Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in *Phenix*

Authors

Dorothee Liebschner[a], Pavel V. Afonine[a], Matthew L. Baker[b], Gabor Bunkoczi[cd], Vincent B. Chen[e], Tristan I. Croll[c], Bradley Hintze[ef], Li-Wei Hung[g], Swati Jain[eh], Airlie J. McCoy[c], Nigel W. Moriarty[a], Robert D. Oeffner[c], Billy K. Poon[a], Michael G. Prisant[e], Randy J. Read[c], Jane S. Richardson[e], David C. Richardson[e], Massimo D. Sammito[c], Oleg V. Sobolev[a], Duncan H. Stockwell[c], Thomas C. Terwilliger[gi], Alexandre G. Urzhumtsev[jk], Lizbeth L. Videau[e], Christopher J. Williams[e] and Paul D. Adams[al]*

[a]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

[b]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX, 77030, USA

[c]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, CB2 0XY, UK

[d]Current address: Astex Therapeutics Ltd, Cambridge, CB4 0QA, UK

[e]Department of Biochemistry, Duke University, Durham, NC, 27710, USA

[f]Current address: Duke Institute for Health Innovation, Duke University Medical Center, Durham, NC, 27701, USA

[g]Los Alamos National Laboratory, Los Alamos, NM, 87545, USA

[h]Current address: Department of Chemistry, New York University, New York, NY, 10003, USA

[i]New Mexico Consortium, Los Alamos, NM, 87544, USA

[j]Centre for Integrative Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS–INSERM–UdS, Illkirch, 67404, France

[k]Faculté des Sciences et Technologies, Université de Lorraine, BP 239, Vandoeuvre-lès-Nancy, 54506, France

[l]Department of Bioengineering, University of California Berkeley, Berkeley, CA, 94720, USA

Correspondence email: PDAdams@lbl.gov

Synopsis Recent developments of the *Phenix* software package are described in the context of macromolecular structure determination with X-rays, neutrons and electrons.

Abstract Diffraction (X-ray, neutron, electron) and electron cryo-microscopy are powerful methods to determine three-dimensional macromolecular structures that are required to understand biological processes and to develop new therapeutics against diseases. The overall structure solution workflow is similar for these techniques, but nuances exist because the properties of the reduced experimental data are different. Software tools for structure determination should therefore be tailored for each method. *Phenix* is a comprehensive software package for macromolecular structure determination that handles data from any of these techniques. Tasks performed with *Phenix* include data quality assessment, map improvement, model building, the validation/rebuilding/refinement cycle and deposition. Each tool caters to the type of experimental data. The design of *Phenix* emphasizes automation of procedures where possible to minimize repetitive and time-consuming manual tasks, while default parameters are chosen to encourage best practice. A graphical user interface provides access to many command-line features of *Phenix* and streamlines the transition between programs, project tracking and re-running of previous tasks.

Keywords: Phenix, Automation, Macromolecular crystallography, Cryo-EM, X-ray, Neutron, Diffraction, Python, cctbx

1. Introduction

Macromolecules are essential for biological processes within organisms engendering the need to understand their behavior to explain the fundamentals of life. The function of macromolecules correlates with their three-dimensional structure, i.e. how the atoms of the molecule are arranged in space and how they move over time. Two major methods to obtain macromolecular structures are diffraction (usually using X-rays, but also neutrons or electrons) and electron cryo-microscopy (cryo-EM[1]) (Fig. 1), both of which are handled by *Phenix*. The following subsections describe some concepts underpinning each method for the benefit of readers who are not experts in each of these areas.

1.1. X-ray diffraction

X-ray diffraction relies on the interaction of X-rays with the electron cloud of atoms in a crystal. As the atomic core electron density dominates the electron density distribution, major peaks equate to atomic positions and can be used to determine the structure. An exception is the hydrogen atom

[1] A consensus for the name is not yet established, both the terms “cryo-electron microscopy” or “electron cryo-microscopy” can be found in the literature (Jensen, 2010).

because it possesses only one valence electron whose distribution is shifted towards its covalent bond partner. The electron density in the unit cell is related to the Fourier transform of the amplitude and phase of the scattered X-rays. As only intensities of the waves can be measured, the phase information is lost and has to be inferred by various methods (section 4.1).

Of the models deposited in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Burley *et al.*, 2019), 89 % originate from X-ray crystallography. Since the first protein structures were determined in the 1950's (Kendrew *et al.*, 1958; Perutz *et al.*, 1960), the method has experienced many methodological and technological developments and is now considered quite mature (Wlodawer *et al.*, 2013). Nevertheless, structures determined at low resolution (for example, worse than 3 Å) remain challenging and could benefit from some of the new developments for cryo-EM that target similar resolution ranges.

1.2. Neutron diffraction

Neutron diffraction, which relies on the same formalism as X-ray diffraction, is based on the interaction of neutrons with atomic nuclei and therefore yields actual atomic positions directly. As the neutron scattering cross section varies by element (or isotope) in a nonlinear fashion, the scattering lengths of light atoms such as hydrogen and deuterium atoms (D) are similar to those of the heavier atoms (C, O, N). It is therefore possible to locate H (or D) atoms and deduce protonation states; this knowledge helps in understanding catalytic mechanisms and ligand binding (Yamaguchi *et al.*, 2009; Bryan *et al.*, 2013; Knihtila *et al.*, 2015). Furthermore, the neutron scattering length can be positive or negative (it is always positive for X-rays). For example, H has a negative scattering length, whose magnitude is about half of the scattering length of carbon. The nuclear scattering length density can therefore cancel out for groups such as CH_2, which occur frequently in macromolecules. To avoid negative scattering of hydrogen atoms, hydrogen can be partially or fully exchanged with deuterium, by soaking the crystal in deuterated buffer solutions or by performing protein expression in fully deuterated reagents, respectively.

The number of structures determined by neutron crystallography (0.1 % of models deposited in the PDB) is small compared to the number of X-ray structures (89 %). Neutron diffraction is not used to solve the structure of a macromolecule *de novo* as it requires considerable effort to prepare deuterated crystals suitable for the experiment. Instead, neutron diffraction provides complementary information because it enables locating hydrogen or deuterium atoms.

1.3. Cryo-EM

Cryo-EM relies on the interaction of electrons with the electrostatic field of the atoms in the sample. The method comprises many techniques, such as electron tomography, electron single-particle microscopy or electron crystallography. Single-particle analysis is a commonly used variant that

combines 2D projection images of macromolecules into a 3D reconstruction (electrostatic potential map or cryo-EM map). This is in contrast to diffraction experiments, where phase information is lost (in the absence of experimental phases, electron density maps thus have to be calculated using a model). While being visually similar to electron density maps from X-ray diffraction, a cryo-EM map exhibits some differences, such as negative peaks from negatively charged nucleic acids (Wang & Moore, 2017). Furthermore, the reconstruction process and motion or heterogeneity of the sample can lead to blurring of cryo-EM maps; high-resolution details can be revealed by operations such as map-sharpening (section 4.2).

Cryo-EM was traditionally employed to investigate large protein and nucleic acid complexes, filaments and viruses but was often limited to resolutions worse than 5 Å. Technological advances, such as the development of direct electron detectors (Li *et al*., 2013) and improvements in image processing (Bai *et al*., 2015) have transformed the method, leading to greatly improved resolution of cryo-EM maps. More recently, 3D reconstructions routinely attain resolutions significantly better than 4 Å, allowing for atomic model interpretation and solving structures *de novo*. Cryo-EM has thus become another principal method of macromolecular structure determination (2 % of models deposited in the PDB). For large molecules and structures determined at low resolution, annual cryo-EM model depositions now outnumber X-ray models (Figs. 2 and 3).

1.4. Other techniques

Another method to determine macromolecular structures is NMR (Nuclear magnetic resonance, 9 % of models deposited in the PDB), which uses quantum properties of atomic nuclei. *Phenix* does not have tools for structure determination with NMR data, so it is not addressed here.

Electron diffraction on nearly single-layer crystals is an emerging technique to determine high-resolution structures of macromolecules. It accounts for a slightly smaller number of models in the PDB than neutron diffraction.

1.5. Phenix

Phenix (Adams *et al*., 2002, 2010) is a software suite that uses reduced data from X-ray diffraction, electron diffraction, neutron diffraction or cryo-EM to determine macromolecular structures. Each method has a different approach to derive structural information, with *Phenix* offering specific tools to address the unique properties of the experimental data. Emphasis is put on the automation of all procedures to avoid burdening the user with repetitive, time-consuming and often error-prone tasks. Another important feature is the user-friendly design making the program accessible to novice users while keeping it flexible for experts. New tools are regularly developed or enhanced to improve the structure solution workflow. For example, a series of programs, which focuses on the analysis of

cryo-EM maps and models, was recently created to answer the emerging needs of the cryo-EM community.

This paper describes the structure determination process of three methods (X-ray diffraction, neutron diffraction and cryo-EM), summarizes major tools and reports recent developments in *Phenix*.

2. Steps in the structure solution process

Figure 4 shows the steps of the structure solution process for X-ray/neutron crystallography and cryo-EM. Due to the different nature of the interactions, there are nuances for each structure determination method (Figs. 5 and 6), but the overall procedure to obtain a molecular model is similar. The first step consists of analyzing the derived experimental data to detect any anomalies that can complicate or even prevent structure determination (section 3). The second and third steps are to obtain the best possible map (section 4) so that a model can be built (section 5). The fourth step focuses on iteratively improving the model by cycles of local rebuilding, refinement and validation (sections 6 and 7). The following sections elaborate on the steps and explain similarities and differences for X-ray/neutron crystallographic and cryo-EM data. *Phenix* tools that perform the corresponding steps are described.

3. Data quality assessment

Data quality should be analyzed carefully because unusual features can thwart structure solution. If the data have anomalies, it is not guaranteed that they can be addressed at later stages, in which case it might be necessary to perform new experiments or re-analyze the raw data.

3.1. Crystallography

3.1.1. Xtriage

Macromolecular crystals are prone to pathologies and rarely achieve perfect order, as the molecules interact weakly with each other. For example, a crystal is called “twinned” if two or more crystals (domains) are intergrown in such a way that their orientations are related by a specific geometrical operation (twin operation) (Hahn & Klapper, 2006). The overlap of diffraction spots adds noise to the measurements and reduces the information content of the data. Translational non-crystallographic symmetry (tNCS) is another pathology that complicates structure determination. This arises when more than one copy of a molecule or assembly is found in a similar orientation in the asymmetric unit of the crystal. Interference effects between diffraction from the copies lead to an overall modulation of the intensities in the diffraction pattern.

The program *Xtriage* (*phenix.xtriage*) identifies twinning, tNCS and other unusual features of diffraction data (Zwart *et al.*, 2005). To detect twinning, the tool examines amplitude and intensity ratios, $\langle|E^2-1|\rangle$ values, the *L*-statistic (Padilla & Yeates, 2003) and N(Z) plots (Howells *et al.*,

1950). The twin fraction is estimated by interpreting the Britton plot[2] (Britton, 1972) and performing the *H*-test (Yeates, 1997). *Xtriage* reveals tNCS using the native Patterson function and uses database-derived Wilson plots to find anomalies in the mean intensity. The tool also analyses reflection merging statistics to detect if the input data symmetry is too low and systematic absences to identify screw axes. A warning is issued if ice rings are detected. Apart from identifying pathologies, *Xtriage* also reports data quality indicators, such as the signal-to-noise ratio and data completeness. Furthermore, the tool estimates anomalous signal strength based on the fraction of statistically significant Bijvoet differences (Zwart, 2005) and the overall anisotropic scale factor using the likelihood formalism described by Popov & Bourenkov (2003).

3.1.2. Planning and assessing a SAD experiment

Before conducting a single-wavelength anomalous diffraction (SAD) diffraction experiment, it is useful to assess its chances of success. "*Plan SAD experiment"* is a new tool for estimating the anomalous signal from a SAD experiment with a particular anomalous scatterer and data quality, and for predicting whether this signal would be sufficient to solve the structure (Terwilliger *et al*., 2016*a*,*b*). The tool provides a summary of the anomalous signal required and what can be expected if the data can be measured with the suggested overall signal-to-noise ratio (I/σ). Once data have been collected and then scaled with the "*Scale and Merge Data"* tool, the "*Anomalous Signal"* tool estimates the amount of signal that has actually been achieved and predicts whether or not this will be sufficient to solve the structure.

3.2. Cryo-EM: *Mtriage*

The sample for a cryo-EM experiment is not crystalline, so many of the problems discussed in the previous section (3.1.1) are not relevant. However, the quality of the reconstruction and therefore the interpretability of a cryo-EM map can deteriorate from many causes, such as structural heterogeneity, radiation damage and beam induced sample movement. The information content of a cryo-EM map is typically expressed by the resolution. While the same term ("resolution") is used in crystallography to describe data quality, its meaning differs: the resolution of a crystallographic data set depends on the largest angle to which diffracted beams were measured or, equivalently, the shortest distance between reciprocal lattice planes (McPherson, 2009). The overall resolution $d_{\rm FSC}$ of a cryo-EM map is usually defined as the maximum spatial frequency at which the information content of the map is reliable (Penczek, 2010). The value is obtained by analyzing the Fourier shell correlation (FSC) for two cryo-EM half-maps binned in resolution shells (van Heel & Harauz, 1986). If the macromolecule is structurally heterogeneous (*e*.*g*. flexible regions in the macromolecule), a single value for the

[2] Number of negative intensities after detwinning as a function of the twin fraction.

resolution is most likely inadequate. A "local resolution" is thus assigned to different map regions (Cardone *et al*., 2013; Kucukelbir *et al*., 2014).

In *Phenix*, the resolution of cryo-EM maps can be estimated with the newly developed tool *Mtriage* (*phenix.mtriage*) (Afonine, Klaholz *et al*., 2018) using several different approaches, some of which fundamentally differ from d_{FSC}. The tool also summarizes map statistics. As the map resolution strongly influences decisions made in subsequent steps, it is important to get a reliable estimate.

3.3. Common map tools

Several new tools are available to analyze cryo-EM maps (or any map). In the context of molecular densities, a map is a 3D grid of density values. The map has an origin and a gridding (the distance between neighboring grid points). A map typically extends only over grid points where the values are non-zero (and a buffer), but it is possible that a majority of map points is zero or very small. Especially for cryo-EM, the molecules can have symmetry (such as viruses) and the map will have the same symmetry. The following tools analyze maps and perform some basic operations:

– *"Show map info"* (*phenix.show_map_info*) lists properties of a map, such as origin, grid points, unit cell and map size.

– *"Map box"* (*phenix.map_box*) cuts out a box from a large map.

– Some molecules, such as viruses, can have high internal symmetry. It can thus be beneficial to reduce the map to the repeating unit. "*Map symmetry"* (p*henix.symmetry_from_map*) finds such symmetries and "*Map box"* can extract the unique part of the map.

– The tool "*Combine focused maps"* (*phenix.combine_focused_maps*) creates a weighted composite map from a set of locally focused maps and associated models, where each part of each map is weighted by its correlation with the corresponding model.

4. Optimizing maps

4.1. X-ray

To calculate maps, the phase information is required. As phases are lost in the diffraction experiment, they have to be recovered by additional experiments or by computational procedures. In *Phenix*, phases can be determined by experimental phasing or by molecular replacement (Adams, Afonine *et al*., 2009). Once an initial set of phases is known, they can be improved by optimizing electron density maps or by optimizing an atomic model from which phases are calculated.

4.1.1. Experimental phasing

Experimental phasing relies on the properties of a few special atoms in the macromolecule. The special properties can be a large number of electrons, anomalous scattering or a combination of both

(reviewed by Dauter & Dauter, 2017). Phasing is then performed in two steps: 1) The properties are exploited to determine the location of the special atoms (substructure). 2) Knowledge of the substructure in one or more crystals is used to deduce phase information for the entire macromolecule.

AutoSol (*phenix.autosol*) is a comprehensive, automatic tool that performs experimental phasing with the MAD, MIR, SIR or SAD methods[3] (Terwilliger *et al.*, 2009). The program locates the substructure, estimates phases, performs density modification, identifies non-crystallographic symmetry (NCS), builds and refines a preliminary model. To carry out the tasks, *AutoSol* uses the *Phenix* tools *HySS* (Hybrid Substructure Search) (Grosse-Kunstleve & Adams, 2003; Bunkoczi *et al.*, 2013), *SOLVE* (Terwilliger & Berendzen, 1999), *Phaser* (McCoy *et al.*, 2007), *RESOLVE* (Terwilliger, 2002), *Xtriage* and *phenix.refine* (Afonine *et al.*, 2012).

4.1.2. Molecular replacement

Molecular replacement (MR) is used to solve structures when a structurally similar model (homologue) is available (Hoppe, 1957; Rossmann, 1972; Blow *et al.*, 2012). Success in MR calculations is determined by how much signal can be extracted from the data using a particular model. This depends on a combination of model quality and completeness, resolution of the data and the number of diffraction observations (Oeffner *et al.*, 2018). For typical cases involving crystals of medium-sized proteins diffracting to moderate resolution, the sequence identity between the molecule and the homologue should be greater than 25-30 % and the r.m.s. deviation between Cα atoms should be less than 2.0 Å (Taylor, 2010). The search model can be enhanced by trimming off parts of the model that are unlikely to be preserved in the target structure, using *Sculptor* (*phenix.sculptor*) (Bunkóczi & Read, 2011). The MR method consists of determining the orientation and position that places each copy of the homologue in the unit cell containing the unknown structure, judged by matching the calculated structure factors to the observed ones. An initial electron density map is then calculated with the phases from the homologue and the observed structure factors (Evans & McCoy, 2008).

Phaser (*phenix.phaser*) applies maximum-likelihood principles (Bayesian probabilities) to crystal structure solution by MR, by single-wavelength anomalous diffraction (SAD) or by a combination of both (MR-SAD) (McCoy *et al.*, 2007). In common with most MR algorithms, it divides the six-dimensional search problem for each copy into a three-dimensional rotation search followed by a three-dimensional translation search. The use of maximum-likelihood accounts for the effects of model imperfections and measurement error in the diffraction observations. In addition, likelihood provides a framework for the use of ensemble models created with *Ensembler* (*phenix.ensembler*) and

[3] MAD = multiple-wavelength anomalous diffraction, MIR = multiple isomorphous replacement, SIR = single isomorphous replacement

for exploiting the placement of one copy to increase the signal of the search for another copy (McCoy *et al*., 2007).

4.1.3. Density modification

As initial phases are often quite inaccurate, they need to be improved by exploiting prior knowledge about electron density distributions in crystals (Podjarny *et al*., 1996). Examples of methods to improve phases are solvent flattening, histogram matching and, if NCS is present, non-crystallographic symmetry averaging.

Several *Phenix* programs carry out density modification: *phenix.density_modification* performs iterative phase improvement with *RESOLVE* (Terwilliger, 2000), including the use of NCS and electron-density distributions. *phenix.multi_crystal_average* improves phases iteratively and averages electron density, both within a crystal and between crystals. *Phenix.ncs_average* can be used to average the electron density for molecules related by NCS.

4.2. Cryo-EM: map optimization

In the crystallographic case map improvement is achieved by manipulation of phase information, with the diffraction intensities (or amplitudes) remaining unchanged[4]. In contrast, cryo-EM maps are improved by methods such as sharpening and blurring, that typically modify the amplitudes of Fourier coefficients, leaving the phases unchanged. Cryo-EM maps can appear smooth and lack a high level of detail (contrast) because high-resolution amplitudes of corresponding Fourier map coefficients decay from causes such as radiation damage, sample movement, sample heterogeneity and errors in the reconstruction procedure. However, sharpening can reveal the high resolution details concealed in a cryo-EM map (Rosenthal & Henderson, 2003; Fernández *et al*., 2008).

The recently developed program *"Autosharpen map"* (*phenix.auto_sharpen*) performs map sharpening by optimizing the detail and connectivity of a cryo-EM map (Terwilliger, Sobolev *et al*., 2018).

5. Obtaining a model that fits the experimental data

To determine the structure of a macromolecule, a model must be built that fits the experimental data. Cryo-EM maps contain phase information, but they often have low resolution (Fig. 7), which makes interpretation difficult. For both techniques, but especially in the case of cryo-EM, the molecules can be very large so that automated procedures are preferred over manual interpretation, wherever feasible.

[4] Exceptions exist: for example, some procedures fill in missing reflections with hypothetically derived values (Sheldrick, 2008).

5.1. X-ray

Automatic model building is performed after phasing because the models from MR or *AutoSol* might be too incomplete to carry out refinement immediately: MR models typically originate from a homologue model which has a different sequence and side-chain conformations; it is thus necessary to build a model according to the sequence of the target molecule. *AutoSol* includes a building step, but in order to optimize runtime, it creates a preliminary model that can be further improved.

AutoBuild (*phenix.autobuild*) is an automated system for model rebuilding and completion (Terwilliger *et al.*, 2008). *AutoBuild* uses *RESOLVE*, *Xtriage* and *phenix.refine* to build an atomic model, refine it and improve it with iterative density modification, refinement and model building.

5.2. Cryo-EM

To obtain a model that fits the cryo-EM map, the following procedures are available.

5.2.1. Docking

A docking procedure is used if the model or a part of the model is already known (Roseman, 2000) but is not yet placed into the map. For example, the cryo-EM map might show a molecular complex assembled from components available from other experiments (such as crystallography). These components are then docked in the cryo-EM map to obtain a model for the entire complex.

The new tool "*Dock in map"* (*phenix.dock_in_map*) docks one or several models into a map (Terwilliger, 2018). The routine uses a convolution-based shape search to find a part of a map that is similar to the model. The shape search applies the following key elements: An initial search, focusing on the overall shape of the molecule, is performed at lower resolution, This initial search is done without rotation to optimize runtime;, it can supplemented optionally by matching the moments of inertia of the model and map. If the placement is satisfactory, it is optimized using real-space rigid-body refinement with the full resolution of the map.

5.2.2. Model building

If the structure of the molecule or of its components is unknown, the model has to be built *ab initio* into the cryo-EM map. This task is challenging, because the molecules are typically very large, chain tracing is difficult at low resolution and effective resolution can be even lower in some regions. Manual interpretation of cryo-EM maps is therefore time-consuming and error-prone, so automatic procedures are desirable.

The recent tool "*Map to model"* (*phenix.map_to_model*) interprets a cryo-EM map and builds an atomic model (Terwilliger, Adams *et al.*, 2018). All steps are performed automatically: First, the map is sharpened with "*Autosharpen map"* (section 4.2). The unique parts of the structure are then identified by taking into account reconstruction symmetry. The procedure also identifies which parts

of the map correspond to protein or RNA/DNA. After atomic models have been generated, they are real-space refined using secondary structure restraints (section 6.1.3). To get optimal building results, the resolution of the map should be 4.5 Å or better.

6. Refinement

Models from the building or phasing steps are approximate and need to be improved; for example, side chains may not fit the density and water molecules and ligands are likely to be missing. Refinement is the process of improving the parameters of a model until the best fit is achieved between experimental and model-calculated data. The parameterization of an atomic model mainly depends on the data quality and the current stage of refinement. Generally, the parameterization is chosen such that a simpler model is used in the beginning (such as rigid-body) and a more complex model is used towards the end. The target function guides the refinement by linking the model parameters to the experimental data and by scoring model-vs-data fit. For reciprocal space refinement, the target function (T) is expressed through structure factors (or diffraction intensities). For real-space refinement, the target is formulated in terms of a map. In both cases, the process alternates automated refinement with validation and either manual or automated model corrections.

6.1. Restraints

Crystallographic and cryo-EM refinements need additional information because there are generally too many model parameters compared to the amount of experimental data (unless the resolution is better than ~1 Å). Restraints introduce information and modify the target function by creating relationships between independent parameters. Using the example of restrained bond lengths, the coordinates of the two atoms are independent while the restraint keeps their distance within a certain target value and imposes a penalty if it deviates too much. Other restraints are imposed typically on bond angles, dihedrals, planes, chirality and coupling of atomic displacement parameters (*ADP*s) between bonded or neighboring atoms (Evans, 2007).

If restraints are used, the target function is a sum of an experimental-data component ($T_{\rm data}$) and a weighted restraints-based component ($w_{\rm restraints}$x$T_{\rm restraints}$):

$$T = T_{\rm data} + w_{\rm restraints}\ \mathrm{x}\ T_{\rm restraints} \qquad (1)$$

6.1.1. Stereochemical restraints

Proteins and nucleic acids (RNA and DNA) are composed of amino acids and nucleotides, respectively. The structures of these components are known from small-molecule crystallography with the assumption that they are similar when they assemble to form a macromolecule. For bond lengths and angles, *Phenix* makes use of the *CCP4* monomer library restraints (Engh & Huber, 1991; Vagin & Murshudov, 2004; Vagin *et al*., 2004) in protein side chains and somewhat modified classic

values for nucleic acids (Clowney *et al*., 1996; Gelbin *et al*., 1996). Planarity, dihedral angles and chirality are also restrained. Recent additions to restraints used in *Phenix* are the Conformation Dependent Library (Berkholz *et al*., 2009; Moriarty *et al*., 2014, 2016), which restrains the protein main chain as a function of the backbone dihedral values. Ribose-pucker and base-type dependent dihedral restraints are available for RNA (Jain *et al*., 2015). Algorithms, as opposed to libraries, are also used to provide stereochemical restraints. Linking including metal and metal cluster coordination (Moriarty & Adams, 2019), covalent bonding of standard and non-standard carbohydrates and other specialized entity specific restraints can be performed automatically.

6.1.2. Ligand restraints

Ligands are small molecules bound covalently or noncovalently to a macromolecule. While ligands can be naturally present, they can be also artifacts from reagents used for sample preparation or they can be introduced to investigate binding properties. Ligands in a model need to be refined and therefore need restraints. Some ligands are very common, so that geometry restraints are available in dictionaries (Vagin *et al*., 2004; Moriarty & Adams, http://sourceforge.net/projects/geostd) which can be obtained and updated from a number of sources (Moriarty & Adams, 2019). Other ligands are rare or novel requiring restraints be generated on a case-by-case basis.

Phenix has several tools for generating and handling ligand restraints. *eLBOW* (*phenix.elbow*) automatically generates geometry restraints for novel ligands or improves restraints for standard ligands (Moriarty *et al*., 2009). *ReadySet* (*phenix.ready_set*) prepares a model for refinement by generating all necessary ligand restraints with *eLBOW* and by updating the model file to reflect atom name changes from the new restraints. The tool *REEL* includes a 3D view of the ligand and a tabular view of the restraints, so that target values and standard deviations can be easily edited (Moriarty *et al*., 2017).

6.1.3. Other restraints

Several other types of restraints are available in *Phenix* tools:

– Secondary structure restraints: When data resolution is low, secondary structure elements (helices, sheets in proteins, base pairs and stacking pairs in nucleic acids) might not correctly maintain their conformation; for example a helix can lose its regular arrangement. Restraining the hydrogen bonds in the secondary structure element can help to maintain the regular structure (Headd *et al*., 2012).

– Ramachandran plot restraints: The backbone dihedral angles can be restrained to stay in the allowed regions of the Ramachandran plot (Oldfield, 2001; Emsley *et al*., 2010; Headd *et al*., 2012). These restraints can prevent the model from degrading at low resolution when the conformation is approximately correct, but should be used with caution because they can result in an incorrect local conformation minimum when the model geometry is poor (Richardson *et al*., 2018).

– Parallelity restraints: molecules may contain planar atom groups that are approximately parallel to each other, such as base pairs and stacking pairs in nucleic acids. Parallelity restraints keep the atom groups parallel (Sobolev *et al*., 2015; Richardson, 2015).

– Rotamer-specific restraints: These restraints lock a particular χ-angle configuration of an amino-acid residue side chain to preserve its valid rotameric state.

– NCS restraints: If the asymmetric unit contains two or more similar copies of the same molecule, torsion- or Cartesian-based NCS restraints can be used. NCS-related atoms can be identified automatically or be defined by the user. Torsion-based restraints are generally preferred because they require little or no manual intervention to account for common features such as domains that are very similar in structure but differ in relative orientation (Headd *et al*., 2014).

– Reference model restraints: If the data have low resolution, it can be helpful to use a related structure determined at higher resolution as reference model to steer refinement (Headd *et al*., 2012).

6.2. X-ray: *phenix.refine*

For crystallographic data, refinement is usually performed in reciprocal space, i.e. the parameters of the model are changed so that model-derived structure factors match experimental structure factor amplitudes or intensities. The refinement target in *Phenix* can be expressed either as a least-squares or as a maximum-likelihood target. Model parameters, which describe the crystal content and its properties, are a combination of 1) atomic parameters, such as coordinates, *ADP*s, occupancies and scattering factors and 2) non-atomic parameters, which describe contributions arising from bulk solvent, twinning and crystal anisotropy.

phenix.refine performs crystallographic structure refinement of atomic and non-atomic model parameters against experimental data (Afonine *et al*., 2012) at low to ultra-high resolutions. Each refinement run begins with bulk-solvent correction and anisotropic scaling (Afonine *et al*., 2005, 2013). The subsequent refinement strategy can be adapted to data resolution. Useful strategies at low resolution (or at initial stages) are rigid-body refinement (Afonine *et al*., 2009), simulated-annealing refinement in Cartesian or torsion-angle space (Grosse-Kunstleve *et al*., 2009) and detection and use of NCS (Kleywegt & Jones, 1995). *ADP* parameterizations include the Translation/Libration/Screw (TLS) model for movement of groups treated as rigid (Schomaker & Trueblood, 1968; Urzhumtsev *et al*., 2016; Afonine, Adams *et al*., 2018) as well as individual isotropic, anisotropic and grouped isotropic *ADP*. At ultra-high resolution (better than 0.7 Å), the interatomic scatterer model can account for residual density from bonding effects (Afonine *et al*., 2007). The program also offers occupancy refinement for any user-defined atoms. Water molecules can be placed and updated automatically, and improbable side chain rotamers can be replaced. In the later stages of refinement it is worthwhile adding hydrogen atoms, since they participate in most inter- and intra-molecular

contacts and their presence enables identification of steric clashes and hydrogen bonds. Hydrogen atoms can be added at nuclear positions or at electron-cloud-center positions (Deis *et al.*, 2013).

phenix.refine is designed to be flexible so that multiple refinement strategies can be combined with each other and applied to any selected part of the model in a single run. As there are several hundred parameters, protocols can be customized for specific needs. The *phenix.refine* graphical user interface (GUI) is integrated with *Coot* (Emsley *et al.*, 2010) and *PyMOL* (DeLano, 2002), so that refined models and associated maps can be readily displayed and analyzed.

While *phenix.refine* is the main crystallographic refinement program of *Phenix*, the following integrated alternatives exist:

– A recent addition integrates the *Amber* molecular-mechanics force field (Case *et al.*, 2018) for restraints with the functionality of *phenix.refine*. *Amber* uses energy based geometry terms and adds electrostatics and van der Waals attractive/dispersive interactions. *Amber* refinement in *Phenix* has been shown to improve model quality, especially sterics and hydrogen bonding at lower resolutions, and to reduce overfitting (Moriarty *et al.*, in preparation).

– *"Ensemble refinement"* (*phenix.ensemble_refinement*) combines crystallographic refinement with molecular dynamics to produce ensemble models fitted to diffraction data (Burnley *et al.*, 2012). The ensemble models can contain ~50-500 individual copies and simultaneously account for anisotropic and anharmonic distributions.

– *"DEN refinement"* (*DEN* = Deformable elastic network) uses a restraint network to maintain local model geometry while allowing for larger global domain motions over the course of several cycles of simulated annealing. The protocol is particularly useful for low resolution diffraction data (Brunger *et al.*, 2012).

–"*Rosetta refinement"* (*phenix.rosetta_refine*) integrates the *Rosetta* methods for conformational sampling (DiMaio *et al.*, 2013) with the X-ray targets, *ADP* refinement and map generation in *phenix.refine*. This tool is useful at low resolution, where it combines a wide radius of convergence across distinct local minima with realistic geometry. It can also be used to prepare crystal structures for further modelling in *Rosetta*.

6.3. Cryo-EM: *phenix.real_space_refine*

The outcome of the single-particle cryo-EM reconstruction is a three-dimensional map, so it is natural to perform refinement of the model in real-space. Phases are experimentally determined and are not improved by the procedure.

The recently developed tool, *phenix.real_space_refine,* was specifically designed to perform refinements in real-space (Afonine, Poon *et al.*, 2018). The algorithm uses a simplified refinement

target function that makes calculations faster, so that optimal data-restraint weights can be identified with little runtime cost. In addition to standard restraints on covalent geometry, *phenix.real_space_refine* makes use of secondary-structure, Ramachandran plot and rotamer-spacific restraints as well as internal molecular symmetry constraints. The default mode performs gradient-driven minimization of the entire model, but optimization can also be performed using simulated annealing, morphing (Terwilliger *et al.*, 2013), rigid-body refinement and systematic side-chain improvement (Oldfield, 2001). As is the case for reciprocal-space refinement, real-space refinement should be alternated with validation and manual corrections. The real-space refinement procedure is robust and works at resolutions from 1 to 6 Å.

6.4. Tools for neutron crystallography

6.4.1. Adding H/D atoms

Crystals used for neutron diffraction experiments contain H, D or both H and D atoms. If both isotopes are present, some sites (labile or exchangeable sites) can be shared by H and D, i.e. some molecules have D at a particular site, while the others have H.

ReadySet (*phenix.ready_set*) adds H or D atoms to a model file using the *REDUCE* algorithm (Word, Lovell, Richardson *et al.*, 1999). In particular, the tool can add H/D at exchangeable sites of protein amino acids and H or D to water molecule O atoms. At labile sites, hydrogen atoms are placed in alternative location "A" and the corresponding deuterium atoms are placed in "B".

6.4.2. Joint refinement in *phenix.refine*

Due to fairly prohibitive experimental demands, neutron diffraction data from macromolecules typically have low data completeness and a low signal to noise ratio. Furthermore, the model contains H (or D) atoms as independent parameters, increasing the number of variables significantly. As an X-ray structure is usually available before a neutron experiment is conducted, it is possible to refine a single model of a macromolecule simultaneously against X-ray and neutron data. This strategy, called joint X-ray and neutron refinement (joint XN refinement), ameliorates the data-to-parameters ratio by increasing the amount of experimental data used in refinement, leading to more complete and accurate models (Coppens, 1967; Orpen *et al.*, 1978; Wlodawer, 1980; Wlodawer & Hendrickson, 1982).

Macromolecular models can be refined with *phenix.refine* using neutron data or X-ray and neutron data simultaneously (Adams, Mustyakimov *et al.*, 2009; Afonine *et al.*, 2010). The program automatically detects exchangeable H/D sites in the model and ensures that the sum of occupancies is equal to one. The position of H (or D) atoms can be refined with a ‘riding model’ (Busing & Levy, 1964; Sheldrick & Schneider, 1997) or individually. All standard tools available for X-ray refinement (section 6.2) are also available for refinement using neutron data.

7. Validation

Validation indicates good parts and highlights problems in macromolecular models and should guide corrections throughout the structure solution process. In particular, the refinement stage benefits from validation: the process consists of cycles of validation, rebuilding (either manual or automated) and automated refinement, repeated until a satisfactory model is obtained. As problems are corrected, the model quality, refinement behavior and even the density map quality (for X-ray and neutron) all improve.

Validation addresses data, model and model-vs-data quality. This section covers model and model-vs-data quality as data quality was already described in section 3. There are many well-established metrics with new ones being developed to cover emerging needs. Some metrics are global (such as R_{free}), others are local (such as a Ramachandran outlier), but each local measure is usually also collected into a global score (such as the clashscore, Word, Lovell, LaBean *et al*., 1999). The most diagnostic and reliable validation criteria are those not used in the refinement target, providing independent direction for rebuilding.

In *Phenix*, validation for crystallographic or cryo-EM data and models is performed with the respective "*Comprehensive validation"* GUIs, or on the command line. The underlying principles behind model validation are the same for any experimental method. A good model should make chemical sense and be consistent with empirical statistics for high-quality prior structures. The most useful validation criteria depend on the resolution of the data. Model-vs-data validation depends on the type of experimental data and requires that the model explains its own data well. Generally, the goal is not zero outliers, but as few outliers as feasible (Richardson, Williams, Hintze *et al*., 2018). Ideally, each outlier should be explainable by reference to its environment (e.g. hydrogen bonding and/or steric packing stabilizing a rotamer outlier) and/or by the experimental data.

7.1. Model validation

In *Phenix*, model validation is provided in the "*Comprehensive validation"* GUI. Overall model statistics are presented in a summary chart with local scores as graphic plots and as tables that list the outliers on each criterion. Model validation tasks are essentially identical to the MolProbity web service (molprobity.biochem.duke.edu) (Chen *et al*., 2010; Richardson, Williams, Hintze *et al*., 2018; Williams, Headd *et al*., 2018).

The "*Comprehensive validation"* tool uses bond lengths and angle target values for proteins, nucleic acids and ligands from the same libraries applied for refinement restraints (section 6.1). Conformational, steric and some special-purpose metrics use the algorithms developed for *MolProbity* (Williams, Headd *et al*., 2018) and implemented in the *cctbx* (section 9). Cβ deviations diagnose sidechain-backbone incompatibility around the Cα tetrahedron (Lovell *et al*., 2003) except when

covalent geometry restraints need to be so tight at low resolution that a Cβ atom cannot deviate from ideal even if its position is incorrect.

Conformational validation relies on MolProbity's smoothed, multi-dimensional distributions for dihedral-angle combinations from quality-filtered reference data (Chen *et al*., 2010; Williams, Headd *et al*., 2018). Ramachandran backbone scores use six ϕ,ψ distributions that have quite different outlier contours: general, Ile/Val, Gly, pre-Pro, *trans*-Pro and *cis*-Pro (Read *et al*., 2011). Figure 8a shows the underlying pre-Pro data distribution and 8b the pre-Pro plot for a query model as shown in the validation GUI. Sidechain rotamer distributions were recently updated (Hintze *et al*., 2016). Omega distributions flag *cis* or twisted peptides. RNA ribose pucker outliers are diagnosed by a simple relationship between the well-fit 3' phosphate and glycosidic bond direction (Richardson *et al*., 2008), which also enables pucker-specific geometry targets in refinement (Adams *et al*., 2010).

Steric validation is accomplished by adding and optimizing hydrogen atoms with *REDUCE* and calculating their all-atom contacts with *PROBE* (Word, Lovell, Richardson *et al*., 1999; Word, Lovell, LaBean *et al*., 1999). An overall measure, called clashscore, is the number of serious clashes (non-H-bond overlap ≥ 0.4 Å) per 1000 atoms. *REDUCE* can also correct Asn/Gln/His “flips” and suggest His protonation. Clashes flag problems at any resolution if they occur, but it is possible that the clashscore can be artificially reduced due to tight non-bonded distance restraints. At high resolution, clashes flag incompatibilities within each alternate-conformation model or in disordered regions where geometry and steric restraints have been down-weighted or removed.

The CaBLAM analysis (Williams, Headd *et al*., 2018) was recently developed to validate protein backbone conformations in models determined at low resolution (2.5 to 4 Å), where it is difficult to determine peptide orientations as carbonyl oxygens cannot be discerned in density maps (this applies for crystallography and for cryo-EM). CaBLAM uses virtual Cα dihedrals to determine the local chain trace along with a virtual CO dihedral to diagnose where a peptide orientation is incompatible with it. CaBLAM outliers are not subject to overfitting, so they provide a less biased quality indicator. Most but not all CaBLAM outliers at the 1 % level flag a real problem, which usually requires changing the peptide orientation considerably rather than tweaking it across a contour boundary. They can often be corrected manually by modifying peptide orientations or by regularizing the local secondary structure.

Several rare but serious problems are now flagged if they occur, such as *cis*-nonPro peptides which are genuine (and typically with a clear structural role) for only one in 3000 residues and have been grossly overused especially at low resolution (Croll, 2015; Williams, Videau *et al*., 2018). *Cis*-nonPro peptides cannot be justified by experimental data at resolutions lower than 2.5 Å and should be modeled only if known from other resources (Richardson, Williams, Videau *et al*., 2018). Hydrogen and deuterium atoms are now analyzed if they are present, summarizing relevant properties and

flagging issues with H, D or exchanged sites, such as missing atoms, unusual geometry and unlikely occupancies. This is of particular use for models determined with neutron diffraction (Liebschner *et al*., 2018).

In the *Phenix* GUI, the results from all specific validations are seamlessly integrated with the graphics programs *Coot* and *PyMol* (Fig. 9). Outliers of any type are listed as a table with clicking on an outlier will recenter the graphics window on that atom or residue. If experimental data were supplied, maps will be displayed as well. The KiNG Java-based viewer (Chen *et al*., 2009) set up by *phenix.kinemage*, displays all model validation outliers in 3D to highlight local clusters of outliers around single serious problems. Generally, the integration of validation results with graphics programs reduces the effort required to fix problems. An extensive guide to the interpretation and use of model validation is available from the 2017 CCP4 Study Weekend (Richardson, Williams, Hintze *et al*., 2018).

7.2. Model vs. data validation

7.2.1. Comprehensive validation: Crystallography

Overall agreement between the model and the diffraction data is measured by *R*-factors, which evaluate the difference between observed (F_{obs}) and calculated (F_{model}) structure factor amplitudes:

$$R=\Sigma| |F_{obs}|-|F_{model}| |/\Sigma|F_{obs}| \qquad (2)$$

The R_{work} value is calculated on the large subset of the diffraction data used for refinement. For cross-validation, R_{free} is calculated on a subset (typically about 2000 reflections) that are not used in refinement (Brünger, 1992). If R_{free} increases while R_{work} decreases, the model might be incorrectly parameterized or the refinement strategy needs to be revised (Kleywegt & Brünger, 1996).

phenix.refine reports R_{work} and R_{free} values after refinement; *R* factor plots show how the values changed during the refinement run. The distribution of *R* factors across resolution shells can be used to pin-point anomalies such as ice rings, saturated reflections or problems with bulk-solvent modeling.

While *R* factors are calculated in reciprocal space, real-space correlation coefficients measure how the model fits the density map locally, for example at the chain, residue or atom level. In *Phenix* validation, residues with low real-space correlation coefficients are listed in a table linked to graphics programs, allowing recentering on the residue in question to enable analysis and correction. This is useful because the correlation may be sometimes misleading.

The *Polygo*n tool combines six diverse measures (average *ADP*, *rmsd* bonds and angles, clashscore, R_{work} and R_{free}) to visualize the refinement outcome in radial, one-dimensional histograms (Urzhumtseva *et al*., 2009). Lines connecting the scores form a hexagonal polygon that should be small and

approximately symmetric. This is similar to the wwPDB slider graphic[5] that conveys model quality at a glance.

7.2.2. Comprehensive validation: Cryo-EM

While validation methods in crystallography have had decades to mature, cryo-EM only recently evolved into a routine technique for near-atomic resolution models (Kühlbrandt, 2014). The search for appropriate metrics to assess model and model-to-data fit is therefore still ongoing (Afonine, Klaholz *et al*., 2018). One measure of agreement between a model and a map is the model-map correlation coefficient (map CC) (Brändén & Jones, 1990; Jones *et al*., 1991). In reciprocal space, model-vs-data agreement is assessed by curves of the Fourier shell correlation (*FSC*) as a function of resolution. The map-model *FSC* is the correlation between the Fourier coefficients computed by Fourier transformation of the 3D reconstruction and of a model-based map (Rosenthal & Henderson, 2003).

The *Phenix* "*Comprehensive validation"* tool (Fig. 8) for cryo-EM reports newly developed model and map-to-model quality indicators (Afonine, Klaholz *et al*., 2018), such as CaBLAM outliers, the *CC* using the entire map (CC_{box}), the *CC* within a mask (CC_{mask}) and the map-model *FSC* curve. Plots of average *CC*s per chain and per residue help identify problematic regions of the macromolecule.

The *EMRinger* (*phenix.emringer*) score quantifies how well the model backbone puts side chains in density peaks that are consistent with rotameric conformations (Barad *et al*., 2015).

8. Other tools

8.1. X-ray

8.1.1. Electron density maps

Electron density maps are routinely used to guide manual model building of crystallographic structures. Below is a selection of tools for map calculations in *Phenix*. The *Phenix* documentation includes a more complete list (www.phenix-online.org/documentation/).

– "*Polder maps"* (*phenix.polder*) uncover weak difference densities by locally excluding bulk solvent (Liebschner *et al*., 2017). They are useful for ligands and residues protruding into the solvent area.

– "*Composite OMIT maps"* (*phenix.composite_omit_maps*) are generated by combining omit maps of specific regions to obtain a map covering the entire contents of the unit cell (Brünger *et al*., 1998). This map is relatively bias-free without severely compromising phase quality.

[5] https://www.wwpdb.org/validation/2017/XrayValidationReportHelp#overall_quality

– "*Feature-enhanced maps"* (FEM, *phenix.feature_enhanced_map*) modify a $2mF_{obs}$-DF_{model} sigmaA-weighted map so that weak signals are strengthened while model bias and noise are reduced (Afonine *et al*., 2015).

New metrics to compare crystallographic contour maps are available in the tool "*Map Sigma Level Comparison"* (*phenix.map_comparison*)(Urzhumtsev *et al*., 2014).

8.1.2. Structure comparison

It is common to study near-identical protein structures, such as mutants, proteins with different ligands or NCS-related copies. Oftentimes, it is useful to compare the structures to find differences and similarities.

The "*Structure comparison"* tool (Moriarty *et al*., 2018) validates and analyzes similar protein models (> 80 % sequence identity). The GUI displays validation outliers and conformational differences between chains in a table, linked to graphics windows (*Coot* and *PyMOL*). Analyses include ligands, persistent ions and water molecules, rotamers, Ramachandran angles, missing atoms, secondary structure, water locations, omega angles and *ADPs*. The extracted chains and electron density maps can be superimposed onto a common frame of reference. The chains may subsequently be edited in *Coot* to ensure consistency and/or fix errors and recovered in their original orientations for further refinement and rebuilding.

8.1.3. Ligand fitting

The goal of a crystallographic study is often to understand the interaction of a small-molecule ligand with a macromolecule. It is also common to discover density for an unanticipated small molecule. In both cases, it is necessary to fit the ligand into the electron density to complete the atomic model.

Phenix has several tools to investigate ligands:

– "*LigandFit"* (*phenix.ligand_fit*) identifies difference density peaks in a map and tries to place a user defined ligand in the density (Terwilliger *et al*., 2006, 2007).

– *"Guided ligand replacement"* (*phenix.guided_ligand_replacement*) uses prior knowledge about ligand binding in a protein to assist the fitting of a similar ligand into the same or a similar protein (Klei *et al*., 2014). This tool helps study a series of compounds for the same or related macromolecular targets.

– *"Ligand identification"* (*phenix.ligand_identification*) analyzes difference density peaks to reveal which ligand is likely to be present (Terwilliger *et al*., 2006, 2007). The tool uses a library of 180 most frequently observed ligands in the PDB and ranks each molecule by density fit and chemical interactions with the macromolecule.

8.2. Using other programs within *Phenix*

Several programs from external developers can be executed in *Phenix*. Most require separate installation.

– *MR-Rosetta* (*phenix.mr_rosetta*) uses homology modeling in the *Rosetta* program to improve a model before and/or after MR (DiMaio *et al*., 2011; Terwilliger *et al*., 2012). Once a potential solution is obtained, *Rosetta* fills in missing sections and rebuilds the model to improve the fit to the electron density map and the phases for map interpretation or automated model building. This approach is helpful in cases where the MR model differs greatly from the target.

– *ERRASER* (Chou *et al*., 2013) improves RNA backbone conformations by combining the *MolProbity* clash analysis, *Phenix* refinement and a pruned enumeration and optimization in *Rosetta*.

– Conventional restraints may not capture the influences of intermolecular covalent and non-bonded interactions, metal coordination or solvation. The semiempirical quantum mechanics engine *DivCon* is integrated into *phenix.refine* to create gradients for a region of interest (Borbulevych *et al*., 2014).

– *CryoFit* uses molecular dynamics to perform flexible fitting of a model to a cryo-EM map (Kirmizialtin *et al*., 2015; Kim *et al*., 2019). The approach produces a new conformational model with optimized atomic coordinates and preserved stereochemistry and secondary structure.

– Quantum refinement (*Q|R*) is a method to refine macromolecular models with restraints derived from quantum chemistry instead of library-based restraints (Zheng, Moriarty *et al*., 2017; Zheng, Reimers *et al*., 2017). The *Q|R* source code (https://github.com/qrefine/qr-core) uses the *cctbx* (section 9.1) to construct a refinement protocol resembling *phenix.refine*.

- *Isolde* (Croll, 2018) is a plugin to UCSF ChimeraX (Goddard *et al*., 2018) for improving low-resolution cryoEM or crystal structures. It performs interactively guided simulation with molecular-dynamics flexible fitting against a map and real-time validation; it will soon be possible to launch from the *Phenix* GUI.

- Cryo-EM model building with *Pathwalker* (Baker *et al*., 2012; Chen *et al*., 2016) will soon be available within *Phenix*. *Pathwalker* can construct protein backbone models directly from near-atomic resolution cryo-EM density maps using a modified approach to Traveling Salesman Problem solvers. When coupled with *Phenix* tools, such as *phenix.pulchra* (Rotkiewicz & Skolnick, 2008) and real-space refinement, complete atomistic models can be generated within a few minutes for individual proteins or complexes.

8.3. Tools for model deposition

Several tools are available to facilitate deposition process of macromolecular structures:

- *"Generate table 1 for journal"* (*phenix.table_one*) is a tool for generating the standard table of crystallographic statistics required by most scientific journals for structure solutions. It summarizes validation statistics and calculates merging statistics for crystallographic data. For cryo-EM structures, the "*Comprehensive validation* (CryoEM)" GUI has a button to generate a similar table containing the model validation statistics as well as map resolution estimates.

- *"Prepare model for PDB deposition"* (*mmtbx.prepare_pdb_deposition*) adds the sequence information to the model file. In particular, the tool creates model files in PDBx/mmCIF format, which is mandatory for crystallographic depositions to the PDB since July 1st 2019 (Adams *et al*., 2019).

- *"Get PDB validation report"* (*phenix.get_pdb_validation_report*) retrieves the validation report from the wwPDB through their web interface. By providing the model and optional data in mmCIF format, the validation report can help users identify problems before starting the deposition process.

9. Infrastructure

9.1. Architecture

Phenix is built on the Computational Crystallography Toolbox (*cctbx*), which is an open-source library (https://github.com/cctbx/cctbx_project) of reusable software components for macromolecular structure determination (Grosse-Kunstleve *et al*., 2002). The *cctbx* components are written both in a compiled language (C++) and a flexible scripting language (Python, Lutz & Ascher, 1999). This approach is very efficient because high-level algorithms (such as refinement protocols) can be rapidly developed in the scripting language while computationally intensive algorithms can be implemented in the compiled language. The Boost.Python Library (http://www.boost.org/) is used to expose the C++ interfaces, classes and functions to Python (Grosse-Kunstleve & Abrahams, 2003).

The GUI is scripted through Python and produces a 'native' look[6] on each operating system. The current *Phenix* release (version 1.16) includes GUIs for all major programs. Embedded graphs are computed with the free matplotlib Python library (Hunter, 2007). A simple 3D graphics viewer can be used to display the molecule or to pick atom selections interactively.

9.2. Documentation

The *Phenix* GUI provides more than 175 tools, with even more programs available on the command line (~500). The extensive online manual covers about 180 separate HTML pages and describes GUI and command line versions of individual programs and includes tutorials and FAQs. The *Phenix* Tutorials Youtube channel (https://www.youtube.com/c/phenixtutorials, Fig. 10) currently provides 29

[6] The application takes on the convention, gestures and aesthetics of their operating system.

tutorial videos. Each video introduces a *Phenix* tool, summarizes the input files and parameters, explains how to run the program and discusses the results.

10. Conclusion

The *Phenix* software for macromolecular structure determination handles data from three experimental methods: cryo-EM, X-ray diffraction and neutron diffraction. All steps in the structure solution process are addressed by programs that are tailored for the type of experimental data, but share algorithms where appropriate. Procedures are automated to minimize repetitive and time-consuming manual tasks as far as feasible. For the future, the improvement of automated model building, refinement and validation at low resolution (worse than 3 Å) remains a priority; another area of development to help the structural biology community is the automated identification, fitting and refinement of ligands, ions and water.

Many challenging opportunities still exist in crystallography and cryo-EM owing to advances in light sources and instrumentation making it possible to go beyond structure determination by single crystal diffraction and single particle cryo-EM. For example, free electron lasers (FELs) and serial synchrotron crystallography (Chapman *et al*., 2011; Rossmann, 2014; Diederichs & Wang, 2017; Standfuss & Spence, 2017; Schlichting, 2015) have opened up new approaches to studying the dynamics of macromolecules. Therefore, methods are needed to extract models of molecular motion from time resolved diffraction experiments. Diffuse X-ray scattering (Wall *et al*., 2014, 2018) can also reveal molecular motions and lattice disorder, but methods to exploit the information contained in diffuse scatter are still scarce. Micro-electron diffraction (MicroED) (Liu *et al*., 2017; Gruene *et al*., 2018; Jones *et al*., 2018; Nannenga & Gonen, 2018) is an emerging technique to determine high-resolution structures of macromolecules. Current procedures for diffraction data will require fine-tuning to treat MicroED data adequately. Similarly, while cryo-EM is typically currently applied to large molecules and complexes, it is becoming increasingly possible to look at smaller molecules because of improvements in instrumentation and data processing. Furthermore, the use of focused refinement of cryo-EM data (von Loeffelholz *et al*., 2017; Natchiar *et al*., 2017) can generate much improved local reconstructions, but it remains to be seen how these can be best combined for model generation and subsequent model refinement. Finally, cryo-tomography, which is a type of electron microscopy that can probe entire cells and thus enable visualization of molecules *in situ*, nowadays produces reconstructions better than 10 Å, in some cases significantly better. There will be an increasing need to accurately and effectively combine such lower resolution information with results from high-resolution crystallographic or cryo-EM experiments.

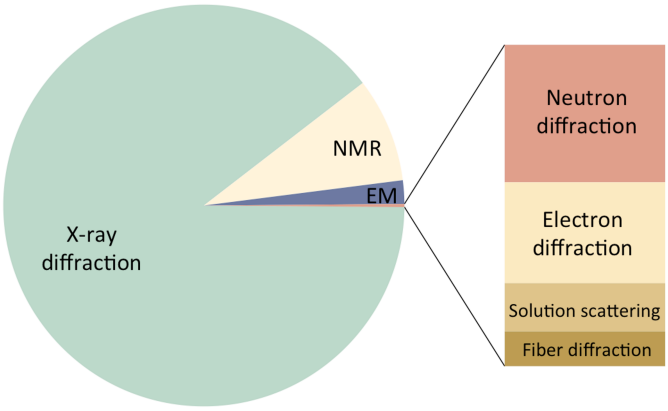

Figure 1 Experimental methods used to determine macromolecular structures that are deposited in the PDB. The predominant method is X-ray diffraction, followed by Nuclear Magnetic Resonance (NMR), cryo-EM and neutron diffraction.

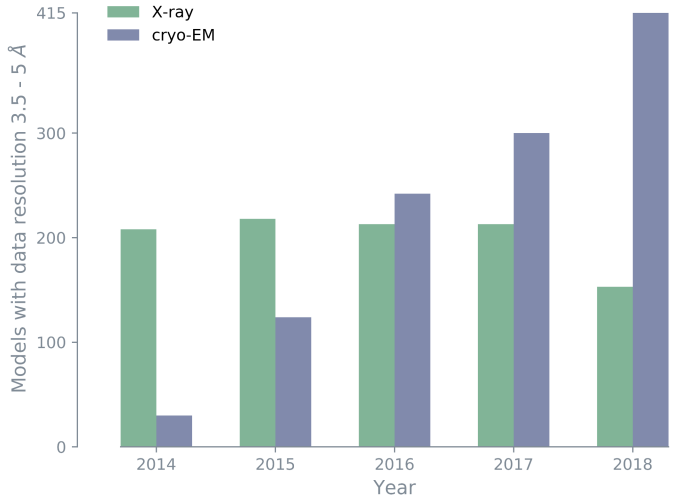

Figure 2 Annual cryo-EM model deposition now outnumbers X-ray model deposition in the resolution range 3.5 - 5 Å.

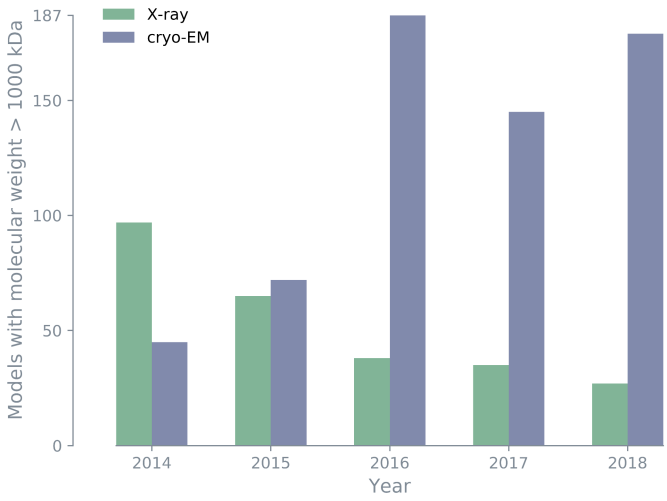

Figure 3 Since 2015, cryo-EM depositions account for the majority of large macromolecular structures.

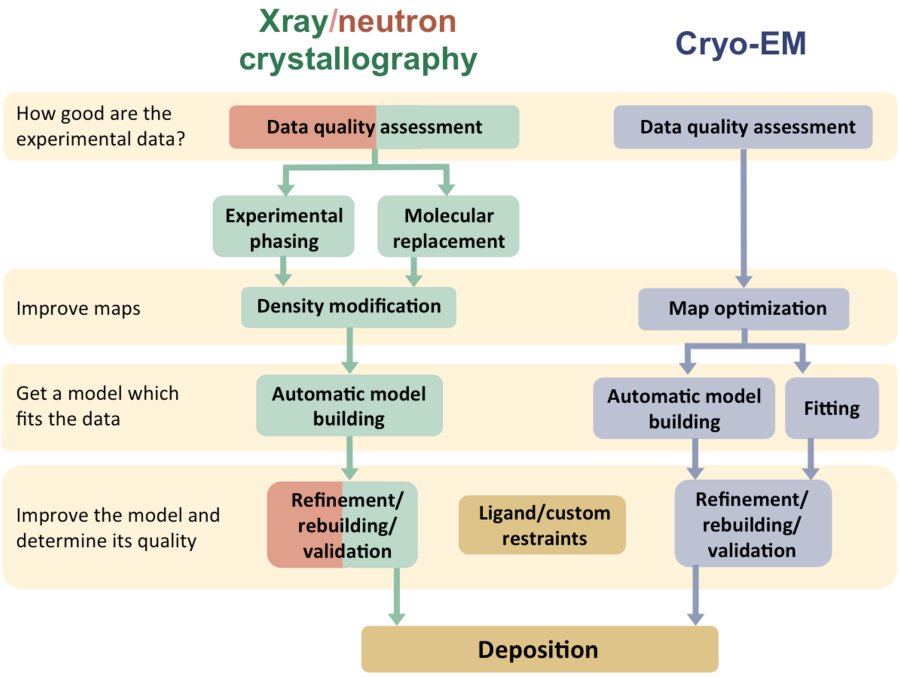

Figure 4 The structure solution steps for X-ray/neutron crystallography and cryo-EM have nuances for each technique, but the overall workflow is similar. Color code: cryo-EM = grey, X-ray crystallography = green, neutron crystallography = red. As neutron diffraction experiments are typically performed with samples whose structure is known, the phasing, density modification and model-building steps are not part of the workflow.

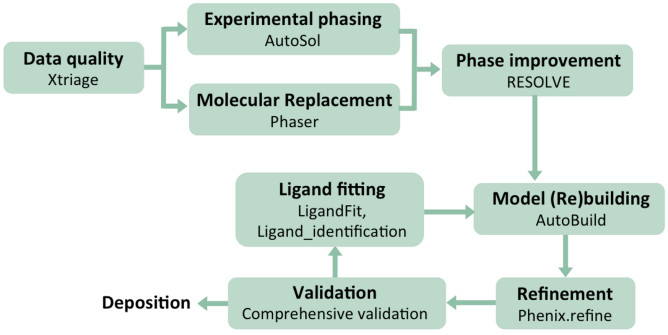

Figure 5 The primary tools for X-ray crystallography in *Phenix*.

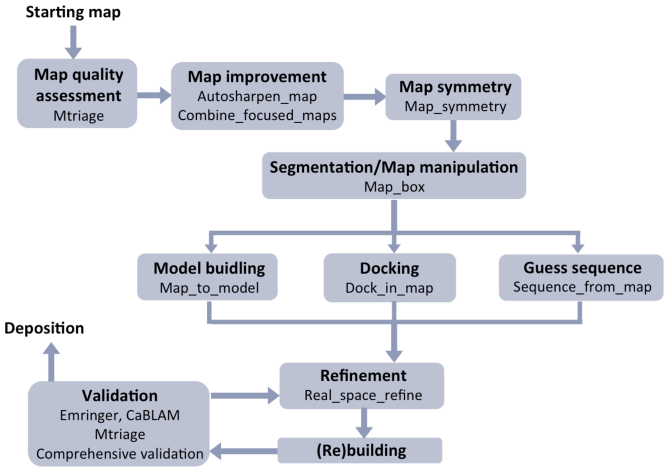

Figure 6 The primary tools for cryo-EM in *Phenix*.

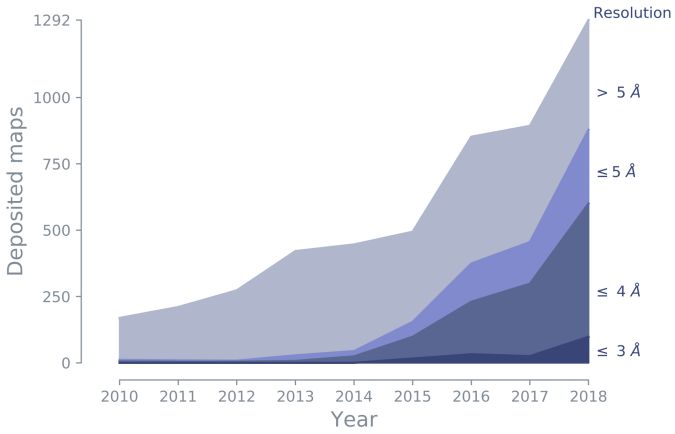

Figure 7 Cryo-EM maps deposited per year for different resolution ranges: better than 3 Å, 3 - 4 Å, 4 - 5 Å, worse than 5 Å.

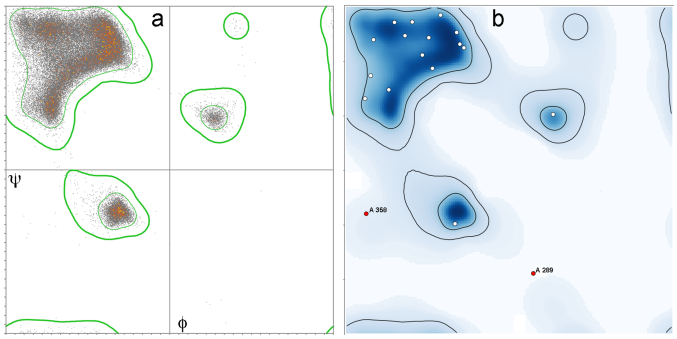

Figure 8 Ramachandran ϕ,ψ plots for the pre-Pro case. a) The reference distribution of 60,000 well-determined pre-Pro residues, with contours that enclose the favored 98% of the data (thin green) and that exclude the outliers (heavy green). b) A pre-Pro Ramachandran plot in the *Phenix* GUI for a query structure, showing two labeled outliers in red. Note that pre-Pro is very different from a general-case Ramachandran plot.

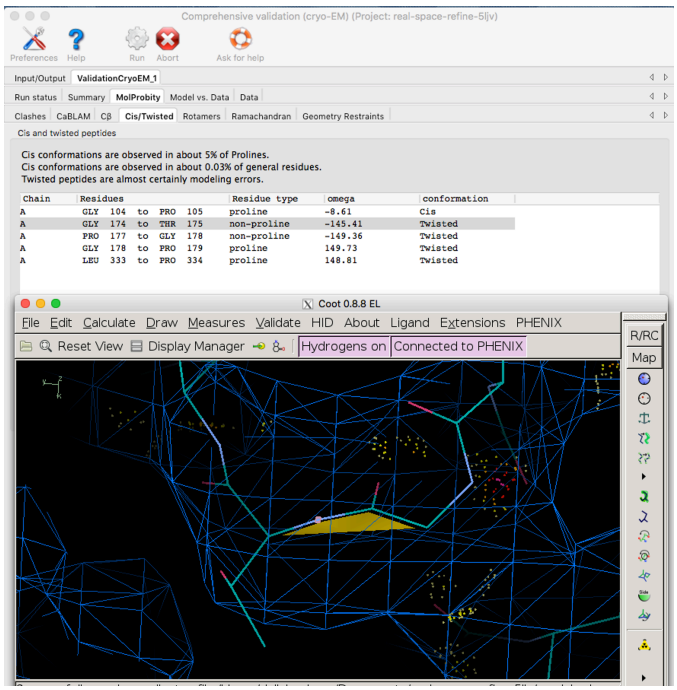

Figure 9 Screenshots of the cryo-EM "*Comprehensive validation*" tool and a *Coot* window. Clicking on the item in the table of cis/twisted peptides (highlighted in grey) recenters the *Coot* window on that peptide.

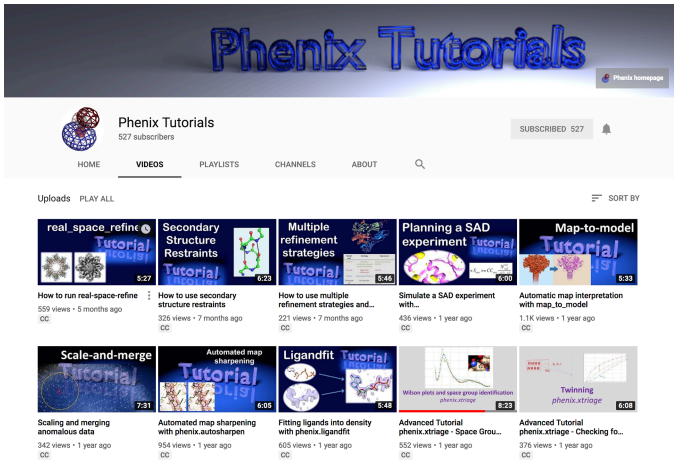

Figure 10 The videos on the *Phenix* tutorials YouTube channel cover main *Phenix* programs, refinement strategies and lectures.

Acknowledgements This research is supported by the US National Institutes of Health (NIH) (grant P01GM063210 to P.D.A., J.S.R., R.J.R. and T.C.T.), the *Phenix* Industrial Consortium and the NIH–funded (1R01GM071939) Macromolecular Neutron Consortium between Oak Ridge National Laboratory and Lawrence Berkeley National Laboratory. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH1123. Support from the Wellcome Trust (Principal Research Fellowship to R.J.R, grant number 209407/Z/17/Z) is gratefully acknowledged. M.D.S. gratefully acknowledges fellowship support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant (number 790122). A.U. acknowledges the support and the use of resources of the French Infrastructure for Integrated Structural Biology FRISBI ANR-10-INBS-05 and of Instruct-ERIC.

References

Adams, P. D., Afonine, P. V., Baskaran, K., Berman, H. M., Berrisford, J., Bricogne, G., Brown, D. G., Burley, S. K., Chen, M., Feng, Z., Flensburg, C., Gutmanas, A., Hoch, J. C., Ikegawa, Y., Kengaku, Y., Krissinel, E., Kurisu, G., Liang, Y., Liebschner, D., Mak, L., Markley, J. L., Moriarty, N. W., Murshudov, G. N., Noble, M., Peisach, E., Persikova, I., Poon, B. K., Sobolev, O. V., Ulrich, E. L., Velankar, S., Vonrhein, C., Westbrook, J., Wojdyr, M., Yokochi, M. & Young, J. Y. (2019). *Acta Cryst. D.* **75**, 451–454.

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.

Adams, P. D., Afonine, P. V., Grosse-Kunstleve, R. W., Read, R. J., Richardson, J. S., Richardson, D. C. & Terwilliger, T. C. (2009). *Curr. Opin. Struct. Biol.* **19**, 566–572.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst. D.* **58**, 1948–1954.

Adams, P. D., Mustyakimov, M., Afonine, P. V. & Langan, P. (2009). *Acta Cryst. D.* **65**, 567–573.

Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2018). *Acta Cryst. D.* **74**, 621–631.

Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *Acta Cryst. D.* **61**, 850–855.

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D., Lunin, V. Y. & Urzhumtsev, A. (2007). *Acta Cryst. D.* **63**, 1194–1197.

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. (2013). *Acta Cryst. D.* **69**, 625–634.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst. D.* **68**, 352–367.

Afonine, P. V., Grosse-Kunstleve, R. W., Urzhumtsev, A. & Adams, P. D. (2009). *J. Appl. Cryst.* **42**, 607–615.

Afonine, P. V., Klaholz, B. P., Moriarty, N. W., Poon, B. K., Sobolev, O. V., Terwilliger, T. C., Adams, P. D. & Urzhumtsev, A. (2018). *Acta Cryst. D.* **74**, 814–840.

Afonine, P. V., Moriarty, N. W., Mustyakimov, M., Sobolev, O. V., Terwilliger, T. C., Turk, D., Urzhumtsev, A. & Adams, P. D. (2015). *Acta Cryst. D.* **71**, 646–666.

Afonine, P. V., Mustyakimov, M., Grosse-Kunstleve, R. W., Moriarty, N. W., Langan, P. & Adams, P. D. (2010). *Acta Cryst. D.* **66**, 1153–1163.

Afonine, P. V., Poon, B. K., Read, R. J., Sobolev, O. V., Terwilliger, T. C., Urzhumtsev, A. & Adams, P. D. (2018). *Acta Cryst. D.* **74**, 531–544.

Bai, X., McMullan, G. & Scheres, S. H. W. (2015). *Trends in Biochemical Sciences*. **40**, 49–57.

Baker, M. R., Rees, I., Ludtke, S. J., Chiu, W. & Baker, M. L. (2012). *Structure*. **20**, 450–463.

Barad, B. A., Echols, N., Wang, R. Y.-R., Cheng, Y., DiMaio, F., Adams, P. D. & Fraser, J. S. (2015). *Nat. Methods*. **12**, 943–946.

Berkholz, D. S., Shapovalov, M. V., Dunbrack, R. L. & Karplus, P. A. (2009). *Structure*. **17**, 1316–1325.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *Journal of Molecular Biology*. **112**, 535–542.

Blow, D. M., Navaza, J., Tong, L., Arnold, E. & Rossmann, M. G. (2012). *International Tables for Crystallography, Vol. F*, Vol. edited by E. Arnold, D.M. Himmel & M.G. Rossmann, pp. 333–366. Chichester: Wiley & Sons.

Borbulevych, O. Y., Moriarty, N. W., Adams, P. D. & Westerhoff, L. M. (2014). *Computational Crystallography Newsletter*. **5**, 26–30.

Brändén, C.-I. & Jones, T. A. (1990). *Nature*. **343**, 687–689.

Britton, D. (1972). *Acta Cryst A*. **28**, 296–297.

Brünger, A. T. (1992). *Nature*. **55**, 472–475.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst. D*. **54**, 905–921.

Brunger, A. T., Das, D., Deacon, A. M., Grant, J., Terwilliger, T. C., Read, R. J., Adams, P. D., Levitt, M. & Schroeder, G. F. (2012). *Acta Cryst. D*. **68**, 391–403.

Bryan, T., González, J. M., Bacik, J. P., DeNunzio, N. J., Unkefer, C. J., Schrader, T. E., Ostermann, A., Dunaway-Mariano, D., Allen, K. N. & Fisher, S. Z. (2013). *Acta Cryst. F*. **69**, 1015–1019.

Bunkoczi, G., Echols, N., McCoy, A. J., Oeffner, R. D., Adams, P. D. & Read, R. J. (2013). *Acta Crystallogr. Sect. D-Biol. Crystallogr*. **69**, 2276–2286.

Bunkóczi, G. & Read, R. J. (2011). *Acta Cryst D*. **67**, 303–312.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlić, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M. & Zardecki, C. (2019). *Nucleic Acids Res*. **47**, D464–D474.

Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. (2012). *ELife*. **1**, e00311.

Busing, W. R. & Levy, H. A. (1964). *Acta Cryst*. **17**, 142–146.

Cardone, G., Heymann, J. B. & Steven, A. C. (2013). *J. Struct. Biol*. **184**, 226–236.

Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, III, T. E., Cruzeiro, V. W. D., Darden, T. A., Duke, R. E., Ghoreishi, D., Gilson, M. K., Gohlke, H., Goetz, A. W., Greene, D., Harris, R., Homeyer, N., Izadi, S., Kovalenko, A., Kurtzman, T., Lee, T. S., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Mermelstein, D. J., Merz, K. M., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, A., Pan, F., Qui, R., Roe, D. R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C. L., Smith, J., Salomon-Ferrer, R., Swails, J., Walker, R. C., Wang, J., Wei, H., Wolf, R. M., Wu, X., Xiao, L., York, D. M. & Kollman, P. A. (2018). Amber18 University of California, San Francisco.

Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., Doak, R. B., Maia, F. R. N. C., Martin, A. V., Schlichting, I., Lomb, L., Coppola, N., Shoeman, R. L., Epp, S. W., Hartmann, R., Rolles, D., Rudenko, A., Foucar, L., Kimmel, N., Weidenspointner, G., Holl, P., Liang, M., Barthelmess, M., Caleman, C., Boutet, S., Bogan, M. J., Krzywinski, J., Bostedt, C., Bajt, S., Gumprecht, L., Rudek, B., Erk, B., Schmidt, C., Hömke, A., Reich, C., Pietschner, D., Strüder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G., Schopper, F., Soltau, H., Kühnel, K.-U., Messerschmidt, M., Bozek, J. D., Hau-Riege, S. P., Frank, M., Hampton, C. Y., Sierra, R. G., Starodub, D., Williams, G. J., Hajdu, J., Timneanu, N., Seibert, M. M., Andreasson, J., Rocker, A., Jönsson, O., Svenda, M., Stern, S., Nass, K., Andritschke, R., Schröter, C.-D., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton, J. M., Barends, T. R.

M., Neutze, R., Marchesini, S., Fromme, R., Schorb, S., Rupp, D., Adolph, M., Gorkhover, T., Andersson, I., Hirsemann, H., Potdevin, G., Graafsma, H., Nilsson, B. & Spence, J. C. H. (2011). *Nature*. **470**, 73–77.

Chen, M., Baldwin, P. R., Ludtke, S. J. & Baker, M. L. (2016). *J. Struct. Biol*. **196**, 289–298.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst. D*. **66**, 12–21.

Chen, V. B., Davis, I. W. & Richardson, D. C. (2009). *Protein Sci*. **18**, 2403–2409.

Chou, F.-C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. & Das, R. (2013). *Nat. Methods*. **10**, 74–76.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *JACS*. **118**, 509–518.

Coppens, P. (1967). *Science*. **158**, 1577–1579.

Croll, T. I. (2015). *Acta Cryst D*. **71**, 706–709.

Croll, T. I. (2018). *Acta Cryst. D*. **74**, 519–530.

Dauter, M. & Dauter, Z. (2017). *Methods Mol Biol*. **1607**, 349–356.

Deis, L. N., Verma, V., Videau, L. L., Prisant, M. G., Moriarty, N. W., Headd, J. J., Chen, V. B., Adams, P. D., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2013). *Computational Crystallography Newsletter*. **4**, 9–10.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. Palo Alto, California, USA: DeLano Scientific LLC.

Diederichs, K. & Wang, M. (2017). *Methods Mol. Biol*. **1607**, 239–272.

DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D. & Baker, D. (2013). *Nat. Methods*. **10**, 1102–1104.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwaï, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature*. **473**, 540–543.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst. D*. **66**, 486–501.

Engh, R. A. & Huber, R. (1991). *Acta Cryst. A*. **47**, 392–400.

Evans, P. & McCoy, A. (2008). *Acta Cryst. D*. **64**, 1–10.

Evans, P. R. (2007). *Acta Cryst. D*. **63**, 58–61.

Fernández, J. J., Luque, D., Castón, J. R. & Carrascosa, J. L. (2008). *J. Struct. Biol*. **164**, 170–175.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *JACS*. **118**, 519–529.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. (2018). *Protein Sci.* **27**, 14–25.

Grosse-Kunstleve, R. W. & Abrahams, D. (2003). *C/C++ Users Journal*. **21**, 29–36.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst. D.* **59**, 1966–1973.

Grosse-Kunstleve, R. W., Moriarty, N. W. & Adams, P. D. (2009). *Proceedings of ASME 2009 International Design Engineering Technical Conferences*, Vol. p. San Diego, California.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.

Gruene, T., Wennmacher, J. T. C., Zaubitzer, C., Holstein, J. J., Heidler, J., Fecteau-Lefebvre, A., De Carlo, S., Müller, E., Goldie, K. N., Regeni, I., Li, T., Santiso-Quinones, G., Steinfeld, G., Handschin, S., van Genderen, E., van Bokhoven, J. A., Clever, G. H. & Pantelic, R. (2018). *Angewandte Chemie International Edition*. **57**, 16313–16317.

Hahn, T. & Klapper, H. (2006). *International Tables for Crystallography*, Vol. *D*, pp. 393–448.

Headd, J. J., Echols, N., Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Moriarty, N. W., Richardson, D. C., Richardson, J. S. & Adams, P. D. (2012). *Acta Cryst. D.* **68**, 381–390.

Headd, J. J., Echols, N., Afonine, P. V., Moriarty, N. W., Gildea, R. J. & Adams, P. D. (2014). *Acta Cryst. D.* **70**, 1346–1356.

van Heel, M. & Harauz, G. (1986). *Optik (Jena)*. **73**, 119–122.

Hintze, B. J., Lewis, S. M., Richardson, J. S. & Richardson, D. C. (2016). *Proteins*. **84**, 1177–1189.

Hoppe, W. (1957). *Acta Cryst. A.* **10**, 750–751.

Howells, E. R., Phillips, D. C. & Rogers, D. (1950). *Acta Cryst. A.* **3**, 210–214.

Hunter, J. D. (2007). *Comput. Sci. Eng.* **9**, 90–95.

Jain, S., Richardson, D. C. & Richardson, J. S. (2015). *Methods Enzymol.* **558**, 181–212.

Jensen, G. J. (2010). *Methods Enzymol.* **483**, xv–xvi.

Jones, C. G., Martynowycz, M. W., Hattne, J., Fulton, T. J., Stoltz, B. M., Rodriguez, J. A., Nelson, H. M. & Gonen, T. (2018). *ACS Cent. Sci.* **4**, 1587–1592.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst. A.* **47**, 110–119.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature*. **181**, 662.

Kim, D. N., Moriarty, N. W., Kirmizialtin, S., Afonine, P. V., Poon, B. K., Sobolev, O. V., Adams, P. D. & Sanbonmatsu, K. Y. (2019). *J. Struct. Biol.* accepted.

Kirmizialtin, S., Loerke, J., Behrmann, E., Spahn, C. M. T. & Sanbonmatsu, K. Y. (2015). *Methods Enzymol.* **558**, 497–514.

Klei, H. E., Moriarty, N. W., Echols, N., Terwilliger, T. C., Baldwin, E. T., Pokross, M., Posy, S. & Adams, P. D. (2014). *Acta Cryst. D.* **70**, 134–143.

Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*. **4**, 897–904.

Kleywegt, G. J. & Jones, T. A. (1995). *Structure*. **3**, 535–540.

Knihtila, R., Holzapfel, G., Weiss, K., Meilleur, F. & Mattos, C. (2015). *Journal of Biological Chemistry*. **290**, 31025–31036.

Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. (2014). *Nat. Methods*. **11**, 63–65.

Kühlbrandt, W. (2014). *Science*. **343**, 1443–1444.

Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., Agard, D. A. & Cheng, Y. (2013). *Nat. Methods*. **10**, 584–590.

Liebschner, D., Afonine, P. V., Moriarty, N. W., Langan, P. & Adams, P. D. (2018). *Acta Cryst. D.* **74**, 800–813.

Liebschner, D., Afonine, P. V., Moriarty, N. W., Poon, B. K., Sobolev, O. V., Terwilliger, T. C. & Adams, P. D. (2017). *Acta Cryst. D.* **73**, 148–157.

Liu, S., Hattne, J., Reyes, F. E., Sanchez-Martinez, S., Cruz, M. J. de la, Shi, D. & Gonen, T. (2017). *Protein Sci.* **26**, 8–15.

von Loeffelholz, O., Natchiar, S. K., Djabeur, N., Myasnikov, A. G., Kratzat, H., Ménétret, J.-F., Hazemann, I. & Klaholz, B. P. (2017). *Curr. Opin. Struct. Biol.* **46**, 140–148.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*. **50**, 437–450.

Lutz, M. & Ascher, D. (1999). Learning Python California, USA: O'Reilly & Associates.

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.

McPherson, A. (2009). Introduction to macromolecular crystallography Wiley-Blackwell.

Moriarty, N. W. & Adams, P. D. (2019). *Acta Cryst. D.* **75**, 16–20.

Moriarty, N. W. & Adams, P. D. GeoStd.

Moriarty, N. W., Draizen, E. J. & Adams, P. D. (2017). *Acta Cryst. D.* **73**, 123–130.

Moriarty, N. W., Liebschner, D., Klei, H. E., Echols, N., Afonine, P. V., Headd, J. J., Poon, B. K. & Adams, P. D. (2018). *Protein Sci.* **27**, 182–194.

Moriarty, N. W., Tronrud, D. E., Adams, P. D. & Karplus, P. A. (2014). *FEBS J.* **281**, 4061–4071.

Moriarty, N. W., Tronrud, D. E., Adams, P. D. & Karplus, P. A. (2016). *Acta Cryst. D.* **72**, 176–179.

Nannenga, B. L. & Gonen, T. (2018). *Emerg Top Life Sci.* **2**, 1–8.

Natchiar, S. K., Myasnikov, A. G., Kratzat, H., Hazemann, I. & Klaholz, B. P. (2017). *Nature*. **551**, 472–477.

Oeffner, R. D., Afonine, P. V., Millán, C., Sammito, M., Usón, I., Read, R. J. & McCoy, A. J. (2018). *Acta Cryst. D*. **74**, 245–255.

Oldfield, T. J. (2001). *Acta Cryst. D*. **57**, 82–94.

Orpen, A. G., Pippard, D., Sheldrick, G. M. & Rouse, K. D. (1978). *Acta Cryst. B*. **34**, 2466–2472.

Padilla, J. E. & Yeates, T. O. (2003). *Acta Cryst. D*. **59**, 1124–1130.

Penczek, P. A. (2010). *Methods Enzymol*. **482**, 73–100.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. (1960). *Nature*. **185**, 416.

Podjarny, A. D., Rees, B. & Urzhumtsev, A. G. (1996). *Crystallographic Methods and Protocols*, Vol. *56*, edited by C. Jones, B. Mulloy & M.R. Sanderson, pp. 205–226. New Jersey: Humana Press.

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst. D*. **59**, 1145–1153.

Read, R. J., Adams, P. D., Arendall, W. B., Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J. S., Sheffler, W. H., Smith, J. L., Tickle, I. J., Vriend, G. & Zwart, P. H. (2011). *Structure*. **19**, 1395–1412.

Richardson, J. S. (2015). *Computational Crystallography Newsletter*. **6**, 47–53.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., Berman, H. M. & RNA Ontology Consortium (2008). *RNA*. **14**, 465–481.

Richardson, J. S., Williams, C. J., Hintze, B. J., Chen, V. B., Prisant, M. G., Videau, L. L. & Richardson, D. C. (2018). *Acta Cryst. D*. **74**, 132–142.

Richardson, J. S., Williams, C. J., Videau, L. L., Chen, V. B. & Richardson, D. C. (2018). *J. Struct. Biol*. **204**, 301–312.

Roseman, A. M. (2000). *Acta Cryst. D*. **56**, 1332–1340.

Rosenthal, P. B. & Henderson, R. (2003). *J. Mol. Biol*. **333**, 721–745.

Rossmann, M. G. (1972). The molecular replacement method New York: Gordon & Breach.

Rossmann, M. G. (2014). *IUCrJ*. **1**, 84–86.

Rotkiewicz, P. & Skolnick, J. (2008). *Journal of Computational Chemistry*. **29**, 1460–1465.

Schlichting, I. (2015). *IUCrJ*. **2**, 246–255.

Schomaker, V. & Trueblood, K. N. (1968). *Acta Cryst B*. **24**, 63–76.

Sheldrick, G. M. (2008). *Acta Cryst. A*. **64**, 112–122.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol*. **277**, 319–343.

Sobolev, O. V., Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2015). *J. Appl. Cryst*. **48**, 1130–1141.

Standfuss, J. & Spence, J. (2017). *IUCrJ*. **4**, 100–101.

Taylor, G. L. (2010). *Acta Cryst. D*. **66**, 325–338.

Terwilliger, T. C. (2000). *Acta Cryst. D*. **56**, 965–972.

Terwilliger, T. C. (2002). *Acta Cryst. D*. **58**, 1937–1940.

Terwilliger, T. C. (2018). *Computational Crystallography Newsletter*. **9**, 51–57.

Terwilliger, T. C., Adams, P. D., Afonine, P. V. & Sobolev, O. V. (2018). *Nat. Methods*. **15**, 905–908.

Terwilliger, T. C., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2007). *Acta Cryst. D*. **63**, 101–107.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Acta Cryst. D*. **65**, 582–601.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst. D*. **55**, 849–861.

Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. & Adams, P. D. (2016*a*). *Acta Cryst. D*. **72**, 346–358.

Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. & Adams, P. D. (2016*b*). *Acta Cryst. D*. **72**, 359–374.

Terwilliger, T. C., DiMaio, F., Read, R. J., Baker, D., Bunkóczi, G., Adams, P. D., Grosse-Kunstleve, R. W., Afonine, P. V. & Echols, N. (2012). *J. Struct. Funct. Genomics*. **13**, 81–90.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst. D*. **64**, 61–69.

Terwilliger, T. C., Klei, H., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2006). *Acta Cryst. D*. **62**, 915–922.

Terwilliger, T. C., Read, R. J., Adams, P. D., Brunger, A. T., Afonine, P. V. & Hung, L.-W. (2013). *Acta Cryst. D*. **69**, 2244–2250.

Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. (2018). *Acta Cryst. D*. **74**, 545–559.

Urzhumtsev, A., Afonine, P. V., Lunin, V. Y., Terwilliger, T. C. & Adams, P. D. (2014). *Acta Cryst. D*. **70**, 2593–2606.

Urzhumtsev, A., Afonine, P. V., Van Benschoten, A. H., Fraser, J. S. & Adams, P. D. (2016). *Acta Cryst. D*. **72**, 1073–1075.

Urzhumtseva, L., Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2009). *Acta Cryst. D*. **65**, 297–300.

Vagin, A. A. & Murshudov, G. N. (2004). *IUCr Comput. Comm. Newsl*. **4**, 59–72.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst D*. **60**, 2184–2195.

Wall, M. E., Adams, P. D., Fraser, J. S. & Sauter, N. K. (2014). *Structure*. **22**, 182–184.

Wall, M. E., Wolff, A. M. & Fraser, J. S. (2018). *Curr. Opin. Struct. Biol*. **50**, 109–116.

Wang, J. & Moore, P. B. (2017). *Protein Sci*. **26**, 122–129.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (2018). *Protein Sci*. **27**, 293–315.

Williams, C. J., Videau, L. L., Hintze, B. J., Richardson, D. C. & Richardson, J. S. (2018). *BioRxiv*.

Wlodawer, A. (1980). *Acta Cryst. B*. **36**, 1826–1831.

Wlodawer, A. & Hendrickson, W. A. (1982). *Acta Cryst. A*. **38**, 239–247.

Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. (2013). *FEBS J*. **280**, 5705–5736.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol*. **285**, 1711–1733.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol*. **285**, 1735–1747.

Yamaguchi, S., Kamikubo, H., Kurihara, K., Kuroki, R., Niimura, N., Shimizu, N., Yamazaki, Y. & Kataoka, M. (2009). *PNAS*. **106**, 440–444.

Yeates, T. O. (1997). *Methods Enzymol*. **276**, 344–358.

Zheng, M., Moriarty, N. W., Xu, Y., Reimers, J. R., Afonine, P. V. & Waller, M. P. (2017). *Acta Cryst. D*. **73**, 1020–1028.

Zheng, M., Reimers, J. R., Waller, M. P. & Afonine, P. V. (2017). *Acta Cryst. D*. **73**, 45–52.

Zwart, P. H. (2005). *Acta Cryst. D*. **61**, 1437–1448.

Zwart, P. H., Grosse Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsletter on Protein Crystallography*. **43**, 10.