

# REdiii: a pipeline for automated structure solution

Markus-Frederik Bohn and Celia A. Schiffer\*

Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, MA 01605, USA. \*Correspondence e-mail: celia.schiffer@umassmed.edu

Received 10 June 2014

Accepted 12 February 2015

Edited by G. J. Kleywegt, EMBL–EBI, Hinxton, England

**Keywords:** REdiii; automated structure solution.

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

High-throughput crystallographic approaches require integrated software solutions to minimize the need for manual effort. *REdiii* is a system that allows fully automated crystallographic structure solution by integrating existing crystallographic software into an adaptive and partly autonomous workflow engine. The program can be initiated after collecting the first frame of diffraction data and is able to perform processing, molecular-replacement phasing, chain tracing, ligand fitting and refinement without further user intervention. Preset values for each software component allow efficient progress with high-quality data and known parameters. The adaptive workflow engine can determine whether some parameters require modifications and choose alternative software strategies in case the preconfigured solution is inadequate. This integrated pipeline is targeted at providing a comprehensive and efficient approach to screening for ligand-bound co-crystal structures while minimizing repetitiveness and allowing a high-throughput scientific discovery process.

## 1. Introduction

### 1.1. Automation in bio-crystallography

Automation of processes in biological research aids in accelerating the scientific discovery process and allows the efficient allocation of resources. Macromolecular crystallography, in particular, has seen tremendous and rapid advances in method automation from the use of robotics in experimental setup and handling to image analysis while monitoring crystal growth and data processing and model-building solutions.

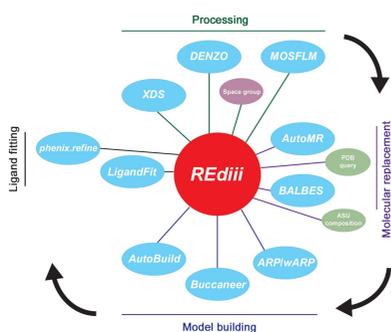
### 1.2. High-throughput crystallography during drug discovery

Especially during the drug-discovery process, crystallography is an essential tool to gain insights into the structural characteristics of drug candidates, their targets and the determinants for effective interactions between them. Positional information with near-atomic resolution has proven to be invaluable for developing specific and reliable therapeutics and has led to modern structure-based drug design (Amzel, 1998).

In cases where obtaining well diffracting crystals of the target protein is no longer the rate-limiting step, but efforts are directed at screening libraries of known binders for co-crystal structures, solving up to hundreds of crystallographic data sets becomes a necessity. Excellent software solutions for data processing, phasing, model building and ligand fitting are available, but they do not readily fall into one comprehensive pipeline.

### 1.3. Current state of automated pipelines

There are existing solutions that attempt to combine the different steps and solve parts of the crystallographic



© 2015 International Union of Crystallography

structure-determination pipeline once data have been collected, including *ACrs* (Brunzelle *et al.*, 2003), *ELVES* (Holton & Alber, 2004) and *SGXpro* (Fu *et al.*, 2005). Similar proprietary solutions developed for structural genomics consortia (Rupp *et al.*, 2002) and *Auto-Rickshaw* (Panjikar *et al.*, 2005) were designed for experimental phasing but also allow molecular replacement. The *autoSHARP* software (Vonrhein *et al.*, 2007) allows direct phasing starting from merged data followed by automatic model building by calling *ARP/wARP* (Langer *et al.*, 2008). Also, the widely used *HKL-3000* (Minor *et al.*, 2006) automates individual steps of structure solution and provides extensive feedback and user control. However, a fully automated pipeline is required for high-throughput crystallography. A recent development, *phenix.ligand\_pipeline* (Echols *et al.*, 2014), allows molecular replacement phasing of merged data followed by a refinement and/or model-building round with *AutoBuild* allowing subsequent ligand placement using *phenix.ligandfit*, all in an automated manner without requiring additional user input during the process.

Although all of these approaches combine existing software into a pipeline with reduced user intervention, and the term ‘automated structure-determination pipeline’ has been used to describe software that perform data manipulation and decision-making during the structure-solution process, most of these software solutions still require the user to learn how to interact with a new software package without knowing how well its design suits the question that the user is asking. The automated structure-determination pipeline described here will only require user intervention once, immediately after initialization of data collection, and will perform all required data manipulation from transfer of the first frame of diffraction data to refinement of ligand-bound structural models.

### 1.4. Necessity of fully automated structure solution

Co-crystallization trials often lead to an abundance of crystals composed of the unliganded protein, and only the complete process of diffracting the crystal, collecting an entire data set, finding a solution and subsequent refinement reveals the crystal composition. To computationally streamline this process, automated software can determine the status of data collection, prepare a search model of the protein, process the data, perform MR phasing, generate solvent and refine the solution. The crystallographer can then determine whether the co-crystallization trial was successful, eliminating the need to manually iterate through the entire procedure for many crystals. Software specifically designed for searching for a bound ligand, for example, can place a large emphasis on speed, with accuracy and completeness not inevitably as crucial (Dauter, 2010). For example, the *Dimple* software (Wojdyr *et al.*, 2014) allows the quick location of potential ligands by comparing maps derived from putatively ligand-bound structures with the corresponding apo structure. No fully automated solution can be adapted to all experimental needs, therefore certain drawbacks are inevitable since certain aspects of the data and model may be unknown at the

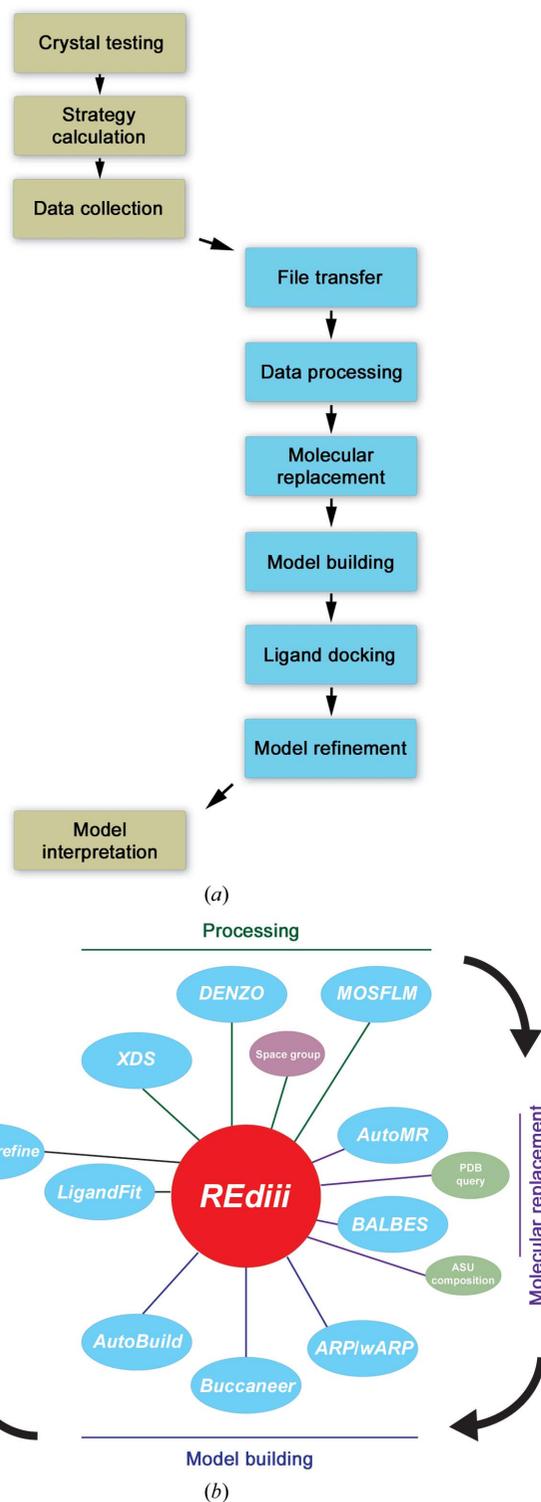
beginning of the experiment. *phenix.ligand\_pipeline* can produce ligand-bound solutions of spectacular quality in a very short time and even provides an automatically generated *Coot* session for convenient access to the results. This pipeline has to rely on the user to provide merged data and large differences between the search model and the target structure, such as domain movements that may accompany ligand binding, can lead to failure. With *REdiii*, we attempt to introduce steps of automated troubleshooting and generation of a software ‘memory’ to overcome some of the obstacles encountered by complete automation. The test cases presented here were all processed with minimal information and no user intervention. The lack of intervention can be especially useful when working with large amounts of similar data, but challenges remain. Sacrificing control by the user has significant drawbacks in validating achieved results, especially at intermediate steps. *REdiii* currently only incorporates tools for molecular-replacement phasing, can run into issues with highly mosaic data and does not check for twinned data. Twinned data can be difficult when determining whether a ligand can be fitted *versus* how many copies of the ligand are present. Since *REdiii* will be an open-source tool with simple code structures, significant improvements will happen over time.

### 1.5. Introducing an automated structure-solution pipeline

Our solution streamlines commonly used software tools, following well established methodologies, to mimic the process of manual data processing and structure solution to yield high-quality structural models without requiring additional manual manipulation. This approach can be used to screen data of modest resolution for protein–ligand complexes with reasonable results or to achieve very high-throughput and high-quality models with good data. This software solution can replace conventional user-guided processing and model building, which can be extremely useful in all scenarios where high throughput is needed. In addition to the requirements posed by this task, we aspired to design a workflow that would satisfy a few additional constraints for functional design: user input should only be required once, when crystallographic diffraction data acquisition is initiated. Besides raw data, any file containing computationally modified data (*e.g.* .mtz, .pdb and .fasta files) can either be generated by *REdiii* or be provided by the user, allowing the user to enter or leave the *REdiii* workflow at any given point and also permitting the user the flexibility to interface *REdiii* with other software if so desired. *REdiii* should not only perform tasks but also initiate automated troubleshooting of common errors, generating a memory of successful runs and learning project-specific parameters that have proven to be successful.

## 2. Methodology

*REdiii*, a completely automated tool that integrates excellent software solutions for the individual steps of the crystallographic structure-solution process into a pipeline with



**Figure 1**  
 Workflow of macromolecular structure determination using *REDiit*. (a) The operations depicted in green boxes are required to be performed manually; the fully automated *REDiit* pipeline is depicted with blue boxes. (b) Layered architecture of *REDiit*. Besides following several predefined pathways to find a suitable solution, *REDiit* can recruit tools to actively manipulate parameters and rerun select processes when necessary (details are given in Fig. 3a). The blue layer represents the available set of crystallographic tools on top of which decisions regarding the composition of the asymmetric unit, the search model can be made, influencing the chosen pathway. A database in which details about successful solutions are being saved can also be queried for parameters (red layer).

additional layers of decision-making, has been developed (see Fig. 1). Written in the Ruby programming language, *REDiit* is platform-independent but is best suited to Unix-based systems. Being user-friendly presupposes being flexible and adaptive to individual needs; therefore, an approach in which multiple tools are combined and little user input is required was chosen. The software architecture allows swapping or appending modules, easily allowing adaptability while maintaining full automation during individual experiments.

## 2.1. User interface

The premise was to require interfacing between the user and the software only once and at the beginning of data collection, at which stage the knowledge of some parameters may be limited. Decisions such as the resolution limit and the composition of the asymmetric unit are left to the software; however, the user is always informed about the progress at the above-mentioned stages and software-decided parameters, allowing the user to monitor the whole process. A command-line interface allows the configuration of all necessary parameters in a single command and facilitates server-based or remote access. For more interactive use, a graphical interface provides a form with preconfigured project or user-specific presets (see Fig. 2). After the successful completion of each key task described above, *REDiit* can send notifications and statistical parameters *via* email. The quality of each solution has to be assessed by the user. Different quality indicators, such as the  $R_{\text{meas}}$  value in the highest resolution shell or the overall completeness, can be input to the software as parameters, and the software will report the statistics table generated by *xia2*, the RFZ and TFZ scores after molecular replacement and  $R/R_{\text{free}}$  after completing chain tracing and ligand fitting, as well as an overall statistics table. However, the software will not autonomously decide whether the solution is final and adequate or whether further refinement is required. *REDiit* will only change parameters that were specified by the user if the crystallographic software that has been called for the specific task reports a failure.

## 2.2. Architecture

The Ruby programming language was chosen to implement the *REDiit* concepts because of its ease of reading syntax, strong metaprogramming (Aerts & Law, 2009) and an abundance of libraries providing bioinformatics (Goto *et al.*, 2010), process-management and web-framework tools. An option parser reads the provided parameters for each run into variables. A search model can be generated by providing a PDB code and chain identifiers, in which case the software will generate a ligand-free and solvent-free model using *PyMOL* (DeLano, 2002). In this case, a *PyMOL* script is written and executed to fetch the *.pdb* file, strip solvent and ligands and save the coordinates and sequence of the desired chain identifier to a *.pdb* and a *.fasta* file, respectively.

After each step during structure solution *REDiit* will check whether the called crystallographic tool ran successfully, but will generally not validate the quality of the result with the

exception of completeness, resolution or, for example,  $R_{\text{meas}}$  cutoffs after data processing. Failure to generate valid output files will trigger the pursuit of alternative options (see also Fig. 3a for the decision-making process).

During a typical experiment, the main script of *REdiii* will call relevant crystallographic software to accomplish the following tasks (Fig. 1a). For processing raw diffraction data frames, only one tool is allowed owing to the strong influence of data quality on different processing strategies. *xia2* (Winter, 2010), an expert system for data processing, allows sophisticated parameterization and workflow automation of indexing, integration and scaling. In *REdiii*, *xia2* is mainly used to provide an easily configurable interface with *XDS* (Kabsch, 2010), allowing automated indexing, integration and scaling. *xia2* was also chosen as the default tool for full automation as it performs well in determining the space group if it is unknown to the user and will likely choose the highest symmetry space group that is in agreement with the unit-cell parameters. Alternatively, the user can specify the space group. As data quality can vary, in order to be able to work with low-quality data while screening crystals, *REdiii* provides interfaces to *iMosflm* and *HKL-2000/3000* and then allows a choice between widely used processing suites within the same framework. The modular architecture of *REdiii* also directly accepts the *.sca* files generated by *HKL-2000/3000* and any *.mtz* file containing  $R_{\text{free}}$  flags, allowing complete user control over the data processing. In this case, *UNIQUEIFY*, *CTRUNCATE* and *SCALEPACK2MTZ* from the *CCP4* program suite (Winn *et al.*, 2011) are called separately to generate the structure-factor file. At this stage, the output *.log* file is parsed for statistical parameters and the user is notified. Depending on previous knowledge of the structure, the molecular-replacement pipelines *AutoMR* from the *PHENIX* suite (Zwart *et al.*, 2008) or *BALBES* (Long *et al.*, 2008) are used to solve the phase problem. The default settings generate an *AutoMR* session for molecular-replacement phasing, but *BALBES* is provided as an alternative option. *REdiii* supports three options for model building and explicit solvent generation: *AutoBuild* from the *PHENIX* suite, *Buccaneer* (Cowtan, 2006) and *ARP/wARP* (Langer *et al.*, 2008). By default, chain tracing and water picking are conducted using *AutoBuild* with the molecular-replacement solution as the starting model and the *PyMOL*-generated *.fasta* sequence as a target. The approach of building against the original input sequence also allows the completion of structural fragments that become visible only after molecular replacement with *BALBES*. These regions of unfitted density can lead to falsely attributed binding sites in the ligand-fitting stage. Alternatively, *Buccaneer* or the *auto\_tracing.sh* and *auto\_solvent.sh* scripts from the *ARP/wARP* suite can be called. In the case of the latter, *PyMOL* is invoked to perform chain splitting and dummy-atom replacement before and after the *ARP/wARP* operations.

Subsequent to rebuilding, ligand-fitting steps can be performed using *ReadySet*, *eLBOW* (the *electronic Ligand Builder and Optimization Workbench*) and *LigandFit* from the *PHENIX* suite with additional refinement cycles of the ligand-

bound complex using *phenix.refine*. If a *.pdb* file containing ligand coordinates has been specified, *phenix.elbow* is invoked to prepare geometry restraints for the ligand and *phenix.ligandfit* is used for subsequent ligand placement. Another *PyMOL* script is written and executed to generate a single coordinate file containing the fitted ligand with a separate chain identifier, which is refined by calling *phenix.refine* to generate the final model and the corresponding statistics to conclude the structure solution. The quality of the resulting model and especially ligand placement has to be assessed by the user and typically requires a resolution of 2 Å or better. This linear workflow allows a decision to be made between the 18 possible pathways of model generation which result from combinations of the available processing, phasing and model-building tools (Fig. 3).

### 2.3. Automated troubleshooting

In some cases not all parameters are known at the beginning of the experiment, but have to be estimated and can only

**REdiii GUI**  
We solve your structures while you sleep

Project:

Crystal:

Number:

PDB code / file name:

Chain IDs in pdb:

Ligand .pdb file:

RMS:

Spacegroup:

Mass:

e-mail:

Transfer files from:

Indexing:

completeness cutoff:

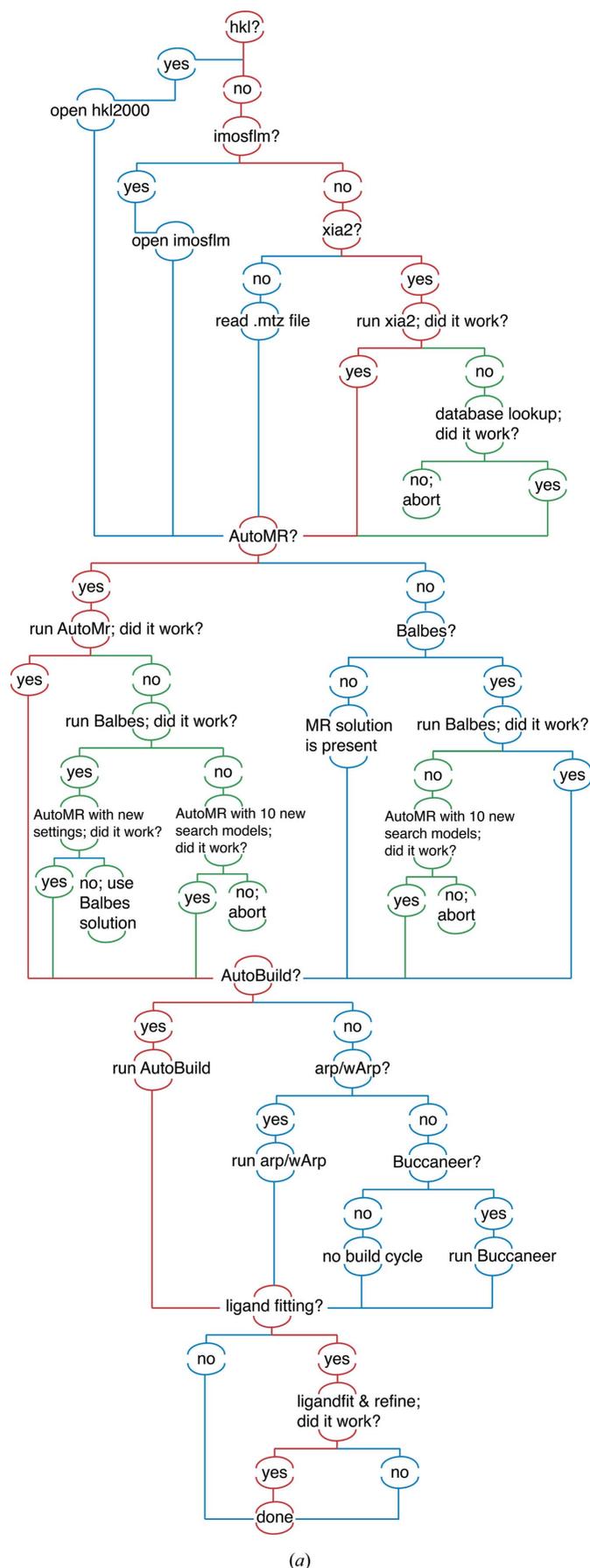
rmeas cutoff:

Molecular Replacement:

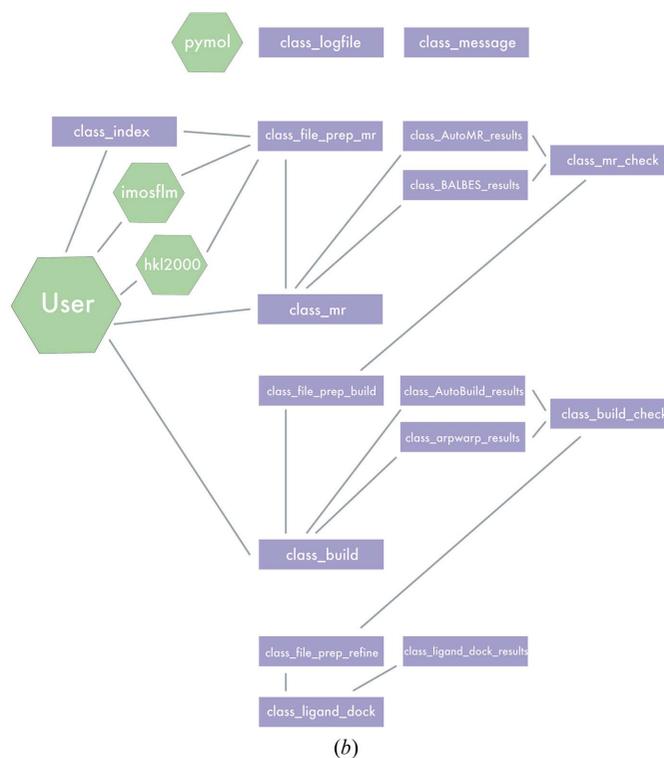
Rebuild:

**Figure 2**

A basic graphical user interface. Based on the cross-platform GTK+ toolkit, the interface allows easy manipulation of a set of parameters. The parameters accessible to the user are hardcoded into the software and can be changed to allow different project-specific or user-specific interfaces. The interface shown here depicts presets for the APOBEC3F C-terminal domain, the default tools for processing, molecular replacement and chain tracing with completeness and  $R_{\text{meas}}$  as variable cutoff criteria during processing.



be determined at a later stage. To address this problem, *REDiii* operates on a second layer of tools working on top of the collection of linear pathways to actively manipulate parameters in the case that the user-defined input does not lead to a successful solution (see Fig. 1). While *REDiii* is not capable of performing true validation after each step, it can realise whether a process finished successfully and pursue alternative strategies in case of failure. A typical scenario involves input of a wrong unit-cell composition, where correct identification requires information that may not yet be available during data collection. *BALBES* was included to resolve this potential problem, as a molecular-replacement tool that does not require the input of molecular weight and copy number, and can automatically determine composition of the asymmetric unit. *REDiii* will automatically use *BALBES* to determine the composition of the asymmetric unit if molecular replacement fails and the user had chosen *AutoMR*, and will rerun the initial setup with the new parameters while informing the user about the change in strategy (Fig. 1*b*, green layer). If molecular replacement still fails, the protein database *RESTful* web services are queried using the search-model sequence to generate an ensemble of up to ten alternative search models from similar available deposited structures. Subsequently,



**Figure 3** Overview of the workflow and class architecture. (a) The points of decision-making during an experiment. The workflow can be divided into three types of pathways: the unadjusted default (red), automatic rerouting in the case where the default fails (green) and optional user-configurable routes (blue). (b) The architecture of individual classes being called. *PyMOL*, *class\_logfile* and *class\_message* are called at multiple stages when the model needs to be modified or output information for the user is being prepared. *PyMOL*, *iMosflm* and *HKL-2000* are either called directly by the user or through separate scripts (highlighted in green boxes).

**Table 1**

Aspects and outcomes of experimental test cases.

Cases were chosen to cover different aspects of typical usage scenarios to test the performance of *REdiii*. The cases span a resolution range from 1.14 to 3.05 Å and include peptide and small-molecule ligand-bound protein structures.

Test case	Resolution (Å)	Space group	Aspect to be validated	Outcome
A3F 1	2.58	<i>P</i> 1	Ability to solve the apo structure of APOBEC3F, a prerequisite for ligand screening during later co-crystallization attempts	Solution has reasonable statistical parameters (see Table 2 <i>a</i> ); the resulting model is identical to the published structure (PDB entry 4iou)
A3F 2	3.05	<i>P</i> 1	Ability to handle data sets from non-ideal diffracting crystals for future optimization	Unit-cell parameters could not be refined without user intervention and <i>HKL</i> -2000, mosaicity and lower resolution data became apparent in the refinement statistics
A3G	2.53	<i>P</i> 12 <sub>1</sub>	Another APOBEC3 family member to show that solutions are model-independent and their quality is solely determined by diffraction data	Refinement statistics are similar to the first test case and provide proof that the quality of the obtained solutions is model-independent
Dengue virus protease	2.07	<i>C</i> 222 <sub>1</sub>	Processing synchrotron-quality data	High-quality model was obtained within less than 2 h computation time on a desktop workstation; refinement parameters are not improved further by manual model building
Hemoglobin	1.14	<i>C</i> 121	High-resolution data from a rotating-anode generator as an example of data obtained from well diffracting crystals using a home source	The search model differs at three residues, which becomes apparent in the refinement statistics; at this resolution the differences are resolved to almost atomic detail, which would be ideal for determining differences between a bound and an unbound state (the typical <i>REdiii</i> scenario)
HCV protease	1.52	<i>P</i> 12 <sub>1</sub>	Comparing an apo with an inhibitor-bound form of the same enzyme, in this case NS3/4A <i>Hepatitis C virus</i> (HCV) protease	Statistical parameters show that the solution does not require manual corrections; the obtained model was used as the search model in the inhibitor-bound case
Inhibitor	1.84	<i>P</i> 2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	HCV protease bound to a covalent inhibitor, demonstrating the reliability of the automated ligand fitting	Processing, model building and fitting of the small molecule work seamlessly and lead to good refinement statistics (see Table 2 <i>b</i> )
Substrate	2.09	<i>P</i> 2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	HIV-1 protease bound to a peptide substrate as a test case for docking of peptide moieties	Particularly difficult data set derived from a heavily quasispherically twinned crystal; despite showing the worst refinement statistics among all test cases, the model is interpretable and the ligand was fitted correctly

molecular replacement can iterate through the alternative models until a solution can be found or this step is considered a failure.

Another common problem is a wrong initial space group. Existing suites for data processing, including those used here, exhibit a strong bias towards high-symmetry space groups during indexing. This problem will often require manual intervention and data-quality assessment. At the end of each successful experiment, *REdiii* writes core parameters into a database file (Fig. 1*b*, purple layer), allowing *REdiii* to later query for entries of the same project group and using a multi-layered perceptron to determine which space groups might be worth trying to rerun the task with altered parameters. The database setup is specific to each installation and comprises a plain-text file of successful parameters for project-specific experiments. Currently, *REdiii* is capable of retrieving space-group information from previous experiments and initially trying to re-index the data using these settings.

### 3. Results

*REdiii* has been tested with unpublished data from six different proteins in apo forms and two data sets with ligand or inhibitor bound to the protein to validate different aspects of typical usage scenarios. Test cases span a resolution range from 1.1 to 3.1 Å. Each test case has been processed and solved with the default settings and the minimum amount of additional information, which consists of a search-model PDB

code, mass, space group and, in ligand-bound examples, a .pdb file containing coordinates of the ligand geometry. The validated aspects and the outcomes for each test case are given in Table 1, where different data sets were used to evaluate the performance of *REdiii*. Table 2 lists the results for these test cases using full automation and default settings and the corresponding statistics. Being designed primarily for the screening of crystals of putatively ligand-bound proteins, two cases were chosen to validate performance in the most common scenarios: a protein bound to a peptide ligand and a protein bound to a small-molecule inhibitor. In addition to novel data, 20 test cases of published data were chosen to validate the reliability of the tool and allow comparison with manual structure solutions. Table 3 shows the results of ten different inhibitor-bound structures of HIV-1 protease and Table 4 shows the corresponding results for inhibitor-bound *Hepatitis C virus* protease.

Structures of APOBEC3F, *Dengue virus* protease, hemoglobin, *Hepatitis C virus* (HCV) protease and APOBEC3G were processed to resolutions between 1.14 and 3.05 Å. The resolution cutoff was determined automatically, with  $R_{\text{meas}}$  in the highest resolution shell being required to be below 0.4. The  $R$  and  $R_{\text{free}}$  values of the final models range between 0.18 and 0.24 and between 0.21 and 0.32, respectively. Ligand-bound cases were solved to 1.8 and 2.1 Å resolution with  $R/R_{\text{free}}$  values of 0.22/0.26 and 0.34/0.39, respectively. CPU time on standard workstations was 1.5–6 h, with mosaic and twinned data being correlated with longer running times.

**Table 2**

Crystallographic statistics for test cases in Table 1.

Statistical parameters that were not generated because *HKL-2000* was used for processing are marked with an asterisk. The HIV protease–substrate complex solution has higher  $R/R_{\text{free}}$  values than expected for a correct solution at this resolution, but the data set contained a significant fraction of twinned data. Despite not accounting for twinning, the solution itself contains the ligand placed correctly in the binding site. Statistics were calculated using *xia2*, *HKL-2000*, *phenix.refine* and *phenix.model\_vs\_data*. Ramachandran quality, rotamer quality and clashscore were calculated using *phenix.table\_one*. R.m.s. deviations were constructed using the bond and angle parameters of Engh & Huber (1991).

(a) Unliganded data sets.

	A3F 1	A3F 2	DEN	Hemoglobin	HCV protease	A3G
Resolution (high) (Å)	2.58	3.05	2.07	1.14	1.52	2.53
Resolution (low) (Å)	32.12	25.84	28.58	29.72	33.24	29.49
Temperature (°C)	−180	−180	−180	−180	−180	−180
Space group	<i>P1</i>	<i>P1</i>	<i>C222<sub>1</sub></i>	<i>C121</i>	<i>P12<sub>1</sub>1</i>	<i>P12<sub>1</sub>1</i>
Unit-cell parameters						
<i>a</i> (Å)	51.86	51.81	60.33	93.15	54.67	61.63
<i>b</i> (Å)	68.75	67.61	61.60	43.29	58.44	67.96
<i>c</i> (Å)	75.52	75.18	114.22	82.27	59.94	63.58
$\alpha$ (°)	110.32	110.22	90	90	90	90
$\beta$ (°)	93.99	94.11	90	122.39	90.06	111.02
$\gamma$ (°)	110.83	110.46	90	90	90	90
Molecules in asymmetric unit	4	4	1	2	2	2
Completeness (%)	84.6	84.7	98.9	87.8	99.5	97.9
Total reflections	46734	414472	97808	260408	246459	68583
Unique reflections	23717	14546	13188	88726	57676	16151
Mean $I/\sigma(I)$	12.9	18.8	10.8	14.5	15.7	17.7
Average multiplicity	2.0	1.6	7.4	2.9	4.3	4.2
$R_{\text{merge}}$	0.056	0.122	0.155	0.038	0.051	0.065
$R_{\text{meas}}$	0.079	*	0.167	0.044	0.059	0.075
$R_{\text{p.i.m.}}$	0.056	*	0.059	0.022	0.028	0.036
CC*	0.995	*	0.994	0.999	0.999	0.996
Wilson <i>B</i> factor (Å <sup>2</sup> )	35	32	24	8	11	22
R.m.s. deviations from ideal values						
Bonds (Å)	0.01	0.01	0.01	0.01	0.01	0.01
Angles (°)	1.44	1.40	1.08	1.36	1.28	1.26
Average <i>B</i> factor (Å <sup>2</sup> )	30	34	26	13	17	12
Ramachandran plot (%)						
Favored	94	90	97	98	98	95
Allowed	6	7	3	2	1	5
Outliers	0	3	0	0	1	0
Rotamer outliers	8	14	3	2	2	4
Clashscore	9	16	4	15	9	9
<i>R</i> factor	0.23	0.23	0.18	0.22	0.20	0.24
$R_{\text{free}}$	0.26	0.32	0.22	0.22	0.21	0.30

In most test cases, solutions with acceptable  $R$  and  $R_{\text{free}}$  values were obtained with default settings and no user intervention. Resolution and data quality (*i.e.* mosaicity and twinning) appear to be limiting factors for the quality of the *REDiii*-generated structure, as seen for the A3F 2 and substrate-bound test cases. A3F 2 required the manual exclusion of highly mosaic frames. The ligand position in the substrate-bound test case was identified correctly, but crystal twinning led to unsatisfyingly high  $R$  values for the resulting model. Nevertheless, all unliganded structures from crystals diffracting to beyond 3 Å were solved without requiring any manual intervention, and the ligand-binding site was successfully identified in all liganded structures. The test cases from published structures show that data were usually processed to higher resolution than those deposited (a deci-

**Table 2 (continued)**

(b) Complexes between protein and a ligand.

	HCV protease + inhibitor	HIV-1 protease + substrate
Resolution (high) (Å)	1.84	2.09
Resolution (low) (Å)	19.22	27.18
Temperature (°C)	−180	−180
Space group	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>
Unit-cell parameters		
<i>a</i> (Å)	55.41	51.36
<i>b</i> (Å)	58.99	60.70
<i>c</i> (Å)	59.82	60.80
$\alpha$ (°)	90	90
$\beta$ (°)	90	90
$\gamma$ (°)	90	90
Molecules in asymmetric unit	1	1
Completeness (%)	90.8	97.6
Total reflections	48641	41261
Unique reflections	15757	11452
Mean $I/\sigma(I)$	22.83	10.5
Average multiplicity	3.1	3.6
$R_{\text{merge}}$	0.032	0.083
$R_{\text{meas}}$	0.037	0.097
$R_{\text{p.i.m.}}$	0.019	0.048
CC*	0.999	0.995
Wilson <i>B</i> factor (Å <sup>2</sup> )	10	25
R.m.s. deviations from ideal values		
Bonds (Å)	0.02	0.01
Angles (°)	2.91	1.60
Average <i>B</i> factor (Å <sup>2</sup> )	12	28
Ramachandran plot		
Favored	98	98
Allowed	2	1
Outliers	0	1
Rotamer outliers	3	9
Clashscore	10	22
<i>R</i> factor	0.22	0.35
$R_{\text{free}}$	0.26	0.40

sion made by *xia2* during indexing), with  $R/R_{\text{free}}$  values usually within a few percentage points of the published values.

#### 4. Discussion

We have described a computational pipeline that automates the process of solving protein structures from crystallographic data. Sets of diffraction data for a variety of proteins with varying data quality and resolution were used to test the performance of *REDiii* and its applicability to different scenarios. In all cases, including the liganded structures, a satisfactory solution was found without the need for any intervention, except when the mosaicity was very high. As may be anticipated, resolution, mosaicity and twinning are limiting factors to the quality of the structure model generated. However, with twinned data a single round of refinement applying the twin law is usually sufficient to resolve issues. Even though unliganded data at 3.0 Å resolution could be treated well, ligand placement is more restricted, especially with small molecules that can be easily misfitted into noise. The test cases show that a resolution of 2.0 Å or higher is desirable.

Building upon previous pipelines mentioned earlier, we show that complete automation of the structure-solution process with current tools is now possible. The commercially

Table 3

Larger ensemble of ligand-bound test cases.

We used an ensemble of inhibitor-bound HIV-1 protease structures (Nalam *et al.*, 2013).

PDB entry	3sa7	3sa6	3sa4	3sa9	3saa	3sac	3o9g	3o9d	3o9b	3o9f
Resolution (high) (Å)	1.43	1.64	1.67	1.49	1.60	1.41	1.53	1.94	1.37	1.50
Resolution (low) (Å)	30.83	23.51	25.48	27.12	39.27	38.24	25.11	28.80	26.21	25.10
Temperature (°C)	−80	−80	−80	−80	−80	−80	−80	−80	−80	−80
Space group	<i>P</i> <sub>2</sub> <sub>1</sub> <sub>2</sub> <sub>1</sub>									
Unit-cell parameters										
<i>a</i> (Å)	50.93	50.84	50.96	50.73	50.85	50.88	50.70	50.77	50.64	50.72
<i>b</i> (Å)	57.78	58.34	50.95	57.55	57.98	57.98	57.82	57.60	57.90	57.76
<i>c</i> (Å)	61.67	61.84	61.83	61.49	61.77	61.56	61.76	61.77	61.70	61.76
Molecules in asymmetric unit	1	1	1	1	1	1	1	1	1	1
Completeness (%)	98.9	92.9	85.2	96.4	92.0	98.9	96.7	99.1	99.2	99.7
Total reflections	231305	139029	130307	212012	246459	250941	176692	88077	282140	207765
Unique reflections	33947	21470	18567	28898	57676	35375	27123	13808	38476	29607
Mean <i>I</i> / $\sigma$ ( <i>I</i> )	19.3	24.7	26.8	17.2	15.7	24.9	30.0	8.0	22.3	20.1
Average multiplicity	6.8	6.5	7.0	7.3	4.3	7.1	6.5	6.4	7.3	7.0
<i>R</i> <sub>merge</sub>	0.063	0.051	0.050	0.103	0.051	0.041	0.041	0.280	0.058	0.068
<i>R</i> <sub>meas</sub>	0.073	0.060	0.059	0.116	0.059	0.047	0.050	0.342	0.067	0.079
<i>R</i> <sub>p.i.m.</sub>	0.028	0.023	0.022	0.042	0.028	0.017	0.019	0.133	0.024	0.030
CC*	0.999	0.999	0.999	0.998	0.999	1.000	1.000	0.959	1.000	0.999
Wilson <i>B</i> factor (Å <sup>2</sup> )	12	15	16	13	15	15	13	19	11	13
R.m.s. deviations from ideal values										
Bonds (Å)	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Angles (°)	1.25	1.20	1.24	1.14	1.21	1.28	1.19	1.50	1.29	1.23
Average <i>B</i> factor (Å <sup>2</sup> )	17	18	19	17	19	20	17	22	15	17
Ramachandran plot (%)										
Favored	99	98	99	99	99	98	99	98	99	99
Allowed	0	1	1	1	1	1	1	1	1	1
Outliers	1	1	0	0	0	1	0	1	0	1
Rotamer outliers	1	0	1	0	1	0	1	1	0	1
Clashscore	7	6	5	6	4	7	6	21	5	10
<i>R</i> factor	0.24	0.19	0.19	0.19	0.20	0.21	0.19	0.24	0.20	0.20
<i>R</i> <sub>free</sub>	0.28	0.21	0.25	0.23	0.24	0.23	0.23	0.28	0.23	0.23

available *HKL-3000* suite (Minor *et al.*, 2006), for example, allows highly automated molecular replacement and model building using *MOLREP* and *ARP/wARP* or *Buccaneer*, respectively. *HKL-3000* has a very intuitive graphical interface for these tools, but does not work autonomously and does not provide automatization beyond what is inherent to those tools, which is very different from fully automatized software such as *phenix.ligand\_pipeline* or *REDiii*. *HKL-3000* is also limited to utilizing a single software solution for molecular replacement and phasing, which is not ideal in scenarios where high throughput is critical, but leads to stellar results especially during indexing because of the vast amount of feedback provided to the user. Because of this, solutions such as *HKL-3000* are at the other end of the spectrum where the user is involved in and during every step to provide as much control as possible. Another solution, the *Auto-Rickshaw* web server, requires scaled and merged .mtz files with *R*<sub>free</sub> flags and can perform molecular replacement and model building without requiring user input, but does not incorporate ligand fitting. However, ligand placement is critical to many crystal screening efforts and the software is proprietary, which does not allow the tool to be refitted to different experimental needs.

Placement of ligands in *REDiii* does still require a .pdb file with reliable ligand coordinates and chemically sound restraints to obtain good geometry of the ligand in the protein–ligand complex, as in most cases X-ray data alone will

not lead to reasonable convergence (Evans, 2007). A number of solutions are available to facilitate the process of generating chemically plausible models (Kleywegt, 2007), but the process most often relies on precalculated restraints from accessible databases. *Ab initio* calculations using quantum mechanics are not feasible when high throughput is desired as they are computationally intensive. The ligand preparation and fitting tools within the *PHENIX* suite appear to be the most efficient and reliable way to incorporate model building of protein–ligand complexes and streamline well with the *REDiii* pipeline.

Applications relying on phasing methods other than molecular replacement or cases requiring multiple search models (such as protein–protein complexes) are not supported within *REDiii* at this point. In such cases, the need for high-throughput crystal screening and automation is not normally a bottleneck, but future developments may incorporate different phasing methods.

*REDiii* encompasses an entire pipeline from data processing to refinement of ligand-bound structures without user intervention and incorporates a wide range of currently used software. In typical cases, a satisfactory solution can be reached within a few hours using a personal computer. The sequential workflow with intelligent decision-making allows the software to reiterate through processes with alternative settings and to optimize the initial setup. *REDiii* is highly modular, and individual components of the pipeline can be added or swapped easily, allowing trouble-free adaptation to a

Table 4

Larger ensemble of ligand-bound test cases.

We used an ensemble of inhibitor-bound HCV protease structures (Romano *et al.*, 2012).

PDB code	3su2	3su1	3su0	3su6	3su5	3su4	3sv7	3sv8	3sv9	3sug
Resolution (high) (Å)	1.27	1.19	1.19	1.19	1.32	2.04	1.46	2.20	1.43	1.70
Resolution (low) (Å)	29.26	27.47	27.64	26.26	29.93	29.56	33.49	23.21	33.39	26.96
Space group	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P6<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P4<sub>2</sub>2<sub>2</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>	<i>P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub></i>
Unit-cell parameters										
<i>a</i> (Å)	55.10	54.95	55.28	54.96	54.87	85.86	55.31	69.57	54.69	54.06
<i>b</i> (Å)	58.53	58.53	58.50	58.46	58.33	85.86	58.80	69.57	58.59	58.29
<i>c</i> (Å)	60.05	59.99	60.51	59.78	59.87	97.51	60.26	78.99	60.74	62.21
Molecules in asymmetric unit	1	1	1	1	1	2	1	1	1	1
Completeness (%)	94.7	100	98.7	99.5	88.1	99.9	99.7	94.9	100	94.6
Total reflections	246782	373050	375789	356504	258979	316945	303713	133754	285413	95711
Unique reflections	49103	63075	62945	62350	39898	26025	34495	9853	36954	20988
Mean <i>I</i> / $\sigma$ ( <i>I</i> )	18.5	8.3	13.4	13.5	8.1	19.9	10.1	21.4	10.2	23.9
Average multiplicity	5.0	5.9	6.0	5.7	6.5	12.2	8.8	13.6	7.7	4.6
<i>R</i> <sub>merge</sub>	0.046	0.113	0.058	0.062	0.114	0.097	0.119	0.115	0.085	0.034
<i>R</i> <sub>meas</sub>	0.059	0.137	0.070	0.076	0.134	0.106	0.135	0.123	0.098	0.045
<i>R</i> <sub>p.i.m.</sub>	0.025	0.076	0.039	0.043	0.070	0.030	0.062	0.033	0.048	0.021
CC*	0.999	0.996	0.999	0.998	0.997	0.999	0.993	0.999	0.998	1.000
Wilson <i>B</i> factor (Å <sup>2</sup> )	10	11	10	8	10	26	14	23	16	23
R.m.s. deviations from ideal values										
Bonds (Å)	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Angles (°)	1.12	1.77	1.70	1.40	1.54	1.60	1.13	1.70	1.35	1.99
Average <i>B</i> factor (Å <sup>2</sup> )	16	19	17	14	16	31	20	32	24	29
Ramachandran plot (%)										
Favored	98	98	97	99	98	93	99	92	99	97
Allowed	2	2	3	1	1	4	1	7	1	3
Outliers	0	0	0	0	1	2	0	1	0	0
Rotamer outliers	1	0	1	1	1	3	3	6	2	1
Clashscore	7	13	10	8	8	9	5	15	9	13
<i>R</i> factor	0.22	0.23	0.23	0.22	0.22	0.22	0.20	0.25	0.21	0.22
<i>R</i> <sub>free</sub>	0.23	0.26	0.24	0.24	0.25	0.27	0.22	0.32	0.22	0.25

variety of needs, which is essential to accommodate the ever-improving and changing crystallographic software environment. The modular layout in combination with the open nature of the code should allow *REDi* to become an easily maintainable community tool for as long as there are active users. We believe that the *REDi* pipeline may be a useful component in the crystallographic toolbox, especially in accelerating scientific discovery through automated bio-crystallography.

### Acknowledgements

We are grateful to Dr William Royer, Dr Shivender Shandilya, Djade Soumana and Kuan-Hung Lin for generously providing crystals to generate test cases and Dr Nese KurtYilmaz for scientific editing. This research was supported by NIH grant P01 GM091743-03.

### References

- Aerts, J. & Law, A. (2009). *BMC Bioinformatics*, **10**, 221.  
 Amzel, L. M. (1998). *Curr. Opin. Biotechnol.* **9**, 366–369.  
 Brunzelle, J. S., Shafaei, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* **D59**, 1138–1144.  
 Cowtan, K. (2006). *Acta Cryst.* **D62**, 1002–1011.  
 Dauter, Z. (2010). *Acta Cryst.* **D66**, 389–392.  
 DeLano, W. L. (2002). *PyMOL*. <http://www.pymol.org>.  
 Echols, N., Moriarty, N. W., Klei, H. E., Afonine, P. V., Bunkóczi, G., Headd, J. J., McCoy, A. J., Oeffner, R. D., Read, R. J., Terwilliger, T. C. & Adams, P. D. (2014). *Acta Cryst.* **D70**, 144–154.  
 Engh, R. A. & Huber, R. (1991). *Acta Cryst.* **A47**, 392–400.  
 Evans, P. R. (2007). *Acta Cryst.* **D63**, 58–61.  
 Fu, Z.-Q., Rose, J. & Wang, B.-C. (2005). *Acta Cryst.* **D61**, 951–959.  
 Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J. & Katayama, T. (2010). *Bioinformatics*, **26**, 2617–2619.  
 Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.  
 Kabsch, W. (2010). *Acta Cryst.* **D66**, 125–132.  
 Kleywegt, G. J. (2007). *Acta Cryst.* **D63**, 94–100.  
 Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nature Protoc.* **3**, 1171–1179.  
 Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* **D64**, 125–132.  
 Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* **D62**, 859–866.  
 Nalam, M. N. L., Ali, A., Reddy, G. S. K. K., Cao, H., Anjum, S. G., Altman, M. D., Yilmaz, N. K., Tidor, B., Rana, T. M. & Schiffer, C. A. (2013). *Chem. Biol.* **20**, 1116–1124.  
 Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* **D61**, 449–457.  
 Romano, K. P., Ali, A., Aydin, C., Soumana, D., Özen, A., Deveau, L. M., Silver, C., Cao, H., Newton, A., Petropoulos, C. J., Huang, W. & Schiffer, C. A. (2012). *PLoS Pathog.* **8**, e1002832.  
 Rupp, B., Segelke, B. W., Krupka, H. I., Lekan, T., Schäfer, J., Zemla, A., Toppani, D., Snell, G. & Earnest, T. (2002). *Acta Cryst.* **D58**, 1514–1518.  
 Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). *Methods Mol. Biol.* **364**, 215–230.  
 Winn, M. D. *et al.* (2011). *Acta Cryst.* **D67**, 235–242.  
 Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.  
 Wojdyr, M., Keegan, R., Winter, G., Ashton, A., Lebedev, A. & Krissinel, E. (2014). *Acta Cryst.* **A70**, C1447.  
 Zwart, P. H., Afonine, P. V., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., McKee, E., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., Storoni, L. C., Terwilliger, T. C. & Adams, P. D. (2008). *Methods Mol. Biol.* **426**, 419–435.