Predicting experimental phasing success for data triaging

Melanie Vollmar[1], Irakli Sikharulidze[1], Gwyndaf Evans[1,2]

[1]Diamond Light Source, Didcot, United Kingdom;
[2]Rosalind Franklin Institute, Didcot, United Kingdom

melanie.vollmar@diamond.ac.uk

Over the recent years there have been large advances in technologies at synchrotron facilities. Photo-counting detectors with high frame rates (several hundred fps) allow for rapid data acquisition. Robotic sample exchangers combined with automated sample centring enable high-throughput sample screening. Fully automated and unattended data collection set-ups offer the possibility to rapidly gather data. Taken together, all these technologies produce vast amounts of data which need to be analysed and stored. Even for an expert crystallographer it can now be very challenging to assess the data gathered during an experimental session. For novel or non-expert users, the data amounts may even feel overwhelming. Additionally, many research groups do not have access to high-performance computing infrastructure or large storage space to keep their data and analyse it and for research facilities like synchrotrons this infrastructure is limited too.

Here we present some initial results for a machine learning-based triaging system which is currently being trialled at Diamond. The aim is to refine the current brute-force experimental phasing pipelines by introducing data driven triage and decision making. The system as it is in place, relies on data fulfilling certain metrics thresholds before being triggered and executing a number of experimental phasing programs in parallel. Each of these programs can run hours and up to a day before producing an output without a guaranteed success. Based on our initial results presented here, we now propose a machine learning-based decision maker which will estimate the chances of successful experimental phasing for the different software packages available within Diamond's automated data analysis pipelines. The outcome of the classification process is then used to execute subsets in the pipelines in a hiearchical fashion.

Keywords: experimental phasing, machine learning, data triaging