Project A : Machine-Learning-Guided Directed Evolution for Multi Objective Protein Engineering

Academic convenor: Prof. Juergen Branke, Warwick Business School (Juergen.Branke@wbs.ac.uk)

External partner: Michael Pearce, ZenithAI

Traditional directed evolution is the process of designing protein sequences of amino acids and or DNA sequences of nucleotide base pairs by successive rounds of mutation and screening in costly wet lab experiments. However, the rise of machine learning and large datasets has led to replacing expensive wet lab experiments with surrogate wet lab prediction models enabling much faster and cheaper sequence screening and optimization [1, 2], see the left figure below. Proteins are frequently designed to maximise a score of interest, eg enzymes are proteins that act as catalysts and must be designed to maximise their catalytic activity [2], or maximising folded structure validity [4], or maximise protein fluorescence [5]. There are multiple challenges with protein optimization, the search space is strings of amino acids, given 20 amino acids, for sequences of length 100, there are 20100 possible protein sequences, further the relationship between sequence and score may be complex and highly non linear. Despite these challenges, past works have shown success in speeding up wetlab experiments with machine learning. When a protein must be engineered for more than one criterion (or score), such as catalytic activity as well as folding structure validity, a directed evolution algorithm must find a range of proteins that satisfy both objectives but with different compromises, that is, find a set of non-dominated proteins, proteins for which no other protein is better in both scores as shown in the right figure below, the Pareto set of proteins. Particularly in biological applications, providing a range of alternative sequences to the wet lab is vitally important as many experiments may fail, “biology is always messy”. In numerical optimization, there exist many machine learning model based and model free multi objective algorithms, in directed evolution there exist many (single objective) sequence optimization algorithms, the focus of this project is to apply ideas from both fields to multi objective protein optimization.

Project B : Fraud, Waste and Abuse in Health Insurance

Academic convenor: Dr. Marya Bazzi, Warwick Mathematics Institute (Marya.Bazzi@warwick.ac.uk)

External partner: Tom Nygren, Kirontech

When you are seen by a medical professional, each procedure, test, diagnosis, and medical item have an associated cost. In the private healthcare sector the cost of these can be covered by a medical insurance company who pay the hospital or doctor for the services they performed. Because of this there is a financial as well as a medical incentive to treatment which leaves the system open to misconduct.

Fraud, Waste and Abuse (FWA) are services and charges that shouldn't really be billed in a medical insurance claim. They are not completely separate concepts but can be described as: Fraud being deliberate deception for monetary gain, Waste being inefficient or unnecessary use of resources, and Abuse being waste with intent to benefit financially rather than providing better medical service. Once a method of abuse has been identified, individuals (e.g. practitioners or hospitals) will repeatedly use it for monetary gain. Detecting and monitoring FWA behaviour as early as possible helps reduce losses and ensure resources are managed in the best way to benefit patients.

At Kirontech we build algorithms and analyse data to identify unwanted behaviour in medical billing. The tools we create help insurers to identify misconduct in their medical claims and we have

sucessfully found individuals who conduct FWA. The aim of this project is to answer the question: given the results we already have, at what point in time has an individual's behaviour changed to being fraudulent? How early could we have detected FWA?

Project C : Analysing heart rhythm irregularity in sports data

Academic convenor: Prof. Colm Connaughton, Warwick Mathematics Institute
(C.P.Connaughton@warwick.ac.uk)

External partner: Ian Green, crickles.casa

Background

Many people participate in sporting activity such as cycling and running while wearing sports watches and chest straps that capture their heart rate as they exercise. Also, it is observed that engaging in regular endurance sports over many years, while conferring very substantial health benefits to participants, appears to be correlated with a measurable rise in the incidence of heart rhythm issues. It would therefore be potentially very valuable if irregularities in heart rate patterns captured from sports watches and straps could be detected that could sometimes provide useful and apt warning cues for participants in endurance sports to prompt them to consult their GP or cardiologist in advance of the development of a potentially serious heart condition.

Some specialised sports watches and other devices already enable their wearers to run ECG's "in the field". These can produce medical-grade reports when or immediately after a sports participant feels that they may be having a heart problem. However, these require:

(a) symptoms to be at the stage where they are already apparent to the athlete;

(b) the athlete to be in possession of such a watch or device, and they are expensive and therefore relatively rarely owned.

It would thus be especially useful if an algorithm could be developed that provided a useful, if less powerful, signal based on data from widely-owned watches and strap and/or that may flag cues that precede the onset of symptoms evident to the wearer.

Work to date

Research that has not been published appears to be able to detect a signal in heart rate data obtained from regular cyclists that is meaningfully associated (p-value of the order 0.0001 to 0.001) with a diagnosis of a heart rhythm issue. This was based on exercise data collected using an app whose users can authorise access to their Strava account. Strava is a popular sports social network to which users upload their sports activity data. Additional information on heart health was collected through a user survey. The analysis was based on exercise and survey data from around 200 participants.

In this research the signal was calculated in two steps:

1. A regular/irregular indicator that is essentially boolean was calculated from the detailed heart rate data recorded for each activity.

2. The frequency with which irregular activities occurred amongst all the recorded activities for each athlete was calculated.

It is the frequency of irregular activities that was meaningfully correlated with a diagnosis of arrhythmia.

The purpose of this project

Irregularity seen in the heart rate data from a sports activity may be “extreme” - for example, the reported heart rate may jump straight from 160 beats per minute (bpm) to 250 bpm - or conversely it may be virtually undetectable to the human eye. Also, it may be due to genuine volatility in the athlete's heart rate or to a recording glitch in the heart rate strap (or watch). The latter question may be hard or impossible to resolve definitively; however, it may, perhaps, be related to the former question: a sudden jump to, say, 250 bpm may indicate an erroneous recording of the heart rate.

The model used in the work to date does not differentiate between extreme and subtle irregularities. It may be fruitful to consider an alternative measure of heart rate irregularity that produces a numeric rather than a boolean value for each activity. This could, perhaps, be arrived at using a stochastic volatility jump diffusion model.

The aim of the project is to see whether such a model would improve on the predictive performance and/or the explanatory power of the current model. There is also scope to look for other improvements to the research to date, possibly before it is published.

Project D : Modelling the spread of SARS-CoV-2 in educational settings

Academic convenor: Prof. Mike Tildesley, Warwick Mathematics Institute and School of Life Sciences (M.J.Tildesley@warwick.ac.uk)

External partner: Stephen Meredith, Head of Higher Education Analysis and Modelling, Department for Education, UK Government

Background

Since late 2019, SARS-CoV-2, the virus that causes COVID-19 infection, has spread worldwide, following its initial emergence in Hubei province, China, to cause a pandemic resulting in over 320 million confirmed cases and over 5.5 million confirmed deaths to date. To reduce the spread of the virus, mitigations have been introduced in many countries across the world, such as restrictions on social mixing, closure of non-essential businesses and closure of educational establishments.

As part of responsive measures introduced across the UK to counter SARS-CoV-2 transmission, schools were closed in March 2020, before some children returned to the classroom in June 2020. In universities, there was a switch to online learning, whilst many exams were cancelled in the summer of 2020. As students returned to university campuses in September 2020, there was concern that significant outbreaks could occur, which might result in spillover to local communities. Mitigations were introduced on university campuses, including restrictions on lecture sizes, the wearing of face coverings and regular testing and isolation of infected students and their contacts.

Prior work

During the 2020/2021 academic year, the Isaac Newton Higher Education working group carried out multiple analyses to investigate the risks associated with students returning to campuses, the potential for onward community transmission and the impact of mass testing and staggering of

student return. This work concluded that there was no clear signal of community spillover, whilst staggering of student return to campuses had a minimal effect upon campus outbreaks (https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.210310).

Project purpose

This project will involve working closely with the Department for Education to analyse the risks associated with the current Omicron wave of SARS-CoV-2 on university campuses and the long-term impact of intervention measures at reducing spread, whilst considering the potential for disruption to educational services. We will also investigate the spatiotemporal variation in risk on university campuses and the potential for onward transmission into the community. The project will be developed in consultation with the participating MathSys students and may involve updating and extending the existing Warwick model of university spread, developing new models and/or analysing the extensive data collected during the 2020/21 and current academic years. Our project partner is Stephen Meredith, Head of Higher Education Analysis and Modelling, who has taken a central role in advising the UK government regarding the impact of COVID-19 on university campuses throughout the pandemic.