EMDA: A Python package for Electron Microscopy Data Analysis

Rangana Warshamanage*, Keitaro Yamashita & Garib N. Murshudov*

Structural Studies, MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, England

* To whom correspondence may be addressed: ranganaw@mrc-lmb.cam.ac.uk, garib@mrc-lmb.cam.ac.uk

Abstract

An open-source Python library EMDA for cryo-EM map and model manipulation is presented with a specific focus on validation. The architecture of the library and its programmable and command-line interfaces are presented. The use of several functionalities in the library is presented through several examples. The utility of local correlation as a metric for identifying model errors and unmodeled regions in maps, and how it is used as a metric of map-model validation is demonstrated. The mapping of local correlation to individual atoms, and its use to draw insights on local signal variations are discussed. EMDA's likelihood-based map overlay is demonstrated by carrying out a superposition of two domains in two related structures. The overlay is carried out globally to bring both maps into the same coordinate frame and then locally to estimate the relative movement of domains. Finally, the map magnification refinement in EMDA is presented with an example to highlight the importance of correcting map magnification in structural comparison studies.

1. Introduction

Single-particle cryo-electron microscopy (cryo-EM) has become an increasingly popular structure determination tool among structural biologists (Faruqi and McMullan, 2011; Kühlbrandt, 2014; Lyumkis, 2019; Lyumkis et al., 2013; Scheres, 2014). The technique has evolved at an unprecedented speed in the past few years as shown by the rapid growth of cryo-EM structure depositions into the Electron Microscopy Data Bank – EMDB (Lawson et al., 2020). As the number of depositions into EMDB increases, it is important to maintain quality standards for both maps and atomic models not only to ensure their reliability, but also to prevent accumulation of errors.

The EM validation task force 2010 (Henderson et al., 2012) has recognized the critical need of validation standards to assess the quality of EM maps, models and their fits. The task force's recommendations for map validation included the tilt-pair experiments for the absolute hand determination (Rosenthal and Henderson, 2003), the raw image to 3D structure projection matching for validating reconstruction accuracy and the data coverage

(Orlova et al., 1996; Tang et al., 2007), statistical tests using map variances for assessing the map quality and interpretability (Ménétret et al., 2007; Penczek et al., 2006), resolution estimation through Fourier Shell Correlation (FSC) using fully independent half data sets (Scheres and Chen, 2012), visual assessment of map features to the claimed resolution, and the identification and validation of the map symmetry where applicable (Reboul et al., 2020).

The task force has identified the model validation in cryo-EM as an area for further research, mainly due to the fact that, at the time, there were few high resolution cryo-EM structures. Thus, the recommendations for model validation included, among others, the assessment of subunits and their interfaces according to the guidelines proposed by the PDB (Read et al., 2011), the assessment of agreement between the model and the map utilizing chemical measures such as chemical properties and atomic interactions and their clashes as employed in EMFIT program (Rossmann, 2000; Rossmann et al., 2001) or statistical measures such as correlation coefficient.

Since the first meeting in 2010, the field has grown by accumulating many methods and tools to address the issue of map validation. Examples include the gold standard FSC to monitor the map overfitting into noise during reconstruction (Rosenthal and Henderson, 2003; Scheres and Chen, 2012), tilt-pair validation to assess the accuracy of initial angle assignment (Wasilewski and Rosenthal, 2014) and the false discovery maps for visual assessment of map features (Beckers et al., 2019).

The progress in atomic model building, refinement and validation has also been substantial. Resolution in cryo-EM reconstructions vary widely, however the progress made in the field of atomic model building encapsulates modelling tools for low, medium to high resolution. Examples include Chimera (Pettersen et al., 2004), DockEM (Roseman, 2000), FlexEM (Topf et al., 2008), COOT(Brown et al., 2015) , DireX (Wang and Schröder, 2012), MDFF (Trabuco et al., 2009), Cryo-Fit (Kim et al., 2019), Rosetta (Wang et al., 2016), MDeNM-EMfit (Costa et al., 2020). The atomic model refinement has also gained a significant progress. Unlike in crystallography, cryo-EM maps contain both amplitudes and phases and the atomic model refinement programs can use a phased likelihood target function as employed in REFMAC5 (Murshudov, 2016; Nicholls et al., 2018) or real space target functions as used in phenix.real_space_refine (Afonine et al., 2018a), ISOLDE (Croll, 2018) and COOT (Brown et al., 2015). There has also been a considerable progress in map and model validation. Examples include various metrics for map-model fit validation (Afonine et al., 2018b; Barad et al., 2015; Brown et al., 2015; Pintilie et al., 2020; Ramírez-Aportela et al., 2021), and chemistry and geometry based tools for model validation (Emsley et al., 2010; Prisant et al., 2020). The EM practitioners can access these methods and tools as parts of stand-alone packages, separate tools in collaborative projects such as CCP-EM (Burnley et al., 2017), or as web based tools such as EMDB validation server (https://www.ebi.ac.uk/pdbe/emdb/validation/fsc/), Molprobity server (http://molprobity.biochem.duke.edu/) etc.

Developing better validation methods and tools in cryo-EM is an active area of research because the goal of validation in cryo-EM is a changing target (Lawson et al., 2020). The metrics for validation should evolve as the field progresses towards the atomic resolution because the methods that are applicable to low and medium resolution may not be equally applicable to atomic resolution data and derived models.

In this paper we present Electron Microscopy Data Analytical toolkit (EMDA) - a new Python package for post reconstruction/atomic model refinement analysis and validation of cryo-EM maps and models. EMDA is a portable Python package with a command line and an Application Programming Interface (API) for Python programmers.

This paper contains three sections. The first section covers the basic design and the infrastructure of EMDA. The second section describes the mathematical framework behind EMDA and the last section contains several representative tools demonstrated through examples. Lastly, the conclusions and outlook followed by information about the package availability are given.

2. EMDA architecture

EMDA is written primarily in Python using Numpy arrays (Harris et al., 2020), Scipy tools (Virtanen et al., 2020) and Matplotlib tools (Hunter, 2007), however, numerically intensive tasks are written in Fortran. F2PY (Peterson, 2009) mediates the communication between Python and Fortran. This combination allows us to integrate powerful numerical calculations with abstraction features in Python.

EMDA code is organized into three layers as shown in Fig. 1. The innermost layer (Layer 3) consists of core and extension modules. The core modules provide basic services such as read & write, format conversion, resampling, binning etc. All higher-level functionalities such as rigid-body fitting, magnification refinement, difference map calculations are provided through extensions. The extension modules use the basic services provided by the core modules. Both core and extension modules are wrapped into another module to form the EMDA Application Programming Interface (API) (Layer 2). API abstracts the underlying complexity of the code into methods and objects providing a simplified mechanism for other developers to gain advantage of EMDA infrastructure. EMDA-API functions are further wrapped to form the EMDA command-line-interface (Layer 1). The users can access the underlying functionalities through the command line. Each functionality is callable with a keyword followed by a set of arguments.

EMDA uses open-source, standalone Python library *mrcfile* (Burnley et al., 2017) for the reading, writing and validating EM files in the standard MRC2014 format (Cheng et al., 2015). Also, EMDA uses *gemmi* (https://gemmi.readthedocs.io) for reading and writing atomic coordinate files, and ProSHADE (Nicholls et al., 2018; Tykac, 2018) for symmetry detection in EM maps.

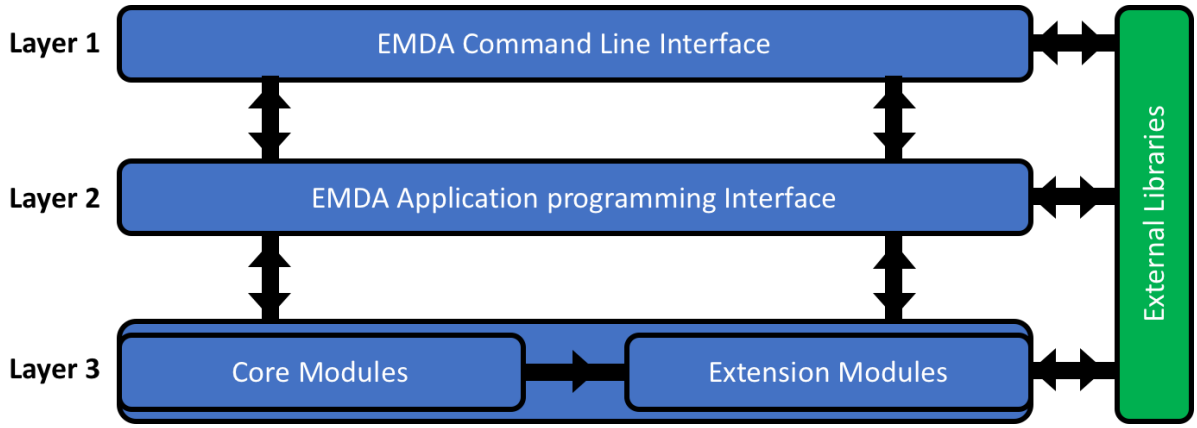

Fig. 1. Architecture of EMDA library. The three Python code layers are shown in blue and the layer of external libraries is shown in green. The black arrows show the data flow and the functional dependencies.

3. Methods

In this section we outline the mathematical framework for local correlation and probability-based methods in EMDA.

3.1. Local correlation in image space

Pearson's product-moment sample correlation coefficient (CC) has been extensively used in various purposes in X-ray crystallography (Karplus and Diederichs, 2012; Tickle, 2012) and in cryo-EM (Van Heel, 1987). The CC depends on the signal and noise levels. If we assume that the noise variance is constant within the masked map then for a given data the CC will be an indicator of the relative signal in the data. Care should be exercised in its interpretation as any systematic behaviour will be considered as signal. Since the CC is calculated using the data, its variance depends on the volume of the data being used.
Local CC in image space can be calculated using the formula for the weighted Pearson's product moment sample correlation coefficient with weights equal to the kernel. The local CC for two maps $\psi_1(x)$ and $\psi_2(x)$ is:

$$CC_{12,m}(x) = \frac{cov_{12,m}(\psi_1(x),\psi_2(x))}{\sqrt{var_{1,m}(\psi_1(x))var_{2,m}(\psi_2(x))}} \quad (1).$$

$cov_{12,m}\left(\psi_1(x), \psi_2(x)\right)$ and $var_{i,m}(\psi_i(x))$ are the local covariance and variances for $\psi_1(x)$ and $\psi_2(x)$ calculated with a kernel $m(x)$. The kernel is normalized such that $\int_{R^3} m(x)dx = 1$. The expression for local covariance is,

$cov_{12,m}(\psi_1(x), \psi_2(x)) = \int_{R^3} m(x-y)\,\psi_1(y)\psi_2(y)dy - \int_{R^3} m(x-y)\,\psi_1(y)dy \int_{R^3} m(x-y)\,\psi_2(y)dy$ (2)

Similar expressions can be written for local variances $var_i(\psi_i(x))$. Note that the equation 2 can be readily evaluated using the convolution theorem (see Appendix A).
Such correlations could be calculated for any pairs of maps. When it is calculated using half maps ($CC_{half,m}(x)$) reconstructed from randomly chosen half of the particles, that correlation indicates the local signal to noise ratio, whereas the local correlation between observed and calculated maps ($CC_{map,model,m}(x)$) indicates local disagreement between atomic model and observed map. Similarly, the local correlation between two different observed maps indicates local common signals between them.
The correlation calculated using half maps is converted to that of full map using the following formula (Rosenthal and Henderson, 2003)

$$CC_{full,m}(x) = \frac{2CC_{half,m}(x)}{1+CC_{half,m}(x)} \quad (3).$$

The full map is usually the combined map derived from raw half maps. Unless explicitly stated otherwise, in the following text "fullmap" stands for the combined map.
The size of the kernel should be chosen such that the number of data points included is sufficient to calculate a reliable statistic. Both too small or too large kernels lead to inaccurate correlations due to insufficient data points or loss of locality, respectively.
Also, the $CC_{half,m}(x)$ (hereafter CC_{half}) depends on the local B value of the map because the magnitude of the signal variance depends on the B value applied on it according to $e^{-Bs^2/2}$, where *s* indicates the inverse resolution and has the unit $Å^{-1}$. The B value (thus, the magnitude of the exponent) determines the degree of blur, where positive B values (negative exponents) induce blurring whereas negative B values (positive exponents) induce sharpening.
Since the $CC_{map,model,m}(x)$ (hereafter $CC_{map,model}$) is calculated between an atomic model and a given map (full map or primary map), it depends on atomic coordinates, occupancies and B values. Therefore, to obtain the best possible $CC_{map,model}$ it should be calculated using a refined model with optimized atomic B values.
In order to compare $CC_{map,model}$ with $CC_{full,m}(x)$ (hereafter CC_{full}) all maps should be weighted appropriately. To achieve this, Fourier coefficients of all maps are normalized and weighted by FSC in resolution bins. All correlation examples discussed in this paper were made with such normalized and weighted maps. Details are given in Appendix C.
A relationship between $CC_{map,model}$ and CC_{full} useful for validation is given in equation 4 (Note that eq. 4 is the image space counterpart of the eq. 7 in (Nicholls et al., 2018)).

$$CC_{map,model,m}(x) = CC_{truemap,model,m}(x)\sqrt{CC_{full,m}(x)} \quad (4)$$

According to equation 4, the $CC_{map,model}(x)$ is equal to $\sqrt{CC_{full,m}(x)}$ only when $CC_{truemap,model,m}(x) = 1$, i.e. perfect model. Since this situation is almost never realised,

the $CC_{map,model,m}(x)$ should always be less than $\sqrt{CC_{full,m}(x)}$. If $CC_{map,model,m}(x)$ is greater than $\sqrt{CC_{full,m}(x)}$, that could be an indication of overfitting.

3.2. Parameter estimation and map calculations: likelihood and posterior distribution

As in any application of Bayesian computations to the data analysis we need two probability distributions: 1) probability distribution of observations given parameters to be estimated – likelihood function and 2) probability distribution of unknown signal given observations and current model parameters – posterior probability distributions. See Appendix B for details.

Likelihood function

The negative log likelihood function in the absence of atomic models and in the presence of multiple related maps is (for details see Appendix B9):

$$LL(\boldsymbol{R},\boldsymbol{t},\boldsymbol{\Sigma}_{o,s}) = \sum_s \left(\left(\boldsymbol{F}_o(\boldsymbol{R}^T\boldsymbol{s})e^{-2\pi\boldsymbol{s}^T\boldsymbol{t}}\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{\mathrm{o,s}} + \boldsymbol{\sigma}^2(\boldsymbol{s})\right)^{-1}(\boldsymbol{F}_o(\boldsymbol{R}\boldsymbol{s})e^{-2\pi\boldsymbol{s}^T\boldsymbol{t}}) + \log\left(det\left(\boldsymbol{\Sigma}_{o,s}(\boldsymbol{s}) + \boldsymbol{\sigma}^2(\boldsymbol{s})\right)\right) \quad (5)$$

Where $\boldsymbol{F_o}(\boldsymbol{s})$ are Fourier coefficients from the observed maps, $\boldsymbol{R}$ and $\boldsymbol{t}$ are rotation and translation parameters, $\boldsymbol{\Sigma}_{\mathrm{o,s}}(\boldsymbol{s})$ is the covariance matrix between "true" maps calculated using observed maps and $\boldsymbol{\sigma}^2(\boldsymbol{s})$ is a diagonal matrix of the noise variances. Note that if we relax the conditions that $\boldsymbol{R}$ is a rotation matrix then this formula allows refinement of magnification parameters also. In cryo-EM, we assume that the magnification is a scalar parameter, thus $\boldsymbol{R}$ becomes a diagonal matrix with equal elements with magnification parameter in magnification only refinements. Since the covariances between signals are calculated using observed maps, they already account for the scale between observed maps. The covariance matrices and transformation parameters are estimated iteratively. Diagonal matrix of the noise variance is estimated using half maps as outlined in (Murshudov, 2016), Yamashita et al (Yamashita et al., 2021).

Posterior distribution:

For map calculations we need the probability distribution of the unobserved "true" maps given the current state of atomic models as well as observations. In the absence of atomic models this distribution is a multivariate Gaussian with mean (see Appendix B2):

$$< \boldsymbol{F}_T > = \boldsymbol{k}^{-1}\boldsymbol{\Sigma}_{o,s}\left(\boldsymbol{\sigma}^2 + \boldsymbol{\Sigma}_{o,s}\right)^{-1}\boldsymbol{\sigma}^{-2}\boldsymbol{F}_o \quad (6)$$

where $\boldsymbol{k}$(s) is the diagonal matrix formed by the blurring parameters, $\boldsymbol{\sigma}^2(\boldsymbol{s})$ is the diagonal matrix formed with the variance of observations, $\boldsymbol{\Sigma}_{o,s}$ is the covariance matrix between "true" maps calculated using observed maps, $\boldsymbol{F}_o$ is a vector of Fourier coefficients of observed maps, $< \boldsymbol{F}_T >$ is a vector of expectation values of the "true" map Fourier coefficients.

Since $\boldsymbol{k}(\mathbf{s})$ is unknown, we replace it with the standard deviation of the signal. After some algebraic manipulation we can write (see Appendix B17):

$$< \boldsymbol{F}_T > = \boldsymbol{\rho}_s \sqrt{\boldsymbol{fsc}}\, \boldsymbol{\rho}_o^{-1}\, \boldsymbol{E}_o \quad (7)$$

where, $\boldsymbol{\rho}_s$ is the correlation matrix between Fourier coefficients of "true" maps, $\boldsymbol{\rho}_o$ is the correlation matrix between Fourier coefficients of observed maps, $\boldsymbol{E}_o$ is a vector of normalised Fourier coefficients of observed maps and $<\boldsymbol{E}_T>$ is a vector of the expectation values of the normalised Fourier coefficients of the "true" map.

It must be noted that if we know the blurring factor then it might be better to use them in the map calculations. However, observations alone do not allow to calculate these quantities, they need to be estimated using different methods.

4. Results and Discussion

In this section, we demonstrate the use of local correlation, map overlay and magnification refinement implemented in EMDA package through examples.

4.1. Examples of use of local correlation

4.1.1. Model errors by local correlation

Local correlation is a useful tool for evaluating the degree of local agreement between an atomic model and a cryo-EM map. To demonstrate this, we used EMD-5623 map with overall resolution 3.3 Å and the corresponding atomic model 3j9i (Li et al., 2013). The atomic model was refined against the fullmap to a resolution of 3.3 Å using REFMAC5 (Nicholls et al., 2018). Local correlations were calculated within a kernel of diameter of 5 pixels (pixel size = 1.22 Å). First, the CC_{full} was calculated using the normalized and weighted EMD-5623 halfmaps (see Appendix A). Similarly, the $CC_{map,model}$ was calculated between the normalized and weighted fullmap and the normalized and the weighted calculated map from the refined model. Fig. 2a shows the density near residues U52-54 of the primary map coloured by the $\sqrt{CC_{full}}$ and the superimposed model. One can easily appreciate a fairly good signal, but the model is outside the density. The same density coloured by $CC_{map,model}$ (Fig. 2b) shows low correlation as a result of misplaced model. Thus, $CC_{map,model}$ can be used to highlight areas with modelling errors. At the same time using $\sqrt{CC_{full}}$ those errors can be validated. Fig. 2c shows the corresponding part of the refined atomic model coloured by $CC_{map,model}$ and that also highlighted those residues whose correlations are low compared to other parts of the model. To conclude, a comparison of $CC_{map,model}$ versus $\sqrt{CC_{full}}$ is useful to pinpoint modelling errors such as missing or misplaced regions in the model. Moreover, colouring the atomic coordinates by $CC_{map,model}$ is an effective way to identify misplaced regions in the model.

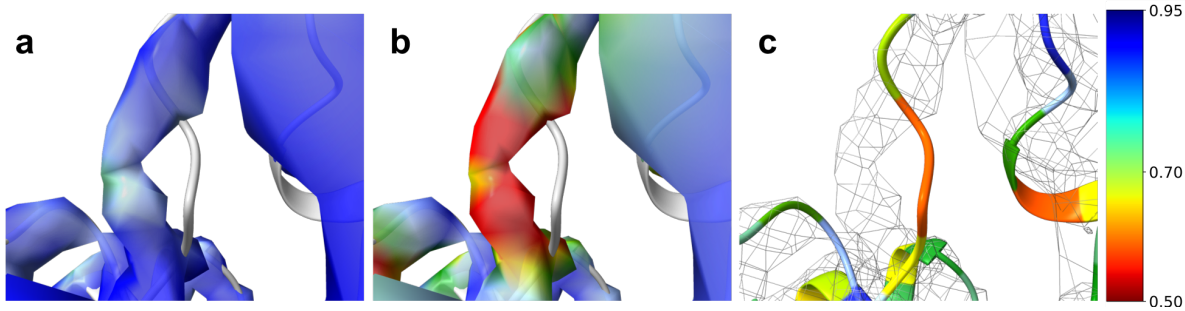

Fig. 2. Identifying model errors by local correlation. a) EMD-5623 primary map density near residues U52-54 of the 3j9i model coloured by $\sqrt{CC_{full}}$. b) the same density coloured by $CC_{map,model}$. c) corresponding atomic model coloured by $CC_{map,model}$ showing low correlation residues U52-54. The figure was made with ChimeraX (Pettersen et al., 2021).

4.1.2. Unmodeled regions by local correlation

In the next example, we use local correlation to highlight an unmodeled ligand in the model 6zge (Wrobel et al., 2020) and EMD-11203 maps. This ligand has been identified as linoleic acid (LA) in the homology model 6zb5 (Toelzer et al., 2020). Both models are SARS-CoV-2 structures of the spike protein. First, the model 6zge was refined against the EMD-11203 fullmap using REFMAC5 (Nicholls et al., 2018) in order to optimise atomic coordinates and the B values. Next, using the normalized and weighted EMD-11203 halfmaps the CC_{full} was calculated within a kernel of diameter of 5 pixels (pixel size = 1.087 Å). The size of the kernel was chosen to maximize the variation of local correlation and minimize the leakage of correlation from the surrounding. The $CC_{map,model}$ was calculated using the normalized and weighted EMD-11203 fullmap and the normalized and weighted model-based map calculated from the refined 6zge model up to 2.6 Å. Then EMD-11203 primary map was coloured by $\sqrt{CC_{full}}$ and $CC_{map,model}$, and their comparison highlighted an unmodeled structured densities located near all receptor binding domains. One such density coloured by $\sqrt{CC_{full}}$ and $CC_{map,model}$ are shown in Fig. 3a and 3b, respectively. The high $\sqrt{CC_{full}}$ implies the density is real, but low $CC_{map,model}$ implies there is no model to explain the density. Next, the homology model 6zb5 with LA was fitted on the EMD-11203 fullmap and refined using REFMAC5 (Nicholls et al., 2018), and the $CC_{map,model}$ was recalculated. The improved $CC_{map,model}$ for the ligand region is shown in Fig. 3c and this improvement is due to the presence of LA in the model.

Fig. 3d shows the refined LA molecule whose atoms are coloured according to the atomic B values. Overall trend shows that B values in the hydrophobic tail are relatively small, but they gradually increase towards the hydrophilic carboxyl group. In Fig. 3e, atomic $\sqrt{CC_{full}}$ and $CC_{map,model}$ are plotted in blue and orange, respectively, along with the atomic B values in grey. The atomic correlation values were obtained from correlation maps by interpolating at atomic positions.

The $\sqrt{CC_{full}}$ is close to 1 throughout the molecule, but largest variation is seen in the carboxyl group. The $CC_{map,model}$ is lower than the $\sqrt{CC_{full}}$ in all atoms indicating no

overfitting, and its variation is larger than that of $\sqrt{CC_{full}}$. The difference in atomic $CC_{map,model}$ for carbonyl O1 and O2 is significant despite their similar B values. The carbonyl group is anchored by the neighbouring Arg408 and Gln409 residues through H-bonding with O2 atom (Toelzer et al., 2020, Fig. 3c), but O1 atom does not seem to have close neighbours thus its atomic correlation may be compromised by the surrounding noise. It should also be noted that at some atoms there is a leakage of correlation. This effect is pronounced at atoms O2, C10, C14 and C18 in Fig. 3b.

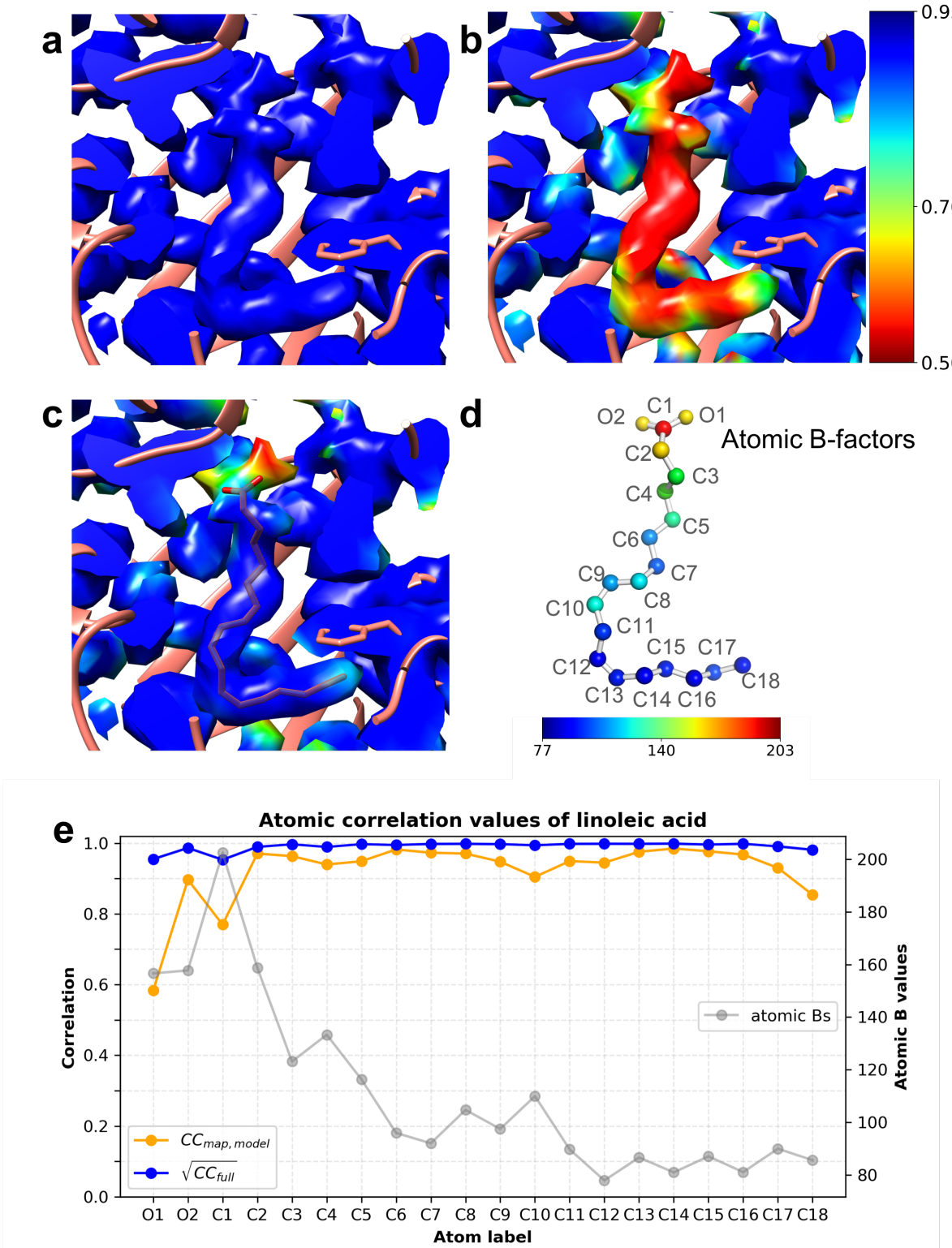

Fig. 3. Use of local correlation to identify unmodeled linoleic acid (LA) in EMD-11203 map. a) unmodeled ligand density in the primary map coloured by the $\sqrt{CC_{full}}$. High correlation indicates the presence of a strong signal. b) the same density coloured by $CC_{map,model}$ calculated between the fullmap and the refined 6zge model using normalized and weighted densities. The correlation in this region is low compared to its surrounding. c) ligand density

coloured by $CC_{map,model}$ calculated between the fullmap and the refined model with LA using normalized and weighted densities. Densities in a, b and c panels were contoured for maximum clarity. Those figures were made with Chimera (Pettersen et al., 2004). d) distribution of atomic B values of refined LA where the atoms are coloured by the B values. This figure was made with PyMOL (Schrödinger, 2020). e) distribution of atomic correlation values at refined LA coordinates. $CC_{map,model}$ and $\sqrt{CC_{full}}$ are shown with orange and blue, respectively. Also, the atomic B values are shown in grey.

A comparison of $\sqrt{CC_{full}}$ versus $CC_{map,model}$ can highlight model errors as shown by previous examples. In model building and refinement, we aim at explaining the full signal in the map by the model, and hence the chance of building model into the noise is unavoidable. In such situations, the local correlation can be a helpful tool to monitor the overfitting. According to Equation 4 the inequality $CC_{map,model} \leq \sqrt{CC_{full}}$ should hold when there is no overfitting, and Fig. 3e shows such a situation.

4.2 Examples of use of map overlay

4.2.1. EMDA map overlay

Map superposition is an important operation in cryo-EM. For instance, in structure comparison studies, this operation brings all maps onto a reference map thereby providing a common coordinate framework for comparison. Also, it is a necessary operation in the difference map calculation and likelihood-based map averaging, where the input maps are required to be pre-aligned. There are several tools available for superposing maps such as Chimera's Fit-in-map (Pettersen et al., 2004), TEMPy2 (Cragnolini et al., 2021) etc. EMDA map overlay is based on the maximisation of the likelihood function given by equation 5 using a quasi-Newton method. To demonstrate the overlay method, we have used cryo-EM maps EMD-21997 and EMD-21999 (Henderson et al., 2020) of SARS-CoV-2 spike protein, whose resolutions are 2.7 Å and 3.3 Å, respectively. The former map is in rS2d locked state in which all three receptor binding domains (RBDs) are down and locked, thus maintaining C3 symmetry. Whereas the latter map is in u1S2q state in which one of the RBDs is open causing the whole structure to C1. In this example, we estimate the movement of one of the down RBDs in EMD-21999 map relative to one of the down RBDs in EMD-21997 map using EMDA overlay operation. We kept the primary EMD-21997 map and the corresponding 6x29 atomic model static, while the primary EMD-21999 map and the corresponding 6x2a atomic model moving. First, we overlaid EMD-21999 map (Fig. 4a(i)) on EMD-21997 map (Fig. 4a(ii)) and the resulting transformation (relative rotation = 8.48°, translation = 4.14 Å) was applied on the 6x2a atomic model. The overlaid maps are shown in Fig. 4a(iii), and they are the starting maps for the subsequent domain overlay (shown by 4a(iv) and (v) for static and moving maps, respectively). Next, a pair of RBDs located proximity to each other on 4a(iii) were extracted within model generated masks. The extracted RBDs are shown in 4a(vi) and 4a(vii) and their superposition before fit optimisation is shown in 4a(viii). Next, their relative

fit was optimized and at the convergence the relative rotation and translation values were 2.54° and 1.03 Å, respectively. The superposed domains after the fit optimisation is shown in 4a(ix). These rotation and translation values indicate the movement of the selected RBD of EMD-21999 map relative to the corresponding RBD of EMD-21997 map in the same coordinate frame. Finally, the estimated transformation between domains was applied on the 6x2a RBD coordinates to bring it on the static model (6x29). Fig. 4a(x) and (xi) present the superposition of models before and after the transformation has been applied, respectively. Fig. 4b shows the FSC curves calculated between the two domains before and after the overlay optimisation.

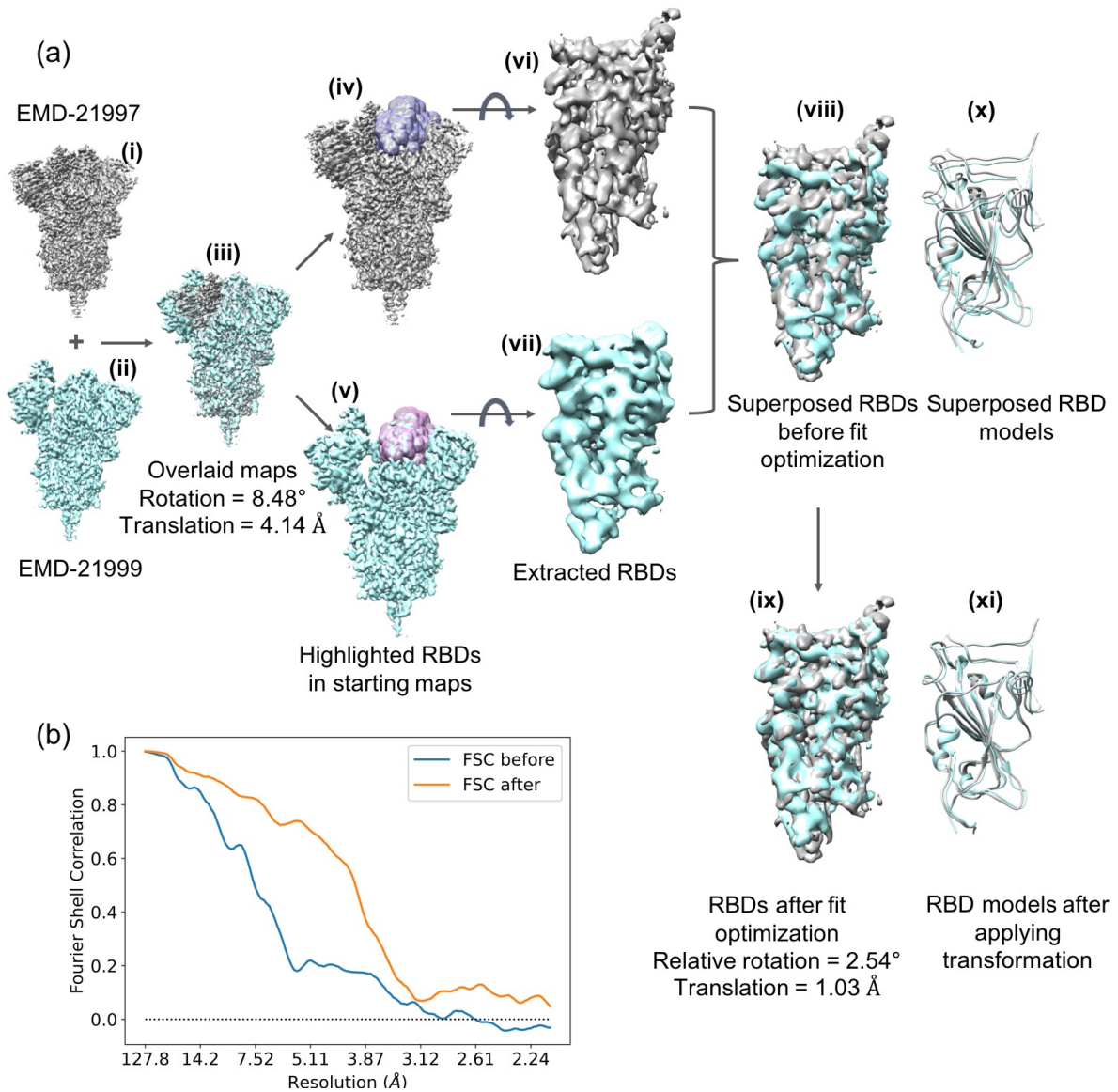

Fig. 4. Map superposition in EMDA illustrated using EMD-21997 and EMD-21999 maps. (a) keeping EMD-21997 map (i) static, EMD-21999 map (ii) was moved to obtain the optimal overlay between them (iii). Starting from the overlaid maps (iv) and (v), RBDs were extracted using masks. The extracted RBDs (vi) and (vii) were superposed (viii) and optimized their overlay (ix) in EMDA. The final values of relative rotation and translation are 2.54° and 1.03 Å, respectively. The same transformation was applied on the model 6x2a of

the moving map. The superposition of 6x29 (static, grey) and 6x2a (moving, cyan) RBD models before (x) and after (xi) the domain transformation being applied. This figure was made with Chimera (Pettersen et al., 2004). (b) FSC between static and moving RBDs before (blue) and after (orange) the overlay optimization.

4.2.2. EMDA magnification refinement

Magnification refinement in EMDA also uses a likelihood function similar to equation 5. Magnification of an EM map is related to the optics of the microscope on which the data has been collected. Thus, merging several data sets collected on different microscopes or on the same microscopy with different optical alignments may require their magnifications to be scaled to a reference (Wilkinson et al., 2019). The reference can be in the form of another map whose accurate pixel size is already known or an atomic model derived independent of those maps. If an atomic model is chosen as the reference, it should not have been refined against the map whose magnification is sought. Magnification refinement involves 1) resampling the target map on the reference grid to make sure both maps refer to the same coordinate system, 2) superposition of the target map on the reference and refining the magnification of the target map, iteratively. To demonstrate the magnification refinement in EMDA, we intentionally introduced a -5% magnification error in one of the half maps of Haemoglobin (half1 of EMD-3651; (Khoshouei et al., 2017)), and let EMDA to refine its magnification. The pixel sizes of the original and the magnification modified maps (initial map, Fig. 5a) are 1.05 and 0.998 Å, respectively. EMDA optimized the magnification of the initial map relative to the original map to yield the rescaled map (Fig. 5a) with the pixel size 1.05 Å. Fig. 5b shows the FSC curves for initial and rescaled maps calculated against the other half map (half2). To validate the accuracy of scaling, the FSC for rescaled map is compared with the half data FSC and they are in very good agreement.

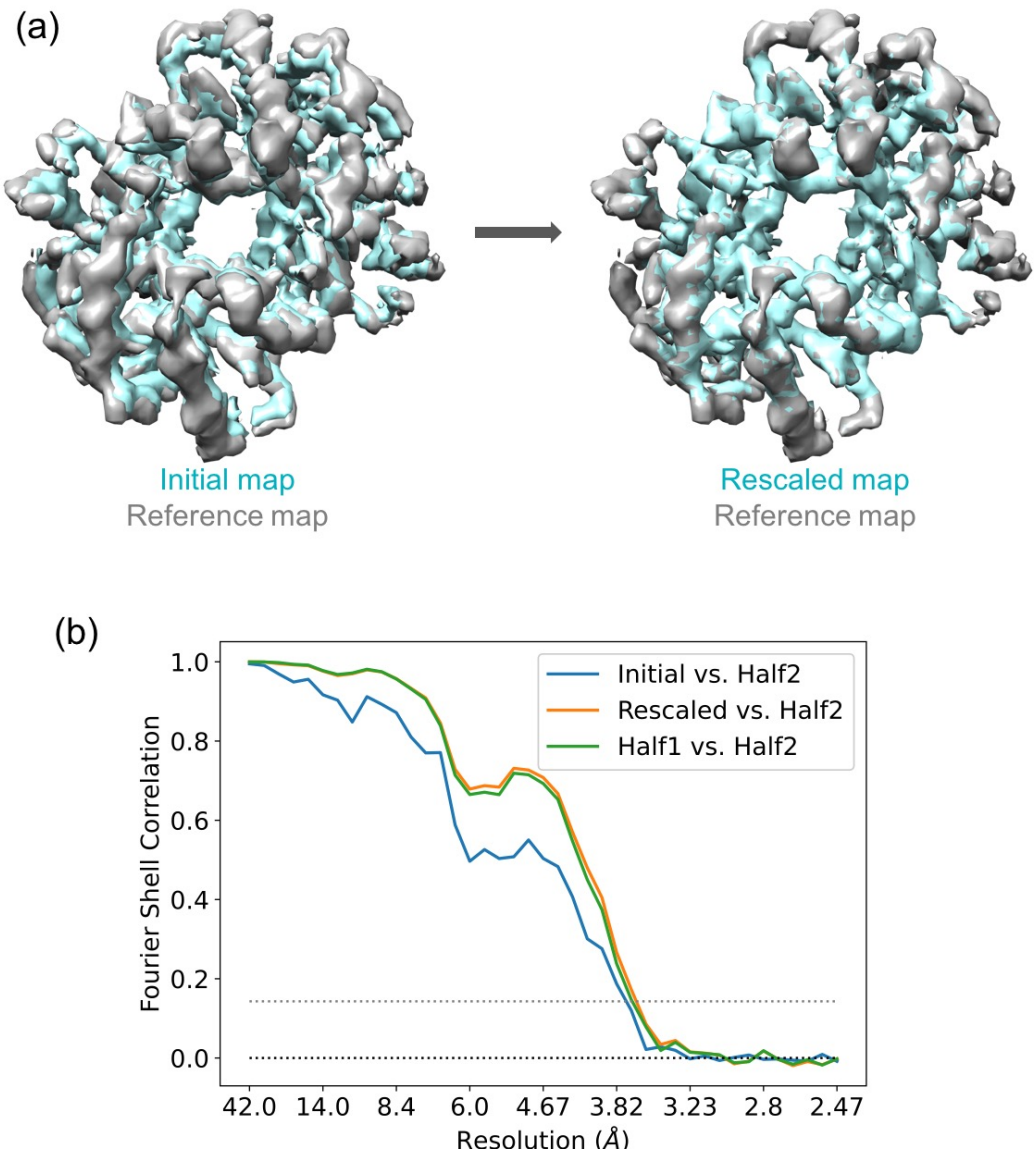

Fig. 5. The magnification scaling in EMDA using Haemoglobin data (EMD-3651). (a) the superposition of the original map (half1/reference map in grey) and the map with -5% magnification error (initial map in cyan) is improved after magnification scaling (rescaled map shown in cyan). This figure was made with Chimera (Pettersen et al., 2004). (b) FSC curves for initial and rescaled maps against the half2 are shown in blue and orange, respectively. The increase of FSC from blue to orange is due to the improved magnification. The green curve is the FSC between the half maps and it serves as the ground truth.

In the next example, we illustrate the estimation of the relative magnification error of two cryo-EM maps of beta-galactosidase [EMD-7770 (Bartesaghi et al., 2018) and EMD-10574 (Saur et al., 2020)] relative to an X-ray crystallography model 3dyp (Juers et al., 2009). The resolution of EMD-7770 and EMD-10574 are 1.9 and 2.2 Å, respectively. The model 3dyp has been derived from X-ray data with resolution 1.75 Å. Both cryo-EM maps and the crystallographic model possess D2 point-group symmetry. Since one of the cryo-EM primary maps (i.e. EMD-10574) has been lowpass filtered, we used fullmaps generated from half

maps for both cryo-EM entries in this analysis. First, all non-polymer atoms of 3dyp model were removed and just the polymers were fitted onto EMD-7770 map in Chimera (Pettersen et al., 2004). Then the model-based map was calculated up to 1.9 Å using REFMAC5 (Nicholls et al., 2018) and it was kept as the crystallographic reference for the subsequent magnification correction analysis. Both the reference map and the EMD-7770 map have the same pixel size 0.637 Å, while EMD-10574 map has 0.68 Å. Thus, the latter map was resampled on the reference to bring all maps on the same coordinate system. Next, a principal component analysis was performed on the variance-covariance matrices of the reference and resampled maps to bring the orientation of the latter approximately matches that of the reference.

Lastly, the fits and the magnifications of EMD-7770 and the resampled EMD-10574 maps were optimized relative to the reference map, iteratively. This resulted in +0.3 % and +1.7 % magnification error in EMD-7770 and EMD-10574 maps relative to the reference, respectively. Fig. 6a(i) and (ii) show the superpositions of EMD-7770 (yellow) and EMD-10574 (cyan) maps on the reference (grey). Their magnified portions enclosed by red rectangles are shown in Fig. 6b on the left two columns. The yellow density overlaid on the grey density does not show an obvious offset discernible to human eye in both center or periphery regions. However, the cyan density shows an offset relative to the grey density. Moreover, this offset increases from the center to periphery; an indication of the magnification problem. Fig. 6a(iii) and (iv) show the magnification corrected EMD-7770 and EMD-10574 maps overlaid on the reference map, respectively. The magnified portions marked by red rectangles are shown in Fig. 6b on the right two columns for center and periphery regions. Both yellow and cyan densities overlay on grey density, and the offset seen in the cyan density before the correction has now disappeared confirming that EMD-10574 map indeed suffers from magnification problem. Furthermore, Fig. 6a(v) and (vi) present the masked FSC curves for EMD-7770 and EMD-10574, respectively, before (blue) and after (orange) the magnification has been corrected. The increase in FSC, especially in (vi) is attributed to the improved magnification.

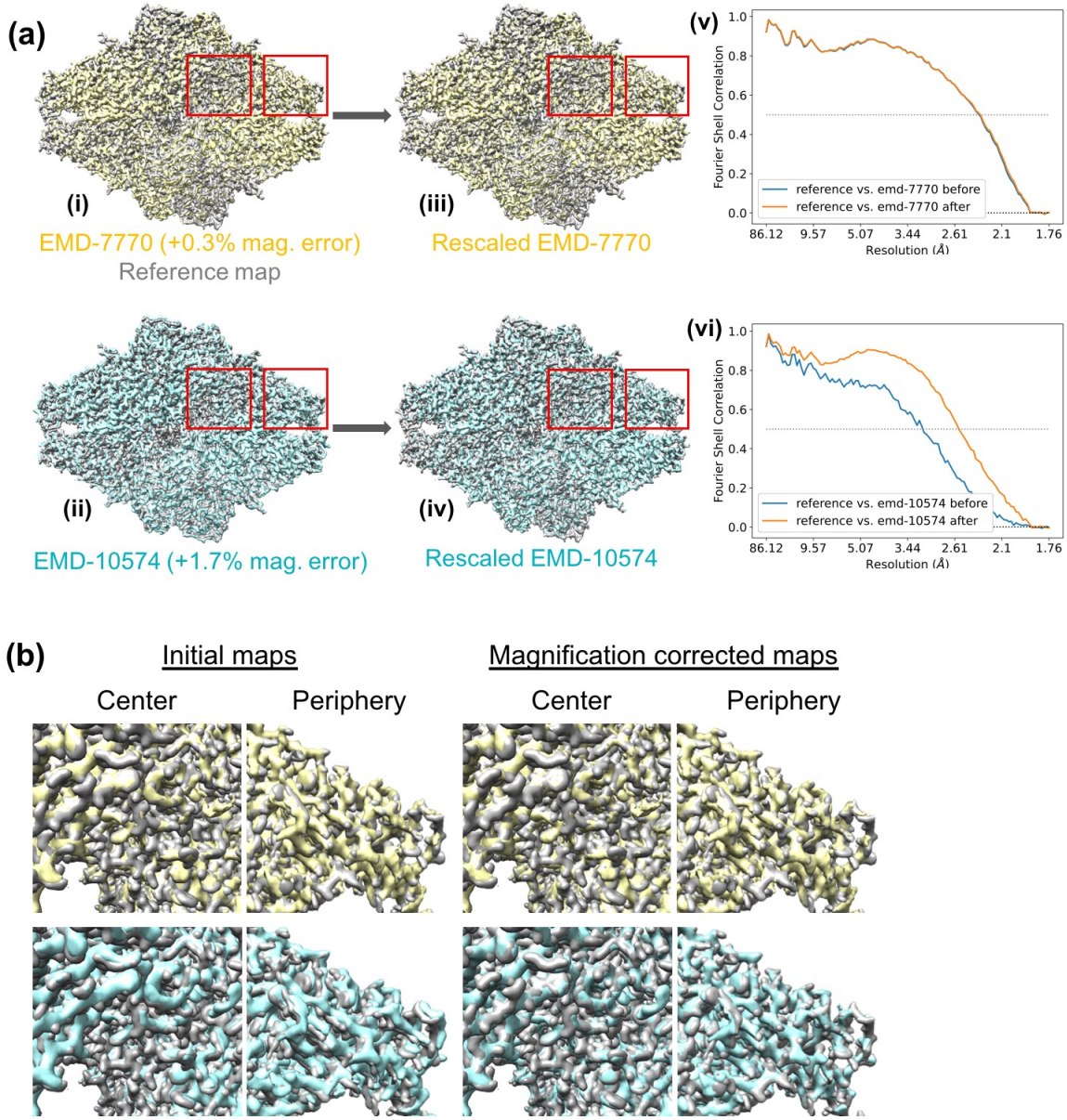

Fig. 6. Magnification correction in EMD-7770, EMD-10574 maps relative to the crystallography model 3dyp. (a) the overlaid EMD-7770 (i, yellow) and EMD-10574 (ii, cyan) maps on the reference map (grey) before the magnification optimisation. (iii) and (iv) are the same maps after the optimisation. The Magnification errors in EMD-7770 and EMD-10574 relative to the reference are +0.3 % and +1.7 %, respectively. The FSC curves for EMD-7770 and EMD-10574 maps against the reference are shown in (v) and (vi), respectively. The blue and orange curves correspond to FSCs before and after the magnification refinement, respectively. The increase in FSC is attributed to the improved magnification. This figure was made with Chimera (Pettersen et al., 2004). (b) comparison of EMD-7770 map (yellow) and EMD-10574 map (cyan) densities against the reference map (grey) in different regions before and after the magnification correction. See text for details.

Even after the magnification correction, some discrepancies in density overlay were apparent in both EMD-7770 and EMD-10574 maps relative to the reference map. We focused on one monomer unit of EMD-7770 map and extracted it using a model generated mask. The corresponding monomer unit of the reference map was also extracted in similar manner. Fig. 7(i) shows the overlaid EMD-7770 map on the reference after the magnification correction. The monomer units chosen is highlighted within the mask. Extracted monomers are shown in Fig. 7(ii), and one can easily appreciate the rotation of the yellow density relative to the reference grey density due to movements between domains. We estimated the relative transformation between those monomer units and that resulted in 1.01° rotation and 0.01 Å translation (similar analysis was performed using monomers from EMD-10574 map and the reference. That resulted in 0.40° rotation and 0.11 Å translation). Fig. 7(iii) and (iv) show the optimized fit of the monomers and the FSCs between them before (blue) and after (orange) the fit optimisation, respectively. The increase in FSC is attributed to the improved fit. The sudden increase in both FSC curves at high resolution is a result of tight masks used to extract monomers.

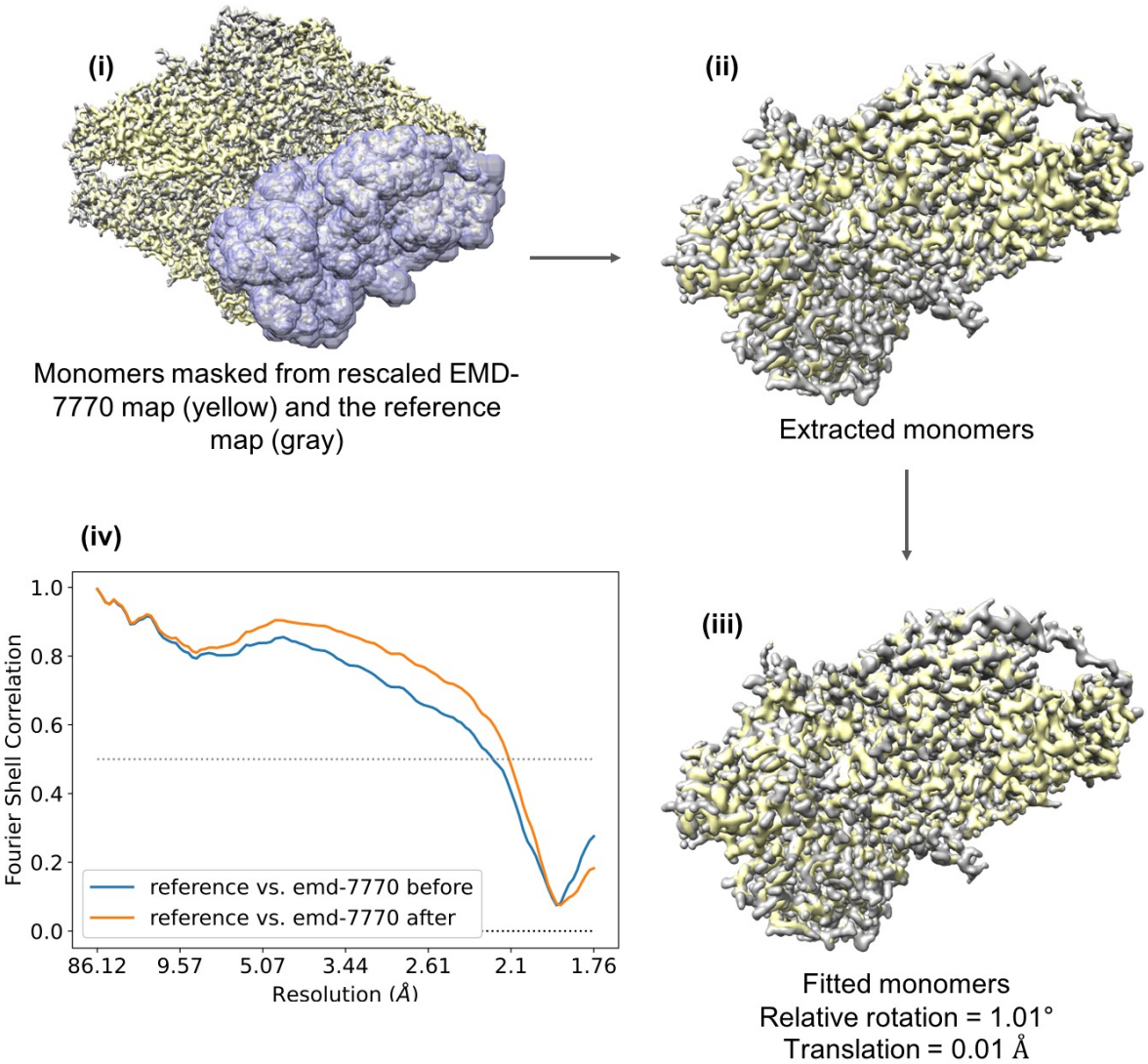

Fig. 7. Movement of one monomer unit of EMD-7770 (yellow) relative to the corresponding monomer unit of the reference map (grey). Selected monomers are highlighted in (i) and those extracted are shown in (ii) before the fit optimisation. (iii) the monomer units after the fit optimisation. (iv) FSCs between monomer units before (blue) and after (orange) the fit optimisation. This figure was made with Chimera (Pettersen et al., 2004).

As illustrated in this example, the map magnification is an important factor to consider during structural comparison studies. It should be refined and make sure all structures have the same magnification before comparing for other structural variations. Internal motions such as domain movements should be estimated and compared to other similar structures only if their magnifications are comparable.

5. Conclusions

We presented the EMDA Python package to serve the need of map and model validation in cryo-EM. The architecture of EMDA contains three layers. The innermost layer contains core and extension modules, while the outermost layer forms the EMDA command line interface. The middle layer abstracts the underlying complex functionalities into an API for further extensions. Methods section contains the mathematical framework of EMDA for local correlation calculation and for probability-based methods in EMDA. In results and discussion section, we illustrated several EMDA functionalities through examples. We showed the use of map-model local correlation to identify residues outside the density or those poorly fitted. Since the full map local correlation gives an indication of the signal level in the map, it can be used to draw insights about the presence of a signal. Moreover, a comparison of map-model local correlation with fullmap local correlation can be used for validating the model-to-map fit. In one of the examples, we used the local correlation to identify an unmodeled ligand in a map, thereby showing its complementary nature to the difference map. The use of local correlation to identify ligands in maps may be superior to the use of difference map because the correlation naturally offers a way to validate the presence/absence of the density as revealed by the half map local correlation. Also, we showed that correlation values mapped into atoms are useful to study the local signal variations. Since, both fullmap and map-model local correlation values resemble the atomic B value distribution, they can be used to derive local B values, and this analysis is underway. Secondly, we presented the likelihood-based map-to-map fitting using an example, where two SARS-CoV-2 structures were first fitted to bring them on the same coordinate frame. Then two receptor binding domains were fitted in the same coordinate frame to estimate their relative movement. Such operations are often useful in structural comparison studies, where the structures are compared on the basis of their commonalities and differences. Thus, density overlay is an essential pre-operation for difference map and/or average map calculation. The last example illustrated the use of likelihood-based fitting for magnification scaling where we corrected the magnifications of two cryo-EM maps relative to an X-ray

derived model. The importance of correcting the relative magnification between structures in structure comparison studies have been highlighted. Other EMDA tools include: FSC calculation and validation using half maps, local correlation calculation in Fourier space to identify preferential orientation of particles, composite map calculation to combine several reconstructions into a common coordinate system, generating masks from half maps, maps and atomic models, difference map and bestmap calculation, and validating symmetry in maps.

Software availability

EMDA is released under the Mozilla Public License Version 2.0 (MPL 2.0) and it is free and open source. The source code is accessible at https://gitlab.com/ccpem/emda. EMDA is distributed as a part of CCP-EM suite and also available via Python Package Installer (`pip`). EMDA's documentation is available at https://emda.readthedocs.io, and we encourage the reader to look at the documentation for most recent functionalities and up-to-date instructions.

Acknowledgements:

RW was supported by the MRC and the Wellcome Trust (grant numbers: MC_UP_A025_1012, 208398/Z/17/Z), GNM and KY were supported by MC_UP_A025_1012, We thank Takanori Nakane for useful discussions, providing test cases and comments on the initial manuscript, Jake Grimmet and Toby Darling for providing scientific computing resources, UKRI, CCP-EM, UK Cryo-EM validation group, the MRC-LMB for creative environment and the users for their valuable feedback.

Appendix A

Local correlation in image space

Let $\psi(x)$ and $m(x)$ be two functions where the latter is normalized

$$\int_{R^3} m(x)dx = 1 \quad \text{(A1)}$$

Then the local averages of $\psi(x)$ with the kernel $m(x)$ can be written as a convolution operation:

$$\psi_m(x) = \int_{R^3} \psi(y)m(x-y)dy \quad \text{(A2)}$$

This can be calculated using the convolution theorem:

$$\psi_m(x) = \mathcal{F}^{-1}[(\mathcal{F}(\psi(x))(\mathcal{F}(m(x))] \quad \text{(A3)}$$

where $\mathcal{F}$ is the Fourier transformation operator and $\mathcal{F}^{-1}$ is its inverse.
Similarly, the local covariance between $\psi_1(x)$ and $\psi_2(x)$ is given by

$$cov_{12,m}(x) = \int_{R^3} m(x-y)\,\psi_1(y)\psi_2(y)dy - \int_{R^3} m(x-y)\,\psi_1(y)dy \int_{R^3} m(x-y)\,\psi_2(y)dy \quad \text{(A4)}$$

The local correlation between the two functions can be written as

$$CC_{12,m}(x) = \frac{cov_{12,m}(x)}{\sqrt{var_{1,m}(x)\,var_{2,m}(x)}} \quad \text{(A5)}$$

Now, let us assume that there are two noisy maps, each has

$$\psi_{i,n}(x) = \psi_i(x) + n_i(x) \quad \text{(A6)}$$

where $\psi_i(x)$ and $n_i(x)$ are the signal and the noise components in i^{th} map.
If the noise components between the maps, and the noise and signal within each map are uncorrelated, then the local variance and covariance can be written:

$$var_{i,m}\left(\psi_{i,n}(x)\right) = var_{i,m}\left(\psi_i(x)\right) + var_{i,m}(n_i(x)) \quad \text{(A7)}$$

$$cov_{12,m}(\psi_{1,n}(x), \psi_{2,n}(x)) = cov_{12,m}\left(\psi_1(x), \psi_2(x)\right) \quad \text{(A8)}$$

And finally, the local correlation can be written

$$CC_{12,n}(x) = \frac{cov_{12,m}(\psi_1(x),\psi_2(x))}{\sqrt{var_{1,m}(\psi_1(x))+var_{1,m}(n_1(x))}\sqrt{var_{2,m}(\psi_2(x))+var_{2,m}(n_2(x))}} \quad \text{(A9).}$$

If two maps are cryo-EM half maps, then they share a common signal. If the noise components have the same variance for both halves then the following relationships hold

$$cov\left(\psi_1(x), \psi_2(x)\right) = var\left(\psi_1(x)\right) = var\left(\psi_2(x)\right) \quad \text{(A10)}$$

$$var\left(n_1(x)\right) = var\left(n_2(x)\right) = 2var\left(n(x)\right) \quad \text{(A11)}$$

where $var\left(n(x)\right)$ is the noise variance in the averaged map.
Thus, the local correlation between half maps is:

$$CC_{half,m}(x) = \frac{var_{1,m}(\psi_1(x))}{var_{1,m}(\psi_1(x))+2var(n(x))} \quad \text{(A12)}$$

The corresponding local correlation in the full map (combined map) according to (Rosenthal and Henderson, 2003) is

$$CC_{full,m}(x) = \frac{2CC_{half,m}(x)}{1+CC_{half,m}(x)} \quad \text{(A13)}$$

Appendix B

Likelihood function

In this appendix, we assume that we have *N* maps and corresponding *N* models. In general, there could be more maps than models, that will be considered in future. We assume that noises between different maps, and noise and signal within maps are uncorrelated,

however underlying signals are related. We also assume that the conditional probability distributions of Fourier coefficients of the maps given the Fourier coefficients of true signals are Gaussian, and the joint probability distributions of Fourier coefficients of all "true" signals given atomic models are multivariate Gaussian distributions.

$$P\big(F_{o,i}(s); F_{T,i}(s)\big) = N_2(k_i(s)F_{T,i}(s), \tfrac{1}{2}\sigma_{n,i}^2(s)) \quad \text{(B1)}$$

I.e, two dimensional normal distributions with mean equal $k_i(s)F_{T,i}(s)$ and variances are equal to $\frac{1}{2}\sigma_{n,i}^2(s)$. Where $k_i(s)$ is the scale parameters for the "true" map number *i* that has an interpretation that "true" signal is blurred with a position independent point spread function before and/or during observations and/or data processing. Under an assumption that blurring is with an isotropic Gaussian kernel then $k_i(s)$ can be expressed in a form of Gaussians parameterised using B values - $e^{(-\frac{B_i|s|^2}{4})}$. We will assume that the "true" signals are on the same coordinate frame, however, observations may have been made for rotated and translated molecules. Then the probability distribution of individual Fourier coefficients will have the form:

$$P\big(F_{o,i}(s); F_{T,i}(s)\big) = N_2(k_i(s)F_{T,i}\big(R_i^T s\big)e^{\iota s^T t_i}, \tfrac{1}{2}\sigma_{n,i}(s)) \quad \text{(B2)}$$

To get the total conditional probability distribution of observed Fourier coefficients all individual components are multiplied. Then, we can transfer transformations to the observed Fourier coefficients. To do this, it is assumed that variances of noise are the same on the surface of each sphere with a radius $|s|$. We also ignore correlation between different Fourier coefficients after rotation. We can write for each Fourier coefficients after transformation of observations:

$$P\Big(F_{o,i}(R_i s)e^{-\iota s^T t_i}; F_{T,i}(s)\Big) = N_2\left(k_i(s)F_{T,i}(s), \tfrac{1}{2}\sigma_{n,i}^2(s)\right) \quad \text{(B3)}$$

The distribution of the Fourier coefficients of "true" signals given Fourier coefficients from atomic model is a multivariate normal distribution.

$$P(\boldsymbol{F_T}(s); \boldsymbol{F_c}) = N_{2N}(\boldsymbol{D}\ \boldsymbol{F_c}, \tfrac{1}{2}\boldsymbol{\Sigma}) \quad \text{(B4)}$$

Where the subscript $2N$ signifies 2N dimensional normal distribution. $\boldsymbol{D}$ is a vector of scale factors between calculated and "true" Fourier coefficients and $\boldsymbol{\Sigma}$ is the matrix of covariances

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1N} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma_{N1} & \Sigma_{N2} & \cdots & \Sigma_{NN} \end{pmatrix} \quad \text{(B5)}$$

With $\Sigma_{ij} =< (F_{T,i} - D_i F_{c,i})(F_{T,j} - D_j F_{c,j})^* >$ to be estimated using data and atomic model. We will denote by $\boldsymbol{\sigma^2}$ a diagonal matrix formed by the variance of the observational noises. In the subsequent treatments we will use the following relationship given, for example, in (Murshudov, 2016):

$$P(\boldsymbol{F_o}, \boldsymbol{F_T}; \boldsymbol{F_c}) = P(\boldsymbol{F_o}; \boldsymbol{F_T}, \boldsymbol{F_c})P(\boldsymbol{F_T}; \boldsymbol{F_c}) = P(\boldsymbol{F_o}; \boldsymbol{F_T})P(\boldsymbol{F_T}; \boldsymbol{F_c}) \quad \text{(B6)}$$

I.e. conditional distribution of observations given "true" and model Fourier coefficients does not depend on model parameters, i.e. if we know "true" signal then we do not need any

model to describe the observations. Now we are ready to derive desired probability distributions.

We need two types of probability distributions: 1) conditional probability distributions given observations – likelihood and 2) conditional probability distribution of "true" Fourier coefficients given observations and model. First one is needed for parameter estimation and the second one is needed for updated map calculations.

Likelihood function:

To derive the likelihood function we need to integrate the joint probability distribution of observation and "true" Fourier coefficients over "true" Fourier coefficients:

$$P\left(\boldsymbol{F}_o(\boldsymbol{R}^T\boldsymbol{s})e^{-\iota\boldsymbol{s}^T\boldsymbol{t}};\boldsymbol{F}_c\right) = \int_{F_T} P(\boldsymbol{F_o};\boldsymbol{F_T})P(\boldsymbol{F_T};\boldsymbol{F_c})d\boldsymbol{F_T} =$$

$$\int_{F_T} N_{2N}\left(\boldsymbol{k}(\boldsymbol{s})\boldsymbol{F_T}(\boldsymbol{R}^T\boldsymbol{s})\boldsymbol{e}^{-\iota\boldsymbol{s}^T\boldsymbol{t}},\boldsymbol{\sigma}^2(s)\right)N_{2N}(\boldsymbol{DF_c},\boldsymbol{\Sigma})d\boldsymbol{F_T} = N_{2N}(\boldsymbol{k}(s)\boldsymbol{D}(s)\boldsymbol{F_c},\tfrac{1}{2}\left(\mathbf{k}^{\mathbf{T}}\boldsymbol{\Sigma}\mathbf{k}+\boldsymbol{\sigma}^{\mathbf{2}}\right)) \quad \text{(B7)}$$

Where $\boldsymbol{\Sigma}$ is the covariance matrix of "true" map without blurring and $\boldsymbol{\Sigma}_{o,s} = \boldsymbol{k}^T\boldsymbol{\Sigma}\boldsymbol{k}$ is the covariance matrix calculated using observed maps and half maps (half maps are used to calculate variances within each map), $\boldsymbol{k}(\boldsymbol{s})$ is a diagonal matrix formed by the scale factors of "true" Fourier coefficient. Since, $\boldsymbol{k}$ is diagonal matrix, it is symmetric. We also note that the covariance matrix $\boldsymbol{\Sigma}$ is a symmetric matrix. We will further assume that there are no atomic models, meaning that $F_c = 0$ and $D = 0$. In this case the probability distribution will have the form:

$$P\left(\boldsymbol{F}_o(\boldsymbol{R}^T\boldsymbol{s})e^{-\iota\boldsymbol{s}^T\boldsymbol{t}};atoms\right) = \frac{C}{\det(\boldsymbol{\Sigma}_{o,s}+\boldsymbol{\sigma}^2)}e^{-\left((\boldsymbol{F}_o(\boldsymbol{R}^T\boldsymbol{s})e^{-\iota\boldsymbol{s}^T\boldsymbol{t}})^T(\boldsymbol{\Sigma}_{o,s}+\boldsymbol{\sigma}^2)^{-1}\left(\boldsymbol{F}_o(\boldsymbol{Rs})e^{-\iota\boldsymbol{s}^T\boldsymbol{t}}\right)\right)} \quad \text{(B8)}$$

where C is a constant, *atoms* signifies that observations are made for a molecule that consists of atoms, but we do not know their positions. We do not know $\boldsymbol{\Sigma}$, and scale factors - $\boldsymbol{k}$ – are absorbed by the covariance matrix. It essentially means that using only observations it is not possible to estimate blurring of the maps. During the estimation of covariances and rotations it is not necessary to scale observations together. However, we should keep in mind that estimated covariance matrix has absorbed scale of observations as well as relative blurring of different observations. In practice, to calculate the elements of the covariance matrices we use half data reconstructions for all maps.

The negative log likelihood function is derived by taking the negative log of the product of all individual probability distributions

$$LL = \sum_s((\boldsymbol{F}_o(\boldsymbol{R}^T\boldsymbol{s})e^{-\iota\boldsymbol{s}^T\boldsymbol{t}})^T\left(\boldsymbol{\Sigma}_{\mathrm{o,s}}+\boldsymbol{\sigma}^2\right)^{-1}\left(\boldsymbol{F}_o(\boldsymbol{Rs})e^{-\iota\boldsymbol{s}^T\boldsymbol{t}}\right)+\log\left(det\left(\boldsymbol{\Sigma}_{o,s}+\boldsymbol{\sigma}^2\right)\right) \quad \text{(B9)}$$

The formula B9 is used in EMDA to estimate rotation and translations of maps into each other as well as for magnification refinement. EMDA uses special case of this, two observation case to fit two maps into each other. In general, R is a rotation matrix. However, if we relax this condition then we can also account for relative magnification of maps. If the only difference between maps is the relative isotropic magnification then R will become a diagonal matrix where diagonal elements are relative magnification parameter.

Posterior probability distribution:

The second probability distribution we need for the Bayesian computation is the conditional probability distributions of "true" given observed and calculated Fourier coefficients. According to the Bayes's theorem:

$$P(\boldsymbol{F}_T; \boldsymbol{F}_o, \boldsymbol{F}_c) = \frac{P(\boldsymbol{F}_o; \boldsymbol{F}_T) P(\boldsymbol{F}_T; \boldsymbol{F}_c)}{P(\boldsymbol{F}_o; \boldsymbol{F}_c)} \quad \text{(B10)}$$

Here we used the fact that if both $\boldsymbol{F}_T$ and F_c are known then $\boldsymbol{F}_o$ does not depend on $\boldsymbol{F}_c$. Since all involved probability distributions are Gaussian distributions then the resulting distribution will also be Gaussian:

$$P(\boldsymbol{F}_T; \boldsymbol{F}_o, \boldsymbol{F}_c) = \frac{C}{\det(\boldsymbol{\Sigma}_{e,T})} e^{-(\boldsymbol{F}_T - <\boldsymbol{F}_T>)^T \boldsymbol{\Sigma}_{e,T}^{-1} (\boldsymbol{F}_T - <\boldsymbol{F}_T>)} \quad \text{(B11)}$$

Where:

$$< \boldsymbol{F}_T > = \left(\boldsymbol{k}\boldsymbol{\sigma}^{-2}\boldsymbol{k} + \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{k}\boldsymbol{\sigma}^{-2}\boldsymbol{F}_o + (\boldsymbol{k}\boldsymbol{\sigma}^{-2}\boldsymbol{k} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{F}_c \quad \text{(B12)}$$

$$\boldsymbol{\Sigma}_{e,T} = (\boldsymbol{k}\boldsymbol{\sigma}^{-2}\boldsymbol{k} + \boldsymbol{\Sigma}^{-1})^{-1} \quad \text{(B13)}$$

In the following treatments we will assume that we have no atomic model and therefore $\boldsymbol{F}_c = 0$ and $\boldsymbol{D} = 0$. Special case of (B12) and (B13) when there is one model and one observation is considered in Yamashita et al (2021).

Using the notation $\boldsymbol{\Sigma}_{o,s} = \boldsymbol{k}\boldsymbol{\Sigma}\boldsymbol{k}$ we can write these formulas as:

$$< \boldsymbol{F}_T >= \boldsymbol{k}^{-1}\left(\boldsymbol{\sigma}^{-2} + \left(\boldsymbol{\Sigma}_{o,s}\right)^{-1}\right)^{-1} \boldsymbol{\sigma}^{-2}\boldsymbol{F}_o = \boldsymbol{k}^{-1}\boldsymbol{\Sigma}_{o,s}^{-1}\left(\boldsymbol{\Sigma}_{o,s} + \boldsymbol{\sigma}^2\right)^{-1}\boldsymbol{F}_o \quad \text{(B14)}$$

$$\boldsymbol{\Sigma}_{e,T} = \boldsymbol{k}^{-1}\boldsymbol{\Sigma}_{o,s}\left(\boldsymbol{\sigma}^2 + \Sigma_{o,s}\right)^{-1}\boldsymbol{\sigma}^2\boldsymbol{k}^{-1} \quad \text{(B15)}$$

This can be recognised as a generalised Wiener filter when there are one or more maps. The covariance term is that of the total error: bias towards the weighted maps and the noise covariance combined.

In practice we do not know $\boldsymbol{k}(s)$, it should be estimated using additional information. For current calculations we can replace the elements of $\boldsymbol{k}(s)$ with the standard deviations of the observed signal $k_i(s) = \sqrt{\Sigma_{o,s,ii}}$. If we also denote with $k_1(s)$ as a diagonal matrix formed with $\sqrt{\sigma_i^2 + \Sigma_{o,s,ii}}$ then we can write:

$$< \boldsymbol{F}_T >= \boldsymbol{k}^{-1}\boldsymbol{\Sigma}_{o,s}\left(\boldsymbol{\Sigma}_{o,s} + \boldsymbol{\sigma}^2\right)^{-1}\boldsymbol{k}_1\boldsymbol{k}_1^{-1}\boldsymbol{F}_o = \boldsymbol{k}^{-1}\boldsymbol{\Sigma}_{o,s}\boldsymbol{k}^{-1}\boldsymbol{k}\boldsymbol{k}_1^{-1}\boldsymbol{k}_1\left(\boldsymbol{\Sigma}_{o,s} + \boldsymbol{\sigma}^2\right)^{-1}\boldsymbol{k}_1\boldsymbol{E}_o \quad \text{(B16)}$$

Here we used the notation $E_{o,i} = \frac{F_o}{\sqrt{\Sigma_{o,s,ii} + \sigma_i^2}}$. We recognise that $\boldsymbol{\rho}_s = \boldsymbol{k}^{-1}\boldsymbol{\Sigma}_{o,s}\boldsymbol{k}^{-1}$ is the correlation matrix between "true" maps, $\sqrt{\boldsymbol{fsc}} = \boldsymbol{k}\boldsymbol{k}_1^{-1}$ is the diagonal matrix formed with square roots full map FSCs, $\boldsymbol{\rho}_{s+n} = \boldsymbol{k}_1^{-1}(\boldsymbol{\Sigma}_{o,s} + \boldsymbol{\sigma}^2)\boldsymbol{k}_1^{-1}$ is the correlation matrix between observed maps. Now we can write:

$$< \boldsymbol{F}_T > = \boldsymbol{\rho}_s \sqrt{\boldsymbol{fsc}}\, \boldsymbol{\rho}_{s+n}^{-1}\, \boldsymbol{E}_o \quad \text{(B17)}$$

When there is only one map then this formula gives normalised and *fsc* weighted map. Next, we consider the case when there are only two maps then we can write the formulas in an explicit form:

$$< \boldsymbol{F_T} > = \boldsymbol{\rho_s} \sqrt{\boldsymbol{fsc}}\, \boldsymbol{\rho_o^{-1} E_o} = \frac{1}{1-\rho_o^2} \begin{pmatrix} \sqrt{fsc_1} - \sqrt{fsc_2}\, \rho_s\rho_o & -\sqrt{fsc_1}\rho_o + \sqrt{fsc_2}\rho_s \\ \sqrt{fsc_1}\rho_s - \sqrt{fsc_2}\rho_o & -\sqrt{fsc_1}\, \rho_o\rho_s + \sqrt{fsc_2} \end{pmatrix} \begin{pmatrix} E_{o,1} \\ E_{o,2} \end{pmatrix} \quad \text{(B18)}$$

And

$$< \boldsymbol{F_T} > = \frac{1}{1-\rho_o^2} \begin{pmatrix} \left(\sqrt{fsc_1} - \sqrt{fsc_2}\, \rho_s\rho_o\right)E_{o,1} + \left(-\sqrt{fsc_1}\rho_o + \sqrt{fsc_2}\rho_s\right)E_{o,2} \\ \left(\sqrt{fsc_1}\rho_s - \sqrt{fsc_2}\rho_o\right)E_{o,1} + \left(-\sqrt{fsc_1}\, \rho_o\rho_s + \sqrt{fsc_2}\right)E_{o,2} \end{pmatrix} \quad \text{(B19)}$$

It must be stressed that since correlations are calculated using observed maps and when signal to noise ratio is very small then this estimation can vary dramatically. Therefore, for accurate estimations we may need to improve the estimation of the correlations, especially those between signal components, for example using smoothening or using prior knowledge derived from the PDB.

Appendix C

Use of normalized and weighted maps in local correlation calculation

The best map according to Bayesian interpretation is obtained from eq. B14, and is

$< F_T > = w_1 F_o$ (C1)

where F_o is the observed Fourier coefficients, and $w_1 = \frac{\Sigma_{o,s}}{\Sigma_{o,s}+\sigma^2}\frac{1}{k(s)}$ in which $\Sigma_{o,s}$ and σ^2 are the covariance and the noise variance in the fullmap estimated using half maps in resolution bins. $k(s)$ is a scale factor that associated with distortions of the true signal such as blurring. In the current implementation $k(s)$ is replaced with the standard deviation of the observed signal (i.e. $k(s) = \sqrt{\Sigma_{o,s}}$) to yield

$< F_T > = \sqrt{FSC(full)}E_o$ (C2)

where $E_o = \frac{F_o}{\sqrt{\Sigma_{o,s}+\sigma^2}}$.

While FSC-type weighting dampening down the noise, the normalisation works as a position independent deblurring operation.

Similar to C1, weights can be assigned on the calculated map as follows:

$F_{c,weighted} = w_2 F_c$ (C3)

where F_c are calculated Fourier coefficients.

To be consistent with, we chose $w_1 = w_2$ and scale F_c and F_o by making their variances equal, i.e.,

$$F_{c,scaled} = \frac{\sqrt{\Sigma_{o,s}+\sigma^2}}{\sigma_c} F_c \quad \text{(C4)}$$

where σ_c is the standard deviation of the signal in the calculated map.

Thus,

$$F_{c,weighted} = \sqrt{\frac{\Sigma_{o,s}}{\Sigma_{o,s}+\sigma^2}}\frac{F_c}{\sigma_c} = \sqrt{FSC(full)}E_c \quad \text{(C5).}$$

References:

Afonine, P.V., Poon, B.K., Read, R.J., Sobolev, O.V., Terwilliger, T.C., Urzhumtsev, A., Adams, P.D., 2018a. Real-space refinement in PHENIX for cryo-EM and crystallography. Acta crystallographica. Section D, Structural biology 74, 531-544.

Afonine, P.V., Klaholz, B.P., Moriarty, N.W., Poon, B.K., Sobolev, O.V., Terwilliger, T.C., Adams, P.D., Urzhumtsev, A., 2018b. New tools for the analysis and validation of cryo-EM maps and atomic models. Acta Crystallographica Section D 74, 814-840.

Barad, B.A., Echols, N., Wang, R.Y.-R., Cheng, Y., DiMaio, F., Adams, P.D., Fraser, J.S., 2015. EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy. Nature Methods 12, 943-946.

Bartesaghi, A., Aguerrebere, C., Falconieri, V., Banerjee, S., Earl, L.A., Zhu, X., Grigorieff, N., Milne, J.L.S., Sapiro, G., Wu, X., Subramaniam, S., 2018. Atomic Resolution Cryo-EM Structure of β-Galactosidase. Structure 26, 848-856.e843.

Beckers, M., Jakobi, A.J., Sachse, C., 2019. Thresholding of cryo-EM density maps by false discovery rate control. IUCrJ 6, 18-33.

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, Biological crystallography 71, 136-153.

Burnley, T., Palmer, C.M., Winn, M., 2017. Recent developments in the CCP-EM software suite. Acta crystallographica. Section D, Structural biology 73, 469-477.

Cheng, A., Henderson, R., Mastronarde, D., Ludtke, S.J., Schoenmakers, R.H.M., Short, J., Marabini, R., Dallakyan, S., Agard, D., Winn, M., 2015. MRC2014: Extensions to the MRC format header for electron cryo-microscopy and tomography. Journal of Structural Biology 192, 146-150.

Costa, M.G.S., Fagnen, C., Vénien-Bryan, C., Perahia, D., 2020. A New Strategy for Atomic Flexible Fitting in Cryo-EM Maps by Molecular Dynamics with Excited Normal Modes (MDeNM-EMfit). Journal of Chemical Information and Modeling 60, 2419-2423.

Cragnolini, T., Sahota, H., Joseph, A.P., Sweeney, A., Malhotra, S., Vasishtan, D., Topf, M., 2021. TEMPy2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. Acta crystallographica. Section D, Structural biology 77, 41-47.

Croll, T., 2018. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. Acta Crystallographica Section D 74, 519-530.

Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., 2010. Features and development of Coot. Acta Crystallographica Section D 66, 486-501.

Faruqi, A.R., McMullan, G., 2011. Electronic detectors for electron microscopy. Quarterly Reviews of Biophysics 44, 357-390.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357-362.

Henderson, R., Edwards, R.J., Mansouri, K., Janowska, K., Stalls, V., Gobeil, S.M.C., Kopp, M., Li, D., Parks, R., Hsu, A.L., Borgnia, M.J., Haynes, B.F., Acharya, P., 2020. Controlling the SARS-CoV-2 spike glycoprotein conformation. Nat Struct Mol Biol 27, 925-933.

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. Outcome of the first electron microscopy validation task force meeting. Structure 20, 205-214.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9, 90-95.

Juers, D.H., Rob, B., Dugdale, M.L., Rahimzadeh, N., Giang, C., Lee, M., Matthews, B.W., Huber, R.E., 2009. Direct and indirect roles of His-418 in metal binding and in the activity of beta-galactosidase (E. coli). Protein science : a publication of the Protein Society 18, 1281-1292.

Karplus, P.A., Diederichs, K., 2012. Linking Crystallographic Model and Data Quality. Science 336, 1030.

Khoshouei, M., Radjainia, M., Baumeister, W., Danev, R. 2017. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate, pp. 16099 Nature communications, Vol. 8.

Kim, D.N., Moriarty, N.W., Kirmizialtin, S., Afonine, P.V., Poon, B., Sobolev, O.V., Adams, P.D., Sanbonmatsu, K., 2019. Cryo_fit: Democratization of flexible fitting for cryo-EM. Journal of Structural Biology 208, 1-6.

Kühlbrandt, W., 2014. The Resolution Revolution. Science 343, 1443.

Lawson, C.L., Berman, H.M., Chiu, W., 2020. Evolving data standards for cryo-EM structures. Structural Dynamics 7, 014701.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A., Cheng, Y., 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nature Methods 10, 584-590.

Lyumkis, D., 2019. Challenges and opportunities in cryo-EM single-particle analysis. J Biol Chem 294, 5181-5197.

Lyumkis, D., Julien, J.-P., de Val, N., Cupo, A., Potter, C.S., Klasse, P.-J., Burton, D.R., Sanders, R.W., Moore, J.P., Carragher, B., Wilson, I.A., Ward, A.B., 2013. Cryo-EM Structure of a Fully Glycosylated Soluble Cleaved HIV-1 Envelope Trimer. Science 342, 1484.

Ménétret, J.-F., Schaletzky, J., Clemons, W.M., Osborne, A.R., Skånland, S.S., Denison, C., Gygi, S.P., Kirkpatrick, D.S., Park, E., Ludtke, S.J., Rapoport, T.A., Akey, C.W., 2007. Ribosome Binding of a Single Copy of the SecY Complex: Implications for Protein Translocation. Molecular Cell 28, 1083-1092.

Murshudov, G.N., 2016. Refinement of Atomic Structures Against cryo-EM Maps. Methods in enzymology 579, 277-305.

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, 492-505.

Orlova, E.V., Serysheva, I.I., van Heel, M., Hamilton, S.L., Chiu, W., 1996. Two structural configurations of the skeletal muscle calcium release channel. Nature Structural Biology 3, 547-552.

Penczek, P.A., Frank, J., Spahn, C.M.T., 2006. A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. Journal of Structural Biology 154, 184-194.

Peterson, P., 2009. F2PY: a tool for connecting Fortran and Python programs. Int. J. Comput. Sci. Eng. 4, 296–305.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera - A visualisation system for exploratory research and analysis. Journal of computational chemistry 25, 1605-1612.

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. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein science : a publication of the Protein Society 30, 70-82.

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. Nat Methods 17, 328-334.

Prisant, M.G., Williams, C.J., Chen, V.B., Richardson, J.S., Richardson, D.C., 2020. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics. Protein Science 29, 315-329.

Ramírez-Aportela, E., Maluenda, D., Fonseca, Y.C., Conesa, P., Marabini, R., Heymann, J.B., Carazo, J.M., Sorzano, C.O.S., 2021. FSC-Q: a CryoEM map-to-atomic model quality validation based on the local Fourier shell correlation. Nature Communications 12, 42.

Read, R.J., Adams, P.D., Arendall, W.B., 3rd, 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. A new generation of crystallographic validation tools for the protein data bank. Structure 19, 1395-1412.

Reboul, C.F., Kiesewetter, S., Elmlund, D., Elmlund, H., 2020. Point-group symmetry detection in three-dimensional charge density of biomolecules. Bioinformatics 36, 2237-2243.

Roseman, A.M., 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta crystallographica. Section D, Biological crystallography 56, 1332-1340.

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, 721-745.

Rossmann, M., 2000. Fitting atomic models into electron-microscopy maps. Acta Crystallographica Section D 56, 1341-1349.

Rossmann, M.G., Bernal, R., Pletnev, S.V., 2001. Combining Electron Microscopic with X-Ray Crystallographic Structures. Journal of Structural Biology 136, 190-200.

Saur, M., Hartshorn, M.J., Dong, J., Reeks, J., Bunkoczi, G., Jhoti, H., Williams, P.A., 2020. Fragment-based drug discovery using cryo-EM. Drug Discov Today 25, 485-490.

Scheres, S.H.W., 2014. Beam-induced motion correction for sub-megadalton cryo-EM particles. eLife 3, e03665.

Scheres, S.H.W., Chen, S., 2012. Prevention of overfitting in cryo-EM structure determination. Nature Methods 9, 853-854.

Schrödinger, L., & DeLano W. . 2020. PyMOL. http://www.pymol.org/pymol.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J., 2007. EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol 157, 38-46.

Tickle, I.J., 2012. Statistical quality indicators for electron-density maps. Acta crystallographica. Section D, Biological crystallography 68, 454-467.

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. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science 370, 725.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., Sali, A., 2008. Protein structure fitting and refinement guided by cryo-EM density. Structure 16, 295-307.

Trabuco, L.G., Villa, E., Schreiner, E., Harrison, C.B., Schulten, K., 2009. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. Methods 49, 174-180.

Tykac, M., 2018. Protein Shape Description and its Application to Shape Comparison (Doctoral thesis).

Van Heel, M., 1987. Similarity measures between images. Ultramicroscopy 21, 95-100.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.-L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T.*, et al.*, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17, 261-272.

Wang, R.Y.-R., Song, Y., Barad, B.A., Cheng, Y., Fraser, J.S., DiMaio, F., 2016. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. eLife 5, e17219.

Wang, Z., Schröder, G.F., 2012. Real-space refinement with DireX: from global fitting to side-chain improvements. Biopolymers 97, 687-697.

Wasilewski, S., Rosenthal, P.B., 2014. Web server for tilt-pair validation of single particle maps from electron cryomicroscopy. Journal of structural biology 186, 122-131.

Wilkinson, M.E., Kumar, A., Casanal, A., 2019. Methods for merging data sets in electron cryo-microscopy. Acta Crystallographica Section D 75, 782-791.

Wrobel, A.G., Benton, D.J., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., Skehel, J.J., Gamblin, S.J., 2020. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 27, 763-767.

Yamashita, K., Palmer, C.M., Burnley, T., Murshudov, G.N., 2021. Cryo-EM single particle structure refinement and map calculation using Servalcat. bioRxiv.