Embedding AI in the Protein Crystallography Workflow

RICHARD J. GILDEA, CHRISTIAN M. ORR, NEIL G. PATERSON, AND DAVID R. HALL

Diamond Light Source Ltd., Diamond House, Harwell Science & Innovation Campus, Didcot, Oxfordshire, OX11 0DE, UK

© david.hall@diamond.ac.uk

Historically, solving the structure of a protein required deep knowledge of crystallography and the ability to produce protein crystals of suitable quality to generate high-quality diffraction data. Over the years, as beamline optics, end-stations, detectors, and data collection strategies have improved, it has become more feasible to extract high-quality diffraction data from ever smaller or less perfect protein crystals and from very large arrays of crystals for techniques such as serial synchrotron crystallography and fragment-based drug discovery. At Diamond, these improvements have been coupled with highly integrated automated pipelines for data reduction and structure solution using techniques such as molecular replacement and experimental phasing. This has led to the dichotomy, and benefits, of being able to do increasingly challenging experiments requiring deep crystallographic knowledge with facility staff support at the same time as lowering the barrier to entry where automated structure solution tools of the facility perform this task for those scientists with less experience. This enables users to focus on the science rather than the process.

Diamond Light Source, the UK’s national synchrotron, has a suite of instruments dedicated to solving the 3D structure of large biological molecules, including seven macromolecular crystallography (MX) beamlines. Solved 3D structures are deposited into the publicly available Protein Data Bank (PDB) and the depositions are released on a weekly basis. In 2020, following 13 years of operation, Diamond hit the milestone of 10,000 structures deposited in the PDB. Two years on, this number is now more than 12,000. Thanks to decades of work across the world, there is an ocean of information in the PDB that serves as an invaluable reference when solving the structures of new proteins.

The perfect case for AI

Aside from the PDB’s main remit of being an open repository of structural information for scientists and the wider community, it has been an invaluable resource for solving protein structures providing a reference to assemble new data. Using a technique called molecular replacement, a scientist can borrow information from a sufficiently similar structure in the PDB and apply it to their own data. This then acts as a starting point to help assemble a brand-new structure.

These techniques have allowed structures to be solved that would otherwise have required experimental phasing, a technique to ascertain the information where it cannot be borrowed. If no similar PDB existed and experimental phasing attempts were unsuccessful, there would be no appropriate scaffold on which to build a new structure.

Such a wealth of information in the PDB, coupled with the complexity of solving protein structures, is the perfect problem for artificial intelligence. Almost exactly a year ago, a paper published in Nature described AlphaFold2 [1], an AI system to predict 3D models of proteins.

AlphaFold2 is a state-of-the-art machine learning system developed by Google’s DeepMind. It is open source, meaning that anyone can use it to predict a 3D model of any protein using only sequence data. This means that before a scientist remotely connects to or sets foot into the experimental hall of a synchrotron, they may have a good idea of what the structure of their proteins could be. At Diamond, AlphaFold2 has been embedded into its computational pipelines for academic users to create models specifically based on their target protein sequence (Figure 1), which are used in downstream automated structure solution pipelines running after data collection (Figures 2 and 3).

Will this change the way scientists work at the beamline?

The availability of model prediction tools (not just AlphaFold2 but also the similar RosettaFold [2] procedure) has already changed the way that scientists conceptualise and execute experiments from optimising expression construct design, de novo protein design and in structure solution from data collected at beamlines [3]. While AlphaFold2 is freely available and the AlphaFold Protein Structure Database [4] (https://alphafold.ebi.ac.uk) provides access to a huge number of structure predictions, it was important to make sure that it was as accessible to beamline users as seamlessly as possible by integrating it with tools that were already in their workflow. AlphaFold2 is now integrated into Zocalo, the data analysis infrastructure used at Diamond [5]. Prior to the experiment, users may provide a target protein sequence via SynchWeb [6] (Figure 1), an interface to the ISPyB database and experimental information management system [7]. On providing a sequence, AlphaFold2 will be run via Zocalo, with the resulting predicted models stored in the ISPyB database for use in molecular replacement pipelines after subsequent data collections.

This is crucial, as MX beamline users may collect hundreds of datasets per session, so the more processing and structure solutions that can be provided automatically the better. ISPyB is the interface for providing the starting information needed to do this processing and monitoring of the results. The approach of running AlphaFold2 on user-provided sequences prior to running experiments also sidesteps a downside of prepopulated AlphaFold databases; namely, that users will often not...
News and Views

Figure 1: Example of a user adding a sequence in SynchWeb/ISPyB and the models generated by AlphaFold2 are inserted back into the database and shown at “Associated PDB Files” for use in automated molecular replacement pipelines following data collection.

Figure 2: Refinement statistics from automated MrBUMP [8] molecular replacement pipeline for user experiment. Top graph used a search model from the PDB from a sequence homology search and the bottom graph used an automatically generated AlphaFold2 model, both using the sequence added in SynchWeb/ISPyB as shown in Figure 1. It is clear that, in this case, using the AlphaFold2 model results in lower R factors with a smaller Rwork-Rfree gap than the homology model found in the PDB.
use the full protein. Their constructs are often optimized to a single protein domain, or region, of interest and that is conducive to crystallization. An additional benefit of users providing sequences in ISPyB is that this information can be used for other decision-making processes in data collection and automated pipelines beyond its usage for generating models via AlphaFold2.

Providing AlphaFold2 through integration within SynchWeb/ISPyB allows users to take advantage of the powerful software run on dedicated computers at Diamond so that they don’t have to use their own systems, which are likely not optimized for such a resource-intensive operation.

Current status

AlphaFold2 is a powerful tool for predicting protein structures requiring only the protein sequence as an input to render hypothetical models. However, AlphaFold2 suffers from some of the same issues as more traditional molecular replacement methods in that its accuracy is biased by being based on what is well-represented in the PDB. The AlphaFold2 algorithm uses the experimentally derived structures deposited in the PDB to learn how to generate similar hypothetical models, and consequently the absence of available training data for as-yet experimentally solved novel proteins and folds will make these harder to predict.

For challenging-to-solve protein structures, where AlphaFold2 inevitably fails either to generate a model or the models are not of sufficient quality for structure solution via molecular replacement, experimental phasing is required and Diamond’s suite of MX beamlines offers strong and unique possibilities in this space. Most of the beamlines (I03, I04, I23, 124, VMXi, VMXm) are energy tuneable, enabling experimental phasing using soaked, bound, substituted, or inherent metals coupled with software pipelines to automatically phase and solve structures. Of note is beamline I23, which extends to very low energies, providing access to absorption edges of elements of biological importance such as Ca, K, Cl, S, and P. Recently, the (sub)micro focus beamline VMXm solved a novel structure from just a few selenomethionine derivatized microcrystals mounted on an EM grid. Experimental phasing methods and anomalous contrast can help enormously in solving challenging protein structures as well as unambiguously identifying the position and nature of elements bound to macromolecules.

Organisms whose proteins have been explored relatively less (including some pathogens) can have novel folds which AlphaFold2 has yet to learn. These have to be solved experimentally, firstly to inform the science and secondly to provide well-curated new experimental structures to enhance the PDB training data set. An algorithm is only as robust as the data it learns from, and it would be naive to believe AlphaFold2 has all the answers. A positive feedback loop between experimental data and machine learning is essential to ensure we advance rapidly in structural biology.

The future with AlphaFold2

Ironically, such a powerful and important change to experimentation and data analysis will be mostly invisible to the users. On the whole, AlphaFold2 will run in the background, predicting models automatically to give the users a higher chance of solving their structure. This is a true benefit of AlphaFold2 and the integration with SynchWeb/ISPyB. With good data stewardship, researchers will be able to solve the structures of more and more challenging and biologically important proteins. This will give us better data and a solid framework for doing more in-depth experiments with the proteins that go beyond simply determining the structure, which is often just the start of the process. With AlphaFold2 models assisting some of the heavy data crunching, scientists and beamlines will be
enabled to explore the proteins in greater detail, such as viewing protein movements and scrutinizing chemical environment at large scale. Consequently, we highly recommend that users add their sequence in SynchWeb/ISPyB to automatically generate AlphaFold2 predicted models to use in the already feature-rich environment available for experiments at Diamond MX beamlines.

Acknowledgments

The authors gratefully acknowledge the generous use of unpublished data from Nick Furnham, Xiaojia Liu (London School of Hygiene and Tropical Medicine), Floriano Silva (FioCRUZ), Halina Mikolajek (Diamond Light Source), Lauren Eyssen and Ray Owens (Rosalind Franklin Institute).

References

1. J. Jumper et al., *Nature* **596** (7873), 583 (2021).
2. M. Baek et al., *Science* **373** (6557), 871 (2021). doi:10.1126/science.abj8754
3. M. Edich et al., *Faraday Discuss.* (2022). doi:10.1039/D2FD00072E
4. K. Tunyasuvunakool et al., *Nature* **596** (7873), 590 (2021). doi:10.1038/s41586-021-03828-1
5. M. Gerstel et al., in *Proc. ICALEPCS’19*, New York, NY, USA, Oct 2019, pp. 1031–1035. doi:10.18429/JACoW-ICALEPCS2019-WEMPR001
6. S. J. Fisher et al., *J. Appl. Crystallogr.* **48** (3), 927 (2015). doi:10.1107/S1600576715004847
7. S. Delageniere et al., *Bioinformatics* **27** (22), 3186 (2011). doi:10.1093/bioinformatics/btq555
8. R. M. Keegan et al., *Acta Cryst* **D74** 167 (2018). doi:10.1107/S205979831803455