**Machine Learning and Conformal Prediction applied to Drug Discovery and Cancer Diagnostics**

**Ola Spjuth** (ola.spjuth@farmbio.uu.se), Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, SE  
**Lars Carlsson** (Lars.A.Carlsson@astrazeneca.com), AstraZeneca, Innovative Medicines & Early Development, Quantitative Biology, Mölndal, SE  
**Martin Eklund** (martin.eklund@ki.se), Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SE  
**Andreas Hellander** (andreas.hellander@it.uu.se), Department of Information Technology, Uppsala University, SE  
**Alexander Gammerman** (A.Gammerman@cs.rhul.ac.uk), Department of Computer Science, Royal Holloway University of London, UK

**Project Aims**

The overarching aim of this project is to contribute to theory and applications of machine learning with valid confidence estimates or probabilities. Building on established mathematical theories and the availability of big data frameworks, you will contribute towards improving large-scale predictive modeling in life science with applications including drug discovery and cancer diagnostics.

Machine learning concerns insights and decisions made from data, and is today pervasive with examples including recommendation systems, anomaly and fraud detection, online marketing, speech recognition, image analysis, and artificial intelligence. The large variations and complexity of these problems prohibit the possibility of deriving analytic solutions, requiring methods for “learning from examples”. Growing volumes and varieties of available data, together with cheaper computational processing and affordable data storage allows for the possibility to analyze bigger, more complex data and deliver faster, more accurate results – even on very large scale. However, most contemporary machine learning approaches do not report valid confidence measures or class probabilities. Confidence in predictions is crucially important: we want to know if the machine is confident in its predictions or if it is essentially guessing because it has no knowledge of the object you are trying to derive predictions for. Also, large data sets are in many cases unbalanced, which traditional methods struggle with.

We will develop new theory and methods in machine learning extending the work on confidence-based prediction intervals or probabilities rather than point predictions [Vovk, Gammerman et al., Springer 2005], and that can cope with very large and unbalanced datasets. Within this learning framework predictions are hedged and incorporate a valid indication of their own accuracy and reliability.
Applications in Drug Discovery and Cancer Diagnostics

Your developed methodology will be assessed on applications from computational pharmacology (main advisor Spjuth), drug discovery at AstraZeneca R&D (co-supervisor Carlsson), and cancer diagnostics at Karolinska Institutet (co-supervisor Eklund). In common to the projects is the use of conformal prediction methodology to address challenges to make efficient predictions with valid estimates of confidence or probability, and to feed back from applications to new theory.

**Drug discovery**

In drug discovery, various types of data are used in the decision making process; e.g. what drug candidates should be advanced, what experiments should be performed, how should candidates be optimized etc. The data is in many cases heterogeneous, unstructured, and incomplete and comes from different sources, including drug screening assays, toxicological studies, image-based analysis of cells, and is often fused with data from previous studies as well as public and private databases. The proposed project will advance the use of machine learning in drug discovery projects by developing and applying confidence machines, and also research how to implement this effectively at AstraZeneca R&D.

**Cancer diagnostics**

The overarching aim of co-supervisor Eklund’s research to develop a personalized prostate cancer diagnostic pipeline using a combination of genomics, proteomics and imaging. Personalization requires predictions, and clinical decision making requires known confidence in the predictions. The proposed project will bring confidence machines into clinical decision making and aid with prostate cancer risk assessment based on biomarkers, clinical-, molecular-, and imaging data.

**The research environment**

You will be part of the Spjuth (main advisor) Lab ([https://pharmb.io/](https://pharmb.io/)) at the Department of Pharmaceutical Biosciences and Science for Life Laboratory at Uppsala University. We focus on developing new methods and applications to meet the demands of high-throughput biology and drug discovery, using predictive modeling, high-performance and cloud e-infrastructures as well as big-data analytics frameworks.

Co-supervisor Carlsson has a PhD in Scientific Computing and is leading machine learning initiatives at AstraZeneca R&D to improve decision making in the drug discovery process. Carlsson will support the project with expertise in machine learning, scientific computing, and with applications and data from drug discovery problems at AstraZeneca.

Co-supervisor Eklund ([http://ki.se/en/people/maeklu](http://ki.se/en/people/maeklu)) is Associate Professor in Biostatistics at Karolinska Institutet, focusing on improving diagnostics and treatments of breast and prostate
cancer. Eklund will support the project with expertise in statistics, machine learning, and with applications and data from breast- and prostate cancer diagnostics.

Co-supervisor Hellander ([www.andreashellander.se](http://www.andreashellander.se)) is Associate Professor in Scientific Computing at the Department of Information Technology. His focus is on new computational methods and e-Science, with an emphasis on applications in Life Science. Central themes in the group are methods for stochastic modeling and simulation of reaction-diffusion processes, mathematical modeling in Systems Biology, Applied Cloud Computing and CSE software development.

Professor Gammerman ([http://www.clrc.rhul.ac.uk/people/alex/](http://www.clrc.rhul.ac.uk/people/alex/)) is Founding Director of the Computer Learning Research Centre at Royal Holloway, University of London. His current research interest lies in field of Algorithmic Randomness Theory and its applications to machine learning, the development of inductive/transductive confidence machines and intelligent data analysis, and will strengthen the theoretical part of the project.

It is expected that the successful candidate will do research visits at AstraZeneca R&D, Karolinska Institutet, and Royal Holloway University of London.

**The Successful Candidate**

The applicant should have a solid background in applied mathematics, machine learning, bioinformatics or a related field and ideally experience with techniques for large-scale distributed machine learning such as Apache Spark. The applicant should have excellent analytical skills, possess curiosity and creativity. Excellent oral and written communication skills in English are required.

Further, the applicant should be highly collegial, and be able to work well in a team as well as having the ability to work independently. Further, the applicant should possess exceptional organizational skills, know how to successfully multitask, and should be able to solve unanticipated and complex problems.