MS45 What is inside the black box?

MS45-05
Peering at the data inside the Black Box
R. Oeffner [1], A. McCoy [1], C. Millán [1], T. Croll [1], R. Read [1]
[1]Cambridge Institute for Medical Research, University of Cambridge - Cambridge (United Kingdom)

Abstract
The availability of high-quality predicted models of proteins, from machine-learning tools such as *AlphaFold2* and *RoseTTAFold*, is accelerating the automation of structural biology and making it more accessible to novices. This trend makes it even more important to have intuitive tools that help inexperienced users to peer inside the black box when things go wrong.

With phasing of crystallographic data and subsequent model refinement being greatly simplified by reliable starting models, undiagnosed data pathologies are much more likely to be the remaining barrier to success [1]. Likelihood-based methods, which power the automation machinery, rely on a variety of assumptions about the data. When pathologies that violate those assumptions cannot automatically be identified and mitigated, the user needs help to explore the data. Tables of statistics and one-dimensional plots, such as those produced by *phenix.xtriage* [2], are helpful, but direct visualization of the data provides a new dimension.

The *HKLviewer* [3], which is included in Phenix [4] and CCTBX [5], is a 3D visualisation program for inspecting X-ray diffraction data. It can launch the *xtricorder* tool in *phasertng* [6] to carry out statistical analyses characterizing features such as anisotropy in the data, the information gained by the diffraction measurements, and the potential presence of translational non-crystallographic symmetry (tNCS). Properties of the data can be visualized by varying the sizes and colours of diffraction spots, and axes corresponding to tNCS translations or twin operators can be superimposed. The reflection data can either be examined as an asymmetric unit wedge or quickly expanded to P1 in a computationally lightweight manner. Derived properties (such as the intensity divided by its standard deviation) can be calculated, using simple Python commands invoking CCTBX, and then displayed as well.

The *HKLviewer* has been designed to help raise awareness of potential issues that might be encountered in downstream structure solution software. It can also be deployed as a teaching aid to educate novice crystallographers about the issues that can arise in data collection.

References

[1] McCoy et al. Acta Crystallogr D Struct Biol Crystallogr. 78, 1–13 (2022).
[2] Zwart et al. CCP4 Newsletter 43 (2005). http://legacy.ccp4.ac.uk/newsletters/newsletter43.pdf.
[3] Oeffner et al. Computational Crystallography Newsletter 12, 15-25 (2021).
[4] Liebschner et al. Acta Crystallogr D Struct Biol 75, 861–877 (2019).
[5] Grosse-Kunstleve et al. J. Appl. Cryst. 35, 126-136 (2002).
[6] McCoy et al. (2022). Acta Crystallogr D Struct Biol Crystallogr. 77, 1–10