Preface

Chapter 1 introduces the design-make-test cycle in drug discovery and the promise of machine learning (ML) for accelerating the process.

Chapter 2 gives an overview of molecular featurisation and methods in machine learning which are used in this thesis.

In Chapter 3 starts at the hit-finding stage of drug discovery, and we discuss the usage of unsupervised learning for modelling the 3D distribution of pharmacophores in fragment-protein complexes. This work resulted in the following preprint (manuscript under review):

> William McCorkindale, Ivan Ahel, Haim Barr, Galen J. Correy, James S. Fraser, Nir London, Marion Schuller, Khriesto Shurrush, Alpha A. Lee. Fragment-Based Hit Discovery via Unsupervised Learning of Fragment-Protein Complexes.

In this work, I implemented the model and conducted the computational validation and virtual screening. Dr Ivan Ahel, Dr Haim Barr, Dr Khriesto Shurrush, and Prof Nir London performed bioactivity assays of ligands against SARS-CoV-2 Mpro. Dr Galen Correy and Prof James Fraser obtained X-ray crystallographic structures of ligand-bound structures to SARS-CoV-2 nsp3-Mac1. Dr Marion Schuller performed bioactivity assays of ligands against nsp3-Mac1. Dr Alpha A. Lee supervised the work. I was the principal author of the article.

Chapter 4 brings us to the hit-to-lead stage where modelling bioactivity becomes possible, and we discuss using a model that learns to rank molecules pairwise by activity. This work resulted in the following publication:

> Aaron Morris, William McCorkindale, The COVID Moonshot Consortium, Nir Drayman, John D. Chodera, Savaş Tay, Nir London and Alpha A. Lee. Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model, *Chem. Commun.*, 2021,57, 5909-5912

In this work, I developed the ranking model and constructed the screening library. Aaron Morris evaluated the model and generated compound synthesis routes. Dr John D. Chodera performed docking calculations. Prof Nir London performed bioactivity assays of ligands

against SARS-CoV-2 Mpro. Dr Nir Drayman and Prof Savaş Tay performed OC43 live virus assays. Dr Alpha A. Lee supervised the work.

In Chapter 5 we quantitatively explain predictions from deep learning models used for chemical reactions prediction, revealing model biases due to shortcomings in the training data. This work resulted in the following publication:

> Dávid Péter Kovács, William McCorkindale and Alpha A. Lee. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications* volume 12, Article number: 1695 (2021)

I worked jointly with Dávid Kovács on this work which he completed as part of his MPhil research project under Dr Alpha A. Lee. We contributed equally to model development. Dávid Kovács trained the models, and analysed the model attributions for various reaction classes. I applied reaction templates for data analysis and artificial dataset generation, and investigated model performance under Tanimoto splitting. Dr Alpha A. Lee supervised the work.

Chapter 6 discusses the training of ML models on high-throughput bioactivty measurements from crude reaction mixtures instead of purified compounds. This research was carried out in collaboration with Dr Emma King-Smith, Mihajlo Filep, Prof Nir London, and Dr Alpha A. Lee. In this work, I implemented the random forest model and constructed the screening library. Dr Emma King-Smith implemented the gaussian process model and cleaned the experimental data. Mihajlo Filep performed bioactivity assays against SARS-CoV-2 Mpro. Dr Alpha A. Lee and Prof Nir London supervised the work.

The final chapter summarises the research presented and discusses promising directions for future research.