electronic reprint

JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN: 1600-5767
journals.iucr.org/j

AutoPD: an integrated meta-pipeline for high-throughput X-ray crystallography data processing and structure determination

Xin Zhang, Haikai Sun, Yu Hu, Zengru Li, Zhi Geng, Zengqiang Gao, Quan Hao, Fazhi Qi and Wei Ding

J. Appl. Cryst. (2025). **58**, 746–758

JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

AutoPD: an integrated meta-pipeline for high-throughput X-ray crystallography data processing and structure determination

Xin Zhang,[a,b,d] Haikai Sun,[b,c] Yu Hu,[b,c] Zengru Li,[a,b] Zhi Geng,[b,c] Zengqiang Gao,[b,c] Quan Hao,[b,c]* Fazhi Qi[b,c]* and Wei Ding[a,c,e]*

[a]Beijing National Laboratory for Condensed Matter Physics, Institute of Applied Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, [b]University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China, [c]Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, [d]The University of Hong Kong, Hong Kong SAR, People's Republic of China, and [e]Songshan Lake Materials Laboratory, Dongguan 523808, People's Republic of China. *Correspondence e-mail: haoquan@ihep.ac.cn, qfz@ihep.ac.cn, dingwei@iphy.ac.cn

Received 12 December 2024
Accepted 9 April 2025

Edited by F. Meilleur, Oak Ridge National Laboratory, USA, and North Carolina State University, USA

Keywords: automated meta-pipelines; data processing; structure determination; high-performance synchrotron sources.

Supporting information: this article has supporting information at journals.iucr.org/j

The advent of hybrid pixel array detectors and fully automated data acquisition workflows has revolutionized synchrotron light sources, enabling high-throughput collection of diffraction data from biological macromolecular crystals. However, these advancements have also created an urgent need for efficient and fully automated data processing pipelines. To address this challenge, we introduce *AutoPD*, an open-source high-throughput meta-pipeline for automated data processing and structure determination. Developed for the biological macromolecular crystallography beamline at the High Energy Photon Source in Beijing, *AutoPD* is also accessible to other academic and synchrotron users. By integrating cutting-edge parallel computing strategies, *AlphaFold*-assisted molecular replacement, a direct-method-based dual-space-iteration approach for model building, and an adaptive decision-making strategy that dynamically selects the optimal modeling pathway based on data quality and intermediate results, *AutoPD* streamlines the process from raw diffraction data and sequence files to high-precision structural models. When benchmarked against 186 recently deposited X-ray diffraction datasets from the Protein Data Bank, *AutoPD* successfully determined structures for 92% of cases, achieving map–model correlation values of at least 0.5 between density-modified electron density maps and the generated models. These results highlight the robustness and efficiency of *AutoPD* in addressing the challenges of modern structural biology, setting a new standard for automated structure determination.

1. Introduction

The field of macromolecular crystallography is undergoing a transformative shift driven by rapid technological advancements. The introduction of hybrid pixel array detectors (Henrich *et al.*, 2009; Johnson *et al.*, 2014; Casanas *et al.*, 2016) and the integration of automation and artificial intelligence (AI) into data collection processes have dramatically increased the speed and capacity of data acquisition. These developments have enabled unprecedented levels of data generation but have also created challenges in efficiently processing and analyzing the resulting datasets. In parallel, fields such as drug discovery have witnessed a surge in data generation, exemplified by methods like virtual synthon hierarchical enumeration screening (Sadybekov *et al.*, 2022). Pharmaceutical companies and research institutions now rely on massive datasets encompassing molecular structures, biological pathways and disease mechanisms, yet experimental verification remains a critical bottleneck. The convergence of advanced synchrotron radiation sources and the vast volume

of candidate data has ushered structural biology into an era characterized by massive data processing demands. This paradigm shift necessitates high-throughput data processing pipelines capable of efficiently managing the deluge of data generated by modern light sources. Beyond addressing the requirements of large-scale synchrotrons, academic and industry users alike need data processing solutions that can be deployed locally, ensuring both confidentiality and security. Therefore, there is an urgent demand for an open-source, fully automated, high-throughput data processing pipeline that supports both large-scale synchrotron facilities and localized installations, offering versatility to meet diverse user needs.

The emergence of high-performance parallel computing combined with automated data processing platforms provides a critical solution to these challenges. Several beamline-specific platforms, including *Auto-Rickshaw* (Panjikar *et al.*, 2005; Panjikar *et al.*, 2009), *ISPyB* (Delageniere *et al.*, 2011; Fisher *et al.*, 2015), *DA+* (Wojdyla *et al.*, 2018) and *Aquarium* (Yu *et al.*, 2019), have been developed with automated data processing capabilities for structure determination. However, these platforms typically either function as comprehensive management systems, incorporating broader sample handling and workflow orchestration, or require tight integration with beamline-specific hardware. This complexity significantly restricts their deployment and accessibility in home laboratories and other synchrotron facilities. In contrast, initiatives such as the Gold Standard (Bernstein *et al.*, 2020) for macromolecular crystallography diffraction data offer promising potential for flexible data handling independent of specific beamline platforms. Such efforts pave the way for developing universally accessible data processing software, enabling users to efficiently handle their datasets regardless of their access to dedicated synchrotron resources.

In the phasing stage of structure determination, the aforementioned platforms can perform experimental phasing when anomalous signals are detected. However, limitations arise when anomalous signals are absent, as seen in *Aquarium* at Shanghai Synchrotron Radiation Facility (SSRF), where the process halts after data reduction in such cases. With the continuous expansion of the Protein Data Bank (PDB; https://www.rcsb.org/), molecular replacement (MR) has become the preferred method for phasing and structure determination. Because MR traditionally relies on homologous structures from existing databases, this approach is constrained by the availability of suitable models. The advent of AI-based protein structure prediction tools, such as *AlphaFold2* (Jumper *et al.*, 2021) and *RoseTTAFold* (Baek *et al.*, 2021), has revolutionized the field by providing highly accurate structural predictions that complement traditional homologous models. Studies have demonstrated that *AlphaFold* predictions can serve as effective initial templates for MR, significantly enhancing the success rate of structure determination (McCoy *et al.*, 2022; Terwilliger *et al.*, 2023). However, existing pipelines, such as *MrBUMP* (Keegan & Winn, 2008; Keegan *et al.*, 2018) employed at the Diamond Light Source, which rely on structural models from the PDB and *AlphaFold* databases for MR, are prone to failure when suitable high-quality templates are unavailable, especially in cases involving novel or previously uncharacterized protein structures. To overcome this limitation, we propose integrating structure prediction directly into data processing pipelines, enabling accurate predictions for sequences lacking database models and thereby improving MR success rates.

Model building, a critical stage for refining and completing structures, is absent from most existing pipelines. While AI-based models have revolutionized structure determination, they face challenges in complex cases where MR solutions are unsatisfactory (Moore *et al.*, 2022; Shao *et al.*, 2022; Terwilliger *et al.*, 2024). Recent research by Li *et al.* (2024) demonstrates the potential of direct methods and dual-space iteration, as implemented in *IPCAS* (Ding *et al.*, 2020), to refine and enhance structural models after MR. *IPCAS*, a direct-method-based pipeline for macromolecular structure determination, encompasses processes from phasing to refinement and model building. Its integration of programs like *OASIS* (Hao *et al.*, 2000; Tao *et al.*, 2010) for direct-method phasing and *CCP4* (Agirre *et al.*, 2023) and *Phenix* (Adams *et al.*, 2010) for location of heavy atoms, density modification and model refinement addresses the challenges of complex cases. By incorporating *IPCAS* into our pipeline, we aim to significantly improve its robustness and success rates.

Here, we introduce *AutoPD*, an open-source, high-throughput, fully automated data processing and structure determination meta-pipeline for biological macromolecular crystallography. Developed specifically for the macromolecular crystallography beamline at the High Energy Photon Source (HEPS) in Beijing (Jiao *et al.*, 2018), *AutoPD* is also accessible to other synchrotrons and academic users. The meta-pipeline seamlessly handles datasets from data reduction to refinement and model building, incorporating advanced features to enhance its performance. To address the limitations of traditional MR methods, *AutoPD* integrates *AlphaFold* predictions, providing robust initial models that expand the scope of structure determination. Moreover, the meta-pipeline incorporates *IPCAS* for refining and extending structural models, leveraging direct methods and dual-space iteration to tackle challenging cases. An adaptive decision-making strategy is embedded in the workflow to dynamically select the most suitable modeling pathway according to the data quality and intermediate results, thereby improving robustness and ensuring higher success rates across diverse datasets. By combining these cutting-edge approaches, *AutoPD* not only streamlines the path from data collection to structure determination but also enhances the reliability and efficiency of the final structural output. The meta-pipeline delivers refined structure models optimized for the lowest $R_{\rm free}$ value, exemplifying its ability to integrate innovative methods into a unified efficient workflow.

2. Implementation

AutoPD comprises five specialized modules, each designed to streamline and enhance the process of crystallographic data analysis.

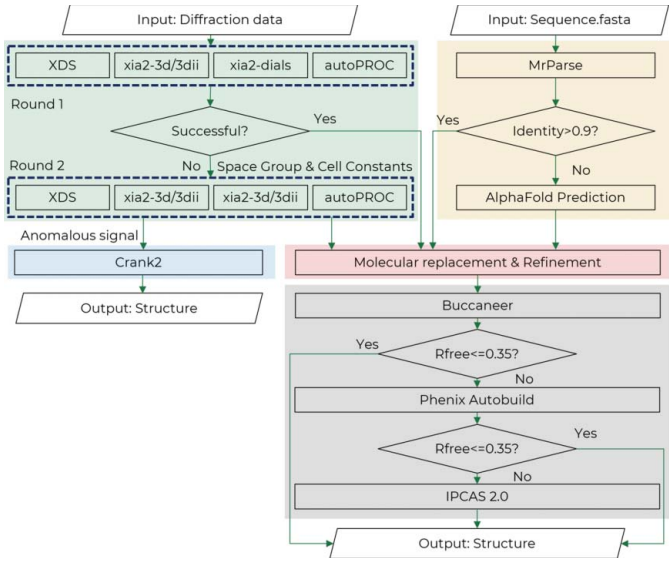


Figure 1
Workflow of *AutoPD*. *AutoPD* is an integrated metapipeline for structure determination using only diffraction data and a sequence file. It includes five modules: data reduction (green), search model generation (buff), molecular replacement (pink), model building and refinement (gray), and experimental phasing (blue). It runs multiple data reduction pipelines in parallel, uses *AlphaFold* prediction when no suitable models are available, and performs iterative model building with *Buccaneer*, *Phenix Autobuild* and *IPCAS*. *Crank2* is triggered if strong anomalous signal is detected.

(i) Data reduction module. This module processes raw diffraction images to extract crystallographic data, providing a solid foundation for subsequent analysis and structure determination.

(ii) Search model generation module. Critical for MR, this module generates the necessary search models, leveraging tools such as *AlphaFold* predictions to address cases where database models are unavailable.

(iii) Molecular replacement module. This module retrieves phase information essential for solving the crystal structure, playing a pivotal role in the structure determination process.

(iv) Refinement and model building module. Following MR, this module constructs and refines the structural model, ensuring improved accuracy and reliability in the final output.

(v) Experimental phasing module. Activated when a strong anomalous signal is detected during data scaling, this module performs phase determination, followed by refinement and model building, offering an alternative pathway to structure determination.

A flowchart of the pipeline is shown in Fig. 1.

2.1. Data reduction module

The data reduction module utilizes a suite of robust data processing software—*XDS* (Kabsch, 2010), *xia2* (Winter, 2010; Winter *et al.*, 2013; Winter *et al.*, 2018; Winter *et al.*, 2022), and *autoPROC* (Vonrhein *et al.*, 2011)—to process raw diffraction data. This module executes a systematic sequence of operations, starting with spot finding and progressing through indexing, integration and scaling, ultimately converting raw diffraction images into a single, cohesive crystallographic file suitable for further analysis.

Using *XDS*, the process begins with indexing, where critical parameters such as beam center, detector distance, rotation axis and incident beam direction are extracted from the `XPARM.XDS` file and updated for subsequent steps. After integration, the space group and unit-cell constants are retrieved from `GXPARM.XDS`, facilitating reindexing. Following this, the beam center, detector distance, rotation axis and incident beam direction are updated again for a final round of integration. The resolution is determined using the *dials.estimate_resolution* tool with default settings, scaling is performed with *AIMLESS* (Evans & Murshudov, 2013), and intensities are converted to structure factors using *CTRUNCATE* (Winn *et al.*, 2011).

For *xia2*, the module employs variations including *xia2-3d*, *xia2-3dii* and *xia2-dials*. The workflow is designed to integrate *xia2-3d* and *xia2-3dii*, beginning with *xia2-3d* as the primary option and switching to *xia2-3dii* in the event of failure or timeout. To prevent stagnation and ensure timely processing, all *xia2* operations are monitored with a configurable time limit, which is set to a default value of 3600 s. This mechanism ensures that any excessively prolonged operations are automatically terminated, maintaining overall workflow efficiency.

For *autoPROC*, most default settings are retained. To enable a fair comparison of data processing statistics across different software packages, the binning in *AIMLESS* is set to 20, consistent with other programs. Certain versions of *autoPROC* can require significantly longer processing times—approximately 1.5 to 3 times more than other programs. Therefore, users may choose to bypass *autoPROC* when time efficiency is a critical consideration.

These processes are executed in parallel using the *GNU parallel* command tool (Tange, 2021) to maximize efficiency. If the initial processing cycle yields incomplete results, a secondary round is initiated to address any unresolved tasks. In this second cycle, the space group and unit-cell parameters derived from the result with the lowest redundancy-independent merging *R* factor ($R_{\rm meas}$) value from the initial run are used as inputs, significantly increasing the likelihood of obtaining valid and comprehensive outcomes. To further enhance flexibility and adaptability, users are also provided with the option to specify key parameters—such as `rotation_axis` (rotation axis), `beam_x` and `beam_y` (beam center), `distance` (crystal-to-detector distance), `space_group` (space group), and `cell` (unit-cell parameters)—to facilitate more accurate and tailored data processing. Once data reduction is complete, all outputs from *XDS*, *xia2* and *autoPROC* are consolidated and seamlessly passed to the next stage of analysis, which may involve either the molecular replacement or experimental phasing module. This workflow ensures efficient and comprehensive processing of diffraction data, maximizing the generation of high-quality crystallographic information for downstream analyses.

2.2. Search model generation module

The search model generation module initiates its process by employing *MrParse* (Simpkin *et al.*, 2022) to retrieve five homologous structures and five *AlphaFold* database models from available databases. To ensure data integrity during validation testing, homologs released after the deposition date of the test entry are excluded. In cases where no *AlphaFold* database models are available, or the top-ranking model shows a sequence identity below 0.9, covers less than 60% of the target sequence or has a pLDDT score[1] lower than 90, a *de novo AlphaFold* prediction is automatically initiated to produce a more accurate and complete structural model specific to the input sequence. The sequence identities of all homologous and *AlphaFold* models are then evaluated and ranked separately, with the highest-ranking homolog model and *AlphaFold* model selected for further analysis. For sequences comprising multiple chains, each chain is processed individually. In cases where a specific chain lacks homologous structures, the *AlphaFold* model for that chain is combined with homologs from other chains to assemble a complete search model for MR.

For operations at HEPS, a local *AlphaFold2* installation configured as a server is used to perform *AlphaFold* predictions efficiently. This setup utilizes software from *ColabFold* (Mirdita *et al.*, 2022) to streamline predictions. The predicted models are then refined using the *Phenix process_predicted_model* tool (Oeffner *et al.*, 2022). This refinement process includes trimming residues with low confidence (pLDDT < 70), converting pLDDT values into estimated atomic displacement parameters and splitting the predicted model into individual chains. These steps ensure that only the most reliable structural predictions are advanced for further analysis. For personal users, *AlphaFold* predictions are performed using the *Phenix PredictAndBuild* tool with the parameter `stop_after_predict = True`. For testing purposes, the parameter `include_templates_from_pdb = False` is applied, excluding PDB templates to ensure unbiased results.

For this module, we provide the parameter `af_predict = true`, which allows users to explicitly invoke *AlphaFold* prediction, regardless of the availability or quality of pre-existing database models. This is particularly useful in cases where users suspect that the database models may not adequately represent the target structure. Additionally, the parameter `pae_split = true` enables automatic splitting of the *AlphaFold*-predicted model into domains based on the predicted aligned error (PAE) matrix. The PAE matrix, generated by *AlphaFold*, provides a residue-wise estimate of the alignment error between pairs of residues. By leveraging this information, the model can be segmented into structurally independent domains, which is especially beneficial when the predicted structure exhibits domain displacements or significant conformational differences relative to the experimentally determined structure. This domain-aware processing improves the robustness and accuracy of the downstream MR step. We do not set `pae_split = true` as the default in *AutoPD* because, in some tests, the model was divided into many small fragments, which did not improve MR performance. In most cases, splitting via the *Phenix process_predicted_model* tool suffices for successful MR.

In sequences with multiple chains, each chain is processed independently, enhancing both efficiency and accuracy. The integration of *MrParse* and *AlphaFold* predictions ensures the generation of highly reliable search models, even for complex cases where database structures are unavailable for certain chains. This modular and robust approach guarantees the creation of precise search models, optimized for downstream MR tasks.

2.3. Molecular replacement module

The molecular replacement module utilizes *Phaser* (McCoy *et al.*, 2007) to perform standard MR analyses, employing all crystallographic files generated by the data reduction module along with two sets of search models prepared by the search model generation module. The *Phenix.Xtriage* tool is used to estimate the number of copies in one asymmetric unit for each ensemble during the MR process.

Initially, our module selected only the crystallographic file with the lowest $R_{\rm meas}$ value from the data reduction module for MR. However, during testing, we observed cases where the lowest $R_{\rm meas}$ file corresponded to a structure with the wrong point group, leading to downstream errors. To address this, we modified our approach to select the best results for each distinct point group according to the lowest $R_{\rm meas}$ value. Nevertheless, this strategy still had limitations, particularly when multiple results shared the same point group but differed significantly in unit-cell parameters. To ensure that potentially correct solutions are not overlooked, the current implementation utilizes all crystallographic files produced by the data reduction module for MR, despite a slight reduction in computational efficiency.

Additionally, we employed two distinct sets of search models—a homolog-based set and an *AlphaFold*-based set—because testing showed that homolog-based models occasionally yield superior results, while *AlphaFold*-based models perform better in other instances. Typically, with four crystallographic files from data reduction and two sets of search models, eight MR jobs run concurrently, systematically exploring all potential space groups within the point group. This parallelized approach ensures comprehensive exploration of possible solutions, enhancing both the thoroughness and success rate of the MR analysis.

Users can also input a crystallographic file using the parameter `mtz_file` or specify the path containing search models using the parameter `pdb_path`. In this case, the corresponding data reduction module or search model generation module will be skipped, and the user-provided files will be directly used in the MR module.

[1] Predicted local distance difference test (pLDDT) (Mariani *et al.*, 2013): a per-residue confidence metric used by *AlphaFold* to estimate the local accuracy of predicted structures.

2.4. Refinement and model building module

Following MR, the resulting models are subjected to automated refinement using *Refmac* (Murshudov *et al.*, 2011; Yamashita *et al.*, 2023) and model building via *Buccaneer* (Cowtan, 2006) within *CCP4i2* (Agirre *et al.*, 2023; Potterton *et al.*, 2018). *Buccaneer* performs iterative cycles of crystallographic rebuilding, refinement and density modification, aiming to generate an improved rebuilt model along with a density-modified map.

If the $R_{\rm free}$ value of the reconstructed model exceeds 0.35, the workflow proceeds to *Phenix Autobuild* for further refinement and model building. If the previous *Buccaneer* job results in an improved $R_{\rm free}$ value, its output model is used as the input for *Phenix Autobuild*; otherwise, the model generated by *Refmac* is used. Should the $R_{\rm free}$ value remain above 0.35 after *Phenix Autobuild*, the process advances to *IPCAS* for additional refinement. Similarly, if the *Phenix Autobuild* step improves the $R_{\rm free}$ value, its resulting model is passed to *IPCAS*; otherwise, the *Refmac* model is reused.

This tiered, multi-step approach ensures the rigorous optimization of model accuracy and module efficiency, progressively refining the structural model through successive stages tailored to address challenging cases and maximize the quality of the final output.

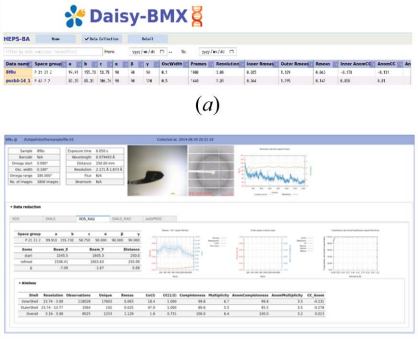
2.5. Experimental phasing module

When a strong anomalous signal is detected during data scaling, all crystallographic files along with the sequence file are submitted to *Crank-2* (Skubák & Pannu, 2013) for experimental phasing. Since *Crank-2* requires a specified heavy-atom type for phasing, sulfur is used as the default when no such information is provided. However, users can override this default by specifying the known heavy-atom type via the `atom` parameter, allowing for more accurate experimental phasing when prior knowledge is available. To enhance efficiency, all *Crank-2* jobs are executed in parallel, with each job utilizing a different crystallographic file. Upon completion, the result with the lowest $R_{\rm work}$ value is selected as the optimal solution. This parallelized approach ensures a thorough exploration of possible phasing outcomes while prioritizing efficiency and accuracy in selecting the best result.

3. Graphical user interface

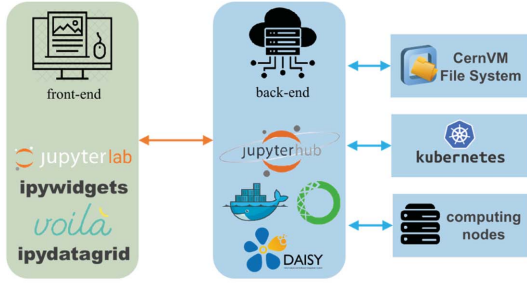
At HEPS, *AutoPD* is integrated into a multi-user graphical web application known as *Daisy-BMX*, which is deployed on an advanced computing platform powered by the *Daisy* framework (Hu *et al.*, 2021*b*). Built atop the *JupyterLab* technology stack (https://jupyter.org/), *Daisy-BMX* offers an intuitive interface for efficient data management and analysis.

The data collection page [Fig. 2(*a*)] provides a comprehensive view of experimental sample results after data processing. Users can sort and filter results by name, collection time or a combination of criteria, facilitating streamlined navigation. By clicking on an entry, users access the detail page [Fig. 2(*b*)], where detailed results and charts generated



(*a*)

(*b*)



(*c*)

```
HEPS-BA
├── date (e.g., 202508)
│   ├── Data
│   │   ├── beamtimeid (e.g., GBA-250809-01)
│   │   │   ├── raw          # Raw detector data (HDF5, CBF)
│   │   │   ├── processed     # Final outputs (MTZ, PDB)
│   │   │   ├── scratch       # Temporary files (e.g., XDS_ASCII.HKL)
│   │   │   └── share         # Intermediate files for collaboration
│   ├── Logs
│   │   └── job_{id}          # Task-specific logs (errors, runtime metrics)
└── Users
    └── {user_id}             # Personal workspaces (JupyterLab notebooks)
```

(*d*)

Figure 2
Daisy-BMX. Web-based user interface for *AutoPD* at HEPS. (*a*) Data collection page: displays sortable and filterable sample results. (*b*) Detail page: integrates *AutoPD* outputs with visualizations and key metrics. (*c*) System architecture: separates *JupyterLab*-based front-end from a containerized back-end managed by Kubernetes, with computing resources accessed via CVMFS. (*d*) Directory structure: supports traceable and scalable data processing, with dedicated spaces for raw data, results, logs and user workspaces.

by *AutoPD* are centrally displayed, offering a clear overview of the processed sample data.

Daisy-BMX employs a modern architecture that separates the front-end and back-end for enhanced flexibility and scalability [Fig. 2(*c*)]. The front-end utilizes *Jupyter-ipywidgets* (https://ipywidgets.readthedocs.io/en/stable/), combining existing and custom-developed widgets for layout and styling. It also integrates *ipydatagrid* (https://github.com/bloomberg/ipydatagrid) and *pandas* (McKinney, 2010) for efficient management and visualization of sample data. *Voila* (https://github.com/voila-dashboards/voila) is used for rendering the interface, ensuring a user-friendly presentation of results and tools.

On the back-end, containerization plays a central role in resource management and task orchestration. Container images for specialized software environments such as *CCP4* and *Phenix* are built using Dockerfiles, managed through Podman and stored in a private Docker registry for deployment via Kubernetes. Open-source applications like *JupyterLab* and *Prometheus* utilize pre-built containers deployed directly from public repositories through Helm charts. The Kubernetes container runtime employs Containerd, enabling efficient compatibility with GPU pass-through via the NVIDIA Container Toolkit and optimized high-performance input/output (I/O) operations.

The directory structure supporting *Daisy-BMX* is designed for effective isolation, traceability and performance [Fig. 2(*d*)]. Directories are organized by beamtime ID and user to prevent file conflicts and ensure data privacy. Essential metadata such as `job_id` systematically links raw detector data, intermediate processing files and final results, enhancing data reproducibility and management. The dedicated scratch directory utilizes high-speed NVMe storage specifically to support intensive I/O operations during data processing.

Given the complexity and potential for failed tasks, jobs are categorized and prioritized into three distinct groups to help users identify the most promising solutions quickly.

(*a*) Highest priority. Jobs successfully solved by MR or single-wavelength anomalous diffraction (SAD) phasing, sorted by model $R_{\rm free}$ values from lowest to highest.

(*b*) Intermediate priority. Jobs where MR fails but data reduction succeeds. These results are sorted by data quality metrics ($R_{\rm meas}$), prioritizing better-quality datasets for further analysis.

(*c*) Lowest priority. Jobs with unsuccessful data reduction due to issues such as poor data quality, incorrect beam center or crystal-to-detector distances, or corrupted image data. These require careful manual examination.

This comprehensive, structured approach ensures that *Daisy-BMX*, powered by *AutoPD*, delivers an accessible, efficient and reliable data processing platform for crystallographic research at HEPS.

4. Computing platform

AutoPD is deployed on the advanced computing platform at HEPS, designed specifically to support high-performance large-scale data processing. This platform integrates Kubernetes orchestration with heterogeneous hardware to optimize resource allocation and task execution, ensuring efficient handling of modern crystallography workflows. A computational workflow diagram illustrating task scheduling, resource allocation and interaction between system components is included to depict this process (Fig. 3).

The computational infrastructure includes GPU servers consisting of 23 nodes equipped with a total of 1472 CPU cores, approximately 1.5 TB of memory per node, and diverse GPU configurations, such as NVIDIA A100 80 GB GPUs dedicated exclusively to *AlphaFold* tasks, and NVIDIA A800 GPUs for general-purpose computations. Additionally, there are 15 CPU servers providing 960 CPU cores with an average memory capacity of 512 GB per node.

Storage infrastructure consists of beamline-specific high-performance storage (`/heps/beamline`) with 1.8 PB capacity and aggregated bandwidth up to 40 GB s^{-1}, centralized shared storage (`/heps/centralfs`) of 14 PB accessible across HEPS beamlines, and a tape-based archival storage system providing 2 PB for secure long-term backups.

Kubernetes orchestration optimizes resource allocation by assigning lightweight front-end tasks, such as user interaction and result visualization, to nodes with fewer CPU cores (*e.g.* four-core pods). Compute-intensive tasks, including MR analyses, *AlphaFold* structure predictions and molecular dynamics simulations, are dynamically scheduled to GPU-

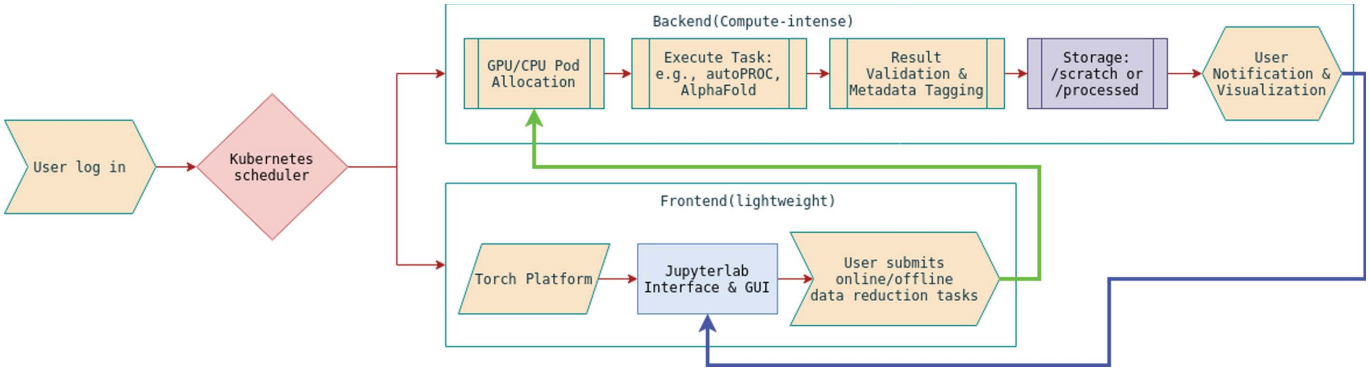


Figure 3
A computational workflow diagram illustrating task scheduling, resource allocation and interaction between system components.

equipped pods or multi-core CPU pods to ensure optimal computational efficiency.

Real-time resource utilization is monitored using *Prometheus* and *Grafana* (https://prometheus.io/), while OMAT (Hu *et al.*, 2021*a*; Hu *et al.*, 2022) is used to coordinate task prioritization and job management.

The computational platform also features dedicated resources for *AlphaFold*, including an exclusive GPU server with two NVIDIA A100 80 GB GPUs and high-speed NVMe SSD storage, capable of supporting concurrent structure prediction tasks for up to four users and managing sequences exceeding 4000 residues. Furthermore, hybrid workload management integrates batch queuing and task distribution through SLURM (Simple Linux Utility for Resource Management) and Ray, complemented by Kubernetes for elastic scaling and efficient resource utilization.

Overall, this robust and flexible computing environment allows *AutoPD* to efficiently handle the computational demands of contemporary crystallography research, providing researchers with a reliable and powerful infrastructure for data processing and structure determination.

The computations described in this work were performed on an *Ubuntu 22.04* operating system, utilizing the following software versions: *CCP4* (version 9.0.004), *Phenix* (version 1.21.2-5419), *XDS* (version 20230630), *DIALS* (version 3.22.1), *autoPROC* (version 20240710) and *IPCAS* (version 2.0).

5. Test and results

5.1. Test data

To evaluate the performance of *AutoPD*, we selected entries from the PDB that met specific criteria, ensuring the collection represented structures determined after the training period of *AlphaFold2*, which utilized data available up to April 2018. The dataset comprises 186 unique protein structures with associated diffraction data, all released between 1 January 2022 and 31 December 2023. The corresponding raw diffraction images were downloaded via the DOI links labeled 'Diffraction Data' provided on the respective PDB entry pages, and the sequences were obtained directly using the sequence download option of each PDB entry page for testing.

To maintain the integrity and relevance of the test set, data with the same space group and similar cell parameters were excluded to avoid redundancy. Additionally, data containing DNA or RNA chains were excluded to focus solely on protein structures. Data involving more than one dataset was also excluded to simplify the evaluation process. This curated dataset provides a diverse and representative sample for assessing the effectiveness and robustness of *AutoPD* in processing modern crystallographic data.

5.2. Overall results

Our results demonstrate the successful application of the comprehensive data processing and structure determination meta-pipeline to 186 deposited datasets. During this process, the data reduction module failed to produce valid solutions—due to incorrect point group or unit-cell parameters, or a complete lack of solution—in seven cases. Additionally, in eight instances, the MR module was unable to generate a solution or produced models with incorrect space groups, leading to the discontinuation of analysis for those datasets. Despite these challenges, *AutoPD* successfully processed the remaining 171 datasets in a fully automated manner, without any manual intervention, generating density-modified electron-density maps (Terwilliger, 2000) and corresponding structural models that accurately interpreted the maps.

To evaluate the accuracy of our results, we calculated the map–model correlation (CC-overall) between density-modified electron-density maps generated by the refinement and model building module and models in PDB depositions. Using a conservative minimum map–model correlation threshold of 0.5 (Oeffner *et al.*, 2013; Terwilliger *et al.*, 2023), and ensuring correct space group and cell parameters as prerequisites, we determined that 171 out of 186 datasets (approximately 92%) were successfully analyzed. The remaining 15 datasets, including the seven unsuccessful in data reduction and the eight that failed in MR, were classified as unsuccessful. Of particular note, seven datasets in the collection were originally solved using SAD methods as reported in the PDB. Remarkably, we were able to solve all of these cases using the MR approach within our pipeline, demonstrating its flexibility and capability in addressing challenging cases.

CC-overall was calculated using the *Phenix get_cc_mtz_pdb* tool for the 171 successfully processed datasets. In cases where multiple structural models were generated, the model with the highest CC-overall value was selected for analysis and is the one presented in Figs. 4(*a*)–4(*c*). However, we recommend that users carefully examine all available models to identify the most suitable starting point for further refinement, as alternative models may offer advantages depending on the specific context of downstream applications.

Fig. 4(*a*) shows the distribution of map–model correlation values CC-overall. For the 171 successfully processed datasets, the map–model correlation values ranged from 0.543 to 0.896, with a mean of 0.812 and a median of 0.828. Over 75% of the datasets achieved correlation values above 0.862, with the majority clustering between 0.8 and 0.9. These results demonstrate strong agreement between the density maps and the rebuilt models, highlighting the accuracy of the pipeline in reconstructing structures.

Fig. 4(*b*) illustrates the relationship between $R_{\rm meas}$ (overall) and CC-overall, serving as an indicator of how data quality correlates with map–model correlation. A slight negative correlation is observed: higher $R_{\rm meas}$ values tend to be associated with lower CC-overall scores. However, this trend is not particularly strong, suggesting that $R_{\rm meas}$ alone does not fully determine the quality of the resulting electron density map. We carefully examined the data points located in the lower-left corner of the plot, where $R_{\rm meas}$ is relatively low but CC-overall is also unexpectedly low. Among these datasets, we identified several contributing factors, including low resolution (worse than 3.0 Å), an MR search model with an r.m.s.d. greater than

1.5 Å or a low MR TFZ value (<10). These findings suggest that, while R_{meas} provides useful insight into data quality, additional factors such as resolution and MR model quality also play critical roles in determining map correlation and should be considered in comprehensive data assessment.

Fig. 4(*c*) illustrates the relationship between resolution and CC-overall. A general trend is observed: higher-resolution data tend to produce higher CC-overall scores, indicating better model-to-map correlation. Most datasets with resolution better than approximately 2.5 Å exhibit CC-overall values above 0.8, while datasets with resolution worse than 3.0 Å often show a notable drop in CC-overall, with some values falling below 0.7. Although this negative correlation is evident, the spread of CC-overall values at similar resolution levels suggests that resolution is not the sole determinant of map quality. Other factors—such as the quality of the MR model and data completeness—likely contribute to the variation in map correlation observed across datasets.

Fig. 4(*d*) illustrates structural completeness, defined as the percentage of Cα atoms in the deposited models that align within 2 Å of those in the models reconstructed by *AutoPD* (Terwilliger *et al.*, 2023). Space-group symmetry was used to include all related copies of chains in the comparison, and the completeness was calculated with *phenix.chain_comparison*. The completeness values were strongly clustered near 100%, with a mean of 95.45% and a median of 98.7%. Over 75% of the datasets achieved a completeness greater than 99.6%, demonstrating the pipeline's ability to produce near-complete structural data. While the minimum completeness observed was 27.4%, the overwhelming majority of datasets fell within the high-completeness range, as indicated by the sharp peak near 100% in the histogram. These findings underscore the pipeline's capability of generating comprehensive crystallographic models.

Fig. 4(*e*) depicts the root-mean-square deviation (r.m.s.d.) values for Cα atoms, calculated using *GESAMT* (Krissinel, 2012) in *CCP4i2*, comparing the coordinates of rebuilt models with those in the deposited structures. The r.m.s.d. values range from 0.059 to 1.433, with a mean of 0.363 and a median of 0.309. Approximately 75% of the datasets had r.m.s.d. values below 0.492, with the majority falling under 1.0. This distribution indicates strong structural agreement between the reconstructed models and the deposited structures, further highlighting the pipeline's accuracy in reproducing detailed structural features.

Fig. 4(*f*) presents R_{work} and R_{free} values derived from the structural models generated by the pipeline. The R_{work} values range from 0.187 to 0.389, with a mean of 0.241 and a median of 0.234. The R_{free} values range from 0.203 to 0.488, with a mean of 0.287 and a median of 0.282. Over 75% of the datasets exhibited R_{work} and R_{free} values below 0.257 and 0.311, respectively, demonstrating strong refinement quality. These results highlight the pipeline's ability to generate structural models with reliable refinement metrics and further validate its robustness in achieving high-quality results.

Fig. 4 thus demonstrates the robustness and reliability of our structural determination pipeline. The high map–model correlation, strong structural completeness, tight r.m.s.d. distributions and low *R* values collectively validate the effectiveness of the pipeline. These results support its applicability

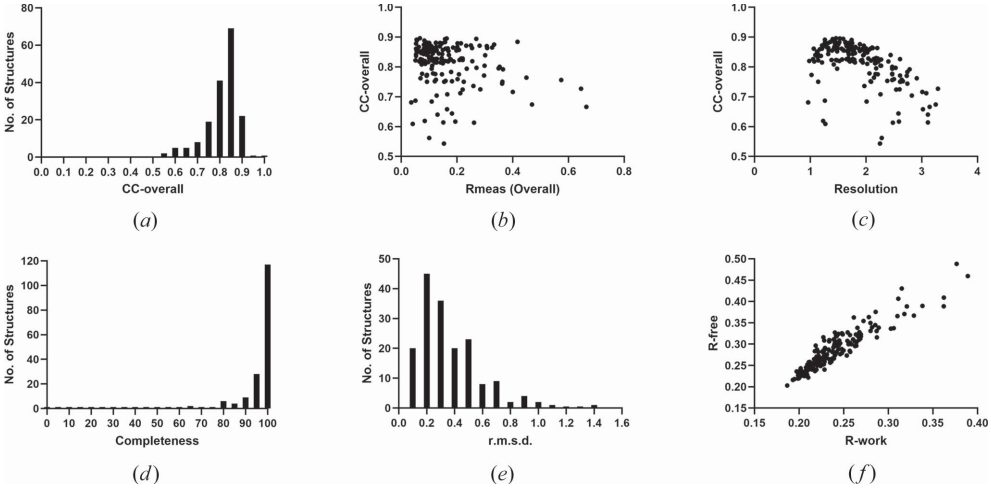


Figure 4
The results of structure redeterminations using *AutoPD* across 171 successful datasets. (*a*) Distribution of map–model correlation values (CC-overall). The majority of datasets exhibit high CC-overall values, indicating strong agreement between the electron density maps and the deposited models. (*b*) Scatter plot of R_{meas} (overall) versus CC-overall. A slight negative correlation is observed, where higher R_{meas} values are generally associated with lower CC-overall scores. However, this trend is weak, and several low-R_{meas} cases still show poor CC-overall due to factors such as low resolution, suboptimal MR models or low-TFZ scores. (*c*) Scatter plot of resolution versus CC-overall. A general trend is observed in which higher-resolution data tend to yield higher CC-overall values, though variation at similar resolution levels suggests that other factors also influence map quality. (*d*) Distribution of structural completeness. Completeness values are strongly skewed toward the high end, demonstrating the pipeline's ability to reconstruct nearly complete models. (*e*) Histogram of r.m.s.d. values for Cα atoms between the rebuilt and deposited models. Most datasets show low r.m.s.d. values, indicating strong structural agreement and precise reproduction of atomic coordinates. (*f*) Scatter plot of R_{work} versus R_{free} values from refined models. A strong positive correlation is observed, with most data points falling within the expected range, reflecting good refinement quality and minimal overfitting. These results collectively demonstrate the robustness, accuracy and completeness of *AutoPD* in fully automated macromolecular structure determination.

as a reliable tool for advancing structural determination in macromolecular crystallography.

5.3. Results of data reduction module

During the data reduction phase, encompassing 186 datasets, seven datasets produced solutions with incorrect point group determinations or unit-cell parameters, leading to erroneous space group or unit-cell parameter assignments in subsequent MR steps. As a result, the overall success rate for this module was 96%.

The data reduction phase employed four distinct programs: *XDS*, *xia2-3d/3dii*, *xia2-dials* and *autoPROC*, with success rates—defined as their ability to generate crystallographic files with correct point group and unit-cell parameter determinations—of 91%, 85%, 84% and 94%, respectively. Of the 171 successful cases, the crystallographic files leading to the best structural models originated from *XDS* (40%), *xia2-3d/3dii* (25%), *xia2-dials* (14%) and *autoPROC* (21%), highlighting the complementary roles of these programs in achieving reliable data reduction.

Fig. 5(*a*) illustrates the distribution of R_{meas} values for the crystallographic files generated during the data reduction process. The R_{meas} values ranged from 0.033 to 0.63, with a mean of 0.149 and a median of 0.116. Over 75% of the datasets exhibited R_{meas} values below 0.187, indicating that the majority of datasets had low redundancy-independent errors. This distribution underscores the effectiveness of *AutoPD* in minimizing errors during data processing, providing high-quality input for downstream analysis.

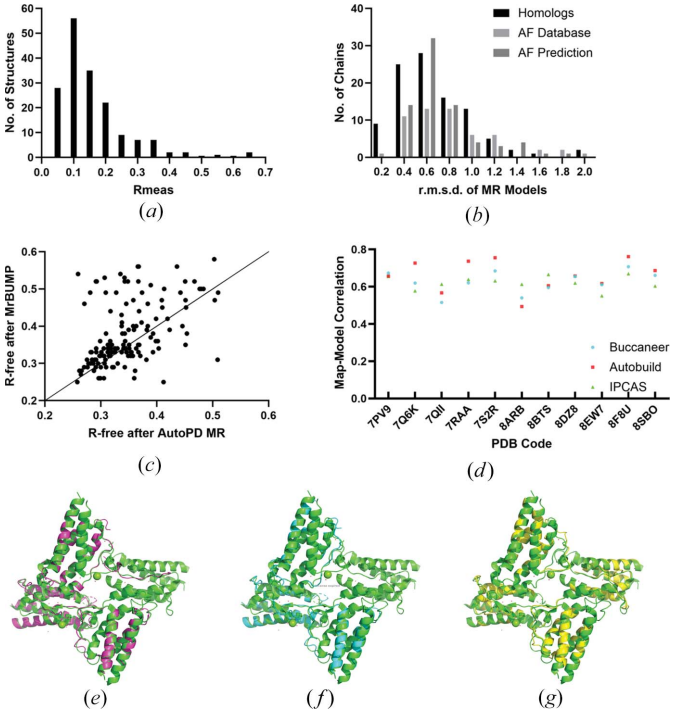


Figure 5
Results across the various modules of *AutoPD*. (*a*) Distribution of R_{meas} values generated by the data reduction module, highlighting data quality for crystallographic files produced during the data reduction process. (*b*) r.m.s.d. values between MR search models and their corresponding deposited PDB structures, stratified by the source of the search models: homologs from the PDB (black bars), *AlphaFold* database models (light gray) and *AlphaFold de novo* predictions (dark gray). (*c*) Comparison of R_{work} values after MR using the *AutoPD* MR module versus *MrBUMP*, with points above the diagonal indicating superior performance by *AutoPD*. (*d*) Map–model correlation coefficients from the refinement and model building module, comparing *Buccaneer* (blue dots), *Phenix Autobuild* (red squares) and *IPCAS* (green triangles) across datasets for which *IPCAS* is triggered. (*e*)–(*g*) Superimposed models of PDB entry 7raa, with the deposited structure in green and rebuilt models from *Buccaneer* (magenta), *Phenix Autobuild* (cyan) and *IPCAS* (yellow). *IPCAS* demonstrates superior structural alignment, successfully resolving regions where *Buccaneer* and *Phenix Autobuild* exhibit significant gaps. These outcomes collectively demonstrate the robust performance and versatility of *AutoPD* across all stages of macromolecular structure determination.

5.4. Results of search model generation module

Within this component of our pipeline, the initial step employs *MrParse* to identify potential homologs and *AlphaFold* models from existing databases. In cases where no *AlphaFold* database models are available, or the top-ranking model shows a sequence identity below 0.9, covers less than 60% of the target sequence or has a pLDDT score lower than 90, a *de novo AlphaFold* prediction is automatically initiated to produce a more accurate and complete structural model specific to the input sequence. In our analysis of 171 successful cases, many included more than one unique protein chain. After excluding chains too short for *AlphaFold* prediction, we identified 229 unique chains across these cases.

This dual-strategy approach effectively integrates both database-derived and prediction-based models to enhance MR. Among the 229 models that contributed to the 171 final structures, 55 were retrieved from the *AlphaFold* database, 100 homologous models were sourced from the PDB and 74 were generated via *de novo AlphaFold* predictions. This comprehensive methodology maximizes the likelihood of identifying accurate and reliable search models, providing robust inputs for successful structure determination.

Fig. 5(*b*) illustrates the r.m.s.d. values between MR search models and their corresponding deposited PDB structures, stratified by the source of the search models: homologs from the PDB (black bars), *AlphaFold* (AF) database models (light gray) and *AlphaFold de novo* predictions (dark gray). Homolog-based models tend to exhibit lower r.m.s.d. values, with a strong concentration between 0.4 and 0.7 Å, indicating close structural similarity to the final structures. *AlphaFold* database models show a slightly broader distribution, with more chains in the range 0.6–1.0 Å, and a modest number extending beyond 1.2 Å. *AlphaFold* predictions exhibit the widest spread, with a notable number of models showing r.m.s.d. values above 1.0 Å, reflecting greater structural variation, but still maintain a substantial portion below 1.0 Å. These results highlight the overall reliability of all three model sources for MR, while also illustrating that homologs generally offer the closest structural match when available. However, *AlphaFold*-based models—both from the database and *de*

novo predictions—provide valuable alternatives, especially when high-quality homologs are lacking.

5.5. Results of molecular replacement module

Among the 179 cases that successfully passed the data reduction stage, four cases failed to yield MR solutions, while four cases were assigned incorrect space groups compared with their deposited models. This resulted in a success rate of 96% for the MR module in the *AutoPD* pipeline. For comparison, MR was also performed using *MrBUMP* with the same crystallographic files and sequence files, yielding a lower success rate of 78%.

Fig. 5(*c*) compares the $R_{\rm free}$ values obtained after MR using *MrBUMP* versus those from the *AutoPD* MR module. Points above the diagonal line represent cases where the *AutoPD* MR module achieved lower $R_{\rm free}$ values, while points below the line indicate instances where *MrBUMP* performed better. The majority of points lie above the diagonal, demonstrating that the *AutoPD* MR module consistently outperformed *MrBUMP* in terms of $R_{\rm free}$ values.

These results highlight the effectiveness of the *AutoPD* MR module in generating accurate initial models for refinement, surpassing the performance of *MrBUMP* in a significant number of cases. The superior $R_{\rm free}$ values underscore the robustness and reliability of the *AutoPD* MR module, validating the strategic integration of database-derived and predictive models to optimize MR. This approach ensures both higher accuracy and greater efficiency in the structure determination process.

5.6. Results of refinement and model building module

In the refinement and model building stage of our pipeline, *Buccaneer* is initially employed due to its rapid processing capabilities, providing a quick baseline for model quality. However, if the output from *Buccaneer* does not meet the predefined quality standard—indicated by an $R_{\rm free}$ value exceeding the default threshold of 0.35—additional refinement is performed using *Phenix Autobuild* and *IPCAS* to improve the model further.

Fig. 5(*d*) presents the map–model correlation (CC-overall) for a set of datasets for which *IPCAS* was triggered, comparing the performance of three model building tools: *Buccaneer* (cyan circles), *Phenix Autobuild* (red squares) and *IPCAS* (green triangles). Each point represents the CC-overall value for a model built by the respective tool for a given PDB entry. The results indicate that all three tools contribute to model building across different datasets, with *Phenix Autobuild* generally yielding high CC-overall values. *IPCAS* also contributes meaningfully in several cases, either matching or exceeding the CC-overall achieved by the other tools. For certain datasets, such as 7qii, 8arb and 8ew7, *IPCAS* provides the highest correlation among the three, highlighting its role in complementing existing model building methods, particularly in cases where initial models may be suboptimal. This comparison illustrates that using multiple model building strategies can improve the robustness of the overall pipeline by providing alternative solutions when standard tools face limitations.

To illustrate the performance of *IPCAS*, we analyzed a specific protein structure (PDB entry 7raa) (Bejger *et al.*, 2021), which contains one unique chain, represented by four copies within a single asymmetric unit. The $R_{\rm free}$ values achieved by *Buccaneer*, *Phenix Autobuild* and *IPCAS* for this structure were 0.4718, 0.4881 and 0.4056, respectively, highlighting the superior accuracy of *IPCAS*.

Figs. 5(*e*), 5(*f*) and 5(*g*) depict the superimposition of the PDB-deposited structure with models reconstructed by *Buccaneer*, *Phenix Autobuild* and *IPCAS*, respectively. These comparisons clearly showcase the superior performance of *IPCAS*. The *IPCAS*-reconstructed model exhibits better structural alignment with the reference structure, showing fewer deviations and greater accuracy in backbone tracing. Notably, in the top-right region of the structure, the *Buccaneer* and *Phenix Autobuild* models exhibit significant structural gaps, whereas *IPCAS* successfully reconstructs these regions without any missing elements. This highlights the robustness of *IPCAS* in addressing difficult regions where other tools fail.

A key advantage of *IPCAS* is its ability to address incomplete or ambiguous regions in the initial model. By leveraging direct-method phasing and iterative dual-space refinement, *IPCAS* excels in extending and refining partial structures, resulting in more complete and accurate final models. This iterative refinement strategy not only fills in missing regions but also improves overall agreement with the deposited structures, as seen in the overlays. These results underscore the ability of *IPCAS* to deliver high-quality models, even for challenging datasets, and its value as a critical tool in crystallographic structure determination.

This analysis underscores a fundamental balance in computational structural biology: the trade-off between rapid processing and meticulous precision. The necessity for a layered approach in model building is evident—where faster algorithms such as *Buccaneer* provide an initial approximation, which can then be refined through more computationally intensive tools like *IPCAS* to achieve higher accuracy. This strategic combination ensures both efficiency and reliability in solving even the most difficult datasets.

5.7. Results of experimental phasing module

Among the 186 datasets, strong anomalous signals were detected in 11 cases. Of these, seven were successfully solved using *AutoPD*, while three of the four failures were attributed to issues in the data reduction module, resulting in a success rate of 87.5% for datasets that proceeded beyond this stage. Notably, three of the successful SAD solutions exhibited higher map–model correlation values than those obtained via MR, underscoring the robustness and effectiveness of SAD-based phasing in suitable cases.

For the successful datasets, the mean $R_{\rm work}$ value was 0.224, indicating reliable refinement quality. The map–model correlation values were consistently high, reflecting strong agreement between the calculated and experimental data. The

r.m.s.d. values were generally low, signifying good structural alignment with the experimental data. Additionally, the completeness for these datasets exceeded 90% in most cases, underscoring the pipeline's capability to generate highly complete models.

These results highlight the effectiveness of the pipeline in solving SAD datasets while also identifying opportunities for further optimization to address the limitations observed in the two unsuccessful cases. Overall, the findings reinforce the reliability and adaptability of the pipeline in leveraging SAD data for accurate structure determination.

6. Discussion

Our study presents *AutoPD*, an open-source high-throughput fully automated data processing and structure determination meta-pipeline, specifically designed to address the challenges introduced by advancements in synchrotron light sources and automated data acquisition technologies. Developed for the macromolecular crystallography beamline at HEPS in Beijing and available for broader use, *AutoPD* represents a transformative step forward, seamlessly integrating processes from data reduction to model building in macromolecular crystallography.

The deployment of *AutoPD* on a dataset of 186 protein structures from the PDB, all determined after the training cut-off date for *AlphaFold2*, achieved a remarkable 93% success rate in generating structural models. This high success rate underscores the robustness and efficiency of the pipeline, as demonstrated by its ability to produce structural models with high map–model correlation values and near-complete reconstruction of atomic structures. By integrating cutting-edge AI tools with traditional crystallographic software in a parallel and tiered workflow, *AutoPD* effectively tackles the complexities of modern structure determination.

The individual modules of *AutoPD* work in concert to optimize speed and accuracy. The data reduction module delivered a 96% success rate, with $R_{\rm merge}$ values reflecting the quality of the processed data. The search model generation module capitalizes on a dual-strategy approach, combining database-derived models with AI-based predictions. This hybrid methodology leverages the strengths of both established and predictive models, significantly improving the likelihood of successful MR. The MR module further reinforces this success, achieving a 97% success rate in contrast to 78% for *MrBUMP*, with superior $R_{\rm work}$ values that underscore its accuracy and reliability.

In the refinement and model building module, *AutoPD* demonstrates the importance of a layered approach. The rapid initial modeling provided by *Buccaneer* is complemented by the advanced precision of *Phenix Autobuild* and *IPCAS*. Notably, *IPCAS* consistently outperforms other tools in challenging cases, delivering superior model quality with higher map–model correlation coefficients and better structural completeness. For example, *IPCAS* successfully resolved regions where *Buccaneer* and *Autobuild* failed. This iterative and robust methodology highlights the capability of *AutoPD* to handle even the most difficult datasets, ensuring accurate and reliable final models.

Overall, the pipeline failed in 15 cases. After examining the *AutoPD* output and consulting PDB deposition information, we found that four of these cases could be resolved with a second run by applying specific input parameters. These include:

(*a*) Blank diffraction images (PDB entry 8sc0). All four pipelines in the data reduction module failed due to the first 90 diffraction images being blank, as indicated by the log files from *xia2-dials* and *autoPROC*. When the problematic images are excluded using the parameters `image_start =` and `image_end =`, *AutoPD* was able to complete successfully, yielding a final model with a CC-overall of 0.877.

(*b*) Missing search model (PDB entry 8u0g). The initial MR solution had a TFZ value of 6.3, suggesting an incorrect placement. Upon investigation, it was found that the search model used by the original authors (PDB entry 7udi) had been missed. Providing 7udi explicitly as a search model allowed the pipeline to generate a correct solution with a CC-overall greater than 0.5. Notably, other tools such as *MrBUMP* also failed to identify this model.

(*c*) PAE-based model splitting (PDB entry 7r3w). The initial MR solution yielded a TFZ value of 6.1 despite using an *AlphaFold* model (AF-Q8ZM00-F1) with a high pLDDT of 90.8. Analysis of the PAE plot indicated that the model consisted of two domains with potentially incorrect relative orientations. By enabling `pae_split = true`, which splits the model on the basis of the PAE matrix, the TFZ improved to 20.8 and a valid model was obtained with CC-overall above 0.5.

(*d*) Incorrect unit-cell parameter (PDB entry 7qsg). The initial run produced a structure with an $R_{\rm free}$ of 0.3411 and a CC-overall of 0.495. Analysis via *Phenix.xtriage* revealed a strong non-origin Patterson peak, suggesting a translation by half the unit cell along the *c* axis. Adjusting the *c* dimension accordingly resulted in a correct structure with improved $R_{\rm free}$ (0.3108) and CC-overall (0.707).

The remaining failures highlight limitations of current automation and the need for further manual intervention. These include large and complex assemblies (*e.g.* 7qij), multi-domain or composite structures (*e.g.* 7z36), and cases with minimal unique features (*e.g.* 8swd). Additionally, complex crystallographic issues such as twinning, translational pseudo-symmetry or incorrect space-group assignment were identified in several cases (*e.g.* 7rox, 7tm4, 8cje, 8dil, 8duy). These findings underscore the current limitations of automation and the importance of expert analysis in resolving difficult cases. Nevertheless, they also provide valuable insights for future development of more robust decision-making strategies.

7. Conclusion

AutoPD is a groundbreaking tool in structural biology, offering an open-source, fully automated, high-throughput meta-pipeline that integrates state-of-the-art AI technologies with crystallographic best practices. An adaptive decision-

making strategy is embedded in the workflow to dynamically select the most suitable modeling pathway according to data quality and intermediate results, thereby improving robustness and ensuring higher success rates across diverse datasets. Its ability to efficiently produce high-quality structural models not only accelerates the process of structure determination but also ensures the reliability of the resulting models. As structural biology continues to advance, tools like *AutoPD* will be instrumental in unraveling the complexities of biological macromolecules, contributing significantly to scientific discovery and medical innovation. By bridging the gap between speed and precision, *AutoPD* sets a new standard for high-throughput structure determination in the era of modern synchrotron light sources.

8. Data and code availability

For our data testing, the input data, including both diffraction data and sequence files, were sourced directly from the PDB. The 186 PDB entries used were as follows: 8v4j, 8bty, 8cqm, 8u1e, 8u1j, 8u0g, 8owm, 8osw, 8dab, 8daa, 8da9, 8da8, 8da6, 8da5, 8da4, 8da3, 8bxt, 8aq8, 8d2z, 8t5t, 8cjd, 8bts, 8su6, 8sqq, 8sqo, 8snj, 7zpf, 8slh, 8slf, 8sld, 8fhj, 8e6h, 8e5s, 8skf, 8g2g, 8arc, 8arb, 8ara, 8sf3, 7uyi, 8sbx, 8sbv, 8sbo, 8sbn, 8sac, 8sa8, 8sa7, 7v0i, 8dz8, 8gca, 8fra, 8fg7, 8f8e, 8cip, 7pho, 7udi, 7u0o, 7pe4, 8g0v, 8g0u, 8g0t, 8g0s, 7pdo, 8fxq, 8fuy, 8bbu, 8ft7, 7r59, 7r3o, 7r3l, 7qsj, 7zb9, 7z3s, 7tsx, 7tsq, 7riz, 8fi4, 8fi3, 7wez, 8cx4, 8a19, 7n2s, 7n2q, 7n2o, 8f8u, 8ey5, 8ew7, 7s47, 7s46, 7qta, 8em8, 7z36, 7r0t, 7r0k, 8ad7, 8egm, 8ek7, 8egn, 7r3w, 7mdc, 7ywj, 8ees, 7t5w, 7t5v, 7t5u, 7qsa, 8dp2, 7u0u, 7u0t, 8dqb, 8dq9, 8dos, 8dor, 8doq, 7qy6, 7nzz, 7fbq, 5soi, 7xc0, 7vid, 7z0r, 8d1x, 8cso, 7v0h, 7qnp, 7q6k, 7q6j, 7qii, 7qih, 7qgf, 7kmj, 7wda, 7wcj, 7uv5, 7s5b, 7pox, 7ulz, 7unn, 7n3t, 7ulh, 7t93, 7r7j, 7mcj, 7s2s, 7s2r, 7ph1, 7raa, 7ra9, 7u5y, 7u5q, 7u5f, 7u56, 7u4h, 7u35, 7qiq, 7tmv, 7tmf, 7tmd, 7tmb, 7tm9, 7tm8, 7tm7, 7tm5, 7ti7, 7pv9, 7pv8, 7f8s, 7f8s, 7f8r, 7tcm, 7bbs, 7rjz, 7rji, 8ent, 8cje, 8swd, 8sng, 8sc0, 7qsg, 7n2p, 7rox, 8duy, 8dop, 8dil, 7qij, 7tm4. To ensure transparency and facilitate reproducibility, the entire collection of output folders, along with a detailed spreadsheet cataloging the raw data and analyses conducted, is accessible at https://docs.google.com/spreadsheets/d/1Xx_4GIsbQ3dc4lGlCgC7WsR7O0Kk5Wo/edit?usp=drive_link&ouid=103695451649651655457&rtpof = true&sd = true.

Additionally, the complete codebase for *AutoPD* is openly available on *GitHub*, offering a valuable resource for further development, review and adaptation in related projects. Interested parties can review and download the code at https://github.com/zhangxinhku/AutoPD, This open-access approach underscores our commitment to advancing research in structural biology through collaboration and shared resources.

Acknowledgements

We extend our gratitude to all authors who generously provided their raw diffraction data for the PDB entries utilized in this study.

Funding information

The following funding is acknowledged: National Natural Science Foundation of China (grant No. 32371280); Guangdong Major Project of Basic and Applied Basic Research (grant No. 2023B0303000003).

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* D**66**, 213–221.

Agirre, J., Atanasova, M., Bagdonas, H., Ballard, C. B., Baslé, A., Beilsten-Edmands, J., Borges, R. J., Brown, D. G., Burgos-Mármol, J. J., Berrisford, J. M., Bond, P. S., Caballero, I., Catapano, L., Chojnowski, G., Cook, A. G., Cowtan, K. D., Croll, T. I., Debreczeni, J. É., Devenish, N. E., Dodson, E. J., Drevon, T. R., Emsley, P., Evans, G., Evans, P. R., Fando, M., Foadi, J., Fuentes-Montero, L., Garman, E. F., Gerstel, M., Gildea, R. J., Hatti, K., Hekkelman, M. L., Heuser, P., Hoh, S. W., Hough, M. A., Jenkins, H. T., Jiménez, E., Joosten, R. P., Keegan, R. M., Keep, N., Krissinel, E. B., Kolenko, P., Kovalevskiy, O., Lamzin, V. S., Lawson, D. M., Lebedev, A. A., Leslie, A. G. W., Lohkamp, B., Long, F., Malý, M., McCoy, A. J., McNicholas, S. J., Medina, A., Millán, C., Murray, J. W., Murshudov, G. N., Nicholls, R. A., Noble, M. E. M., Oeffner, R., Pannu, N. S., Parkhurst, J. M., Pearce, N., Pereira, J., Perrakis, A., Powell, H. R., Read, R. J., Rigden, D. J., Rochira, W., Sammito, M., Sánchez Rodríguez, F., Sheldrick, G. M., Shelley, K. L., Simkovic, F., Simpkin, A. J., Skubak, P., Sobolev, E., Steiner, R. A., Stevenson, K., Tews, I., Thomas, J. M. H., Thorn, A., Valls, J. T., Uski, V., Usón, I., Vagin, A., Velankar, S., Vollmar, M., Walden, H., Waterman, D., Wilson, K. S., Winn, M. D., Winter, G., Wojdyr, M. & Yamashita, K. (2023). *Acta Cryst.* D**79**, 449–461.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. (2021). *Science* **373**, 871–876.

Bejger, M., Fortuna, P., Drewniak-Świtalska, M., Plewka, J., Rypniewski, W. & Berlicki, Ł. (2021). *Chem. Commun.* **57**, 6015–6018.

Bernstein, H. J., Förster, A., Bhowmick, A., Brewster, A. S., Brockhauser, S., Gelisio, L., Hall, D. R., Leonarski, F., Mariani, V., Santoni, G., Vonrhein, C. & Winter, G. (2020). *IUCrJ* **7**, 784–792.

Casanas, A., Warshamanage, R., Finke, A. D., Panepucci, E., Olieric, V., Nöll, A., Tampé, R., Brandstetter, S., Förster, A., Mueller, M., Schulze-Briese, C., Bunk, O. & Wang, M. (2016). *Acta Cryst.* D**72**, 1036–1048.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Delagenière, S., Brenchereau, P., Launer, L., Ashton, A. W., Leal, R., Veyrier, S., Gabadinho, J., Gordon, E. J., Jones, S. D., Levik, K. E., McSweeney, S. M., Monaco, S., Nanao, M., Spruce, D., Svensson, O., Walsh, M. A. & Leonard, G. A. (2011). *Bioinformatics* **27**, 3186–3192.

Ding, W., Zhang, T., He, Y., Wang, J., Wu, L., Han, P., Zheng, C., Gu, Y., Zeng, L., Hao, Q. & Fan, H. (2020). *J. Appl. Cryst.* **53**, 253–261.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.

Fisher, S. J., Levik, K. E., Williams, M. A., Ashton, A. W. & McAuley, K. E. (2015). *J. Appl. Cryst.* **48**, 927–932.

Hao, Q., Gu, Y. X., Zheng, C. D. & Fan, H. F. (2000). *J. Appl. Cryst.* **33**, 980–981.

Henrich, B., Bergamaschi, A., Broennimann, C., Dinapoli, R., Eikenberry, E., Johnson, I., Kobas, M., Kraft, P., Mozzanica, A. & Schmitt, B. (2009). *Nucl. Instrum. Methods Phys. Res. A* **607**, 247–249.
Hu, Q., Wang, L., Zheng, W. & Jiang, X. (2022). *Proceedings of International Symposium on Grids & Clouds 2022*, 011. https://doi.org/10.22323/1.415.0011.
Hu, Q., Zheng, W., Jiang, X. & Shi, J. (2021*a*). *Proceedings of International Symposium on Grids & Clouds 2021*, 021. https://doi.org/10.22323/1.378.0021.
Hu, Y., Li, L., Tian, H., Liu, Z., Huang, Q., Zhang, Y., Hu, H. & Qi, F. (2021*b*). *EPJ Web Conf.* **251**, 04020.
Jiao, Y., Xu, G., Cui, X.-H., Duan, Z., Guo, Y.-Y., He, P., Ji, D.-H., Li, J.-Y., Li, X.-Y., Meng, C., Peng, Y.-M., Tian, S.-K., Wang, J.-Q., Wang, N., Wei, Y.-Y., Xu, H.-S., Yan, F., Yu, C.-H., Zhao, Y.-L. & Qin, Q. (2018). *J. Synchrotron Rad.* **25**, 1611–1618.
Johnson, I., Bergamaschi, A., Billich, H., Cartier, S., Dinapoli, R., Greiffenberg, D., Guizar-Sicairos, M., Henrich, B., Jungmann, J., Mezza, D., Mozzanica, A., Schmitt, B., Shi, X. & Tinti, G. (2014). *J. Instrum.* **9**, C05032.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature* **596**, 583–589.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Keegan, R. M., McNicholas, S. J., Thomas, J. M. H., Simpkin, A. J., Simkovic, F., Uski, V., Ballard, C. C., Winn, M. D., Wilson, K. S. & Rigden, D. J. (2018). *Acta Cryst.* D**74**, 167–182.
Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 119–124.
Krissinel, E. (2012). *J. Mol. Biochem.* **1**, 76–85.
Li, Z., Fan, H. & Ding, W. (2024). *IUCrJ* **11**, 152–167.
Mariani, V., Biasini, M., Barbato, A. & Schwede, T. (2013). *Bioinformatics* **29**, 2722–2728.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
McCoy, A. J., Sammito, M. D. & Read, R. J. (2022). *Acta Cryst.* D**78**, 1–13.
McKinney, W. (2010). *SciPy 2010: proceedings of the 9th Python in science conference*, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-012.
Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. (2022). *Nat. Methods* **19**, 679–682.
Moore, P. B., Hendrickson, W. A., Henderson, R. & Brunger, A. T. (2022). *Science* **375**, 507.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Oeffner, R. D., Bunkóczi, G., McCoy, A. J. & Read, R. J. (2013). *Acta Cryst.* D**69**, 2209–2215.
Oeffner, R. D., Croll, T. I., Millán, C., Poon, B. K., Schlicksup, C. J., Read, R. J. & Terwilliger, T. C. (2022). *Acta Cryst.* D**78**, 1303–1314.
Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.
Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2009). *Acta Cryst.* D**65**, 1089–1097.
Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P. R., Jenkins, H. T., Keegan, R., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S. J., Nicholls, R. A., Noble, M., Pannu, N. S., Roth, C., Sheldrick, G., Skubak, P., Turkenburg, J., Uski, V., von Delft, F., Waterman, D., Wilson, K., Winn, M. & Wojdyr, M. (2018). *Acta Cryst.* D**74**, 68–84.
Sadybekov, A. A., Sadybekov, A. V., Liu, Y., Iliopoulos-Tsoutsouvas, C., Huang, X.-P., Pickett, J., Houser, B., Patel, N., Tran, N. K., Tong, F., Zvonok, N., Jain, M. K., Savych, O., Radchenko, D. S., Nikas, S. P., Petasis, N. A., Moroz, Y. S., Roth, B. L., Makriyannis, A. & Katritch, V. (2022). *Nature* **601**, 452–459.
Shao, C., Bittrich, S., Wang, S. & Burley, S. K. (2022). *Structure* **30**, 1385–1394.e3.
Simpkin, A. J., Thomas, J. M. H., Keegan, R. M. & Rigden, D. J. (2022). *Acta Cryst.* D**78**, 553–559.
Skubák, P. & Pannu, N. S. (2013). *Nat. Commun.* **4**, 2777.
Tange, O. (2021). *GNU Parallel 20210822 ('Kabul')*, https://doi.org/10.5281/zenodo.5233953.
Tao, Z., Yuan-Xin, G., Chao-De, Z. & Hai-Fu, F. (2010). *Chin. Phys. B* **19**, 086103.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C., Afonine, P. V., Liebschner, D., Croll, T. I., McCoy, A. J., Oeffner, R. D., Williams, C. J., Poon, B. K., Richardson, J. S., Read, R. J. & Adams, P. D. (2023). *Acta Cryst.* D**79**, 234–244.
Terwilliger, T. C., Liebschner, D., Croll, T. I., Williams, C. J., McCoy, A. J., Poon, B. K., Afonine, P. V., Oeffner, R. D., Richardson, J. S., Read, R. J. & Adams, P. D. (2024). *Nat. Methods* **21**, 110–116.
Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.
Wojdyla, J. A., Kaminski, J. W., Panepucci, E., Ebner, S., Wang, X., Gabadinho, J. & Wang, M. (2018). *J. Synchrotron Rad.* **25**, 293–303.
Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.
Winter, G., Beilsten–Edmands, J., Devenish, N., Gerstel, M., Gildea, R. J., McDonagh, D., Pascal, E., Waterman, D. G., Williams, B. H. & Evans, G. (2022). *Protein Sci.* **31**, 232–250.
Winter, G., Lobley, C. M. C. & Prince, S. M. (2013). *Acta Cryst.* D**69**, 1260–1273.
Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst.* D**74**, 85–97.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Cryst.* D**67**, 235–242.
Yamashita, K., Wojdyr, M., Long, F., Nicholls, R. A. & Murshudov, G. N. (2023). *Acta Cryst.* D**79**, 368–373.
Yu, F., Wang, Q., Li, M., Zhou, H., Liu, K., Zhang, K., Wang, Z., Xu, Q., Xu, C., Pan, Q. & He, J. (2019). *J. Appl. Cryst.* **52**, 472–477.