**Machine learning for biomolecular assembly**

Dr Matteo Degiacomi and Dr Chris Willcocks

Any organism, from a simple bacterium up to a human body, contains millions of proteins with roles as diverse as sensing, transport, defense, motility or structural support. Proteins are biopolymers that fold into specific three-dimensional shapes, determining their capacity of interacting with designated binding partners such as DNA, drugs or other proteins. While these tightly controlled interactions are key for life as we know it, they are also associated with diseases and disorders. Indeed, pathogens leverage proteins as virulence factors, and genetic mutations lead to proteins with altered shape and capacity of binding with their partners.

Nowadays, the understanding of illnesses and design of new therapies is most often driven by the determination of proteins atomic structure. Yet, >70% of proteins carry out their biological task as part of a complex. Revealing the atomic structures of such complexes is often challenging, and so computational methods are being developed to predict their shape starting from the know atomic structure of their protein subunits. The main difficulties limiting the quality of predictions made is that proteins are flexible molecules that often change conformation when binding to their partners, and computational prediction algorithms are in general not suitable to determine which proteins *shouldn’t* interact. This project will tackle the challenge of protein assembly prediction by combining deep learning and molecular simulation.

Molecular dynamics (MD) simulations can be used to predict the specific movements proteins are capable of. Deep generative neural networks (GNN) can be trained with highly complex datasets, and then used to generate new plausible data. Recently, the Degiacomi and Willcocks groups have jointly devised the first GNN capable of generating protein conformations from examples produced by MD simulations.

******

*Figure: a neural network trained on how proteins move can be used to predict the structure and energy of transition states, forecast the effect of modifications, and predict structure and properties of large complexes.*

In this project, we will create the first GNN capable of simultaneously discovering the most likely conformation of binding partners and their relative arrangement, from samples of their conformations generated by MD. Our experimental collaborators will validate our models via mutagenesis. Depending on the student’s specific interests, further research directions will include enhancing GNN’s training by guiding it with low-resolution data (e.g. known interatomic distances) or by integrating information about the structure of known protein complexes. This is expected to further expand the applicability of GNNs towards the ambitious goal of determining all binding partners within a large pool of proteins (“predicting the interactome”), and how modifications such as genetic mutations can impact protein assembly.