Exploring protein sequence/structure design by combining AI and synthetic biology techniques

The basic idea of the project is to explore the use of generative deep learning AI methods e.g. variational autoencoders for protein design applications in synthetic biology. To date, protein design has relied on the use of very time consuming stochastic simulation methods such as RosettaDesign. Although these methods have transformed the area of protein design in recent years, they are still very far from being reliable or automatic. Many hundreds of designs need to be generated and experimentally validated before even one viable sequence can be found. The exciting prospect we plan to explore in this project is to ask whether a generative AI method could be designed that is capable of greatly reducing the time taken for protein design along with increasing the success rate by intelligently sampling sequence space rather than relying on random search. Initial experiments will focus on variational autoencoders that have already been applied to de novo design of DNA sequences and planning of chemical syntheses. Using simplified fold descriptors and functional descriptors (e.g. ligand-binding) VAEs will be trained on known proteins, along with any available homologous sequences. Once trained, new sequences will be generated by sampling the latent space of the encoder model e.g. sampling variations of folds or variations in binding for existing folds. The collaboration with Dr Savva will allow experimental validation of a limited set of protein sequence designs, the success or failure of which can then be fed back to improve the generative model architecture by selection of alternative models which disfavour failed designs.

Supervisors:

Primary: David Jones (Computer Science, UCL & The Francis Crick Institute)

Secondary: Renos Savva (Biological Sciences, Birkbeck)

Mapping human brain development at new spatial resolutions using machine learning and 7T MRI

MRI is the most important imaging modality for studying human brain development owing to its non-invasive nature and its flexibility to quantify different aspects of brain structure and function. However, MRI has two significant and intertwined limitations for its use in monitoring development. It takes a significant time to acquire an image, and the higher the resolution of the scans, the longer the acquisition. This requires subjects to remain still for long periods - a significant practical and technical challenge when working in paediatric imaging.

To image cortical development, the cortex (1.5-4mm thick) needs sampling yet typical resolutions currently achieved (~1mm) are insufficient. One option currently being actively pursued at King's College London to achieve high resolution scans is to increase the power of the MRI scanners from 3 Tesla (3T) to 7 Tesla (7T), and use the increased sensitivity to increase resolution of brain imaging (Supervisor Carmichael). In parallel, work is being undertaken to characterise, using large population datasets, the trajectory of childhood brain development using MRI, as part of a Wellcome funded project (Supervisor O'Muircheartaigh).

Rapid developments in computer vision based on machine learning, can enable computers to learn the mapping between high and low quality (e.g. resolution) images (Supervisor Alexander).

Our proposed PhD project will aim to develop a high resolution normative model and atlas of the developing brain cortex and white matter utilising existing low resolution

developmental MRI databases and machine learning to enhance its resolution. This will be based on limited cutting edge high resolution data obtained at 7T, while dramatically reducing the requirement for high numbers of long duration costly scans.
Supervisors:
Primary: David Carmichael (Biomedical Engineering, KCL)
Secondary: Daniel Alexander (Computer science, UCL),
Tertiary: Johnathan O'Muircheartaigh (Forensic & Developmental Neurosciences, KCL)

Synthesis of dynamic locomotor behaviours using deep learning applied to realistic musculoskeletal models of bipedal and quadrupedal animals

Legged locomotion is complex and dynamic, involving abrupt contact transitions and uncertainty due to variable terrain, neural delays and sensorimotor noise. To achieve agile locomotion, animals must effectively integrate neural control with the physical properties of the musculoskeletal system [1-2]. To what extent is locomotor behaviour and learning shaped by the physics of the musculoskeletal anatomy? How does morphology influence economy, stability or maximum speed of movement?

The answer to these questions remain elusive, because, historically, the computational tools available to model locomotion were limited to simplified bodies [1-2], highly constrained movement patterns, such as steady walking [3], or both. Recent AI breakthroughs in deep reinforcement learning by DeepMind can now support the synthesis of full-body dynamic locomotor behaviours from scratch [4-5].

The aim of this project is to investigate the relationship between locomotor behaviour and morphology using anatomically realistic musculoskeletal models of exemplar bipedal and quadrupedal species (ostrich, human, dog, horse). We will develop 3D musculoskeletal models using 3D CT and MRI imaging, multi-body physics modelling and deep-reinforcement learning AI tools in collaboration with DeepMind. The simulated behaviours will be compared to measured dynamics of complex locomotor tasks, making use of wearable IMU/GPS sensors [6].

This work will result in a next-generation toolset for continuous simulation and monitoring of animal locomotor behaviour, which will be made freely available for research and innovation in biomedical sciences, robotics, clinical care and health monitoring. Such tools have enormous potential to enhance animal biomedical research, accelerating drug development and reducing numbers of animal experiments. The work is directly relevant to the BBSRC themes Bioscience for Health and Exploiting New Ways of Working.

1. Birn-Jeffery, Hubicki, Blum, Renjewski, Hurst and Daley 2014. Don't break a leg: running birds from quail to ostrich prioritise leg safety and economy on uneven terrain. J Exp Bio, 217:3786.
2. Hubicki , Jones , Daley , and Hurst 2015. Do limit cycles matter in the long run? Stable orbits and sliding-mass dynamics emerge in task-optimal locomotion. IEEE ICRA, 5113.
3. Hutchinson, Rankin, Rubenson, Rosenbluth, Siston and Delp 2015. Musculoskeletal modelling of an ostrich (Struthio camelus) pelvic limb: influence of limb orientation on muscular capacity during locomotion. PeerJ, 3:1001.
4. Heess , Sriram , Lemmon , Merel , Wayne , Tassa , Erez , Wang , Eslami, Riedmiller and Silver 2017. Emergence of locomotion behaviours in rich environments. arXiv:1707.02286.

5. Merel, Tassa, Srinivasan, Lemmon, Wang, Wayne and Heess 2017. Learning human behaviors from motion capture by adversarial imitation. arXiv:1707.02201.
6. Daley, Channon, Nolan and Hall 2016. Preferred gait and walk–run transition speeds in ostriches measured using GPS-IMU sensors. J Exp Bio, 219:3301.

Supervisors:
Primary: Dr Monica A. Daley, Royal Veterinary College
Secondary: Prof. John Hutchinson, Royal Veterinary College
Collaborator: Dr Yuval Tassa, Google DeepMind

Application and validation of machine-learning frameworks on big functional datasets to identify proteins important for cellular ageing.

Integrative genomic research on ageing in model organisms is vital to uncover proteins and processes affecting lifespan and promoting lifelong health and wellbeing. Modern machine learning has tremendous potential by exploiting big datasets to find elaborate patterns, discover high-level features, and provide accurate predictions for complex biological processes such as ageing. In a collaboration between the Bähler and Orengo laboratories, we use fission yeast as a model organism, together with multi-step machine learning, to comprehensively identify cellular processes with fundamental importance for ageing. This project builds on our recent wet and dry work on ageing-associated proteins and gene predictors, and on the complementary expertise of the two groups. We combine our functional-profiling experiments of yeast mutant libraries (large-scale phenotyping and genetic-interaction assays) together with known ageing-associated proteins and multiple heterogeneous network datasets (e.g., protein-interaction or gene-regulatory networks). We are developing kernel-based predictors which integrate these heterogeneous datasets to identify and rank novel ageing proteins in networks seeded by known ageing proteins ('guilt-by-association').

The student will build on these rich existing machine-learning methods for protein-function prediction from the Orengo lab, by leveraging large-scale functional profiling data from the Bähler lab with diverse published genomic and network data to validate predictors for ageing-related proteins. By applying multiple iterations of machine learning and experimental validation of predicted ageing proteins, we will obtain an optimized predictor that will use all available evidence to identify new ageing-associated proteins. The student will also apply and test our deep-learning predictors that can associate genes with particular biological processes relevant to ageing. The experimental part of the project will generate functional genomic and chronological lifespan data of yeast cells to empower and validate the predictions of the machine-learning methods. The student will then benchmark and experimentally test selected predictions of new ageing proteins.

In conclusion, the student will receive broad, inter-disciplinary training in both the acquisition of large-scale genomic data and various yeast assays as well as in the application and validation of machine-learning based predictors. This exciting project will thus uncover ageing-related proteins with cohesive wet and dry approaches to put new players affecting longevity on the map for focused follow-on studies by the student and others.

Supervisors:
Primary: Jürg Bähler, Department of Genetics, Evolution & Environment, UCL
Secondary: Christine Orengo, ISMB, Birkbeck/UCL

Reverse engineering cell competition using automated microscopy and deep learning

The aim of this project is to use state-of-the-art computer vision, machine learning (ML) and automated time-lapse microscopy to determine the underlying rules governing cell competition.

Cell competition is a phenomenon that results in the elimination of less fit cells from a tissue – a critical process in development, homeostasis and disease. The viability of loser cells depends strongly on context: when they are cultured alone, they thrive, but when in a mixed population, they are eliminated by cells with greater fitness.

Despite its physiological relevance, cell competition remains poorly understood -- we do not know the "rules" of how cells interact, or how their biochemical and mechanical environment affects fate. To address this challenge, we recently built the first deep learning and automated single-cell microscopy system to analyse cell competition (Bove et al. Mol. Biol. Cell 2017). We used deep convolutional neural networks to analyse the state and fate of millions of single cells in mechanical competition, including cell division and death (Video: https://youtu.be/EjqluvrJGCg). Remarkably, this revealed that tissue-scale population shifts are strongly affected by cellular-scale tissue organisation.

In this project, we will develop new deep neural network architectures to identify the features of a single-cell's microenvironment over time which predict its eventual fate. We will use this information to gain insight into and model the outcome of competition between two cell types in a co-culture. We will use the full scope of time-series information to determine a basic set of 'rules' of cell competition with relevance to development, disease and stem cell biology.

Supervisors:

Primary: Dr. Alan R. Lowe (ISMB/LCN, University College London)

Secondary: Prof. Guillaume Charras (LCN/CDB, University College London)

Content-Aware AI Driven Driven Super-Resolution Microscopy

This project focuses on the combination of bioimage informatics (bioinformatics + image processing) and Deep Learning AI methods (mathematics + computational sciences), to develop an unprecedented data-driven content aware microscopy modality that directly studies cell behaviour while imaging (biophysics + engineering + microscopy). We will further use this framework to predict cellular behaviour, using deep learning as a way to identify cellular dynamics that can be statistically described using biophysical parameters.

Super-Resolution Microscopy relies on a finely balanced optimization of the optical configuration and analytical data treatment. These 2 parameters need to be correctly adjusted and matched to the properties of the biological sample to be imaged. Deviations from their optimal combination limits the quality of the data via the introduction of artefacts. Up to now, these optimisations have been based on educated guesses by researchers with some help from empirical criteria.

Recently we have established a new Super-Resolution approach named SRRF, capable of achieving unprecedented live-cell imaging capacity at the nanoscale [1]. In tandem, we have developed SQUIRREL, an analytical approach capable of calculating the quality and resolution of images generated in Super-Resolution Microscopy [2]. We further demonstrated for the first time the capacity of Deep Learning to massively improve microscopy data [3], by restoring corrupted or under-sampled cell imaging data – an approach named CARE.

We propose to combine these approaches in a real-time analysis framework, integrated into a microscope itself designed to study long term cell cycle. This ability immediately opens the door to establish an AI approach (using Deep Reinforcement Learning) that adapts the microscope acquisition and analysis settings to maximise resolution and image quality, while extracting a quantitative biological characterisation. This will enable microscopy systems for the first time to: i) learn from the sample about how to improve imaging in real time, and ii) in a dynamic manner adapt to changes in the imaging properties of a living sample.

1. Gustafsson, N., Culley, S., Ashdown, G., Owen, D. M., Pereira, P. M. & Henriques, R. Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations. Nat. Commun. 7, 12471 (2016).
2. Culley, S., Albrecht, D., Jacobs, C., Pereira, P. M., Leterrier, C., Mercer, J. & Henriques, R. Quantitative mapping and minimization of super-resolution optical imaging artifacts. Nat. Methods (2018). doi:10.1038/nmeth.4605
3. Weigert, M. et al. Content-Aware Image Restoration: Pushing the Limits of Fluorescence Microscopy. bioRxiv 236463 (2017). doi:10.1101/236463

Statistical classification of healthy and diseased tissues using Raman images: simultaneously exploiting morphological and biochemical features

Motivation:

Illness has major impacts on the UK economy (lost productivity, opportunities and healthcare costs that could be reduced by earlier, accurate diagnosis and personalised treatment. Many key diagnoses are made through visual inspection of ex vivo tissue by a pathologist using morphological features revealed with two or three coloured stains, supplemented with limited genetic and biochemical information: a highly subjective approach often with limited consensus even between experts. The ambition of digital pathology is to apply statistics, machine learning and artificial intelligence to render faster, less subjective diagnosis but its impact will depend on the fundamentals of the imaging technology and the power of machines to learn the useful image features. Raman microscopy (RM, direct, stain-free chemical imaging) can produce hyperspectral images as 3D arrays, collecting both morphological information on tissue structure and vast amounts of biochemical data from the Raman spectra of chemical bonds (equivalent to hundreds of different colours in each pixel). This data can be used to probe for healthy and diseased tissue and to capture contrasting biological features crucial for early diagnosis.

Standard Raman analytical approaches statistically reduce the spectral dimension and then apply statistical techniques to separate healthy from diseased tissue. Conversely, a pathologist mainly uses morphological spatial information and experience-dependent pattern recognition. Consequently, there is an unmet need for novel statistical methods to combine analysis of both types simultaneously and to report subtle coincident biochemical and morphological divergence from the healthy condition. This type of problem is eminently suitable for machine learning (ML) approaches but there remain computational challenges: 1) Raman hyperspectral images are of a particularly high spectral dimension for which usual statistical methods are ill-suited; 2) images are complicated by complex spatial and spectral correlations; 3) the number of samples labelled by pathologists is small for machine learning and limits training data for models.

Methodology:

Using healthy and diseased RM image datasets (cancers and others) we will investigate statistical ML methods for RM data to improve accuracy for classification (diagnosis). We

will: 1) exploit the characteristics of hyperspectral images to develop a dimension reduction approach which preserves morphological information; 2) develop classification methods which simultaneously exploit morphological and spectral information to improve the discrimination between healthy and unhealthy tissues; 3) investigate semi-supervised classification methods to exploit rich, unlabelled data and ameliorate the issue of limited training labels, while improving classification accuracy.
Supervisors:
Primary: Dr Jinghao Xue (UCL, Statistical Science)
Secondary: Prof Geraint Thomas (UCL, Cell & Developmental Biology)
Collaborator: Dr Ian Bell (Renishaw plc, Spectroscopy Products Division)

Automated High-throughput Analysis of Gene Expression Codes in a Food Sensing Neuroendocrine Network

This project investigates how food affects ageing via conserved neuroendocrine factors. In C. elegans and humans, neuroendocrine factors such as insulin-like peptides, growth factors, and biogenic amines regulate each other in complex networks to modulate ageing, metabolism, and other physiological outputs. However, the information processing mechanisms in these networks for discriminating food inputs are unclear. We address this question by enhancing high-throughput C. elegans experiments with artificial intelligence and machine learning.
Food-responsive changes in the expression of neuroendocrine factors including TGF□, insulin-like peptides, and serotonin-synthesis enzyme can be quantified in specific neurons in C. elegans, where its stereotyped anatomy allows identification of every cell. We previously performed high-throughput imaging of fluorescent transcriptional reporters for these genes across many food levels, temperature, and genotypes. We also measured lifespans under matching conditions and genotypes. These multifaceted datasets reveal how food shapes gene activity at single-cell resolution and map out connections in the gene network that link food to ageing.
These studies have two key limitations: first, the neurons are identified by hand, which is very labour intensive; and second, the need for highly standardised images means many images are discarded. The first issue offers an opportunity to automate cell identification, either by applying supervised machine learning (e.g. random forest classifier) to tens of thousands of annotated image stacks; or by using deep learning so that the algorithm can select the parameters to optimise. The second issue means that there are many unannotated image stacks that can be analysed by unsupervised deep learning and cross-validated against the annotated subset to check performance.
We will then apply decoding analysis and machine learning to relate food inputs, combinatorial patterns of gene expression, and ageing phenotypes. Using these results, we will predict the lifespans of novel mutants based on their effects on gene expression and verify these hypotheses experimentally by exploiting high-throughput imaging and the automated image analysis developed above. This work thus decodes food-sensing gene networks in the nervous system that impact ageing and other health-related outputs.
Supervisors:
Primary: QueeLim Ch'ng (Developmental Neurobiology, King's College London)
Secondary: Susan Cox (Cell and Molecular Biophysics, King's College London)

Inferring gene function for emerging model organisms

The first generation of molecular-genetic research focused on traditional model organisms including mouse, Drosophila, yeast, C. elegans and zebrafish. Today, much research within the BBSRC remit now focuses on a diversity of organisms that are much more relevant models for specific questions across the BBSRC remit, including for priorities in “Lifelong Health”, “Sustainable agricultural systems”, “Food safety and nutrition”. For example, such emerging organisms exhibit attractive phenotypes including 100-fold intra-specific variation in lifespan, resistance to harsh environmental conditions, represent novel animal models for disease, provide crucial ecosystem services, or are key to food security because they are crops or may pollinate them.

A major challenge when working with such “emerging” genetic model organisms is making sense of the “gene lists” resulting from experiments resulting from genome-wide analyses (e.g., of gene expression or genome-wide associations).

Here, we will develop a bioinformatics tool that takes a list of genes or genomic locations from a new species as input, and transparently produces the most relevant functional information describing this list of loci. When presented with data for which no direct functional information exists, the tool will in a first instance identify relationships of orthology to regions of other species. This will create a trail of links to databases in which functional information to orthologous regions does exist. These databases will be interrogated following a hierarchical set of rules, and the results will be ranked using cutting-edge machine learning techniques (commonly referred to as “learning to rank”) and improve over time by tracking user behaviour. The tool hereby makes it possible to extract significant value from largescale datasets that would otherwise be disconnected. Summary data will be returned to the user using visualisations, statistics and tables in a manner that facilitates interpretation. These results will appear asynchronously, i.e., partial results will be displayed as they are being calculated.

We will package our work in a manner that makes it accessible to biologists working with new or existing genomes. Overall, our approach will substantially improve the ability of genome biologists to generate meaningful biological insight when working with new organisms.

Supervisors:

Primary: Yannick Wurm (Department of Organismal Biology & Centre for Computational Biology, Queen Mary University of London)

Secondary: Christophe Dessimoz (Department of Computer Science, University College London)

Collaborators: Dr Romain Studer, Data Scientist, Benevolent.AI

Dr Timothy Hospedales, Reader in Machine Intelligence, University of Edinburgh