Speaker: Arron Gosnell (University of Bath)

Date: 04/05/2021 on Microsoft Teams

Title: Incorporating the Tanimoto coefficient in a Gaussian process model to predict herbicide performance

Abstract:
With the proliferation of screening tools for chemical testing, it is now possible to create vast databases of chemicals easily. On the other hand, the development of rigorous statistical methodology that can be used to analyse these large databases is in its infancy, and further development to facilitate chemical discovery is imperative. At Syngenta, thousands of potential herbicides will undergo a sequence of screening tests and each time ineffective compounds will be discarded, and the remaining are assessed against a more complex set of criteria. Evidently, the data from the early trials will exhibit high uncertainty and subjectivity. Current methods employed to analyse these data fail to incorporate the chemical structure of the herbicide, and as a result, this feature is unaccounted for in the model.
As each herbicide is defined by a unique fingerprint, a binary vector indicating the presence of atomical features, it is possible to assess the "closeness" of herbicides through the Tanimoto coefficient, a metric on the chemical space. Our hypothesis is that two compounds similar in the chemical space will exhibit similar herbicidal properties. By replacing the Euclidean distance in a Gaussian process (GP) with the Tanimoto coefficient, we may account for correlation in the model and predict the performance of any unseen compound.
In this talk, I will discuss how we incorporate the Tanimoto coefficient in a GP model, both in a regression and classification setting, and show that accounting for correlation results in improved model performance over the uncorrelated model. I will discuss the tools used to overcome certain hurdles in developing the GP model and discuss the use of proper scoring rules to evaluate model performance. I will also present the AIC ensemble model, whose predictions may outperform any individual model, as well as the method of stacking to improve predictive performance.