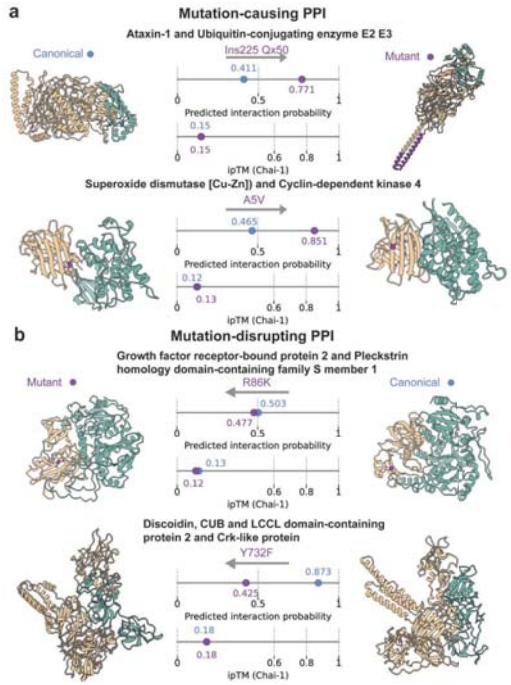

Figure 2. Demonstration of PLM-interact detecting changes in human PPIs associated withmutations. These PPI structures are predicted using Chai-1. Here, the mutated amino acids are highlighted in purple. Prediction interaction probabilities exceeding 0.5 indicate the proteins interact, while below 0.5 indicate non-interact. Chai-1's ipTM scores give the structure prediction confidence where <0.6 indicates failed predictions.

Predicting protein-protein interactions

Current approaches to PPI prediction use protein language models trained on single sequences, which lack the ability to model interactions between proteins. This limitation is akin to models capturing relationships among words but not between paragraphs. To address this, we developed PLM-Interact in collaboration with the Robertson lab (Centre for Virus Research, UoG) and Craig Macdonald (Computing Science, UoG) (D Liu *et al.*, 2024, *bioRxiv*). Trained on protein pairs with labelled interaction status, PLM-Interact achieved a 16%-28% improvement in accuracy (AUPRC) across multiple species, including humans, mice, and yeast. Importantly, it detected interaction-disrupting/causing mutations that eluded advanced tools like AlphaFold3, showcasing its potential for both prediction and mechanistic insight.

Future work

In the future, we will focus on the following directions

1. *AI for histology and spatial deep phenotyping*
   We aim to map all histologic patterns across all mouse models and cancer types generated locally and beyond. On the human side, we will work with NHS GGC to train pathology foundation models on their vast pathology slide archive. The mouse and human models will be used for disease positioning. We are also working on an HPL-like AI model trained with spatial proteomics data to combine molecular and morphological insights for a deeper understanding of tissue organization.

2. *Large language models for biological sequences and their interactions.* We aim to develop a suite of protein and RNA language models that could better predict the effect of mutations that were previously overlooked due to lack of recurrence across patients.

3. *Comprehensive in silico tumour model.*
   Building a digital twin of biological systems is an emerging paradigm that promises to transform experimental research. Our goal is to develop *in silico* tumour models that replicate the biological behaviour of mouse models. These digital twins would enable researchers to simulate experiments virtually, generating high-quality data that mimics real-world observations.

Publications listed on page 129