

An Intelligent Data Analysis System for Biological Macromolecule Crystallography

Hao-Kai Sun^{1,*}, Yu Hu¹, Zhi Geng², Zengqiang Gao², Xin Zhang^{3,4}, and Wei Ding³

¹Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China

²Multi-disciplinary Research Division, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China

³CAS Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing, China

⁴School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong

Abstract. In this work, we design and implement a user-friendly, AI-empowered, auto-pipelining data analysis system for biological macromolecule crystallography. It consists of four modules, (1) data reduction that generates reference reflection files from X-ray diffraction images, (2) structure prediction via database-querying or AlphaFold/OpenFold real-time prediction, (3) molecular replacement and (4) module building and refinement. This data analysis system, currently at Work-In-Progress stage, is based on and developed for High Energy Photon Source initially, aiming at automatic, intelligent, and high-efficiency software and will be open-source for academic research.

1 Introduction

X-ray crystallography is an important technique for revealing the structures of biological macromolecules and understanding related biochemical processes [1]. It plays a crucial role in providing detailed insights into the three-dimensional arrangement of atoms within these molecules, enabling a better understanding of their functions and interactions [2], [3].

With the advancements in light source technology, particularly the construction and operation of fourth-generation light sources like the European Synchrotron Radiation Facility Extremely Brilliant Source (ESRF-EBS) [4], Advanced Photon Source Upgrade (APS-U) [5], Advanced Light Source Upgrade (ALS-U) [6], and High Energy Photon Source (HEPS) [7], significant progress has been made in the field of biological macromolecule crystallography (BMX). These state-of-the-art facilities have facilitated the establishment of advanced BMX beamlines, paving the way for the accumulation of vast amounts of raw experimental data [8].

Among the 14 beamlines of HEPS Phase I currently under construction, the BA beamline is specifically designed for BMX. This beamline, equipped with cutting-edge instrumentation and techniques, aims to further enhance the capabilities of X-ray crystallography in studying biological macromolecules, offering researchers unprecedented opportunities to explore their structures and functions. However, the abundance of data generated by these advanced light sources, coupled with the integration of high-resolution hybrid pixel array detectors, poses

*e-mail: sunhk@ihep.ac.cn

significant challenges to the traditional manual or semi-automatic data processing procedures. The large-scale and excellent-quality data obtained require robust data analysis techniques to extract meaningful information and bridge the gap between experimental data and structural interpretation. Data analysis serves as the key bridge between the “experiment” and the resulting “structure”. Particularly for future advanced light sources that generate high-precision and high-throughput data, efficient and accurate data analysis methods are essential to handle the immense volume of information and extract valuable insights .

In this paper, we aim to introduce an AI-empowered, user-friendly data analysis system for BMX, consisting of four modules. The system operates in two modes: real-time/online analysis, which automatically processes user experimental data in the background using default parameters, and batch mode, where users configure analysis procedures through a graphical user interface (GUI) before processing multiple datasets concurrently for improved performance. All the tools and algorithms within the system are designed as interchangeable plugins, allowing for convenient substitutions.

It should be noted that since this system is a Work-In-Progress (WIP) project, this paper will focus on its design, current status, and future plan.

2 Algorithms and Software

In the field of BMX, significant advancements have been made in methodology over the years. As a result, a plethora of software packages and algorithms have been developed to aid in various aspects of crystallographic data processing and analysis. These software packages excel in specific areas such as indexing, integration, scaling, phasing, model-building and refinement. Some notable examples include XDS, DIALS, HKL-2000/3000, AutoPX, autoPROC, Phaser, Autobuild, Buccaneer, and SHELX.

XDS [9] is a widely used program for processing X-ray diffraction data, providing accurate and efficient integration, scaling and correction of diffraction images. DIALS [10] is a comprehensive software package that focuses on data indexing, integration, and scaling, with an emphasis on high-quality data analysis. HKL-2000/3000 [11] is a popular suite of programs that offers a range of tools for data processing, including scaling, merging, and visualization.

AutoPX [12] and autoPROC [13] are automated pipelines that integrate multiple software tools for data processing and analysis, providing streamlined workflows for crystallographers. Phaser [14] is a widely used program for molecular replacement, a technique used to determine the phase information of a crystal structure. Autobuild [15] and Buccaneer [16] are software packages that assist in automated model-building based on experimental data. SHELX [17] is a suite of programs widely used for crystal structure determination and refinement.

These software packages and algorithms have significantly contributed to the progress in biological macromolecule crystallography, enabling researchers to process and analyze crystallographic data efficiently and accurately. By leveraging the capabilities of these tools, crystallographers can gain valuable insights into the structures and functions of biological macromolecules.

3 Design Goal and Project Architecture

The design goal of this data analysis system consists of three key points,

- Automation

One of the modules of this system achieves automation through parallel-running modules for each data processing step and an integrated pipeline for end-to-end data flow. Each module encapsulates a distinct step, such as data normalization, indexing, prediction, and modeling. The modules run in parallel on distributed compute resources for fast processing of large datasets. The pipeline then automatically sequences the data through each processing stage without manual intervention. This enables one-click execution of complex workflows from raw data to final protein structures.

- **Intelligence**

Intelligent algorithms enhance automation with machine learning capabilities. The framework provides both automatic selection of optimized parameters across modules and integration of AI algorithms like AlphaFold [18] and OpenFold [19] for predictive modeling. Users can validate results and manually select the best outputs, combining human expertise and machine intelligence. The system continuously improves through feedback loops, retraining models on new data.

- **Modularity**

A modular architecture maximizes flexibility, performance, and scalability. Processing modules are built using Docker containers for simplified deployment across environments. The Daisy workflow engine parallelizes and schedules modules while optimizing resource utilization. This high-performance distributed architecture can readily scale from laptops to cloud clusters. The modular design also simplifies integrating new algorithms and software innovations to incrementally enhance the system over time.

By combining automation, machine learning, and modularity, this software framework provides an adaptable platform for efficient processing and analysis of large-scale structural biology datasets. The system reduces manual labor while producing validated, high-quality results. The approach is generalizable to other scientific domains with needs for automated and intelligent data pipelines.

The project architecture can be divided into two stages, with the current emphasis on the initial stage, which involves the implementation of the primary task. This stage primarily focuses on the utilization of advanced X-ray crystallography techniques facilitated by state-of-the-art synchrotron light sources for the analysis of biomolecular structures.

In this context, it is imperative for experiment users to provide the essential raw data files, specifically the diffraction images, along with the corresponding protein sequence(s). This requirement ensures the availability of crucial input data for the subsequent analysis stages. The pipeline, meticulously designed for efficient and accurate analysis, enables parallel execution of the data reduction and structure prediction steps. This parallel processing approach not only expedites the overall analysis workflow but also optimizes the utilization of computational resources, leading to enhanced productivity and reduced processing time.

To further enhance usability and facilitate a seamless user experience, a graphical user interface (GUI) has been designed and implemented. This GUI enables researchers to conveniently navigate through the experimental procedures and swiftly conduct data analysis. The GUI encapsulates the functionalities of data reduction, structure prediction, and result presentation, providing a cohesive and well-designed graphical interactive system. By consolidating these features into a unified interface, the GUI simplifies the complex analysis process, allowing users to efficiently interact with the software and obtain insights from the generated results.

The development of the GUI aims to improve the accessibility and usability of the software, catering to users with varying levels of technical expertise. The intuitive design and smooth operation of the graphical interface enable researchers to seamlessly perform exper-

iment procedures, analyze data, and interpret the outcomes with ease. This integration of a user-friendly GUI with the underlying data reduction and structure prediction algorithms contributes to a comprehensive software package that enables quick and responsive experiment procedures and data analysis.

It is worth noting that this stage of the project focuses on the implementation of the main task, while the subsequent stage may involve additional analyses, such as validation, visualization, and further refinement of the obtained structures, which will contribute to a comprehensive and rigorous analysis of the biomolecular systems under investigation.

4 Current Status

The present status of this system will be delineated individually based on each project module.

- Data Reduction

This module has been successfully implemented, accompanied by the design of an enhanced workflow that operates in a more intelligent manner. The new workflow ensures improved data processing by prioritizing raw data, particularly those with lower quality, to obtain output results with higher reliability and accuracy.

- Structure Prediction

This module has also been implemented and we design two different scenarios for it as illustrated below in **Figure 1** and **Figure 2**.

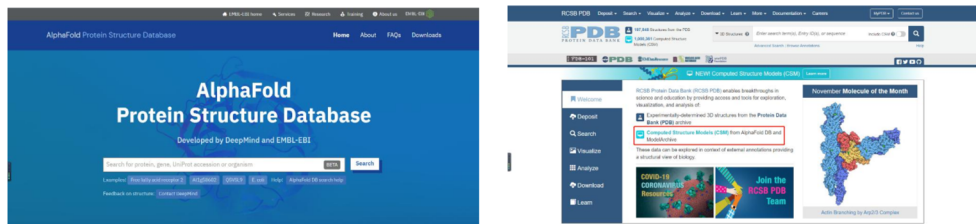


Figure 1. Scenario A: searching predicted structure database which is fast, but not friendly to unconventional sequences.



Figure 2. Scenario B: predicted using AlphaFold2 workstation that is usually with high accuracy but at slow speed.

It is important to highlight that in scenario B, we have taken the initiative to deploy and maintain our own AlphaFold2 prediction server. Additionally, we have localized the genetic databases using full-NVMe SSDs, resulting in a substantial acceleration of the database querying process. To ensure user convenience and seamless integration with GUI, we have encapsulated the execution of AlphaFold2 structure prediction within a readily configurable script, as depicted in **Figure 3**.

```
AF_TEMPLATE="2020-05-14"  
# [monomer, monomer_casp14, monomer_ptm, multimer]  
AF_MODEL="multimer"  
# [full_dbs, reduced_dbs]  
AF_DBPRESET="full_dbs"  
  
queryFile="/path/to/fasta/file"  
outputDir="/path/to/output/folder"
```

Figure 3. Script template for performing structure prediction using AlphaFold2 on our computing platforms.

- Refinement

Additionally, we have devised two distinct scenarios for the refinement module. However, since this module is currently in its early stages of development, a comprehensive discussion of its intricacies is beyond the scope of this paper.

- GUI

This module is highly inspired by the software Aquarium [20]. The development of user interfaces are processing rapidly and now we have achieved a basic framework as illustrated below. In the **Figure 4**, this is the entrypoint of our system’s GUI which consists of three tabs: (1) Home, (2) Data Collection, and (3) Detail.

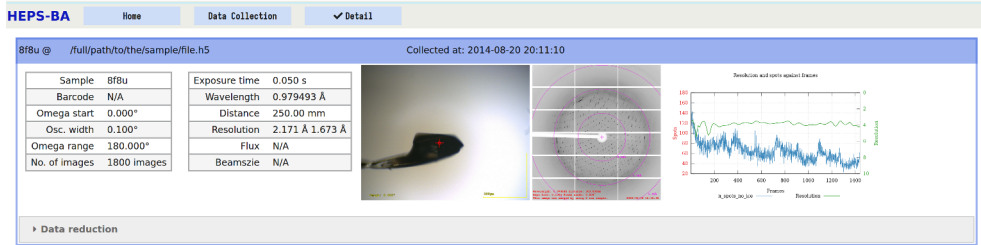


Figure 4. Main frame of the GUI.

And for the Detail tab, all necessary visualization and key values or statistics are shown, for example, in **Figure 5**, and the data reduction can also be expanded as **Figure 6**.

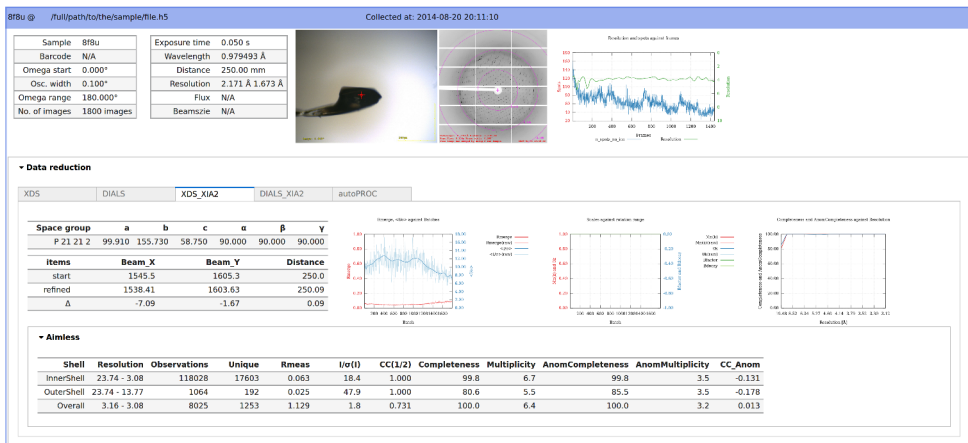
5 Future Plans

In the future, our ongoing efforts will be focused on further advancing the implementation of the main task project, while concurrently updating the system with various enhancements. These improvements encompass augmenting the data reduction performance through the utilization of heterogeneous acceleration techniques and integrating additional algorithms. These developments are being carried out in close collaboration with beamline scientists and experiment users, ensuring that the system is tailored to meet their specific requirements and demands.



© Copyright 2019-2023 IHEP-CC & HEPS-CC & IHEP-PAPS, CAS. All rights reserved.

Figure 5. Detail tab of the GUI,



© Copyright 2019-2023 IHEP-CC & HEPS-CC & IHEP-PAPS, CAS. All rights reserved.

Figure 6. Data reduction expanded under the Detail tab.

Moreover, we are committed to publishing the system, making it readily available for public academic usage at the earliest opportunity. By sharing the system with the scientific community, we aim to facilitate widespread access to advanced tools for biomolecular structure analysis, fostering collaborative research and enabling further discoveries in the field.

Simultaneously, we recognize the importance of developing our own algorithms and software for BMX. This proactive approach will enable us to contribute to the advancement of the field by introducing novel methodologies and innovative solutions that address the unique challenges associated with analyzing complex biomolecular structures. As the field of biomolecular structure analysis continues to evolve, we remain dedicated to staying at the forefront of technological advancements. Through ongoing research and development, in collaboration with experts in the field, we anticipate making significant contributions to the advancement of both experimental techniques and computational algorithms, ultimately facilitating a deeper understanding of biological macromolecules and their functional properties.

References

- [1] Drenth, J., *Principles of Protein X-ray Crystallography* (Springer-Verlag, New York, 2007)
- [2] Chothia, C., Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, **5(4)**, 823-826 (1986)
- [3] Dill, K. A., MacCallum, J. L., The Protein-Folding Problem, 50 Years On. *Science* **338**, 1042-1046 (2012)
- [4] Chaize, J., et al., *Proceedings of the 16th International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS 2017)* (Barcelona, Spain, 2017), 2010-2015 (2018)
- [5] Hettel, R., *Proceedings, 12th International Particle Accelerator Conference (IPAC 2021)* (Online Conference, Brazil, 2021), 7-12 (2021)
- [6] Steier, C., et al., *Proceedings, 10th International Particle Accelerator Conference (IPAC 2019)* (Melbourne, Australia, 2019), W097-100 (2019)
- [7] Jiao, Y., Xu, G., Cui, X. H., et al., The HEPS project. *Journal of Synchrotron Radiation*, **25(6)** 1611-1618 (2018)
- [8] Nogly, P., Kroeger, A., Serial crystallography for membrane protein structure determination. *Current Opinion in Structural Biology*, **66**, 25-31 (2021)
- [9] Kabsch, W., XDS. *Acta Crystallographica Section D: Biological Crystallography*, **66(2)**, 125-132 (2010)
- [10] Winter, G., Waterman, D. G., Parkhurst, J. M., et al., DIALS: Implementation and evaluation of a new integration package. *Acta Crystallographica Section D: Structural Biology*, **74(2)**, 85-97 (2018)
- [11] Otwinowski, Z., Minor, W., Processing of X-ray diffraction data collected in oscillation mode. *Methods in Enzymology*, **276**, 307-326 (1997)
- [12] Wang, L., Yun, Y., Zhu, Z and Niu, L., AutoPX: a new software package to process X-ray diffraction data from biomacromolecular crystals. *Acta Crystallogr D Struct Biol* **78(7)** 890-902 (2022)
- [13] Vonrhein, C., Flensburg, C., Keller, P., et al., Data processing and analysis with the autoPROC toolbox. *Acta Crystallographica Section D: Biological Crystallography*, **67(4)**, 293-302 (2011)
- [14] McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., et al., Phaser crystallographic software. *Journal of Applied Crystallography*, **40(4)**, 658-674 (2007)
- [15] Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., et al., Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D: Biological Crystallography*, **64(1)**, 61-69 (2008)
- [16] Cowtan, K., The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallographica Section D: Biological Crystallography*, **62(9)**, 1002-1011 (2006)
- [17] Sheldrick, G. M., Crystal structure refinement with SHELXL. *Acta Crystallographica Section C: Structural Chemistry*, **71(1)**, 3-8 (2015)
- [18] Jumper, J., Evans, R., Pritzel, A., et al., Highly accurate protein structure prediction with AlphaFold. *Nature*, **596(7873)**, 583-589 (2021)
- [19] Ahdriz G., Bouatta N., Floristean C., et al., OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization, *bioRxiv* 2022.11.20.517210
- [20] Yu, F., Wang, Q., Li, M., et al., Aquarium: an automatic data-processing and experiment information management system for biological macromolecular crystallography beam-

lines. *Journal of Applied Crystallography*, **52(2)**, 472-477 (2019)