Acta Crystallographica Section D
Structural Biology
ISSN 2059-7983

Cryo-EM single particle structure refinement and map calculation using *Servalcat*

Keitaro Yamashita,[a]* Colin M. Palmer,[b] Tom Burnley[b] and Garib N. Murshudov[a]*

[a]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, CB2 0QH, UK, and [b]Scientific Computing Department, UKRI Science & Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0FA, UK

Received
Accepted
Online

In 2020, cryo-EM single particle analysis achieved true atomic resolution. The number of high resolution reconstructions continues to grow, increasing the importance of accurate determination of atomic coordinates. Here, a new Python package and program called *Servalcat* is presented that is designed to facilitate structure refinement. *Servalcat* implements a refinement pipeline, using *REFMAC*5, and map calculation for manual and automatic inspection. The unsharpened and unweighted maps from independent half-set reconstructions are taken as inputs and the noise in the Fourier coefficients is estimated. If a point group symmetry has been applied to the map during reconstruction, the asymmetric unit model is refined with appropriate symmetry constraints. *Servalcat* calculates a weighted $F_o - F_c$ difference map which guides manual model rebuilding in real space, as is common practice in crystallography. The $F_o - F_c$ map helps visualisation of weak features including hydrogen densities. Although hydrogen densities are weak, they are stronger than in electron density maps produced by X-ray crystallography, and some hydrogen atoms are even visible at ~ 1.8 Å resolution.

1. Introduction

Atomic model refinement is the optimisation of parameters against observed data. Atomic parameters typically include coordinates, atomic displacement parameters (ADPs), and occupancies. In crystallography, refinement is crucial because of the phase problem, where the accuracy of density maps relies on the accuracy of phases that are calculated from the model (Tronrud, 2004). More accurate maps may be obtained as the model becomes more accurate through the refinement. In single particle analysis (SPA) there is no phase problem, although the Fourier coefficients can be noisy, especially at high resolution. Accurate atomic model determination is becoming more and more important due to the "resolution revolution" that happened in the field of cryo-EM SPA following the introduction of direct electron detectors and new data processing methods (Bai *et al.*, 2015). As of April 2021, there are more than 2500 SPA entries with resolution better than 3.5 Å deposited in the Electron Microscopy Data Bank (EMDB). This improvement in resolution has accelerated the development of methods for model building, refinement and validation. Automatic model building programs developed originally for crystallography are now being adapted for cryo-EM SPA maps (Terwilliger *et al.*, 2018*a*; Hoh *et al.*, 2020; Chojnowski *et al.*, 2021). Density modification and local map sharpening can help interpret the map (Jakobi *et al.*, 2017; Terwilliger *et al.*, 2018*b*; Ramírez-Aportela *et al.*, 2019; Terwilliger *et al.*, 2020) but care must be exercised when using such data transformation techniques. Bias towards incorrect assumptions might lead to misinterpretation of the maps. Full-atom refinement can be done either

© 0000 International Union of Crystallography

in real space (Afonine *et al.*, 2018) or in reciprocal space (Murshudov, 2016). After refinement, the model should be validated; the model should have a reasonable geometry and should describe the map without severe overfitting. *MolProbity* is the most widely used geometry validation tool, which includes analyses of clashes, rotamers, and the Ramachandran plot (Chen *et al.*, 2010). The map-model quality is assessed using map-model Fourier shell correlation (FSC) and real space local correlations (Murshudov, 2016; Cragnolini *et al.*, 2021). An $F_\mathrm{o}-F_\mathrm{c}$ map, which highlights unmodelled features and errors in the current model, is almost always used in crystallography, and some tools already exist for SPA (Joseph *et al.*, 2020). Cross-validation using a small subset of Fourier coefficients as in crystallography cannot be used in SPA because Fourier coefficients are strongly correlated with each other. Instead, a half-map based approach has been proposed (Brown *et al.*, 2015).

In 2020, cryo-EM SPA achieved atomic resolution – according to Sheldrick's criterion (Wlodawer & Dauter, 2017) – in structural analyses of apoferritin, which were reported by two groups (Nakane *et al.*, 2020; Yip *et al.*, 2020). Nakane *et al.* (2020) observed hydrogen atom densities at 1.2 and 1.7 Å resolutions using $F_\mathrm{o}-F_\mathrm{c}$ maps calculated by *REFMAC*5. Hydrogen density is relatively weak so the $F_\mathrm{o}-F_\mathrm{c}$ map was essential. An electron in a hydrogen atom is usually shifted towards the parent atom from the nucleus position. In electron microscopy, both the electrons and the nucleus contribute to scattering, and this offset results in a shift of hydrogen density peaks from both positions. The Mott-Bethe formula (Mott & Bragg, 1930; Bethe, 1930; Murshudov, 2016), which conveniently takes this fact into account, is used for calculation of atomic scattering factors by default in *REFMAC*5.

Instead of space group symmetry as found in crystallography, SPA structures may have point group symmetries. Approximately half of the SPA entries in the EMDB have non-$C1$ point group symmetry according to their associated metadata. Such symmetry is advantageous and helps to reach higher resolution because it increases the effective number of particles. If the map is symmetrised, the downstream analyses should be aware of it and the structural model must follow the symmetry. As in crystallography, it is natural to work in a single asymmetric unit. The `MTRIX` records in the PDB format or `_struct_ncs_oper` in the mmCIF format can be used to encode the symmetry information. Currently, there are only a few depositions of such asymmetric unit models to the PDB (excepting viruses). We recommend refining and depositing an asymmetric unit model. It should be noted that validation tools must be aware of any applied symmetry operators, but results should be reported for the asymmetric unit only.

Here we present *Servalcat*, a Python package and standalone program for the refinement and map calculation of cryo-EM SPA structures. *Servalcat* implements a refinement pipeline using *REFMAC*5, which uses a dedicated likelihood function for SPA (Murshudov, 2016). It takes unsharpened and unweighted half maps of the independent reconstructions as inputs. After the refinement, a sharpened and weighted $F_\mathrm{o}-F_\mathrm{c}$ map is calculated, derived from Bayesian statistics. If the map has a point group symmetry, a user can give an asymmetric unit model and a point group symbol, and it outputs a refined asymmetric unit model with symmetry annotation as well as a symmetry-expanded model. The non-crystallographic symmetry (NCS) constraint function in *REFMAC*5 has been updated to consider non-bonded interactions and ADP similarity restraints between symmetry copies.

Servalcat is freely available as a standalone package and also as part of *CCP-EM* (Burnley *et al.*, 2017), where the *REFMAC*5 interface has been updated to use *Servalcat*.

2. Map calculation and sharpening using signal variance

Here we define the observed structure factor $F_\mathrm{o} = (F_\mathrm{o1} + F_\mathrm{o2})/2$, where F_o1 and F_o2 are the Fourier transforms of the two unweighted and unsharpened half maps from independent reconstructions. Let us assume that F_o is the result of a position independent blurring $k(s)$ of the true Fourier coefficients F_T with an independent Gaussian noise σ_n. That is,

$$p(F_\mathrm{o};F_\mathrm{T}) = N_2(k(s)F_\mathrm{T},\sigma_\mathrm{n}^2/2), \tag{1}$$

$$\mathrm{var}(F_\mathrm{o}) = k^2(s)\cdot\mathrm{var}(F_\mathrm{T}) + \sigma_\mathrm{n}^2 = k^2(s)\sigma_\mathrm{T}^2 + \sigma_\mathrm{n}^2, \tag{2}$$

where $N_2(\mu,\sigma^2)$ is a 2D Gaussian distribution with the mean μ and the variance σ^2. The variance of the noise (σ_n^2) can be calculated from the half maps in resolution bins (Murshudov, 2016).

$$\sigma_\mathrm{n}^2(s) = \mathrm{var}(F_\mathrm{o1} - F_\mathrm{o2})/4. \tag{3}$$

We will later use the relationship of σ_n^2 and $k^2(s)\sigma_\mathrm{T}^2$ to the FSC (Rosenthal & Henderson, 2003).

$$\mathrm{FSC_{half}} = \mathrm{CC}(F_\mathrm{o1},F_\mathrm{o2}) = \frac{k^2(s)\sigma_\mathrm{T}^2}{k^2(s)\sigma_\mathrm{T}^2 + 2\sigma_\mathrm{n}^2}, \tag{4}$$

$$\mathrm{FSC_{full}} = \frac{k^2(s)\sigma_\mathrm{T}^2}{k^2(s)\sigma_\mathrm{T}^2 + \sigma_\mathrm{n}^2} = \frac{2\mathrm{FSC_{half}}}{\mathrm{FSC_{half}}+1}. \tag{5}$$

Let us also assume that the model error follows a Gaussian (Luzzati, 1952),

$$p(F_\mathrm{T};F_\mathrm{c}) = N_2(DF_\mathrm{c},\sigma_\mathrm{U,T}^2/2), \tag{6}$$

where D is a resolution dependent scale factor and $\sigma_\mathrm{U,T}^2$ is the variance of the unexplained signal. Then, we derive a posterior distribution of F_T given F_o and F_c, as discussed previously (Murshudov, 2016).

$$p(F_\mathrm{T};F_\mathrm{o},F_\mathrm{c}) = \frac{p(F_\mathrm{o};F_\mathrm{T})p(F_\mathrm{T};F_\mathrm{c})}{p(F_\mathrm{o};F_\mathrm{c})}. \tag{7}$$

This also becomes a 2D Gaussian distribution with the mean and variance

$$\langle F_\mathrm{T}\rangle = \frac{k^2(s)\sigma_\mathrm{U,T}^2}{k^2(s)\sigma_\mathrm{U,T}^2+\sigma_\mathrm{n}^2}\frac{F_\mathrm{o}}{k(s)} + \frac{\sigma_\mathrm{n}^2}{k^2(s)\sigma_\mathrm{U,T}^2+\sigma_\mathrm{n}^2}DF_\mathrm{c}, \tag{8}$$

$$\mathrm{var}(F_\mathrm{T}) = \frac{\sigma_\mathrm{n}^2\sigma_\mathrm{U,T}^2}{k^2(s)\sigma_\mathrm{U,T}^2+\sigma_\mathrm{n}^2}. \tag{9}$$

Coefficients for an $F_\mathrm{o}-F_\mathrm{c}$-type difference map can be derived as

$$F_\mathrm{diff} = \langle F_\mathrm{T}\rangle - DF_\mathrm{c} = \frac{k^2(s)\sigma_\mathrm{U,T}^2}{k^2(s)\sigma_\mathrm{U,T}^2+\sigma_\mathrm{n}^2}\left(\frac{F_\mathrm{o}}{k(s)} - DF_\mathrm{c}\right). \tag{10}$$

D and $k^2\sigma^2_{\mathrm{U,T}}$ are obtained in each resolution bin i by a maximum likelihood estimation.

$$D = \frac{1}{k(s)} \frac{\sum_{s\in i} F_{\mathrm{o}}(s) F_{\mathrm{c}}{}^*(s)}{\sum_{s\in i} |F_{\mathrm{c}}(s)|^2}, \quad (11)$$

$$k^2(s)\sigma^2_{\mathrm{U,T}} = \max\left(0, \sum_{s\in i} \frac{|F_{\mathrm{o}}(s) - D_i k(s) F_{\mathrm{c}}(s)|^2}{N_i} - \sigma^2_{\mathrm{n},i}\right). \quad (12)$$

The remaining unknown variable is $k(s)$. If we approximate $k(s)$ using the square root of the variance of the observed signal,

$$k(s) \sim \frac{k(s)\sigma_{\mathrm{T}}}{\sqrt{k^2(s)\sigma^2_{\mathrm{T}} + \sigma^2_{\mathrm{n}}}} \sqrt{k^2(s)\sigma^2_{\mathrm{T}} + \sigma^2_{\mathrm{n}}} = \sqrt{\mathrm{FSC}_{\mathrm{full}}} \sqrt{\langle |F_{\mathrm{o}}|^2 \rangle}. \quad (13)$$

This approximation should give the "best" sharpening for the region having an average B value. Then, letting E be the normalized structure factor, that is a structure factor divided by the square root of its variance in resolution bins,

$$F_{\mathrm{diff}} = \frac{k^2(s)\sigma^2_{\mathrm{U,T}}}{k^2(s)\sigma^2_{\mathrm{U,T}} + \sigma^2_{\mathrm{n}}} \frac{1}{\sqrt{\mathrm{FSC}_{\mathrm{full}}}} (E_{\mathrm{o}} - \mathrm{FSC}_{\mathrm{model}} E_{\mathrm{c}}), \quad (14)$$

where

$$\mathrm{FSC}_{\mathrm{model}} = \frac{\sum F_{\mathrm{o}} F_{\mathrm{c}}{}^*}{\sqrt{\sum |F_{\mathrm{o}}|^2} \sqrt{\sum |F_{\mathrm{c}}|^2}}. \quad (15)$$

Servalcat calculates an $F_{\mathrm{o}} - F_{\mathrm{c}}$ map using (14). The significance of difference map peaks is usually defined by r.m.s.d. (sigma) level in crystallography. However, in SPA the box size is arbitrary and the voxels outside the molecular envelope lead to underestimation of the r.m.s.d. value (see the Appendix A). When a mask file is given, *Servalcat* writes a map file after normalizing values within the mask. (Otherwise only the $F_{\mathrm{o}} - F_{\mathrm{c}}$ structure factors are written, in mtz format.) Note that the $F_{\mathrm{o}} - F_{\mathrm{c}}$ map is only sensible when the ADPs are properly refined; otherwise we will see peaks due to incorrect ADPs.

The map from the estimated true Fourier coefficients (8) may be useful, but there is a risk of model bias because of the contribution from F_{c}. To avoid this, *Servalcat* provides the following as a default map for manual inspection. This is a special case of (8) in the absence of model, that is, with $D = 0$,

$$\langle F_{\mathrm{T}} \rangle = \frac{k^2(s)\sigma^2_{\mathrm{T}}}{k^2(s)\sigma^2_{\mathrm{T}} + \sigma^2_{\mathrm{n}}} \frac{F_{\mathrm{o}}}{k(s)} = \sqrt{\mathrm{FSC}_{\mathrm{full}}} E_{\mathrm{o}}. \quad (16)$$

This is equivalent to *EMDA*'s "best map" (`https://emda.readthedocs.io`).

3. Refinement procedure

In this section the refinement and map calculation procedures are described. Everything other than *REFMAC*5 itself is implemented in *Servalcat* using the *GEMMI* library (`https://github.com/project-gemmi/gemmi`). Fig. 1 summarizes the procedure.

3.1. Map choice

The optimal map depends on the purpose. For a manual inspection, optimally sharpened and weighted maps should be used so that the best visual interpretability is achieved. In general, this does not mean the best signal to noise ratio, but it does mean that the details of structural features are visible in the map. On the other hand, unsharpened and unweighted maps are preferred in the refinement. If a sharpened map is used, some atoms may need to be refined to have negative ADPs, but the ADPs are constrained to be positive in the refinement resulting in suboptimal atomic models. An unweighted map is preferred because many properties including noise variance and the best weighted maps after refinement can be calculated (see section 2). Users should therefore be aware that the ADPs in the model are not refined against the map that is used for visual inspection. Cross-validation (Brown *et al.*, 2015) can also be carried out throughout refinement and model building if both half maps are readily available. Therefore, unsharpened and unweighted half maps from two independent reconstructions are considered optimal inputs for structure refinement followed by map calculation.

3.2. Masking and trimming

The reconstructed maps consist of macromolecules, solvent, and masked regions. Currently we do not have a model for electron scattering from the bulk solvent, so we simply mask out the region outside the macromolecule. The remaining region can be fitted into a smaller box, which reduces the number of Fourier coefficients and speeds up calculations. Half maps are first sharpened, masked at a radius of 3 Å (default) from the atom positions, and then blurred with the same factor. Sharpening before masking is important to avoid masking out the signal because raw half maps are blurred and the signal is spread out. The optimal sharpening would differ depending on the region, but here we use $k(s)$ (13) as an overall sharpening factor. The sharpened-masked-unsharpened half maps are then averaged to make a full map that is used as the refinement target in *REFMAC*5.

3.3. Point group symmetry

If the maps were symmetrised, the user can specify a point group symbol and give the coordinates only in the asymmetric unit. Symmetry operators are calculated from the symbols (*Cn*, *Dn*, *O*, *T*, and *I*) assuming *RELION*'s axis convention (Scheres, 2012). It is also assumed that the centre of the box is the origin of symmetry. This requires translation for each rotation R_j, which can be calculated as $c - R_j c = c(I - R_j)$ where c is the origin of symmetry. The rotation of the axes and position of the origin are arbitrary in general, and in future will be determined using *ProSHADE* (Nicholls *et al.*, 2018; Tykac, 2018) and *EMDA*. The model in the asymmetric unit is expanded when creating a mask and performing the map trimming. Rotation matrices are invariant to changing the box sizes and shifts of the molecule. The translation vectors in the symmetry operators are recalculated for the shifted model.

*REFMAC*5 internally generates symmetry copies when calculating F_c and restraint terms. For anisotropic ADPs, the B_{aniso} matrix in the Cartesian basis is transformed by $R_j B_{\text{aniso}} R_j^T$. This anisotropic ADP transformation is also implemented in *GEMMI*.

During the refinement, non-bonded interaction and ADP similarity restraints are evaluated using the whole molecule, and the gradients are calculated for the model in the asymmetric unit.

If atoms are on special positions (e.g. on a rotation axis), they are restrained to sit on the special position and anisotropic ADPs are constrained. First, atoms are identified as being on a special position if the following condition is obeyed for any of the symmetry operators j:

$$|x - (R_j x + t_j)|^2 < \epsilon^2 \tag{17}$$

where ϵ is a tolerance that can be modified by users. The default value is 0.25 Å. If an atom is on a special position then the program makes sure that symmetry operators for this position form a group that is a subgroup of the point group of the map. Once the elements of the subgroup for this atom have been identified, the atom is forced to be on that position by simply replacing its coordinates with:

$$x_{\text{sym}} = \frac{1}{N_{\text{sym}}} \sum_j^{N_{\text{sym}}} (R_j x + t_j) \tag{18}$$

At every cycle, the positions of these atoms are restrained to be on their special positions by adding a term to the target function:

$$\frac{1}{\sigma_x^2} \left| x - \frac{1}{N_{\text{sym}}} \sum_j^{N_{\text{sym}}} (R_j x + t_j) \right|^2 \tag{19}$$

where the summation is performed over all subgroup elements of the special position and σ_x is a user controllable weight parameter for special positions. The occupancy of the atom is adjusted based on the multiplicity of the position.

Anisotropic ADPs of atoms on the special positions are also forced to obey symmetry conditions by replacing the anisotropic tensor with:

$$B_{\text{sym}} = \frac{1}{N_{\text{sym}}} \sum_j^{N_{\text{sym}}} R_j B R_j^T \tag{20}$$

After this, similarly to the positional parameters, at every cycle restraints are applied to the anisotropic tensor of the atoms on special positions to avoid violation of the symmetry condition for the ADP:

$$\frac{1}{\sigma_B^2} \left| B - \frac{1}{N_{\text{sym}}} \sum_j^{N_{\text{sym}}} R_j B R_j^T \right|^2 \tag{21}$$

where σ_B is a user controllable weight parameter for B values on special positions. Here, the distance between anisotropic tensors is a Frobenius distance $|B_1 - B_2|^2 = \sum_{i,j} |B_{1,i,j} - B_{2,i,j}|^2$.

3.4. Hydrogen atoms

Hydrogen electrons are usually shifted towards the parent atoms by 0.1–0.2 Å (Williams *et al.*, 2018). This must be accounted for when calculating structure factors from the atomic model (F_c). *REFMAC*5 and *Servalcat* (*GEMMI*) use the Mott-Bethe formula (Mott & Bragg, 1930; Bethe, 1930; Murshudov, 2016) which conveniently takes this fact into account.

The atomic scattering factor for an atom with a shifted nucleus is

$$f_e(s) = \frac{me^2}{8\pi h^2 \varepsilon_0} \frac{Ze^{-2\pi i s^T \Delta x} - f_X(s)}{|s|^2}, \tag{22}$$

where Δx is the positional shift of the nucleus with respect to the centre of the electron density. The hydrogen density peak in real space is shifted with respect to both nuclear and electron positions and varies depending on ADP and resolution cutoff (Nakane *et al.*, 2020). The expected peak position may be calculated by the Fourier transform of (22). The new *CCP4* monomer library includes nucleus bond distances (_chem_comp_bond.value_dist_nucleus). They were generated by *AceDRG* (Long *et al.*, 2017) using QM calculations and neutron diffraction data from the Crystallography Open Database (COD).

3.5. Refinement

*REFMAC*5 performs a maximum-likelihood refinement against a sharpened-masked-unsharpened map (see 3.2) using a dedicated likelihood function for SPA. No solvent model is used. After the refinement, *Servalcat* shifts the model back to the original box and adjusts the translation vectors of the symmetry operators if needed. It also generates an mtz file of map coefficients including the sharpened and weighted $F_o - F_c$ map (14) and the "best" map (16).

3.6. User interface

Servalcat has a command-line interface. A graphical interface will be available in *CCP-EM*, where the *REFMAC*5 interface has been updated and is now based on *Servalcat*.

From the user's point of view, the main difference in setting up a refinement job is that the default input is now a pair of half maps. (Refinement from a single input map is still possible but is no longer the default option.) The user is also offered more control over options for refinement weight, symmetry and handling of hydrogen atoms. At the end of refinement, the $F_o - F_c$ difference map from *Servalcat* is made available along with the other output files in the *CCP-EM* launcher.

4. Methods and results

4.1. $F_o - F_c$ map for ligand visualisation

$F_o - F_c$ omit maps are widely used to convincingly demonstrate the existence of ligands in crystallography. They are also useful for this purpose in SPA. Fig. 2 shows an example of an $F_o - F_c$ omit map for the ligand density from EMD-22898 (Kern *et al.*, 2021), clearly showing support for the presence of

the ligand. To generate this map, chain A of the atomic model from PDB entry 7kjr was refined against the half maps under $C2$ symmetry constraints. After the refinement, the ligand and water atoms were omitted and the $F_o - F_c$ map was calculated. Map values were normalized within a mask. Since a suitable mask was not available in the EMDB, one was calculated from half map correlation using *EMDA*.

4.2. $F_o - F_c$ map for detecting model errors

In crystallography, $F_o - F_c$ maps are almost always used for manual and automatic model rebuilding. Strong negative density usually indicates that parts of the model should be moved away or removed while strong positive density implies there are unmodelled atoms. The $F_o - F_c$ map is typically updated after every refinement session, and refinement may be stopped when there are no significant strong peaks.

The same refinement practice is possible in SPA. Fig. 3 illustrates the use of the $F_o - F_c$ map for detecting model errors using EMD-0919 and PDB 6lmt (Demura *et al.*, 2020). Chain A of the model was refined against the half maps under $C8$ symmetry constraints. After refinement, the $F_o - F_c$ map was calculated and normalized using the standard deviation of the region within the EMDB-deposited mask. In this example it is clear from the red and green difference peaks that the tryptophan and methionine side chains should be repositioned.

4.3. Hydrogen density analysis

Nakane *et al.* (2020) reported convincing densities for hydrogen atoms in apoferritin and $GABA_AR$ maps by cryo-EM SPA at 1.2 and 1.7 Å resolutions respectively. There is a higher chance of observing hydrogen density in electron microscopy than in X-ray crystallography, because of the increased contrast for the lighter elements (Clabbers & Abrahams, 2018).

It is natural to ask what is the lowest resolution at which hydrogen atoms can be seen in cryo-EM SPA using currently available computational tools. Here, we analyzed apoferritin maps from the EMDB to see if and when hydrogen densities could be observed. There are 25 mouse or human apoferritin entries at resolutions better than 2.1 Å, of which 19 had half maps and were used in the analysis (Table 1). Chain A of each model was refined against the half maps under O symmetry constraints. If there was no corresponding PDB code, 7a4m or 6z6u were placed in the map using *MOLREP* (Vagin & Teplyakov, 1997) followed by jiggle-fit in *Coot* (Brown *et al.*, 2015) before full atomic refinement. After 10 cycles of refinement with *REFMAC*5, an $F_o - F_c$ map was calculated and normalized within the mask. Riding hydrogen atoms were used in the refinement but they are omitted for $F_o - F_c$ map calculation. Peaks $\geq 2\sigma$ and $\geq 3\sigma$ were detected using *PEAKMAX* from the *CCP4* package (Winn *et al.*, 2011), and associated with hydrogen positions if the distance from the peak was less than 0.5 Å. Hydrogen atoms having multiple potential minima (like those in hydroxyl, sulfhydryl or carboxyl groups) were ignored in the analysis. The ratios of the number of hydrogen peaks to the number of hydrogen atoms in the model are plotted in Fig. 4a. The result shows that 1.25 Å data gave the highest ratio of $\sim$ 70% hydrogens detected (Fig. 5a). Even at 1.84 Å approximately 17% of the hydrogen atoms may be found (Fig. 5b) while at 2.0 or 2.1 Å only a few hydrogen atoms are visible in the map (Fig. 5c). Note that there may be false positives due to, for example, alternative conformations or inaccuracies in the model.

In addition, $F_o - F_c$ maps were made from the 1.2 Å data (PDB 7a4m; EMD-11638) using several different resolution cutoffs. These were analysed in the same way (Fig. 4c), along with F_c maps calculated from the 7a4m model at the same resolutions (Fig. 4d). Fig. 4c,d show that if the cryo-EM experiment and atomic model refinement are carried out carefully, with due attention to ADPs, then hydrogen atoms can be seen even at 2.0 Å resolution.

For comparison, we performed the same analysis using X-ray crystallographic data of (apo)ferritins deposited at PDB. 51 re-refined coordinate models available at *PDB-REDO* (Joosten *et al.*, 2012) were downloaded, $mF_o - DF_c$ maps were calculated using *REFMAC*5 and density peaks for hydrogen atoms were analysed as just described. The result (Fig. 4b) confirms that, as expected, hydrogen atoms are more visible in EM than X-ray.

5. Conclusions

A new program *Servalcat* for refinement and validation of atomic models using cryo-EM maps has been developed. The program controls the refinement flow and does difference map calculations. A weighted and sharpened $F_o - F_c$ map was derived, obtained from the posterior distribution of F_T and an approximation of an overall blurring factor calculated from the variance of the signal. It was shown to be useful to visualize hydrogen atoms and model errors as in crystallography. We assumed the blurring factor $k(s)$ was position independent. However in reality, blurring of maps is position and direction dependent, e.g. due to varying mobility of different domains and/or uncertainty in the particle alignments. If we can estimate (anisotropic) local B values, then we can use them for visual improvement of maps.

We showed that many hydrogen atoms may be observed in the difference maps, even up to a resolution of 2 Å. We would expect that they should also be visible in electron diffraction (MicroED) experiments. However, high accuracy would be needed in the experiment, data analysis, and model refinement in both MicroED and cryo-EM SPA to achieve this experimentally. For example, electron dose in cryo-EM experiments is often high enough to cause radiation damage (Hattne *et al.*, 2018); hydrogen atoms are known to suffer from radiation damage (Leapman & Sun, 1995). They naturally come and go even in normal solutions. Lower dose experiments might be needed for better hydrogen detection, even at the expense of resolution.

Symmetry is widely used in cryo-EM SPA. When symmetry is imposed in the reconstruction, it should be used throughout the downstream analyses, and all software tools should be aware of the symmetry. The asymmetric unit model should be refined under symmetry constraints, and it should be deposited to the

PDB with correct annotation of the symmetry. The PDB and EMDB deposition system will need to validate the symmetry of both model and map. We hope this will become a common practice in future. The same practice should be established for helical reconstructions, in which symmetry is described by the axial symmetry type (*Cn* or *Dn*), twist, and rise (He & Scheres, 2017). *Servalcat* will support helical symmetry in future.

Servalcat is freely available under an open source (MPL-2.0) licence at `https://github.com/keitaroyam/servalcat`. Features described in this paper have been implemented in *REFMAC*5.8.0272, *CCP-EM* XXX and *Servalcat* XXX (which requires *GEMMI* 0.4.5).

Appendix A
Variance of a masked map

We consider a masked map containing n points in total, where m points are within the mask and thus the values for $n-m$ points are zero. If we calculate the mean value of the whole data:

$$\mu_{\text{total}} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{m} d_i}{n} = \frac{\sum_{i=1}^{m} d_i}{m}\frac{n}{m} = \mu_{\text{mask}}\frac{n}{m}. \quad (23)$$

Thus, to calculate the mean within the mask we can calculate the total mean and then use the formula for correction:

$$\mu_{\text{mask}} = \mu_{\text{total}}\frac{m}{n}. \quad (24)$$

For the variance,

$$\begin{aligned}\text{var}_{\text{total}} &= \frac{\sum_{i=1}^{n} d_i^2}{n} - \mu_{\text{total}}^2 = \frac{\sum_{i=1}^{m} d_i^2}{m}\frac{m}{n} - \frac{m^2}{n^2}\mu_{\text{mask}}^2 \\ &= \frac{m}{n}\text{var}_{\text{mask}} + \frac{m(n-m)}{n^2}\mu_{\text{mask}}^2. \quad (25)\end{aligned}$$

From here we can calculate var_{mask} if we know $\text{var}_{\text{total}}$ and μ_{total}. If we denote $f = m/n$ then we can write:

$$\text{var}_{\text{total}} = f\text{var}_{\text{mask}} + f(1-f)\mu_{\text{mask}}^2. \quad (26)$$

If the mean inside the mask is zero then there is a simple relationship between total variance and variance within the mask. This explains the dependence between the box size and the r.m.s.d. of a cryo-EM SPA map.

If we assume that the map consists of signal and noise, and there is no correlation between them then we can claim that $\text{var}_{\text{total}} = \text{var}_{\text{signal}} + \text{var}_{\text{noise}}$. Now, in addition, if we assume that we have modelled the map fully with an atomic model (or two maps have almost perfect overlap of signals) then the difference maps should consist almost entirely of noise. Therefore, $\text{var}_{\text{diffmap}} = \text{var}_{\text{noise}}$. This variance should be calculated within the mask to make sure that we do not have variance reduction because of systematically low values outside the region occupied by the macromolecule. If we want to increase the reliability of these variances for a region of interest then we may also mask out other regions where there might be signal that is not fully accounted for by the current model.

This work was supported by the Medical Research Council, as part of UK Research and Innovation (MC_UP_A025_1012 to K.Y. and G.N.M.; MR/V000403/1 to C.M.P. and T.B.). The authors are grateful to Marcin Wojdyr for the implementation of F_c calculation for EM in the *GEMMI* library; Takanori Nakane for critical reading of the manuscript; computational structural biology group members for discussion; Jake Grimmett and Toby Darling from the MRC-LMB Scientific Computing Department for computing support and resources.

References

Afonine, P. V., Poon, B. K., Read, R. J., Sobolev, O. V., Terwilliger, T. C., Urzhumtsev, A. & Adams, P. D. (2018). Real-space refinement in *PHENIX* for cryo-EM and crystallography. *Acta Crystallographica Section D*, **74**(6), 531–544.

Bai, X.-c., McMullan, G. & Scheres, S. H. W. (2015). How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, **40**(1), 49–57.

Bethe, H. (1930). Zur Theorie des Durchgangs schneller Korpuskularstrahlen durch Materie. *Annalen der Physik*, **397**(3), 325–400.

Brown, A., Long, F., Nicholls, R. A., Toots, J., Emsley, P. & Murshudov, G. (2015). Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallographica Section D*, **71**(1), 136–153.

Burnley, T., Palmer, C. M. & Winn, M. (2017). Recent developments in the *CCP-EM* software suite. *Acta Crystallographica Section D*, **73**(6), 469–477.

Chen, V. B., Arendall, III, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, **66**(1), 12–21.

Chojnowski, G., Sobolev, E., Heuser, P. & Lamzin, V. S. (2021). The accuracy of protein models automatically built into cryo-EM maps with *ARP/wARP*. *Acta Crystallographica Section D*, **77**(2), 142–150.

Clabbers, M. T. B. & Abrahams, J. P. (2018). Electron diffraction and three-dimensional crystallography for structural biology. *Crystallography Reviews*, **24**(3), 176–204.

Cragnolini, T., Sahota, H., Joseph, A. P., Sweeney, A., Malhotra, S., Vasishtan, D. & Topf, M. (2021). *TEMPy*2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. *Acta Crystallographica Section D*, **77**(1), 41–47.

Danev, R., Yanagisawa, H. & Kikkawa, M. (2019). Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. *Trends in Biochemical Sciences*, **44**(10), 837–848.

Demura, K., Kusakizako, T., Shihoya, W., Hiraizumi, M., Nomura, K., Shimada, H., Yamashita, K., Nishizawa, T., Taruno, A. & Nureki, O. (2020). Cryo-EM structures of calcium homeostasis modulator channels in diverse oligomeric assemblies. *Science Advances*, **6**(29), eaba8105.

Fislage, M., Shkumatov, A. V., Stroobants, A. & Efremov, R. G. (2020). Assessing the JEOL CRYO ARM 300 for high-throughput automated single-particle cryo-EM in a multiuser environment. *IUCrJ*, **7**(4), 707–718.

Guo, H., Franken, E., Deng, Y., Benlekbir, S., Singla Lezcano, G., Janssen, B., Yu, L., Ripstein, Z. A., Tan, Y. Z. & Rubinstein, J. L. (2020). Electron-event representation data enable efficient cryoEM file storage with full preservation of spatial and temporal resolution. *IUCrJ*, **7**(5), 860–869.

Hattne, J., Shi, D., Glynn, C., Zee, C.-T., Gallagher-Jones, M., Martynowycz, M. W., Rodriguez, J. A. & Gonen, T. (2018). Analysis of Global and Site-Specific Radiation Damage in Cryo-EM. *Structure*, **26**(5), 759–766.e4.

He, S. & Scheres, S. H. (2017). Helical reconstruction in RELION. *Journal of Structural Biology*, **198**(3), 163–176.

Hoh, S. W., Burnley, T. & Cowtan, K. (2020). Current approaches for automated model building into cryo-EM maps using *Buccaneer* with *CCP-EM*. *Acta Crystallographica Section D*, **76**(6), 531–541.

Jakobi, A. J., Wilmanns, M. & Sachse, C. (2017). Model-based local density sharpening of cryo-EM maps. *eLife*, **6**, e27131.

Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *PDB_REDO*: constructive validation, more than just looking for errors. *Acta Crystallographica Section D*, **68**(4), 484–496.

Joseph, A. P., Lagerstedt, I., Jakobi, A., Burnley, T., Patwardhan, A., Topf, M. & Winn, M. (2020). Comparing Cryo-EM Reconstructions and Validating Atomic Model Fit Using Difference Maps. *Journal of Chemical Information and Modeling*, **60**(5), 2552–2560.

Kato, T., Makino, F., Nakane, T., Terahara, N., Kaneko, T., Shimizu, Y., Motoki, S., Ishikawa, I., Yonekura, K. & Namba, K. (2019). CryoTEM with a Cold Field Emission Gun That Moves Structural Biology into a New Stage. *Microscopy and Microanalysis*, **25**(S2), 998–999.

Kern, D. M., Sorum, B., Mali, S. S., Hoel, C. M., Sridharan, S., Remis, J. P., Toso, D. B., Kotecha, A., Bautista, D. M. & Brohawn, S. G. (2021). Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *bioRxiv*.

Leapman, R. D. & Sun, S. (1995). Cryo-electron energy loss spectroscopy: observations on vitrified hydrated specimens and radiation damage. *Ultramicroscopy*, **59**(1), 71–79.

Long, F., Nicholls, R. A., Emsley, P., Gražulis, S., Merkys, A., Vaitkus, A. & Murshudov, G. N. (2017). *AceDRG*: a stereochemical description generator for ligands. *Acta Crystallographica Section D*, **73**(2), 112–122.

Luzzati, V. (1952). Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallographica*, **5**(6), 802–810.

Mott, N. F. & Bragg, W. L. (1930). The scattering of electrons by atoms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **127**(806), 658–665.

Murshudov, G. N. (2016). *Refinement of Atomic Structures Against cryo-EM Maps*, vol. 579, pp. 277–305. Academic Press.

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). Single-particle cryo-EM at atomic resolution. *Nature*, **587**, 152–156.

Naydenova, K., Peet, M. J. & Russo, C. J. (2019). Multifunctional graphene supports for electron cryomicroscopy. *Proceedings of the National Academy of Sciences*, **116**(24), 11718–11724.

Nicholls, R. A., Tykac, M., Kovalevskiy, O. & Murshudov, G. N. (2018). Current approaches for the fitting and refinement of atomic models into cryo-EM maps using *CCP-EM*. *Acta Crystallographica Section D*, **74**(6), 492–505.

Pintilie, G., Zhang, K., Su, Z., Li, S., Schmid, M. F. & Chiu, W. (2020). Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nature Methods*, **17**(3), 328–334.

Ramírez-Aportela, E., Vilas, J. L., Glukhova, A., Melero, R., Conesa, P., Martínez, M., Maluenda, D., Mota, J., Jiménez, A., Vargas, J., Marabini, R., Sexton, P. M., Carazo, J. M. & Sorzano, C. O. S. (2019). Automatic local resolution-based sharpening of cryo-EM maps. *Bioinformatics*, **36**(3), 765–772.

Rosenthal, P. B. & Henderson, R. (2003). Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology*, **333**(4), 721–745.

Scheres, S. H. W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, **180**(3), 519–530.

Tan, Y. Z. & Rubinstein, J. L. (2020). Through-grid wicking enables high-speed cryoEM specimen preparation. *Acta Crystallographica Section D*, **76**(11), 1092–1103.

Terwilliger, T. C., Adams, P. D., Afonine, P. V. & Sobolev, O. V. (2018*a*). A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. *Nature Methods*, **15**(11), 905–908.

Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. (2018*b*). Automated map sharpening by maximization of detail and connectivity. *Acta Crystallographica Section D*, **74**(6), 545–559.

Terwilliger, T. C., Sobolev, O. V., Afonine, P. V., Adams, P. D. & Read, R. J. (2020). Density modification of cryo-EM maps. *Acta Crystallographica Section D*, **76**(10), 912–925.

Tronrud, D. E. (2004). Introduction to macromolecular refinement. *Acta Crystallographica Section D*, **60**(12 Part 1), 2156–2168.

Tykac, M. (2018). *Protein Shape Description and its Application to Shape Comparison*. Ph.D. thesis, University of Cambridge.

Vagin, A. & Teplyakov, A. (1997). *MOLREP*: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography*, **30**(6), 1022–1025.

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 III, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, **27**(1), 293–315.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). Overview of the *CCP*4 suite and current developments. *Acta Crystallographica Section D*, **67**(4), 235–242.

Wlodawer, A. & Dauter, Z. (2017). 'Atomic resolution': a badly abused term in structural biology. *Acta Crystallographica Section D*, **73**(4), 379–380.

Wu, M., Lander, G. C. & Herzik, M. A. (2020). Sub-2 Angstrom resolution structure determination using single-particle cryo-EM at 200 keV. *Journal of Structural Biology: X*, **4**, 100020.

Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature*, **587**, 157–161.

Zivanov, J., Nakane, T., Forsberg, B. O., Kimanius, D., Hagen, W. J., Lindahl, E. & Scheres, S. H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, **7**, e42166.

Table 1
Test data for hydrogen peak analysis. *No PDB entry was assigned and the code in parenthesis was used for refinement (7a4m from mouse and 6z6u from human).

EMDB	PDB	resolution (Å)	reference
EMD-11638	7a4m	1.22	Nakane *et al.* (2020)
EMD-11103	6z6u	1.25	Yip *et al.* (2020)
EMD-30683	(7a4m)*	1.31	no publication
EMD-30685	(7a4m)*	1.35	no publication
EMD-30684	(7a4m)*	1.43	no publication
EMD-30686	(7a4m)*	1.43	no publication
EMD-9865	(7a4m)*	1.54	Kato *et al.* (2019)
EMD-11121	6z9e	1.55	Yip *et al.* (2020)
EMD-11122	6z9f	1.56	Yip *et al.* (2020)
EMD-9599	(7a4m)*	1.62	Danev *et al.* (2019)
EMD-0144	(6z6u)*	1.65	Zivanov *et al.* (2018)
EMD-20026	(6z6u)*	1.75	Pintilie *et al.* (2020)
EMD-21024	6v21	1.75	Wu *et al.* (2020)
EMD-10101	6s61	1.84	no publication
EMD-10675	(7a4m)*	1.86	Fislage *et al.* (2020)
EMD-21951	6wx6	2.00	Tan & Rubinstein (2020)
EMD-22351	(6z6u)*	2.07	Guo *et al.* (2020)
EMD-4905	6rjh	2.10	Naydenova *et al.* (2019)
EMD-20521	6pxm	2.10	no publication

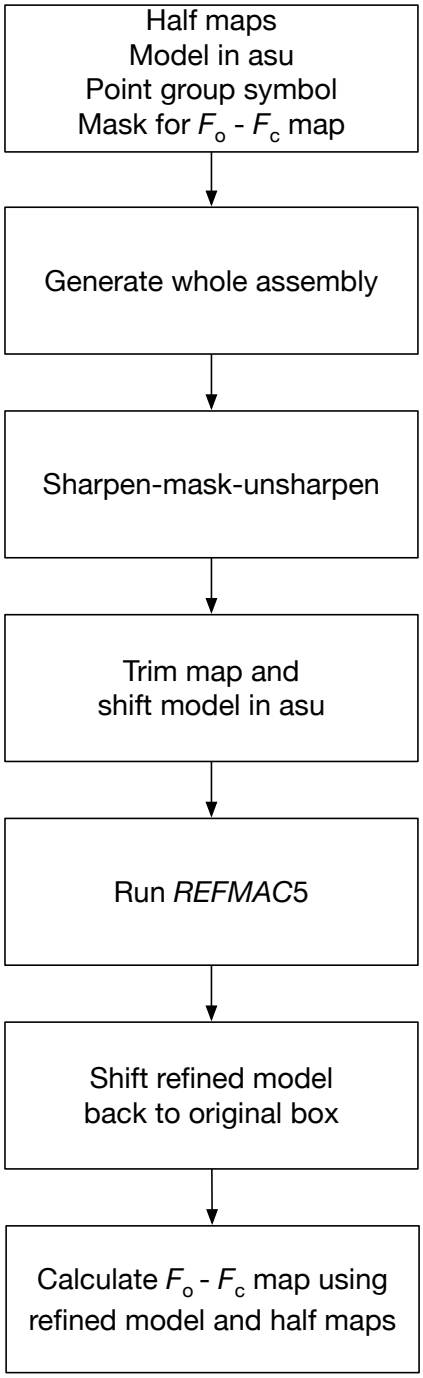

Figure 1
The workflow of *Servalcat* for the refinement of SPA structures.

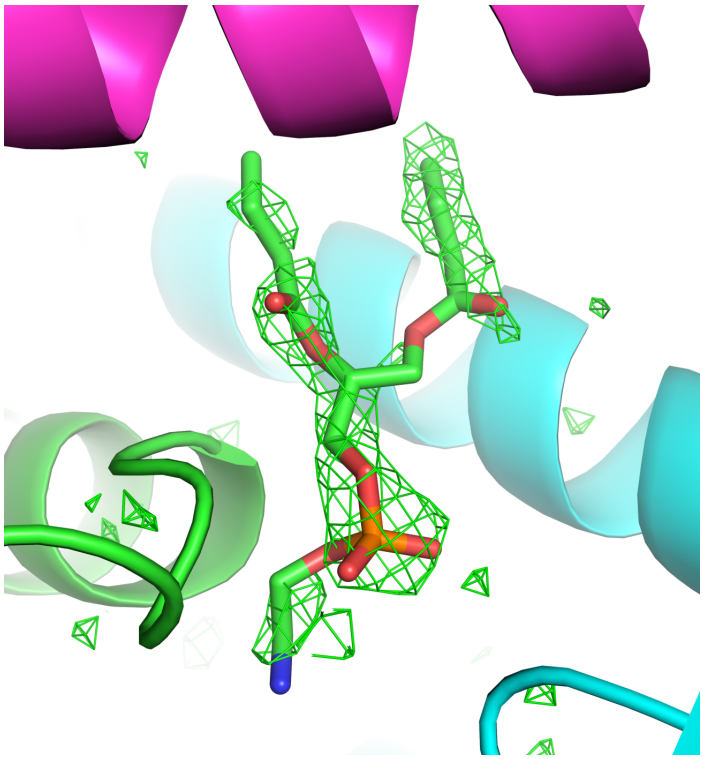

Figure 2
An example of an $F_o - F_c$ omit map for visualization of ligand density. The resolution is 2.08 Å and the $F_o - F_c$ omit map is contoured at 3σ.

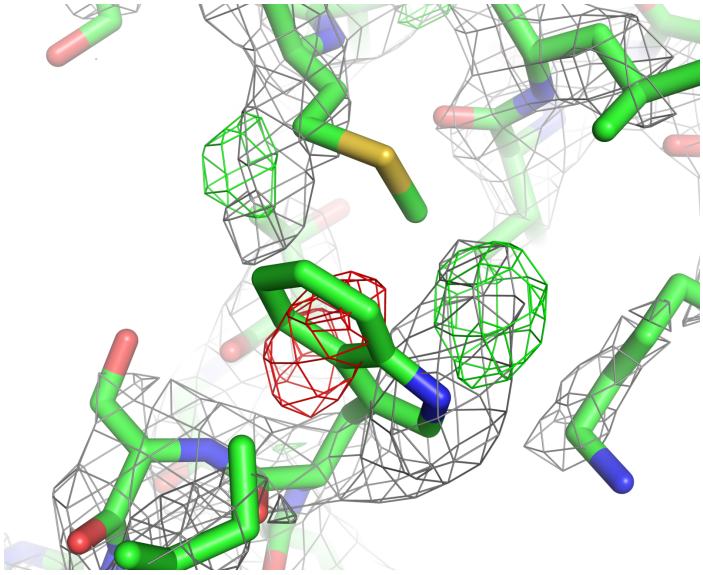

Figure 3
An example of an $F_o - F_c$ map for detecting model error, in this case mispositioned tryptophan and methionine side chains. The resolution is 2.66 Å and the $F_o - F_c$ map is contoured at $\pm 4\sigma$. Green and red meshes represent positive and negative maps, respectively. Grey mesh is the weighted and sharpened F_o map.

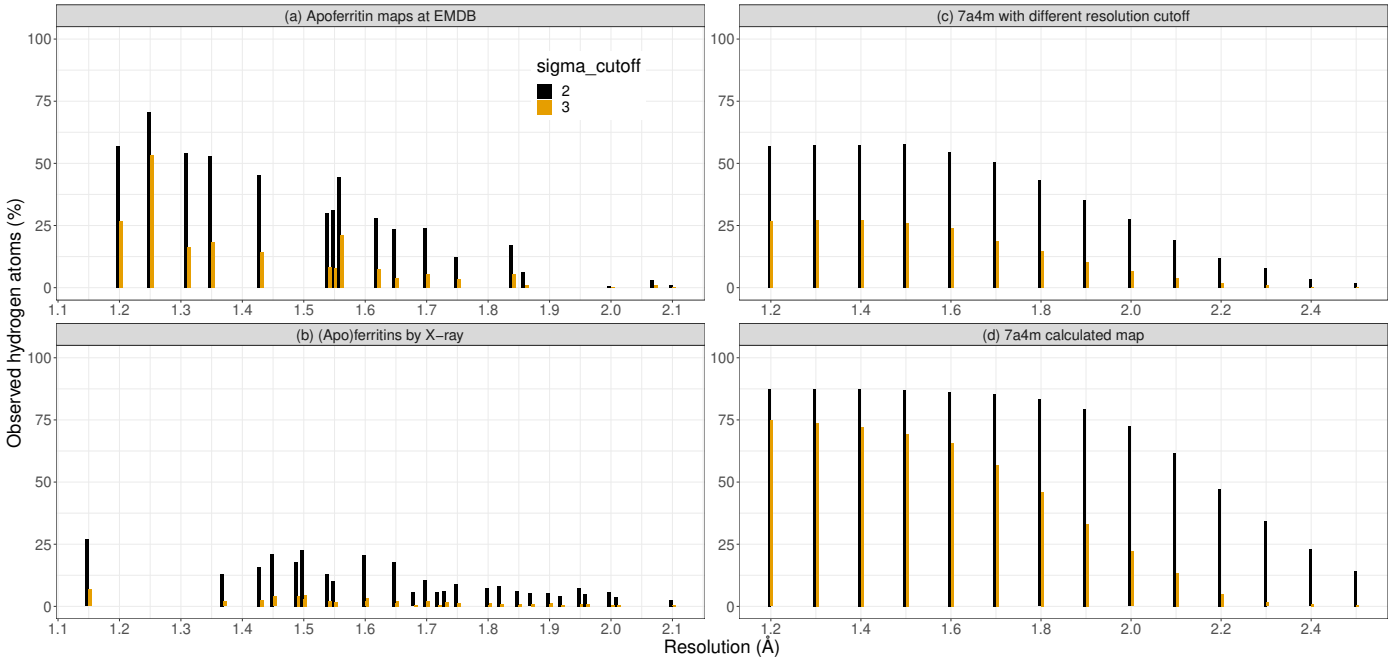

Figure 4
Hydrogen peak numbers divided by the number of hydrogen atoms in the model. (a) different apoferritin cases (see Table 1). (b) X-ray cases using PDB codes 2v2p, 2v2s, 6gxj, 5erj, 5mij, 2cih, 2w0o, 7bd7, 3f37, 2v2n, 1h96, 2chi, 2zg8, 2v2m, 2z5p, 3h7g, 3f34, 2zg7, 3f32, 3f33, 3f36, 2gyd, 3o7s, 1xz1, 1xz3, 2cn7, 2zg9, 3f38, 2cei, 2iu2, 3fi6, 6env, 3f39, 5ix6, 2v2o, 2v2l, 2v2r, 3o7r, 3rav, 3u90, 3f35, 1aew, 5mik, 2g4h, 2v2i, 3rd0, 5erk, 6ra8, 1gwg, 2clu, 2z5q. (c,d) apoferritin cases calculated at different resolutions from the same map and model, 7a4m/EMD-11638 determined at 1.22 Å. (c) shows detection of hydrogen atoms in $F_o - F_c$ maps and (d) in calculated F_c maps.

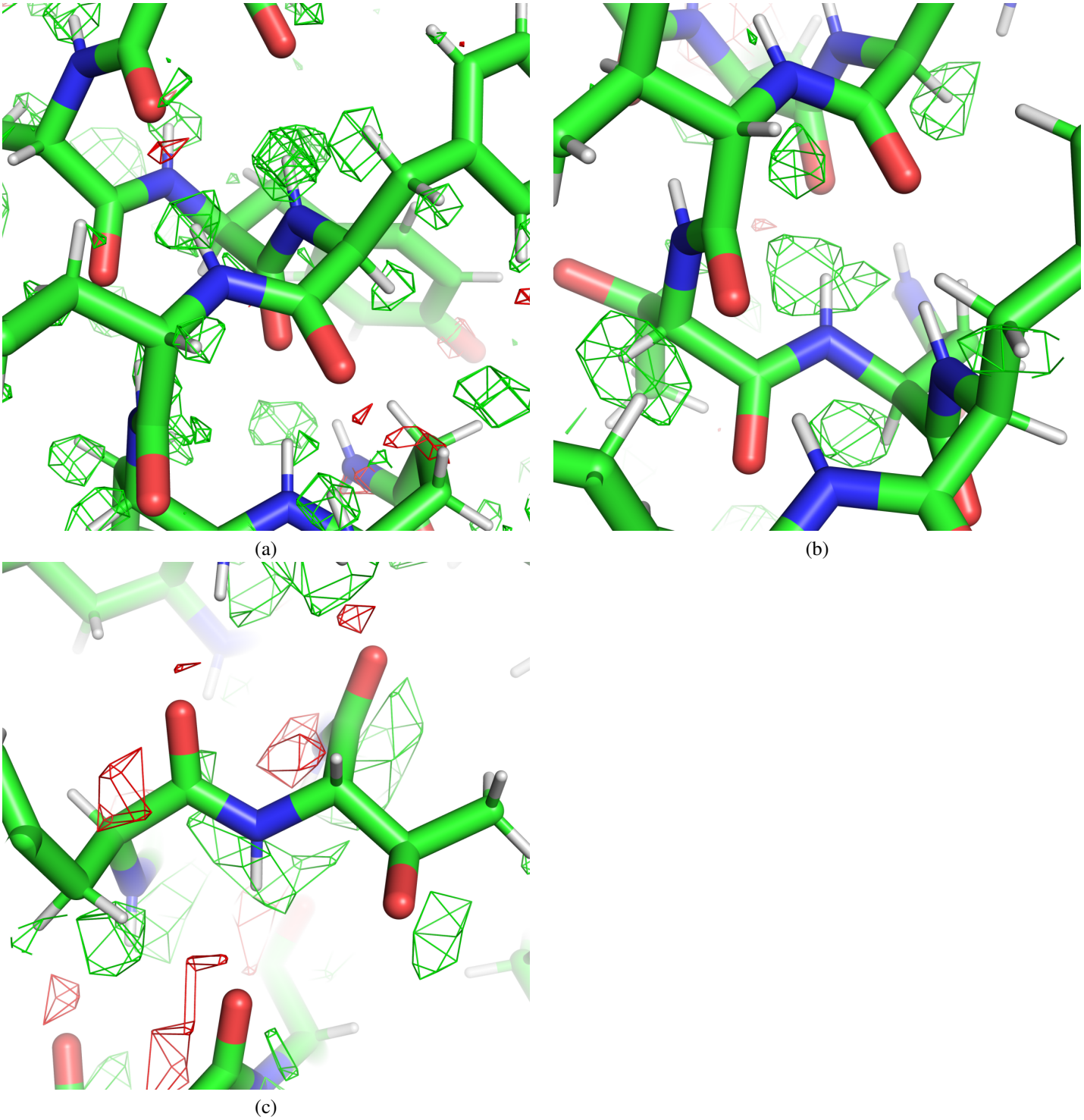

Figure 5
$F_o - F_c$ maps with different resolutions using (a) 1.25 Å data (PDB 6z6u/EMD-11103) (b) 1.84 Å data (PDB 6s61/EMD-10101) and (c) 2.00 Å map (PDB 6wx6/EMD-21951). Green and red meshes represent positive and negative $F_o - F_c$ maps contoured at $\pm 3\sigma$, respectively.