DPhil Computational Discovery

Project Booklet 2020/21

P1: Mapping the protein "diffusome" of bacteria by high-throughput single-molecule tracking and advanced data analysis

Academic supervisor: Prof Achillefs Kapanidis (DPhil in Condensed Matter Physics); IBM co-supervisor: Dr Simon Colgate

Protein mobility and spatial distribution inside single living bacteria can be explored by single-molecule tracking, uncovering novel information about modes of interactions of proteins in the crowded bacterial cytoplasm. We propose to massively expand our analysis to >1000 bacterial proteins with diverse functions. IBM will support the project via their know-how in algorithms, data analytics, and machine learning.

P2: New paradigms for AI-based multidimensional biomedical big-data exploration

Academic supervisor: Prof Ilan Davis (DPhil in Biochemistry); IBM co-supervisor: Dr Flaviu Cipcigan

This project aims to build a new standard for holding multi-dimensional and multimodality informatics and bio-imaging data on GPUs, as a model for clinical genetic and diagnostic data for near instant access and exploration. We aim to develop new paradigms for user interface and conduct exploration of these integrated data sets using the latest cloud computing, gaming technologies and machine learning methods.

P3: Combining MD with machine learning to explore cyclic peptides

Academic supervisor: Prof Philip Biggin (DPhil in Biochemistry); IBM co-supervisor: Dr Flaviu Cipcigan

Many diseases cannot be drugged using current small molecule therapeutics and therefore new approaches are urgently needed. One such approach that shows promise is to use cyclic

peptides. The aim of this project is to understand the conformational landscape of short cyclic peptides and devise design principles using machine learning for improving pharmacokinetic properties.

P4: Development of training algorithms for predicting gene expression outcomes from the distribution of RNA polymerase on genes

Academic supervisor: Prof Jane Mellor (DPhil in Biochemistry); IBM Supervisor: Prof Lior Horesh

The processing, nuclear export and stability of the RNA molecules are pre-determined before transcription. We will use simulated profiles for nascent transcription over human genes to train an algorithm to predict mechanisms and parameters of transcription and gene expression, allowing all the downstream processes to be predicted without experimental determination of each stage.

P5: Atomistic modelling of condensed matter on a quantum compute

Academic supervisor: Prof Dieter Jaksch (DPhil in Condensed Matter Physics) and Prof Charlotte Deane (DPhil in Statistics); IBM co-supervisor: Prof Jason Crain and Dr Vadim Elisseev

The description of multipolar quantum fluctuation is a high-dimensional intrinsically complex many-body problem. In this project, we will develop quantum computing algorithms to start tackling such problems and implement them on the current generation of IBM quantum computers. Hardware specific solutions that utilize the available hardware architectures optimally will be developed.

P6: Revolutionising chemical synthesis with machine learning

Academic supervisor: Prof Fernanda Duarte (DPhil in Organic Chemistry); IBM co-supervisor: TBC

This project aims to accelerate the discovery of complex drugs and materials by combining organic chemistry principles, quantum chemistry, and machine learning. It will focus on the

development of interpretable models that can not only lead to the discovery of new drugs but also provide an understanding of challenging processes leading to their formation and mode of action.

P7: Optimization methods for machine learning

Academic supervisor: Prof Coralia Cartis (DPhil in Mathematics); IBM co-supervisor: Prof Lior Horesh, Dr Soumyadip Ghosh, Dr Songtao Lu

Optimization problems, of huge scale, form the modelling and numerical core of machine learning and statistical methodologies. A grand challenge in this area is the need to augment stochastic gradient optimization methods with inexact second-order derivative information, so as to obtain more efficient methods especially in the nonconvex case of deep learning. In this project, we will investigate ways to approximate second-order information in the finite-sum structure of ML optimization problems.

P8: Compiling chemistry / optimisation problems to higher-level anharmonic oscillators

Academic supervisor: Niel de Beaudrap (DPhil in Computer Science); IBM co-supervisor: TBC

This project aims to develop techniques to break down optimisation problems and other problems, using building-blocks consisting of higher-level interactions in superconducting quantum systems.

P9: Optimising qubit control with machine learning

Academic supervisor: Dr Natalia Ares (DPhil in Materials); IBM co-supervisor: Dr Ali Javadi-Abhari

This projects aims to achieve automated tuning of semiconductor qubits encoded in gate-defined quantum dots and to use a machine learning approach to enable the tuning of large quantum circuits.

P10: New paradigms for dissecting those “black-box” AI models

Academic supervisor: Prof. Jared Tanner (DPhil in Mathematics); IBM co-supervisor: Dr. Payel Das and Dr. Pin-Yu Chen

Understanding deep learning (DL) models is of critical importance, as a framework to explain black-box models will chart a path to trustworthy AI. Neural network training relies on the ability to find “good” minimizers of highly non-convex loss functions. There is an urgent need to develop new theoretically grounded framework that can automatically and efficiently decipher the structure-performance (both in terms of generalization and robustness) of NN models. A similar framework is also needed to understand the impact of hyperparameter settings on model performance, as it is known that well- chosen training parameters (batch size, learning rate, optimizer) produce minimizers that generalize better. Finally, the framework will allow developing models capable of multi-task and multi-environment learning.

P11: New paradigms for combining AI and Molecular Simulations for Accelerated Discovery

Academic supervisor: Prof Philip Biggin (DPhil in Biochemistry); IBM co-supervisor: Dr. Payel Das

Data-driven approaches including Machine learning (ML) and deep learning (DL) have shown incredible promises for accelerating scientific discovery in domains such as biology, chemistry, and material science. On the other hand, molecular simulations have come up a long way and being routinely used as a complement or supplement to experiments for validating predictions and providing mechanistic insights to complex processes. There is an urgent need to develop new technology paradigms for automatically and efficiently (1) including the feedback from simulations in order to improve the ML model, (2) optimizing the need for additional expensive simulations, and (3) identify hidden features from the cheaper simulations that can be a good proxy to learning from expensive simulations.