Image processing to interpret protein crystal structures

Our understanding of the function of proteins, DNA, RNA and other biological macromolecules, as well as the design of new drug molecules, rely strongly on the possibility to obtain atomic-resolution structures by X-ray and neutron crystallography. Currently, almost 150 000 such structures are freely available in the protein databank. In Lund, data collection for such structures can be performed at the Max IV laboratory and when ESS is running, it will be possible to collect data for neutron structures at an unprecedented speed.

However, the way from the experimental raw data (i.e. the reflection intensities) to the detailed atomistic models is quite long and involves much computation, model building and model optimisation. In particular, the lack of accurate experimental phases is a serious problem. After some processing, an electron- (for X-rays) or nuclear-density (for neutrons) map can be obtained, into which the atomistic model is built. This involves many important but often hard and somewhat subjective choices. One important example is to tell what density peaks correspond to water molecules, i.e. to discern water molecules from other molecules and from random noise. In this choice, both crystallographic and chemical information should be considered.

Traditionally, such choices are performed by the crystallographer more or less by hand. Therefore, structures refined by different crystallographers may not be the same, even if they used the same program. This is quite unsatisfactory. In this project we will investigate whether this problem can be solved by image-processing techniques. The problem of identifying peaks in a density map is similar to other problems treated by image processing.

To start with, we will restrict the project to identify water molecules in X-ray crystallographic maps. To train the model, we will employ a set of (~1000) curated maps where standard methods clearly identify the presence or absence of a water molecule. Additional data can easily be generated, both from existing crystal structures or from simulated molecular data.
One way to approach this problem is to train an artificial neural network (ANN) based on the available training data. In order to do this the representation of data obtained from X-ray or neutron crystallography has to be carefully considered. Also, different architectures of the network has to be tried out, e.g. deep neural networks (DNN). Of course, completely different techniques may also be used and tested.

The project is intended to lead to a scientific article. It is a collaboration between the image-processing group, crystallographers and theoretical chemists.