Acta Cryst D
STRUCTURAL BIOLOGY

ISSN 2059-7983

Received 1 November 2021
Accepted 23 March 2022

Edited by T. Burnley, Rutherford Appleton Laboratory, United Kingdom

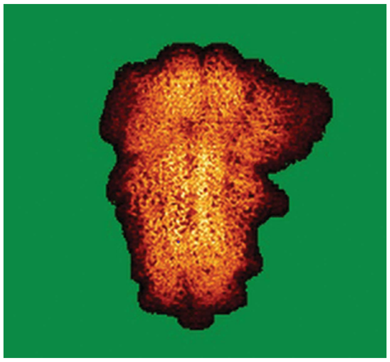
Keywords: cryo-EM; validation; archiving; Electron Microscopy Data Bank; Protein Data Bank.

Validation analysis of EMDB entries

Zhe Wang, Ardan Patwardhan and Gerard J. Kleywegt*

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL–EBI), Wellcome Genome Campus, Hinxton CB10 1SD, United Kingdom. *Correspondence e-mail: gerard@ebi.ac.uk

The Electron Microscopy Data Bank (EMDB) is the central archive of the electron cryo-microscopy (cryo-EM) community for storing and disseminating volume maps and tomograms. With input from the community, EMDB has developed new resources for the validation of cryo-EM structures, focusing on the quality of the volume data alone and that of the fit of any models, themselves archived in the Protein Data Bank (PDB), to the volume data. Based on recommendations from community experts, the validation resources are developed in a three-tiered system. Tier 1 covers an extensive and evolving set of validation metrics, including tried and tested metrics as well as more experimental ones, which are calculated for all EMDB entries and presented in the Validation Analysis (VA) web resource. This system is particularly useful for cryo-EM experts, both to validate individual structures and to assess the utility of new validation metrics. Tier 2 comprises a subset of the validation metrics covered by the VA resource that have been subjected to extensive testing and are considered to be useful for specialists as well as nonspecialists. These metrics are presented on the entry-specific web pages for the entire archive on the EMDB website. As more experience is gained with the metrics included in the VA resource, it is expected that consensus will emerge in the community regarding a subset that is suitable for inclusion in the tier 2 system. Tier 3, finally, consists of the validation reports and servers that are produced by the Worldwide Protein Data Bank (wwPDB) Consortium. Successful metrics from tier 2 will be proposed for inclusion in the wwPDB validation pipeline and reports. The details of the new resource are described, with an emphasis on the tier 1 system. The output of all three tiers is publicly available, either through the EMDB website (tiers 1 and 2) or through the wwPDB ftp sites (tier 3), although the content of all three will evolve over time (fastest for tier 1 and slowest for tier 3). It is our hope that these validation resources will help the cryo-EM community to obtain a better understanding of the quality and of the best ways to assess the quality of cryo-EM structures in EMDB and PDB.



1. Introduction

Structural biology has been revolutionized by cryogenic electron microscopy techniques (cryo-EM), which produce Coulomb potential maps of biomolecules, complexes and assemblies that have often proved difficult to resolve by X-ray crystallography (Kühlbrandt, 2014). Improvements in microscopy, detection and computational methods have in favourable cases even enabled atomic resolution (~1.2 Å) structure determination (Yip *et al.*, 2020; Nakane *et al.*, 2020). These developments have led to a rapid increase in the number of cryo-EM structures being determined and deposited in the Electron Microscopy Data Bank (EMDB; Tagari *et al.*, 2002; Lawson *et al.*, 2016), as shown in Fig. 1, and in many cases allow atomic models to be built and deposited in the Protein Data Bank (PDB; wwPDB Consortium, 2019). While single-particle averaging (SPA) methods provide the bulk of these structures, there is an increasing interest in structure determination *in situ.* Here, multiple copies of an object of interest (for example ribosomes, virus particles or nuclear pore complexes) are identified in and extracted from cryogenic electron tomograms and then improved in an iterative process of superposition and averaging (Briggs, 2013; Schur *et al.*, 2015).

As always with structure determination based on experimental data, there is a need to validate the data, the model constructed based on the data and the fit of the model and the data (Kleywegt, 2000; Gore *et al.*, 2017). Fortunately, the cryo-EM field can build on the wide experience with validation methods gained by protein crystallographers over the past three decades, which has resulted in community-wide agreement regarding sensible validation checks to apply to new X-ray structures upon deposition in the PDB (Read *et al.*, 2011). Much of the software and many of the methods developed, in particular for model validation [for example *O* (Jones *et al.*, 1991), *WHAT_CHECK* (Hooft *et al.*, 1996) and *MolProbity* (Davis *et al.*, 2007; Williams *et al.*, 2018)], can also be used for models derived by other methods such as cryo-EM (Henderson *et al.*, 2012) and NMR spectroscopy (Montelione *et al.*, 2013). Such methods assess the adherence of models to known geometric, physical, stereochemical and conformational criteria (for example bond lengths and angles, non-bonded distances, chirality of amino acids and nucleic acids, and preferred combinations of main-chain and side-chain torsion angles). However, because of the different nature of the underlying experimental data, methods for data validation and model/data-fit assessment in cryo-EM have to be developed from scratch, although some methods used in the X-ray field can be adapted (for example various residue-based model/data-fit criteria; Lawson *et al.*, 2021). In cryo-EM, the raw experimental data are only available for a minority of structures (in the EMPIAR archive; Iudin *et al.*, 2016) and there are no established methods to assess the fit of a model to the underpinning raw data. Hence, in practice the quality is assessed instead of the map derived from the raw data as well as of the fit between the model and that map.

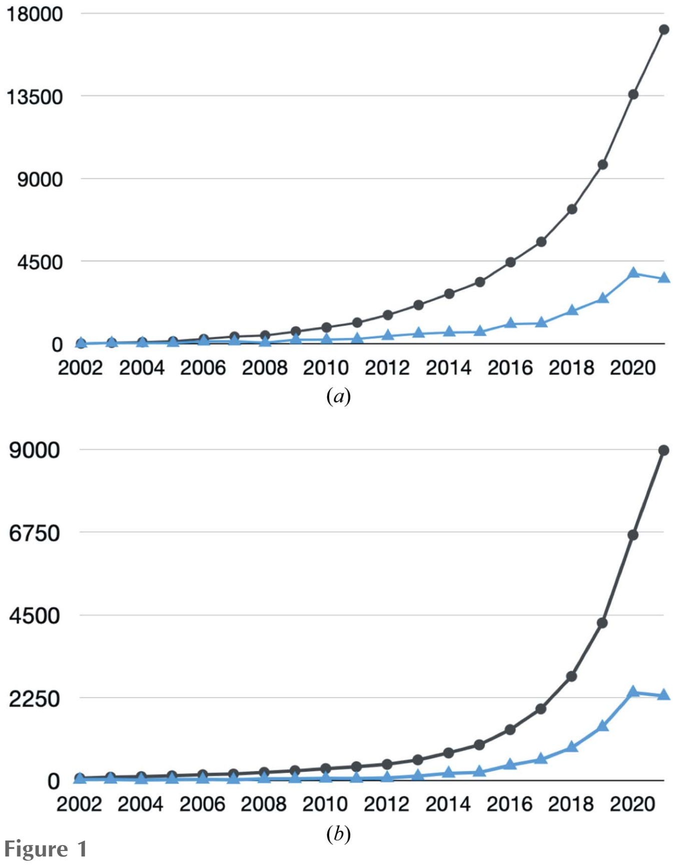


Figure 1

Trends in archiving of cryo-EM structures. (*a*) The number of released EMDB entries per year (blue) and the cumulative number of entries (black) as of October 2021 (data from https://www.ebi.ac.uk/emdb/statistics). (*b*) The annual (blue) and cumulative (black) number of EM-based structures in the PDB as a function of year as of October 2021 (data from https://www.rcsb.org/stats/growth/growth-em).

The EMDB and Worldwide Protein Data Bank Consortium (wwPDB; Berman *et al.*, 2003) established a Validation Task Force for cryo-EM (EM-VTF) in 2010, which published its recommendations two years later (Henderson *et al.*, 2012). Inspired by this, EMDB developed so-called Visual Analysis pages for every EMDB entry (Lagerstedt *et al.*, 2013). These pages contained such elements as orthogonal projections of maps, orthogonal surface views of maps and models, atom-inclusion plots, FSC curves *etc.* However, both the first EM-VTF meeting and the Visual Analysis functionality predated the 'resolution revolution' (Kühlbrandt, 2014), and with the rapidly increasing number of moderate- to high-resolution cryo-EM structures being determined there was a need to reassess the state of the field and to update the recommendations regarding the archiving and validation of these structures. This happened at the wwPDB single-particle cryo-EM data-management workshop in early 2020 (Kleywegt *et al.*, to be published), again organized jointly by EMDB and wwPDB. This meeting resulted in a large number of recommendations regarding the validation criteria that should be used by the archives, and also identified a number of areas where there is no community agreement yet regarding which method or software is best suited to validate a certain aspect of the map or map/model fit (Kleywegt *et al.*, to be published).

Prior to the wwPDB single-particle cryo-EM data-management workshop, refactoring of the Visual Analysis functionality had begun to use up-to-date technologies and add some functionality. The new system is referred to as the Validation Analysis (VA) resource (https://emdb-empiar.org/va), and it continues to present its results on a single webpage for each EMDB entry. Following the recommendations of the 2020 workshop, we have developed a three-tiered strategy to implement, test and disseminate validation results on cryo-EM structures (with or without a model): a development version (tier 1, the VA resource), a production version (tier 2, part of the EMDB entry pages) and a version incorporated into the wwPDB validation-pipeline (tier 3).

The VA resource (tier 1) is the full development version, aimed at specialist users, offering a rich and evolving selection of validation data. New validation methods are implemented here first and run on the entire archive. This enables archive-wide analyses to assess the usefulness, robustness, reliability

Table 1
Description of data types that can be deposited in EMDB.

Category (number)	Explanation or examples
Primary map or tomogram (or tomogram average) (1): mandatory	EM map or tomogram that is shown and described in the associated literature publication
Raw map (0–1)†	Unfiltered, unsharpened and unmasked raw map on which the primary map is based
Half-maps (0 or 2)‡	Unfiltered, unsharpened and unmasked raw half-maps for single-particle analysis, single-particle-based helical reconstructions or subtomogram averaging
Masks (0 or more)	Primary/raw map masks, segmentation/focused refinement masks and half-map masks
Segmentations (0 or more)§	Segmentations of the map or tomogram (3D regions of interest; note that *masks* are computational tools, whereas *segmentations* embody interpretation in terms of biological entities/objects)
Additional cryo-EM maps (0 or more)	Examples include difference maps, maps showing alternative conformations and/or compositions, maps with different processing (for example filtering, sharpening and masking)
Auxiliary files (0 or more)	Examples include author-determined FSC curves (half-maps, map–model, . . .), structure factors and layer lines

† Raw maps can currently be deposited as unspecified ‘additional maps’. In the future it will become possible to deposit them as a separate type of map. ‡ As of February 2022, the deposition of half-maps is mandatory for some modalities if they were used to estimate resolution. § This will become possible in the future.

etc. of these methods and allows individual specialists to assess how they perform in specific cases. An example URL for a VA page (for EMDB entry EMD-11145; Toelzer *et al.*, 2020) is https://www.ebi.ac.uk/emdb/va/EMD-11145. Once community consensus has been reached regarding the utility of certain validation criteria, these can be added to the production version (tier 2). On the other hand, if there are questions regarding the soundness of any criterion, we may either continue to expose it in tier 1 to give the community more time to experiment with and assess it, or if the concerns are major we may drop it from tier 1 altogether.

Tier 2 is a scaled-down version of the VA resource containing validation components that are well tested and generally considered to be valuable, robust, informative and well understood. These pages are accessible in a separate tab (labelled Validation) of the EMDB entry pages for each EMDB structure. The tier 2 page for the same example entry as before can be found at https://www.ebi.ac.uk/emdb/EMD-11145?tab=validation.

Finally, following agreement with wwPDB (as of 1 January 2021 EMDB is a part of the wwPDB Consortium), some of the most informative criteria will be implemented in tier 3, the validation pipeline that is part of the OneDep deposition, annotation and validation system (Young *et al.*, 2017) and its validation servers. Any validation components in tier 3 are thus applied to every newly deposited cryo-EM volume in EMDB and accompanying model in the PDB, and these reports are also made available for the entire EMDB archive and all EM structures in the PDB. Many of the recommendations from the wwPDB single-particle cryo-EM data-management workshop have already been implemented in one or more of the validation tiers.

Here, we describe the current state of the VA resource (tier 1). Some of its elements are also part of the production version (tier 2) and even the OneDep validation pipeline (tier 3).

2. Validation Analysis resource

It is useful to understand what types of data may be archived as part of a single EMDB entry; this is summarized in Table 1.

Table 2
Validation components available as part of the three tiers of the EMDB validation resource (see text for details).

Note that the tier 3 functionality is a subset of the tier 2 functionality, which in turn is a subset of the tier 1 components. (Yes) indicates validation components that have been integrated into the VA pipeline but whose results are not yet shown on the tier 1 webpages.

Components	Tier 1	Tier 2	Tier 3
Orthogonal projections	Yes	Yes	Yes
Orthogonal maximum-value projections	Yes	Yes	
Orthogonal standard-deviation projections	Yes	Yes	
Central slices	Yes	Yes	Yes
Largest variance slices	Yes	Yes	Yes
Orthogonal surface views	Yes	Yes	Yes
Map–model overlay views	Yes	Yes	Yes
Assembly map–model overlay views	Yes	Yes	Yes
Map and mask overlay views	Yes	Yes	Yes
Atom-inclusion plot	Yes	Yes	Yes
Residue-inclusion analysis	Yes	Yes	Yes
Voxel-value distribution	Yes	Yes	Yes
Volume-estimation graph	Yes	Yes	Yes
RAPS plot	Yes	Yes	Yes
Half-map FSC plot	Yes	Yes	Yes
Map–model FSC plot	Yes		
Q-score	Yes		
3*D-Strudel*	Yes		
SMOC	(Yes)		
CCC	(Yes)		
3*DFSC*	(Yes)		
EMRinger	(Yes)		

EMDB accommodates several cryo-EM modalities including SPA, tomography, subtomogram averaging, helical reconstruction and electron diffraction. Any accompanying models must be deposited in the PDB, while raw data (micrographs, particle stacks, tilt series *etc.*) can optionally be deposited in EMPIAR (Iudin *et al.*, 2016). The exact information shown in the VA resource for a specific entry depends on the modality and the presence or not of one or more models, masks and segmentations. Note that pure model validation is not included; for this, users are referred to the wwPDB validation reports (Gore *et al.*, 2017).

The VA page for an entry in the most general case (SPA, with a model, half-maps, masks *etc.*) will include the sections shown as examples in Fig. 2. The full contents are discussed in more detail below. Table 2 summarizes which validation components are currently implemented in each of the three tiers.

2.1. 3D volume analysis

2.1.1. 2D slices through and projections of the 3D volume. This section allows the inspection of internal details of a map or tomogram and may also be useful for identifying any artefacts (for example streaking or 'imprints' of masks used during processing). All slices and projections are calculated and shown along the orthogonal X, Y and Z axes. In addition to standardized views of 300 × 300 pixels (1200 × 1200 for tomograms), versions with the native pixel dimensions are also provided (accessible by clicking on the standardized images). Images included are the following.

(i) Central slices, showing the central plane in each direction.

(ii) Largest-variance slices, showing the plane (in each direction) whose voxel values have the highest variance. The rationale for this is that these might be 'interesting' planes.

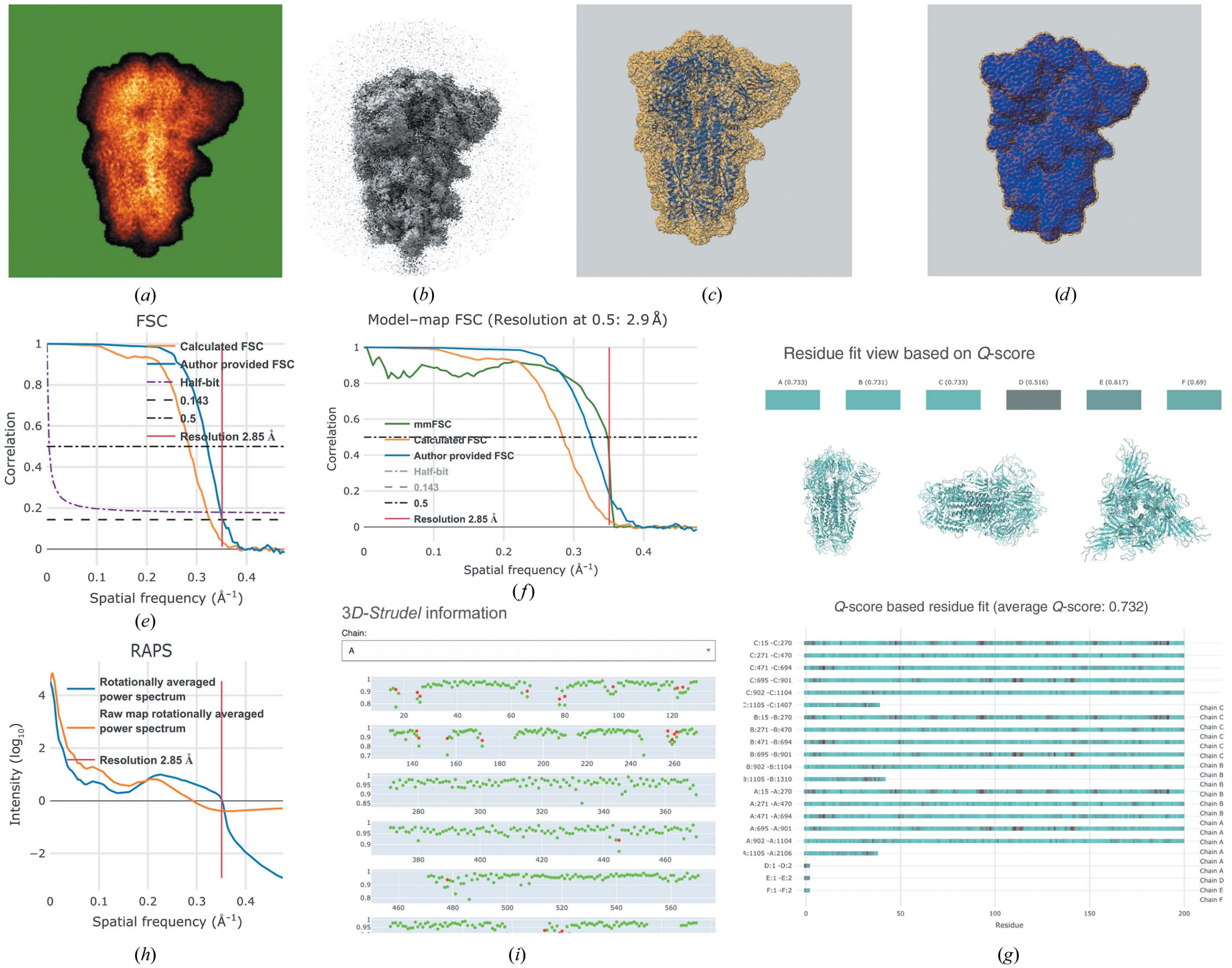


Figure 2
Examples of validation components in VA pages for EMDB entries (see text for details). The examples are for EMDB entry EMD-11145, which is a SARS-CoV-2 spike protein (Toelzer *et al.*, 2020). (*a*) Orthogonal standard-deviation projection (shown with false colour) along the X axis. The lowest values are mapped to solid green, which may help to identify if and how masking was applied. (*b*) Orthogonal surface view of the raw map (calculated by averaging the two half-maps) along the X axis. (*c*) Overlay of the primary map (yellow, semi-transparent) and fitted model (blue) viewed along the X axis. (*d*) Overlay of the primary map (yellow, semi-transparent) with a mask (blue) viewed along the X axis. (*e*) FSC plot combining the author-provided curve (cyan) and that calculated from the two half-maps (orange) as well as various criteria used by the community to estimate resolution. The red vertical line indicates the resolution reported by the author. (*f*) Map–model FSC plot. FSC values are calculated between the primary map and a map calculated from the model using *REFMAC* (Murshudov *et al.*, 2011). (*g*) Overview of map/model-fit analysis using the Q-score (Pintilie *et al.*, 2020) both on a per-residue basis (3D view and lower part of the plot) and averaged per chain (coloured blocks at the top). The overall average score is shown in the legend (0.732 in this case). Colour ramping is used, with cyan indicating higher Q-scores (better fit). (*h*) RAPS plot of the primary (cyan) and raw (orange) maps. The red vertical line indicates the author-provided resolution of the map. (*i*) Overview of map–model fit analysis using 3*D-Strudel* (Istrate *et al.*, 2021). Each row of the plot shows the 3*D-Strudel* score for up to 100 residues. Red dots indicate residues where the map motif of a different amino-acid type to that in the model is most similar to the local map.

(iii) Orthogonal projections, obtained by averaging voxel values along one axis for all pixels in the plane orthogonal to that axis (for example the average value of all voxels along Z for each pixel in the X, Y plane).

(iv) Standard-deviation projections, calculated in the same way but showing the standard deviation rather than the average value along each projection axis.

(v) Maximum-value projections, calculated in the same way but showing the maximum value along each projection axis.

All images are shown in grayscale rendering, but the maximum-value and standard-deviation projections are also shown using false colour (Fig. 2*a*) using the 'glow' lookup table (https://github.com/fiji/fiji/blob/master/luts/glow.lut) from *Fiji* (Schindelin *et al.*, 2012; Schneider *et al.*, 2012). This occasionally reveals features that are impossible or difficult to assess from a grayscale image (for example 'ghosts' of masks used during processing). If half-maps are available, a raw map will be computed by averaging them, and several of the above images will also be provided for that map. To make these images comparable to those derived from the primary map, the raw map is scaled to have the same average value and standard deviation.

2.1.2. 3D views

Except for tomograms, where there is no sensible contour level, a number of orthogonal surface views (along the X, Y and Z axes) are generated using *ChimeraX* (Goddard *et al.*, 2018; Pettersen *et al.*, 2021), including the following.

(i) The surface of the primary map, using the recommended (usually by the depositor) contour level.

(ii) The surface of the raw map (Fig. 2*b*). As there is no recommended contour level for the computed raw map, it will be approximated as the level at which its surface encompasses the same volume as the primary map does at its recommended contour level. Although not always perfect, this provides a simple and consistent choice of contour level.

(iii) Surface views of each mask superimposed on the primary map (which is shown as a semi-transparent surface; Fig. 2*d*).

In the future, EMDB will add support for segmentations and these will then be shown in the same fashion as masks.

2.1.3. Plots and graphs

Some plots and graphs are calculated both for maps and tomograms, whereas others can only be meaningfully computed for maps. They include the following.

(i) Voxel-value histogram. The range of voxel values in a map or tomogram is divided into 128 bins and the number of voxels in each bin is plotted on a log scale. A high peak near zero is indicative of masking. If the raw map is available (see above), its voxel-value distribution will be shown in the same plot.

(ii) Enclosed-volume estimate. The volume of space enclosed by a map (in nm^3) is plotted as a function of contour level. The volume of the imaged object is estimated using the recommended contour level (shown as a vertical line). If an estimate of the molecular weight (MW) is available, then the predicted volume (using an average density of $1.5\ \mathrm{g\ cm}^{-3}$) is also indicated (by a horizontal line). Ideally, the curve and both lines intersect, but there can be many causes for this not being the case; for example, the MW may have been provided for a single unit which is repeated many times in the sample, or conversely for a larger unit than is in the map (for example an entire fibre). Other factors that can contribute include staining, heavier or lighter than average material in the sample or an inaccurate contour level.

(iii) Rotationally averaged power spectrum (RAPS; Crowther & Amos, 1971; Rosenthal & Henderson, 2003) of the primary and raw maps (if available), *i.e.* the distribution of intensity (on a log scale) versus spatial frequency (Fig. 2*h*). This may provide insight into the data-processing steps leading to the deposited primary map, in terms of CTF (contrast-transfer function) correction, low- or high-pass filtering, masking artefacts and temperature-factor corrections.

(iv) FSC curves (Fig. 2*e*). The FSC curve between two maps calculated from independently processed half-data sets is often used to provide an estimate of the resolution limit to which both maps are still correlated (Saxton & Baumeister, 1982; van Heel & Stöffler-Meilicke, 1985; Harauz & van Heel, 1986). Here, the FSC curve calculated from the half-maps is plotted as well as that provided by the authors (if available). Curves for several resolution-estimation criteria are also shown, as is the author-provided resolution estimate.

2.1.4. Symmetry analysis

Symmetry analysis of a map can be useful to check that the user-supplied point-group symmetry information is correct and that standard symmetry conventions for different point groups have been correctly followed. The VA resource uses *ProSHADE* (Nicholls *et al.*, 2018) for this purpose. It produces a list of symmetries detected in the map, including the symmetry axes and a score (values smaller than ~0.95 typically indicate partial or pseudo-symmetry), a list of all symmetry elements and finally a list of alternative symmetries (often subgroups of the main symmetries). If symmetry was applied (and reported) this will be shown for comparison.

2.2. Map plus model assessment

There may be one, multiple or no PDB entries associated with an EMDB entry. If there are one or more models, then the sections below are provided separately for each of them.

2.2.1. 3D views (Fig. 2c)

The model is shown as a blue ribbon superimposed on a semi-transparent rendering of the primary map and viewed along the X, Y and Z axes. If point-symmetry information is available for a model (for example for viruses), a fully assembled model will also be generated and shown separately, also overlaid on the primary map.

2.2.2. Atom inclusion

An atom-inclusion graph shows the fraction of atoms that lie entirely inside the map as a function of contour level (separately for all atoms and just the backbone atoms). This plot may reveal if an inappropriate contour level was selected or if the side chains tend not to be well contained in the map, for example in a low-resolution study.

2.2.3. Map/model-fit analyses

At present this includes information about residue inclusion (Lagerstedt *et al.*, 2013), *Q*-score (Pintilie *et al.*, 2020), 3*D-Strudel* (Istrate *et al.*, 2021) and model–map FSC (van Heel *et al.*, 2000), but this section will be expanded considerably with the inclusion (in 2022) of additional methods such as *EMRinger* (Barad *et al.*, 2015), *SMOC* (Joseph *et al.*, 2016), *CCC* (Warshamanage *et al.*, 2022) and 3*DFSC* (Tan *et al.*, 2017).

(i) Residue-inclusion analysis (Lagerstedt *et al.*, 2013). This is similar to the atom-inclusion metric discussed above but is calculated on a per-residue basis (amino-acid residue, nucleotide or ligand). An interactive viewing panel shows the score for every residue in each molecule (in batches of 200 residues in the case of longer chains). The colour of each residue depends on its inclusion score and varies from red (zero inclusion) to cyan (all atoms inside the map at the recommended contour level). In addition, three orthogonal views of the model, coloured the same way as in the panel, are shown. A significant weakness of atom- and residue-inclusion scores is that they depend directly on the selected contour level. This can be chosen to be unrealistically low so as to obtain high inclusion scores, which is clearly undesirable.

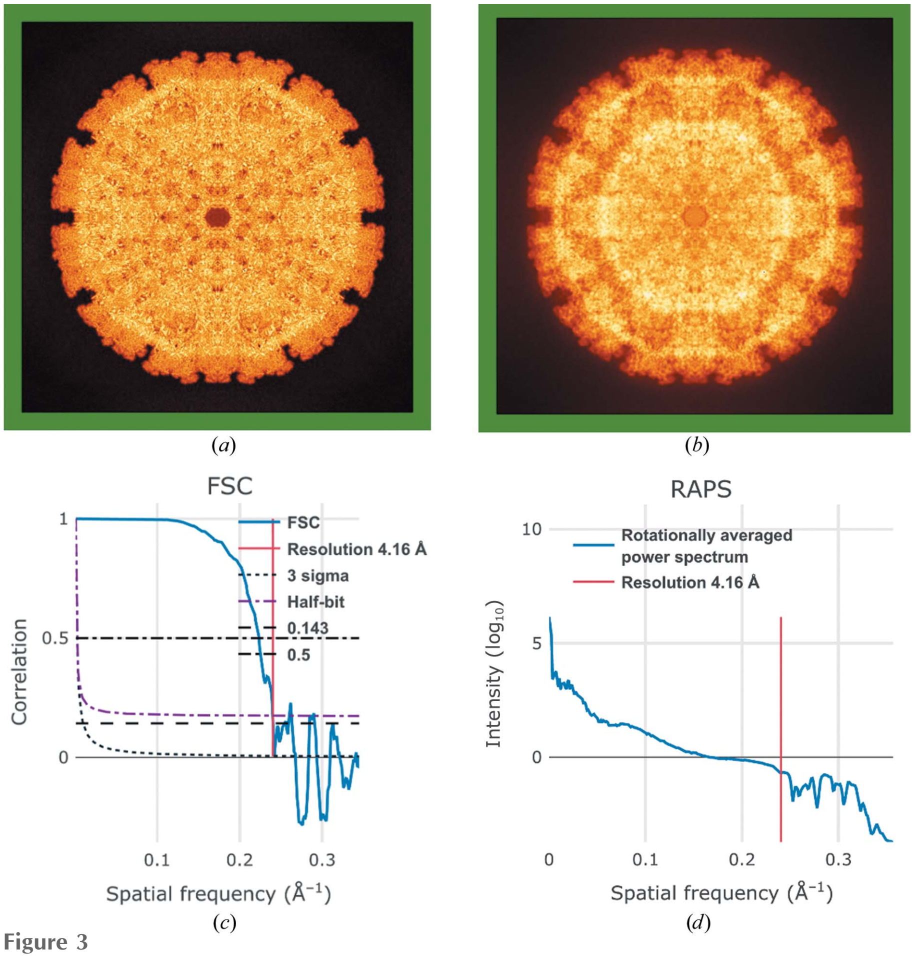
(ii) *Q*-score (Pintilie *et al.*, 2020; Fig. 2*g*). A more recent method, which has been calibrated using high-resolution structures and does not require any subjective parameter choices, is the *Q*-score, a quantitative metric of resolvability in a cryo-EM map. It is calculated for individual atoms and then averaged, for example for a residue, molecule or complex. The presentation of the *Q*-score analysis is identical to that of the residue-inclusion analysis.

(iii) Map–model FSC (van Heel *et al.*, 2000; Fig. 2*f*). The Fourier-shell correlation is calculated and plotted between the primary map and a simulated map based on the model. The graph shows the resolution limit to which the model explains the map (using a cutoff value of 0.5 since the model is not independent of the map).

(iv) 3*D-Strudel* score (Istrate *et al.*, 2021; Fig. 2*i*). 3*D-Strudel* is a model-based map-validation method which entails calculating the linear correlation coefficient between an amino-acid- and rotamer-specific map motif from the 3D-Strudel library and the cryo-EM map values around a residue of interest (after optimal superposition). The 3D-Strudel motif library was mined from EMDB maps by averaging large numbers of map fragments that correspond to each specific amino-acid rotamer in a number of resolution bands.

As explained in Section 1, the validation resources will evolve and will contain different components depending on the imaging modality and the tier. New components will be added to tier 1 (the EMDB VA resource) and, after extensive testing and evaluation by experts, some of them will be incorporated into tier 2 (EMDB entry pages). Finally, following broad community consensus and approval by wwPDB, some metrics will be included in tier 3, *i.e.* the wwPDB validation software pipeline, reports and servers. In the case of tier 1, some components may also be removed if they turn out to be less useful, informative or consistent, or if they provide information that is essentially redundant with other methods that are already incorporated.

Tier 1, 2 and 3 validation information is made available for every cryo-EM structure upon its release in EMDB (and PDB for models) and is therefore always complete and up-to-date. To reflect the evolving functionality of tier 1, its pipeline is occasionally re-run on the entire EMDB archive to make the information for all entries up to date.


(*a*)

(*b*)

(*c*)

(*d*)

Figure 3
Examples of artefacts in the FSC and RAPS plots possibly due to masking or interpolation effects (see text for details). (*a*) Maximum-value projection along the *X* axis of a map (shown using false colour). (*b*) Standard-deviation projection along the *X* axis of the same map (also using false colour). (*c*) FSC plot and (*d*) RAPS plot of the same map, revealing unexpected oscillations beyond the resolution cutoff.

3. Example applications

Validation-analysis components can play different roles in validating maps and any fitted models. Below we discuss some examples of issues that can be detected and diagnosed. Ideally, such issues should be addressed before deposition takes place. The depositors receive, and must approve, a preliminary validation report before being able to submit, and they are strongly recommended to run the wwPDB validation server before even commencing a deposition. The validation reports are also used by wwPDB biocurators to check entries and flag issues to the depositor prior to public release. Despite the best efforts, occasionally an issue slips through the net into the public archive. The VA resource then plays a key role for EMDB staff and the cryo-EM community to check entries and rectify them after release. A few examples of this are discussed below.

Fig. 3 shows an example of a map which has good internal detail but has artefacts in the FSC and RAPS plots (correlation and intensity oscillations, respectively) which would merit following up by the depositors. They could be due to the application of a hard mask in the periphery of the map or to some other interpolation effect.

If there is a fitted model, the various visualizations of the map and model will provide some understanding of how well the model explains the map, allowing several fairly trivial issues to be identified and diagnosed. Fig. 4 shows an example in which the map and the model are misaligned. The problem is easy to identify visually from the orthogonal map/model views and the model obviously shows zero atom inclusion. Sometimes the issues are more subtle than in this case, but even then the overlaid images and other map/model analyses often facilitate their detection and diagnosis.

Fig. 5 shows an example where there is a discrepancy in the relative scaling of the map and model. Whereas the determination of the physical scaling of a model is an intrinsic part of the model-determination process, the same is not the case for SPA maps. It is not uncommon for errors in microscope magnification calibration to propagate to the deposited voxel size, leading to such scaling errors.

Upon deposition in EMDB, the depositor provides the recommended contour level for viewing and rendering the primary map. It is also used to assess the fit of the model to the map using the atom- and residue-inclusion scores. Fig. 6 shows what happens if the contour level selected is considerably too high or low. In Fig. 6(*a*) the contour level is set too high so that the map fails to cover the model properly, as reflected in the 3D view and the atom-inclusion plot (the map covers only ~30% of the atoms). In Fig. 6(*b*) the contour level is instead set too low. This leads to a deceptively 'good' atom-inclusion plot (100% of all atoms inside the map), but clearly the map looks unrealistic in the 3D view. The *Q*-score and 3*D-Strudel* measures are independent of the choice of contour level and hence do not suffer from inappropriate choices.

Fig. 7 shows examples of two of the issues that can be identified with 3*D-Strudel*. In one case a phenylalanine residue has been modelled with an incorrect side-chain conformation that does not fit the map (also indicated by a low *Q*-score value of 0.31). In the other case, a stretch of residues has been built in a location where there is no support for them in the map. Again, low *Q*-scores (between −0.18 and 0.27) confirm that there are issues with this part of the model. Both 3*D-Strudel* and *Q*-score are independent of contour level, so these

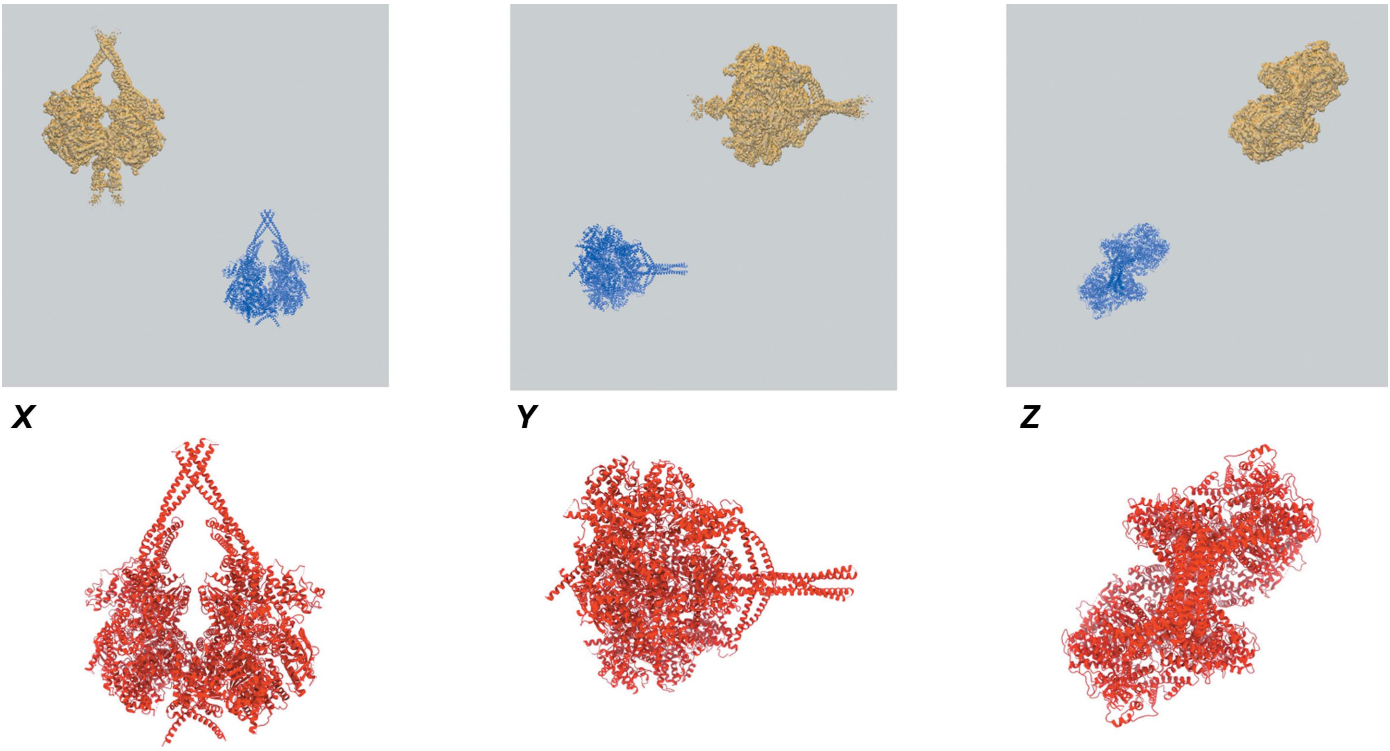


Figure 4
Example of a map-origin error, leading to a misaligned (and in this case completely non-overlapping) map and model. The top row shows the map–model overlay views and the bottom row shows the model coloured by residue-inclusion score (viewed along the same axes), where the red colour of the entire model reflects the fact that not a single residue is inside the map.

are not issues that are caused by an incorrect choice of this level. *3D-Strudel* can in addition be useful in detecting sequence-register errors between the map and model (not shown; see Istrate *et al.*, 2021).

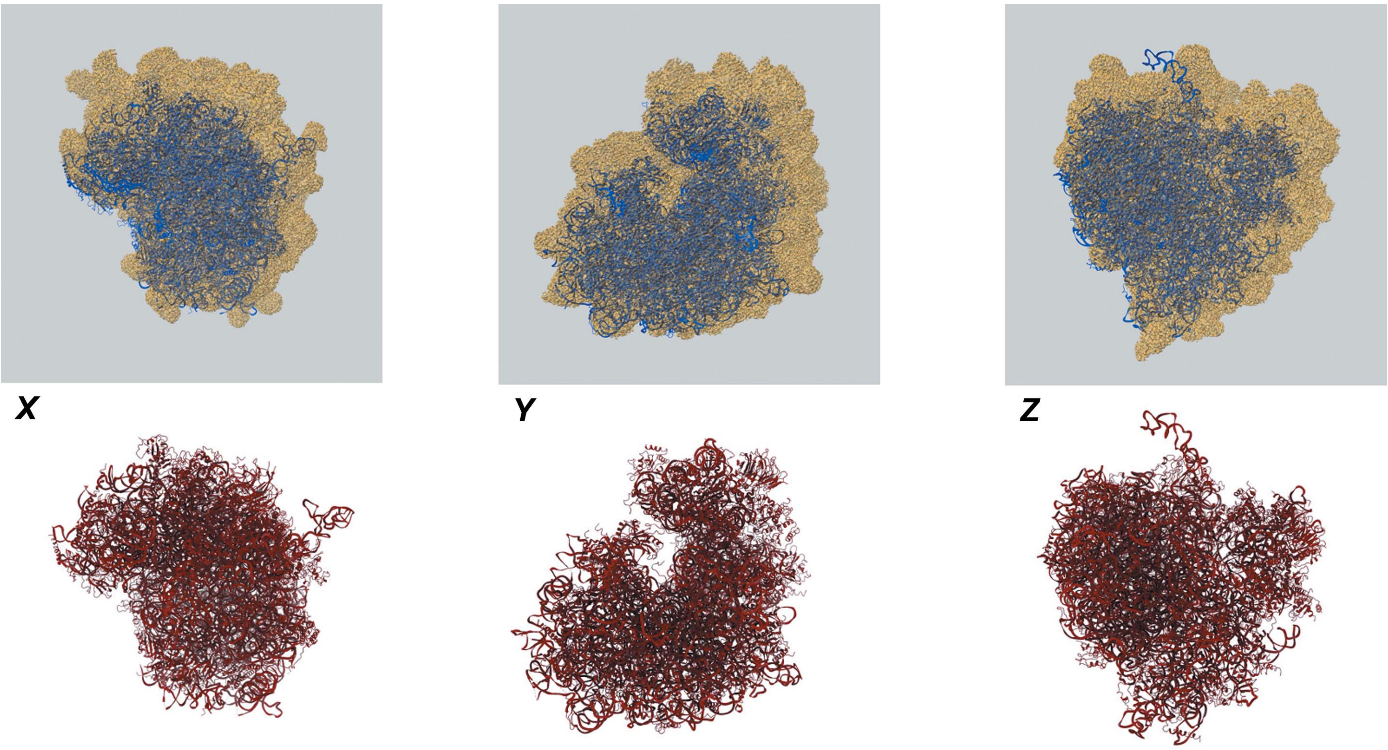


Figure 5
Example of a pixel-sampling error, leading to an incorrect relative scaling of the map and model. The top row shows the map–model overlay views and the bottom row shows the model coloured by residue Q-score (viewed along the same axes).

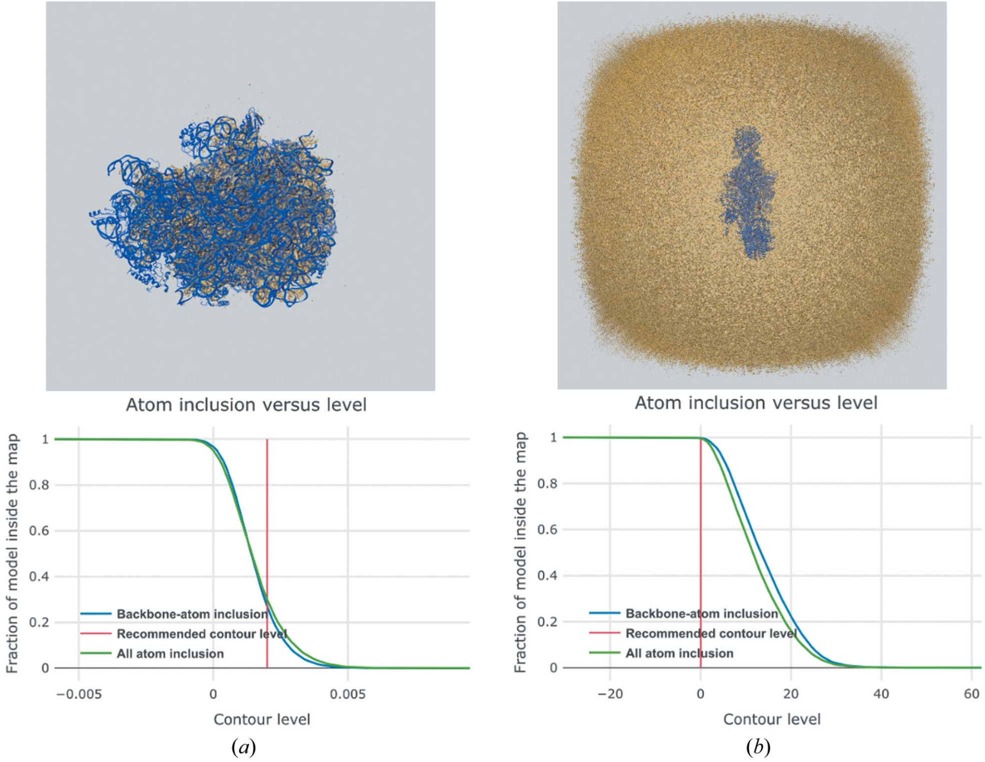


Figure 6
Effects of an inappropriate choice of contour level (see text for details). (*a*) The atom-inclusion graph and map–model overlay view for a case where the recommended contour level has been set too high. (*b*) The same components shown for a case where a contour level has been chosen that is too low.

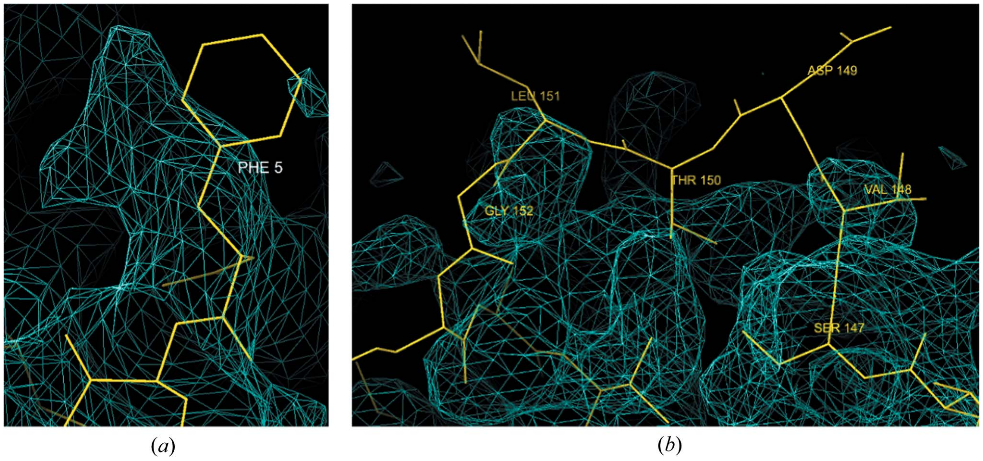


Figure 7
Examples of two map/model fit issues highlighted by *3D-Strudel* (see text for details). (*a*) A phenylalanine residue that has been modelled with an incorrect side-chain conformation. (*b*) A stretch of residues (147–153) that have been built in the absence of supporting map features.

4. Conclusions and plans

The primary goal of this work is to facilitate the assessment and validation of cryo-EM data (maps, tomograms and models) by specialists and nonspecialists alike, through the use of intuitive 2D and 3D visualizations and well understood metrics, plots and graphs. This is achieved in particular with the validation tools in tiers 2 and 3 of our strategy. Tier 1 provides additional validation information suitable for specialists but also has a partly different purpose, namely to allow experimentation at scale with new validation approaches and metrics. For some validation tasks (for example, the assessment of local resolution or map/model fit analysis) there are multiple competing methods whose relative merits are not yet understood. For example, we plan to integrate a number of tools to calculate local resolution, including *ResMap* (Kucukelbir *et al.*, 2014), *BlocRes* (*Bsoft*; Cardone *et al.*, 2013), *MonoRes* (Vilas *et al.*, 2018) and *DeepRes* (Ramírez-Aportela *et al.*, 2019), as well as tools to assess anisotropy and angular coverage such as *CryoEF* (Naydenova & Russo, 2017) and *MonoDIR* (Vilas *et al.*, 2020). Archive-wide analysis and comparison of such metrics, as well as inspection of the results for individual structures by experts, will hopefully bring more clarity and eventual community consensus about which tools are most appropriate to apply for a certain validation task. We are also developing a server with which the tier 1 pipeline can be run on data and models that are not yet in the public archives, and by incorporating the VA pipeline into the *CCP-EM* package (Wood *et al.*, 2015) this can also be accomplished in-house. EMDB will continue to engage with developers of validation tools and include promising new methods in the VA pipeline.

The cryo-EM community has played a crucial role in this ongoing project. We are grateful to all software providers who have allowed us to use their tools in the pipeline, and the input from the wwPDB single-particle cryo-EM data-management workshop (Kleywegt *et al.*, to be published) has been invaluable in guiding our work on all three tiers of the validation resources. The workshop focused exclusively on the validation of molecular maps and models, and these are therefore covered most extensively at present. However, many of the 3D volume analyses are also applied to tomograms by all three tiers.

We expect that future improvements in EMDB and PDB metadata collection, more explicit file typing, the mandatory deposition of certain types of data *etc.* will also improve our ability to validate cryo-EM structural data. Simultaneously, the field is still experiencing rapid growth and expansion as well as the development of new or improved approaches to structure determination and analysis, many of which will impact the way in which we validate the data. We look forward to continuing to address all of these challenges in collaboration with the cryo-EM community.

5. Software dependencies and availability

The VA pipeline is a Python-based program that uses several standard Python packages (NumPy, SciPy). In addition, there are several external dependencies for EM-specific calculations, including *CCP-EM* (Wood *et al.*, 2015; Burnley *et al.*, 2017), *TEMPy* (Joseph *et al.*, 2017), *REFMAC* (Murshudov *et al.*, 2011), *ProSHADE* (Nicholls *et al.*, 2018), *EMDA* (Warshamanage *et al.*, 2022), *3D-Strudel* (Istrate *et al.*, 2021), *Q*-score (Pintilie *et al.*, 2020) and *ChimeraX* (Goddard *et al.*, 2018). Several other metrics have already been integrated but their results are not yet presented on the VA webpages. These include *CCC* (Warshamanage *et al.*, 2022), *3DFSC* (Tan *et al.*, 2017), *SMOC* (Joseph *et al.*, 2016) and *EMRinger* (Barad *et al.*, 2015), which in turn relies on *cctbx* (Grosse-Kunstleve *et al.*, 2002).

The VA pipeline will be integrated into the *CCP-EM* package and the VA Python programming package is also available from PyPI (https://pypi.org/project/va/). Tier 3 validation is further accessible as part of the wwPDB validation resources through the OneDep validation server (https://www.wwpdb.org/validation/validation-servers) and the OneDep Python API (https://www.wwpdb.org/validation/onedep-validation-web-service-interface).

Acknowledgements

We are grateful to our EMDB colleagues A. Istrate for many discussions, code reviews and support for *3D-Strudel* integration, O. Salih and J. Turner for providing example applications, R. Pye and S. Abbott for integrating the tier 3 components into the wwPDB validation pipeline and N. Fonseca for help with integrating the tier 2 components into the EMDB entry pages. We thank P. Korir and A. Iudin (EMPIAR) for technical discussions and J. Berrisford (PDBe) for help with the mmCIF format. We thank the wwPDB biocurators and partners for valuable feedback and bug reports. We owe a debt of gratitude to all participants in the 2020 single-particle cryo-EM data-management workshop. Their feedback and recommendations have led to substantial improvements in the contents of all three tiers, and indeed it was their suggestion to develop a three-tiered approach in the first place. We thank the *CCP-EM* staff for software support and ongoing integration of the VA pipeline into *CCP-EM*. We further thank our collaborators in the Wellcome Trust-funded UK EM Validation Network for fruitful discussions: M. Topf and T. Cragnolini (also for help with *TEMPy*); G. Murshudov, R. Warshamanage and M. Tykac (also for help with *REFMAC*, *EMDA*, *CCC* and *ProSHADE*); P. Rosenthal, M. Winn, E. Orlova and A. Roseman. We thank Y. Z. Tan, P. Baldwin and D. Lyumkis for help with the implementation of the *3DFSC* software, the *Phenix* team for software support, J. Fraser and B. Barad for *EMRinger* support, G. Pintilie for help with the *Q*-score software and T. Goddard and G. Couch for help with *ChimeraX*. Open access funding enabled and organized by Projekt DEAL.

Funding information

During this work, ZW was funded by the Wellcome Trust (grant 208398/Z/17/Z to P. Rosenthal). Work on EMDB at EMBL–EBI is further supported by the Wellcome Trust (grant 212977/Z/18/Z to AP and GJK) and EMBL with funding from its member states.

References

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.

Berman, H., Henrick, K. & Nakamura, H. (2003). *Nat. Struct. Biol.* **10**, 980.

Briggs, J. A. G. (2013). *Curr. Opin. Struct. Biol.* **23**, 261–267.

Burnley, T., Palmer, C. M. & Winn, M. (2017). *Acta Cryst.* D**73**, 469–477.

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

Crowther, R. A. & Amos, L. A. (1971). *J. Mol. Biol.* **60**, 123–130.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *Nucleic Acids Res.* **35**, W375–W383.

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.

Gore, S., Sanz García, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Mading, S., Mak, L., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Peisach, E., Sahni, G., Sekharan, M. R., Sen, S., Shao, C., Smart, O. S., Ulrich, E. L., Yamashita, R., Quesada, M., Young, J. Y., Nakamura, H., Markley, J. L., Berman, H. M., Burley, S. K., Velankar, S. & Kleywegt, G. J. (2017). *Structure*, **25**, 1916–1927.

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

Harauz, G. & van Heel, M. (1986). *Optik*, **73**, 146–156.

Heel, M. van, Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. & Patwardhan, A. (2000). *Q. Rev. Biophys.* **33**, 307–369.

Heel, M. van & Stöffler-Meilicke, M. (1985). *EMBO J.* **4**, 2389–2395.

Henderson, R., Sali, A., Baker, M. L., Carragher, B., Devkota, B., Downing, K. H., Egelman, E. H., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S. J., Medalia, O., Penczek, P. A., Rosenthal, P. B., Rossmann, M. G., Schmid, M. F., Schröder, G. F., Steven, A. C., Stokes, D. L., Westbrook, J. D., Wriggers, W., Yang, H., Young, J., Berman, H. M., Chiu, W., Kleywegt, G. J. & Lawson, C. L. (2012). *Structure*, **20**, 205–214.

Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature*, **381**, 272.

Istrate, A., Wang, Z., Murshudov, G. N., Patwardhan, A. & Kleywegt, G. J. (2021). *bioRxiv*, 2021.12.16.472999.

Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. (2016). *Nat. Methods*, **13**, 387–388.

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

Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. (2017). *J. Struct. Biol.* **199**, 12–26.

Joseph, A. P., Malhotra, S., Burnley, T., Wood, C., Clare, D. K., Winn, M. & Topf, M. (2016). *Methods*, **100**, 42–49.

Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.

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

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

Lagerstedt, I., Moore, W. J., Patwardhan, A., Sanz-García, E., Best, C., Swedlow, J. R. & Kleywegt, G. J. (2013). *J. Struct. Biol.* **184**, 173–181.

Lawson, C. L., Kryshtafovych, A., Adams, P. D., Afonine, P. V., Baker, M. L., Barad, B. A., Bond, P., Burnley, T., Cao, R., Cheng, J., Chojnowski, G., Cowtan, K., Dill, K. A., DiMaio, F., Farrell, D. P., Fraser, J. S., Herzik, M. A. Jr, Hoh, S. W., Hou, J., Hung, L.-W., Igaev, M., Joseph, A. P., Kihara, D., Kumar, D., Mittal, S., Monastyrskyy, B., Olek, M., Palmer, C. M., Patwardhan, A., Perez, A., Pfab, J., Pintilie, G. D., Richardson, J. S., Rosenthal, P. B., Sarkar, D., Schäfer, L. U., Schmid, M. F., Schröder, G. F., Shekhar, M., Si, D., Singharoy, A., Terashi, G., Terwilliger, T. C., Vaiana, A., Wang, L., Wang, Z., Wankowicz, S. A., Williams, C. J., Winn, M., Wu, T., Yu, X., Zhang, K., Berman, H. M. & Chiu, W. (2021). *Nat. Methods*, **18**, 156–164.

Lawson, C. L., Patwardhan, A., Baker, M. L., Hryc, C., Garcia, E. S., Hudson, B. P., Lagerstedt, I., Ludtke, S. J., Pintilie, G., Sala, R., Westbrook, J. D., Berman, H. M., Kleywegt, G. J. & Chiu, W. (2016). *Nucleic Acids Res.* **44**, D396–D403.

Montelione, G. T., Nilges, M., Bax, A., Güntert, P., Herrmann, T., Richardson, J. S., Schwieters, C. D., Vranken, W. F., Vuister, G. W., Wishart, D. S., Berman, H. M., Kleywegt, G. J. & Markley, J. L. (2013). *Structure*, **21**, 1563–1570.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P. M. G. E., Grigoras, I. T., Malinauskaite, L., Malinauskas, T., Miehling, J., Uchański, T., Yu, L., Karia, D., Pechnikova, E. V., de Jong, E., Keizer, J., Bischoff, M., McCormack, J., Tiemeijer, P., Hardwick, S. W., Chirgadze, D. Y., Murshudov, G., Aricescu, A. R. & Scheres, S. H. W. (2020). *Nature*, **587**, 152–156.

Naydenova, K. & Russo, C. J. (2017). *Nat. Commun.* **8**, 629.

Nicholls, R. A., Tykac, M., Kovalevskiy, O. & Murshudov, G. N. (2018). *Acta Cryst.* D**74**, 492–505.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H. & Ferrin, T. E. (2021). *Protein Sci.* **30**, 70–82.

Pintilie, G., Zhang, K., Su, Z., Li, S., Schmid, M. F. & Chiu, W. (2020). *Nat. Methods*, **17**, 328–334.

Ramírez-Aportela, E., Mota, J., Conesa, P., Carazo, J. M. & Sorzano, C. O. S. (2019). *IUCrJ*, **6**, 1054–1063.

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.

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

Saxton, W. O. & Baumeister, W. (1982). *J. Microsc.* **127**, 127–138.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P. & Cardona, A. (2012). *Nat. Methods*, **9**, 676–682.

Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. (2012). *Nat. Methods*, **9**, 671–675.

Schur, F. K. M., Dick, R. A., Hagen, W. J. H., Vogt, V. M. & Briggs, J. A. G. (2015). *J. Virol.* **89**, 10294–10302.

Tagari, M., Newman, R., Chagoyen, M., Carazo, J. M. & Henrick, K. (2002). *Trends Biochem. Sci.* **27**, 589.

Tan, Y. Z., Baldwin, P. R., Davis, J. H., Williamson, J. R., Potter, C. S., Carragher, B. & Lyumkis, D. (2017). *Nat. Methods*, **14**, 793–796.

Toelzer, C., Gupta, K., Yadav, S. K. N., Borucu, U., Davidson, A. D., Kavanagh Williamson, M., Shoemark, D. K., Garzoni, F., Staufer, O., Milligan, R., Capin, J., Mulholland, A. J., Spatz, J., Fitzgerald, D., Berger, I. & Schaffitzel, C. (2020). *Science*, **370**, 725–730.

Vilas, J. L., Gómez-Blanco, J., Conesa, P., Melero, R., de la Rosa-Trevín, J. M., Otón, J., Cuenca, J., Marabini, R., Carazo, J. M., Vargas, J. & Sorzano, C. O. S. (2018). *Structure*, **26**, 337–344.

Vilas, J. L., Tagare, H. D., Vargas, J., Carazo, J. M. & Sorzano, C. O. S. (2020). *Nat. Commun.* **11**, 55.

Warshamanage, R., Yamashita, K. & Murshudov, G. N. (2022). *J. Struct. Biol.* **214**, 107826.

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, J. S. (2018). *Protein Sci.* **27**, 293–315.

Wood, C., Burnley, T., Patwardhan, A., Scheres, S., Topf, M., Roseman, A. & Winn, M. (2015). *Acta Cryst.* D**71**, 123–126.

wwPDB Consortium (2019). *Nucleic Acids Res.* **47**, D520–D528.

Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. (2020). *Nature*, **587**, 157–161.

Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., Sen, S., Gutmanas, A., Armstrong, D. R., Berrisford, J. M., Chen, L., Chen, M., Di Costanzo, L., Dimitropoulos, D., Gao, G., Ghosh, S., Gore, S., Guranovic, V., Hendrickx, P. M. S., Hudson, B. P., Igarashi, R., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Liang, Y., Mading, S., Mak, L., Mir, M. S., Mukhopadhyay, A., Patwardhan, A., Persikova, I., Rinaldi, L., Sanz-Garcia, E., Sekharan, M. R., Shao, C., Swaminathan, G. J., Tan, L., Ulrich, E. L., van Ginkel, G., Yamashita, R., Yang, H., Zhuravleva, M. A., Quesada, M., Kleywegt, G. J., Berman, H. M., Markley, J. L., Nakamura, H., Velankar, S. & Burley, S. K. (2017). *Structure*, **25**, 536–545.