Improvement of cryo-EM maps by density modification

Thomas C. Terwilliger[1,2], Steven J. Ludtke[3], Randy J. Read[4], Paul D. Adams[5,6], Pavel V. Afonine[5]

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

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

[3]Baylor College of Medicine, Houston, TX 77030 USA

[4]Cambridge Institute for Medical Research, Cambridge, CB2 0XY, UK

[5]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720-8235, USA

[6]Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA

Abstract

An iterative density modification procedure for improving maps produced by single-particle electron cryo-microscopy is presented. The theoretical basis of the method is identical to that of maximum-likelihood density modification, previously used to improve maps from macromolecular X-ray crystallography. Two key differences from applications in crystallography are that the errors in Fourier coefficients are largely in the phases in crystallography but in both phases and amplitudes in electron cryo-microscopy , and that half-maps with independent errors are available in electron cryo-microscopy . These differences lead to distinct approaches for combination of information from starting maps with that obtained in the density modification process. The applicability of density modification theory to electron cryo-microscopy was evaluated using half-maps for apoferritin at a resolution of 3.1 Å and a matched 1.8 Å reference map. Error estimates for the map obtained by density modification were found to closely agree with true errors as estimated by comparison with the reference map. The density modification procedure was applied to a set of 54 datasets where half-maps, a full map and a model all had been deposited. The procedure improved map resolution (the

resolution at which the map-model Fourier shell correlation dropped to ½) by an average of 0.26 Å (standard deviation of 0.12 Å) and increased the visibility of details in the maps. The procedure requires only two unmasked half-maps and a sequence file or other source of information on the volume of the macromolecule that has been imaged.

Single-particle electron cryo-microscopy (cryo-EM) is rapidly becoming the dominant technique for determination of large three-dimensional structures of macromolecules and their complexes[1]. The result of a cryo-EM analysis is a three-dimensional map reflecting the electric potential of the macromolecule[2] and which has map values and an appearance closely related to maps obtained from X-ray crystallography[3]. In both cryo-EM and in macromolecular crystallography, the accuracy of the map, and in particular, accuracy as a function of resolution, is a crucial characteristic. Typically both cryo-EM and crystallographic maps have high accuracy at low resolution, then the accuracy decreases at higher resolution. The resolution at which the map-model correlation (the correlation between Fourier coefficients representing the map and true Fourier coefficients) falls off to about ½ is often used as a measure of the "resolution" of a cryo-EM map[4]. Typically, the higher the resolution of a map (the smaller the value of the resolution in Å), the better the interpretability of the map[5].

In macromolecular crystallography, the amplitudes of Fourier coefficients are generally measured accurately and the phases are poorly estimated. It is common practice in that field to carry out a procedure known as density modification to reduce the errors in the phases and thereby improve the resulting map[6-9] The source of new information in crystallographic density modification is prior knowledge about expected values in a map. For example, the probability distribution of map values may be known, or there may be knowledge about specific features in the map such as a flat solvent region. Information about the true density in part or all of the map can be used to obtain improved estimates of the phases, and these improved phases lead to an improved map everywhere, not just where the information was applied[8].

In cryo-EM a form of density modification may be applied during the process of image reconstruction. The macromolecule typically occupies only a small part of the volume of the reconstruction, and during reconstruction noise is removed from the part of the map that is outside the macromolecule[10,11]. This can improve the map in the region of the macromolecule and is related to the "solvent flattening" aspect of crystallographic density modification[12]. Though the overall process of density modification as implemented in a crystallographic setting

is thought to be inappropriate for cryo-EM[13], it has been suggested that the general concept could be adapted and applied[10,13]. Here we show that a version of density modification with the same theoretical basis as crystallographic density modification but with key differences reflecting the differences between crystallographic and cryo-EM maps can be used to improve cryo-EM maps.

There are several frameworks for density modification that could be applied to cryo-EM maps[6]. Here we use maximum-likelihood density modification as it makes a clear distinction between information coming from the original data and information that comes from expectations about the features in the map[7]. The process for map improvement by maximum-likelihood density modification involves identifying how the current Fourier coefficients can be changed so as to increase the plausibility of the map (expressed as a likelihood), while retaining compatibility with the original experimental map (see Methods).

Cryo-EM maps differ in fundamental ways from crystallographic maps, and the actual process of density modification cannot be applied in the same way in the two situations. As detailed in Methods, one key difference is that in crystallography only the amplitudes of Fourier coefficients are measured, while in electron cryo-microscopy both phase and amplitudes are directly available from experiment. Another is that half-maps with relatively independent errors are available in electron cryo-microscopy [5] but not in crystallography.

We tested the applicability of density modification theory to cryo-EM by applying it to a map of apoferritin at a reported resolution of 3.1 Å (EM Data Bank entry EMD-20028[14]. We used a matched 1.8 Å reference map (EMD-20026) to evaluate the error estimates that make up a key part of the density modification process. The density modification procedure for the 3.1 Å map was carried out using Fourier coefficients to a resolution of 2.5 Å, so we checked the accuracy of the 1.8 Å map up to this resolution by calculating the Fourier shell correlation (FSC) of independent half-maps[4,5]. Fig. 1A shows that the 1.8 Å half-maps have an FSC value above 0.97 at all inverse resolutions of up to 0.4 $Å^{-1}$ (corresponding to a resolution of 2.5 Å). Half-dataset

FSC values can be used to estimate the expected correlation of Fourier coefficients for a map to Fourier coefficients representing the true map (C_{ref}) using the formula[4], $C_{ref} = [2\ FSC / (1 + FSC)]^{1/2}$. An FSC value of 0.97 corresponds to a value of C_{ref} =0.99, indicating (aside from systematic errors affecting both half-maps) that up to at least a resolution of 2.5 Å, the 1.8 Å map closely matches a perfect map of this structure.

Fig. 1B illustrates estimated resolution-dependent map accuracy[4] (C_{ref}) based on half-dataset FSC values for the 3.1 Å map (orange dots) and shows that they are very similar to actual map accuracy (the Fourier correlation between the 3.1 Å and reference 1.8 Å maps, blue line). From the C_{ref} estimates in Fig. 1B, the resolution of the "3.1 Å" map is 2.97 Å, and based on the correlation to the reference 1.8 Å map, it is 2.98 Å.

For the density-modified map, values of resolution-dependent map quality C_{ref} can be estimated from resolution-dependent error estimates based on correlations among the two original half maps and two map-phasing half maps (see Methods). Fig. 1B displays estimates of C_{ref} for the density-modified map (purple triangles) and shows that they are very similar to the actual Fourier shell correlation values for the density-modified map and the 1.8 Å reference map (green line). Based on C_{ref} values, the density modified map is estimated to have a resolution of 2.77 Å, and its actual resolution estimated from correlation to the reference map is similar (2.79 Å).

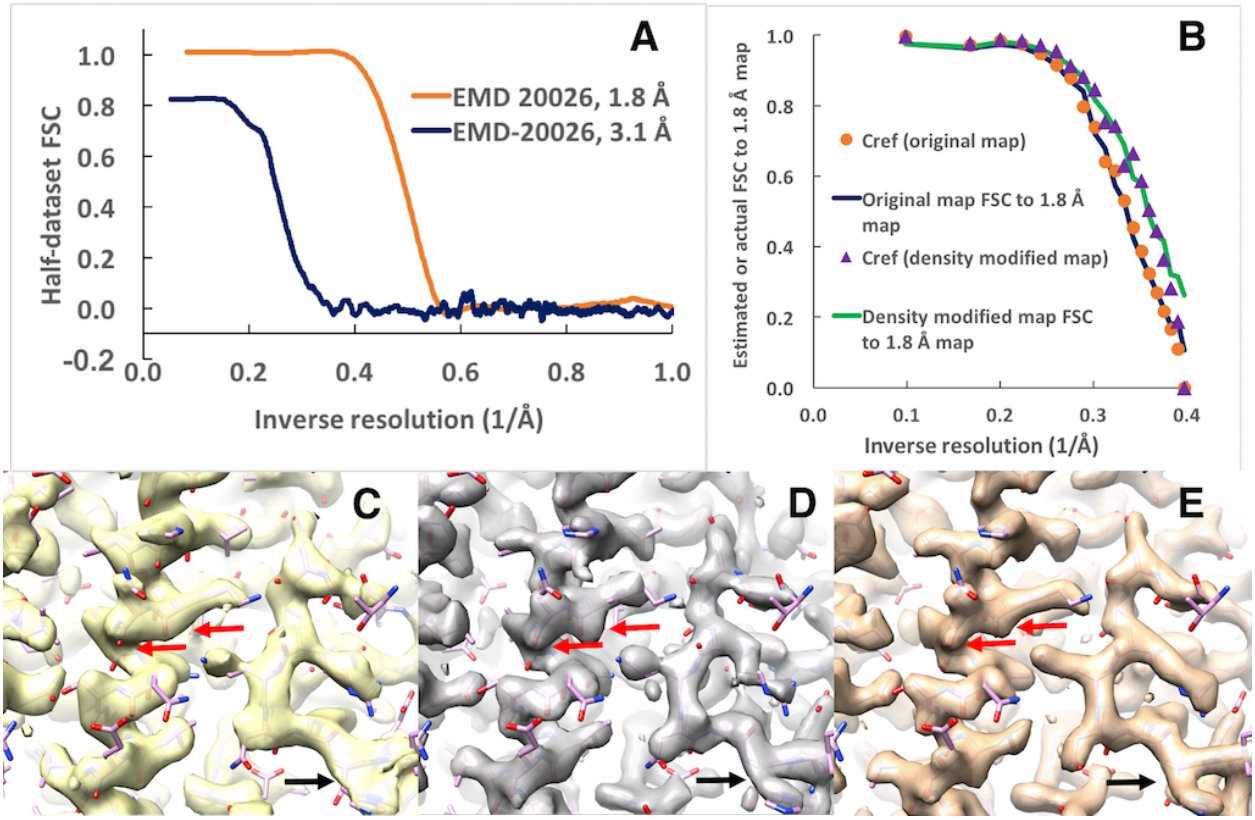

Figure 1. Density modification of apoferritin 3.1 Å map and evaluation using apoferritin 1.8 Å map. A. Fourier shell correlation (FSC) curves for EMD-20028 (3.1 Å) and EMD-20026 (1.8 Å). B. Orange circles are estimated resolution-dependent map quality (C_{ref}) of original 3.1 Å map (see text). Blue line is the FSC of the original 3.1 Å map to the 1.8 Å map. Purple triangles are C_{ref} of density-modified 3.1 Å map, and the green line is the FSC of the density-modified 3.1 Å map to 1.8 Å map. C. Apoferritin 3.1 Å map (EMD-20028), sharpened using X-ray structure 3ajo[15] docked into and refined against the 1.8 Å EMD-20026 map. D. Apoferritin density-modified 3.1 Å map, sharpened using the same refined, docked model. E. Apoferritin 1.8 Å map, blurred with an overall B-value of 22 Å^2 to match the overall B-value of the density-modified map and low-pass filtered at a resolution of 2.5 Å to represent an ideal map (removing Fourier terms between 1.8 Å and 2.5 Å). Contours in C, D, and E are drawn to enclose equal volumes[16]. Black arrows indicate the location of histidine 13 where the distinction between the main-chain and histidine side chain is clear in D and E but not in C. Red arrows indicate the carbonyl oxygen atoms of residue 108 and 109 which can be seen as clear protrusions from the main-chain in D and E but not in C.

This analysis and Fourier shell correlation with the reference 1.8 Å map indicate that the error estimates produced by our density modification procedure are accurate and that the density modification procedure has improved the resolution of this map by about 0.2 Å. Figs. 1C-E illustrate this improvement in resolution visually. Fig. 1C shows the original map, Fig. 1D shows the density-modified map, and Fig. 1E shows the reference 1.8 Å map, low-pass filtered at a resolution of 2.5 Å. To make the visual comparison of original and density-modified maps as unbiased as possible, the maps were all sharpened automatically[17] based on an X-ray structure of apoferritin (3ajo[15]) docked into and refined using the 1.8 Å map[18], and contours are drawn that enclose equal volumes in the two maps[16,19]. As anticipated based on the improvement in resolution, the density-modified map in Fig. 1D shows the features found in the reference map in Fig. 1E with considerably more detail than the original map in Fig. 1C. For example, the distinction is clear in the density-modified map between the main-chain and a histidine side chain (black arrows) and the locations of carbonyl oxygen atoms can be clearly distinguished (red arrows). This improvement in clarity is exactly the type of improvement that can make a difference in map interpretation because clear visualization of side chains and main-chain carbonyl oxygen atoms provide strong constraints on how a model should be built.

Taken as a whole, Fig. 1 indicates that application of density modification to the apoferritin 3.1 Å map improves the map in a significant way that is in close agreement to the improvement expected from the error model for density modification described in Methods.

We next tested the generality of maximum-likelihood density modification of cryo-EM maps by applying it to 54 sets of half-maps and full maps available in the EM Databank (EMDB[14]). We evaluated each map using Fourier shell correlation[5] to the deposited atomic model from the Protein Data Bank (PDB[20]) after re-refining the model the map to be evaluated (see Methods). This purpose of this approach is to create similar biases for the analysis of original and density modified maps and yield a comparison that is fair. The map-model FSC values were used to

estimate the resolution of the map (the resolution at which the map-model FSC is ½) and the average FSC, a measure of overall quality of the map.

Fig. 2A illustrates the estimated resolution for original maps (blue dots) and density-modified maps (orange dots), plotted as a function of the stated resolution of the original maps. The estimated resolutions of the 54 maps were improved by a mean of 0.26 Å by density modification, with relatively consistent improvement (the standard deviation of the improvement was 0.12 Å). Fig 2B shows the overall quality of each original map (blue dots) and density-modified map (orange dots) as reflected in the mean map-model Fourier shell correlation, again plotted as a function of stated resolution. Both the improvement in resolution and improvement in overall map quality appear to be relatively independent of the resolution of the maps.

To examine the visual effects of density modification, Fig. 2 panels C-H show matched pairs of deposited maps and corresponding density modified maps. To make the visual comparison as fair as possible, each map was sharpened automatically using the model refined based on that map, and contours for matched maps were chosen to enclose equal volumes[16] as in Fig. 1. Figs. 2C and 2D show sharpened deposited and density modified maps for β-galactosidase (EMDB 2984[21], PDB entry 5a1a, stated resolution 2.2 Å, resolution improvement of 0.16 Å). Figs 2E and 2F show TrpV1 (EMDB 8118[22], PDB entry 5irz, 3.3 Å, improvement of 0.27 Å), and Figs. 2G and 2H show TrpM8 (EMDB 7127[23], PDB entry 6bpq, 4.1 Å, improvement of 0.36 Å). In each case the density modified map shows substantially more detail than the original map.

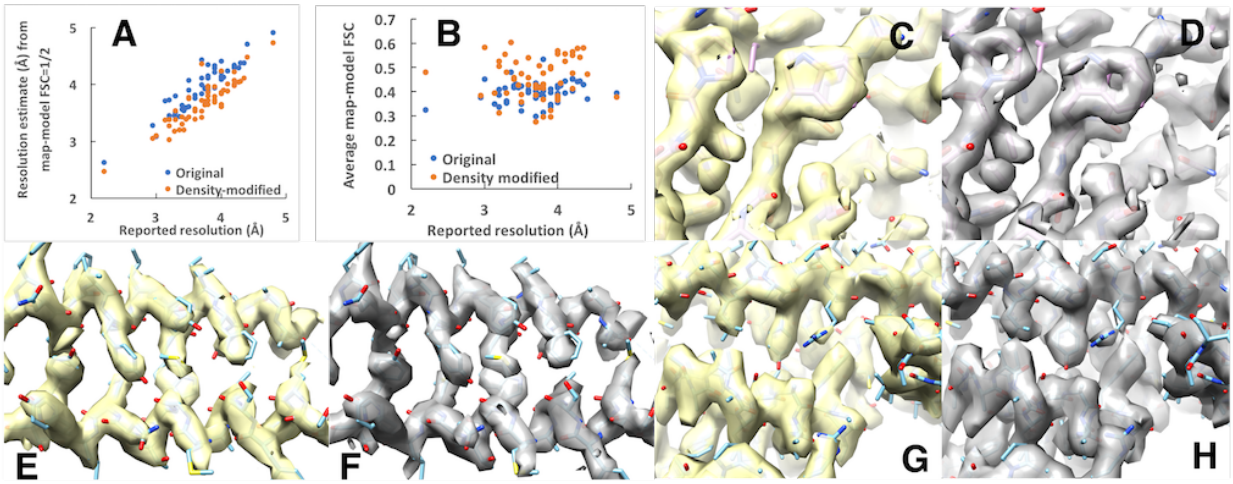

Figure 2. Application of density modification to maps from the EM Data Bank. A. Estimated resolution for original maps (blue dots) and density-modified maps (orange dots) based on map-model FSC using models refined against the maps being examined. B. Mean map-model FSC for original maps (blue dots) and density-modified maps (orange dots), calculated up to inverse resolution of corresponding to 5/6 the stated resolution of the map. C, D. Deposited and density-modified maps for β-galactosidase (2.2 Å, EMDB 2984), sharpened and contoured as described in the text. E,F, TrpV1 (3.3 Å, EMDB 8118), and G,H, TrpM8 (4.1 Å, EMDB 7127), as in C and D.

A limitation in our current implementation of density modification for cryo-EM is the assumption of relatively uniform noise levels throughout the region outside the macromolecule, while actual maps typically have noise levels that vary with distance from the macromolecule. As described in Methods, using a sub-volume from a map in density modification reduces the non-uniformity in noise levels but does not eliminate it. We investigated whether a reconstruction method that produced more uniform noise would improve density modification. We processed a subset of images available for β-galactosidase (EMPIAR-10061[21,24]) with two different procedures using EMAN2[25], yielding maps with resolutions of about 3.9 Å. The first procedure was a standard reconstruction with default parameters including a Gaussian kernel with a resolution-dependent width except that no final masking was applied to the half-maps. A full map with masking was also generated from these

half maps and used for comparison. The second procedure was a reconstruction with a fixed-width Gaussian kernel, expected to yield a more uniform noise distribution in real-space, to test whether the non-uniformity in noise in a cryo-EM map is a limiting factor. The maps obtained from each procedure were density modified, and the resolution of each map was estimated from Fourier shell correlation to the deposited 2.2 Å map for β-galactosidase (EMD 2984[21]), superimposed on the reconstructions.

The full map generated with the standard reconstruction had a resolution of 3.9 Å based on a comparison to the deposited 2.2 Å map. Simple averaging of the half-maps from this reconstruction gave a map with estimated resolution of 4.0 Å, and density modification of these maps gave a map with an estimated resolution of 3.8 Å. Simple averaging of the half-maps with more uniform noise yielded a map with a slightly higher resolution than the standard reconstruction (3.9 Å) which was improved by density modification to a resolution of 3.7 Å.

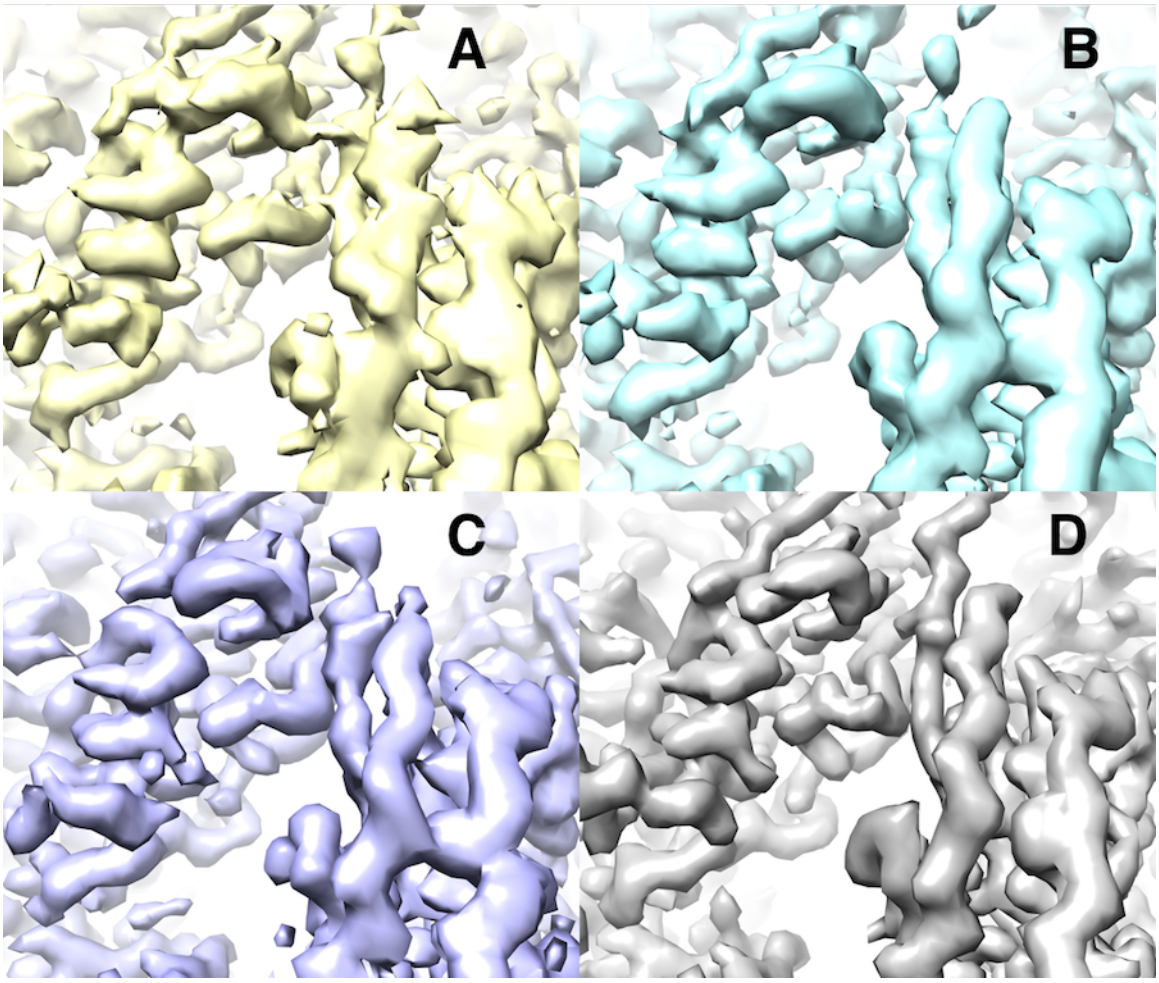

Figure 3. Effect of a reconstruction procedure yielding more uniform noise on the outcome of density modification. A. Standard reconstruction procedure applied to subset of images of for β-galactosidase (EMPIAR-10061). B. Density-modified version of the map in A. C. Density modified version of reconstruction designed to improve the uniformity of noise in the map. D. Deposited 2.2 Å map of β-galactosidase, superimposed on the map in A and low-pass filtered at a resolution of 3.5 Å. Each map was automatically sharpened using model-based sharpening with the deposited model for the 2.2 Å map of β-galactosidase (5a1a) superimposed on the map in A. Contours were chosen to enclose equal volumes in all four maps[16].

A portion of the initial full map is shown in Fig. 3A. For comparison, Fig. 3D shows the deposited high-resolution (2.2 Å) map, low-pass filtered at a resolution of 3.5 Å. The density-modified

maps (Fig. 3B and 3C) both are more similar to the high-resolution, low-pass filtered map than the initial full map (Fig. 3A), with the map obtained by density modification of the half-maps with more uniform noise being the clearer of the two maps (Fig 3C), consistent with the slightly better estimated resolution of this map. As density modification starting with the two reconstruction procedures improved the map by similar amounts (0.2 Å in each case), this test suggests that the variation in noise levels in the map does not dramatically affect density modification, but the observation that the reconstruction protocol with relatively uniform noise produced the best map indicates that such a protocol may be particularly well-suited for density modification.

The density modification procedure described here is fully automatic and requires only two half-maps, a nominal resolution, and information about the molecular volume of the macromolecule (such as a sequence file or molecular mass). The 54 datasets analyzed in Fig. 2 took from 4 to 245 cpu-minutes (average of 22 minutes) on 2.3 GHz AMD processors. In addition to a final density-modified map, the procedure yields two density-modified half-maps. As noted in Methods, these maps may have some correlations introduced by the masking effects of density modification, but with this caveat they can in principle be further processed with local sharpening[26], weighted combination of half-maps and other methods for optimizing the final map.

There are numerous extensions to the methods that are described here that could improve the outcome of density modification. In particular, the analysis could include information about the macromolecule from other sources such as models built using the maps or fitted into them[27,28]. Density modification including symmetry not used in the reconstruction process could be carried out as well[29]. The procedure could allow starting half-maps that have errors that are correlated with each other or that have different expected errors, and errors that do not follow Gaussian distributions. It could be carried out using just one map or more than two 'half-maps". Maps could be density modified without boxing by introducing a location-dependent expectation for map values outside the macromolecule[6]. Errors could be estimated in regions of

reciprocal space or anisotropically rather than in shells of resolution. The density modification step could also be carried out by other methods, for example solvent flipping[30]. The molecular composition could be calculated by analysis of the map, such as using local histograms of map values, allowing the identification of unexpected components. More generally, the entire analysis described here could be extended to any situation where a map of any dimensionality has errors that are at least partially independent in the Fourier domain and in which some information about expected values in the map is available.

Methods

Maps and models

The maps used to generate Fig. 1 are apoferritin maps EMD 20026, 20027 and 20028 and their associated half-maps. These maps have reported resolutions of 1.75 Å, 2.32 Å and 3.08 Å, respectively and we refer to them as the "1.8 Å", "2.3 Å" and "3.1 Å" apoferritin maps. The estimates of resolution for these maps based on comparison of masked half-maps obtained in this work are slightly different, presumably due to different masking procedures, with values of 1.93 Å, 2.36 Å and 2.97 Å, respectively. The model shown in Fig 1 is derived from PDB entry 3ajo[15] and has been superimposed on the 1.8 Å map and refined with *phenix.real_space refine*[18].

The sets of data used in Fig. 2 were all those available from the EMDB based on resolution (5 Å or better), the presence of half-maps matching the full deposited map, and the presence of a model matching the map available in the PDB (a total of 123 datasets satisfied these criteria). Maps in which more than 1% of map values were identical (indicating masking) were removed (28 datasets, most of these had more than 50% of map values identical). Maps for which the model represented only a small portion of the map were removed (3 datasets, such as one protein bound to a ribosome that was not included in the model), as was one map where the model had no sequence (all residues marked as unknown). Maps for which the starting

estimate of resolution based on model-map correlation differed by more than 1.5 Å from the reported resolution (33 datasets) were also removed (this difference indicated that the model-map correlation may not be an accurate measure of resolution for these datasets). This procedure yielded 58 datasets for analysis. Two of these failed in the density modification step (the automatic procedure for finding a mask around the molecule did not yield a result) and two failed in the evaluation step (automatic refinement of the model from the PDB against the original or density-modified map failed), resulting in the 54 datasets reported in Fig. 2.

The data for Figures 1-3 are available as an Excel worksheet in Supplementary Data I.

Procedure for evaluation of map quality

We used an automated procedure to evaluate map resolution and to choose matching map contours for display so that map comparisons would be as fair as possible. For evaluation of map resolution we calculated Fourier shell correlations between a map and an atomic model refined against that map. The rationale for this procedure is that the atomic models available from the PDB are normally already refined against the deposited map. This necessarily biases the map-model FSC calculation. To make a comparison with a new map, the model is refined against the new map before FSC calculation and a similar bias is introduced, leading to a relatively fair comparison between maps. Similarly, the model is re-refined against the original map before analysis of the original map.

Our largely automated procedure for evaluation and display of one map was then (1) refinement of the corresponding model from the PDB using that map, (2) boxing the map with a rectangular box around the model with soft edges, (3) calculation of map-model FSC, (4) sharpening the map based on the map-model FSC[4,17], (5) calculation of estimated resolution of the map from the map-model FSC, and (6) calculation of average map-model FSC[31] up to a resolution of 5/6 the stated resolution of the map (i.e., $0.83\ d_{min}$). Then to compare a density-modified and original map visually, the maps were visually examined and a region of the map

and contour level for the density-modified map were chosen where differences from the original map were clear. The contour level for the original map was then chosen automatically to yield the same enclosed volumes in the two maps[16]. This contour level for the original map was always close to that obtained by simply adjusting it to make about half the surface the color of the original map and half the color of the density-modified map when the two maps are displayed at the same time in Chimera[32]. Finally, keeping the same contour levels, the maps in Figs. 2 and 3 were masked 3 Å around the atoms in the region to be displayed to make it easier to see the region of interest.

Errors in Fourier coefficients representing cryo-EM maps

We assume that the distribution of errors in Fourier coefficients representing cryo-EM maps can be represented by a two-dimensional Gaussian in the complex plane[10]. This assumption is evaluated in Fig. 4 which compares Fourier coefficients for apoferritin from the 3.1 Å and 1.8 Å maps analyzed in Fig. 1. Fourier coefficients for the shell of resolution from 3.0 Å to 3.1 Å were calculated for each map after boxing the maps around the fitted model used in Fig. 1. The Fourier coefficients for the 1.8 Å map were treated as perfect values. These values were multiplied by the correlation coefficient between the two sets of Fourier coefficients and subtracted from the Fourier coefficients from the 3.1 Å map to yield estimates of the errors in the 3.1 Å map. Fig. 4 shows histograms of these errors along directions parallel and perpendicular to the Fourier coefficients from the 3.1 Å map. In each case a Gaussian distribution is fitted to these histograms and is shown as well. It can be seen that the errors are not quite Gaussian and are not quite the same in the two directions, but that a Gaussian is a good first approximation. In this example, the normalized errors perpendicular to the Fourier coefficients from the 3.1 Å map have a mean of zero and a standard deviation of 0.63, while those parallel have a mean of 0.1 and a standard deviation of 0.70.

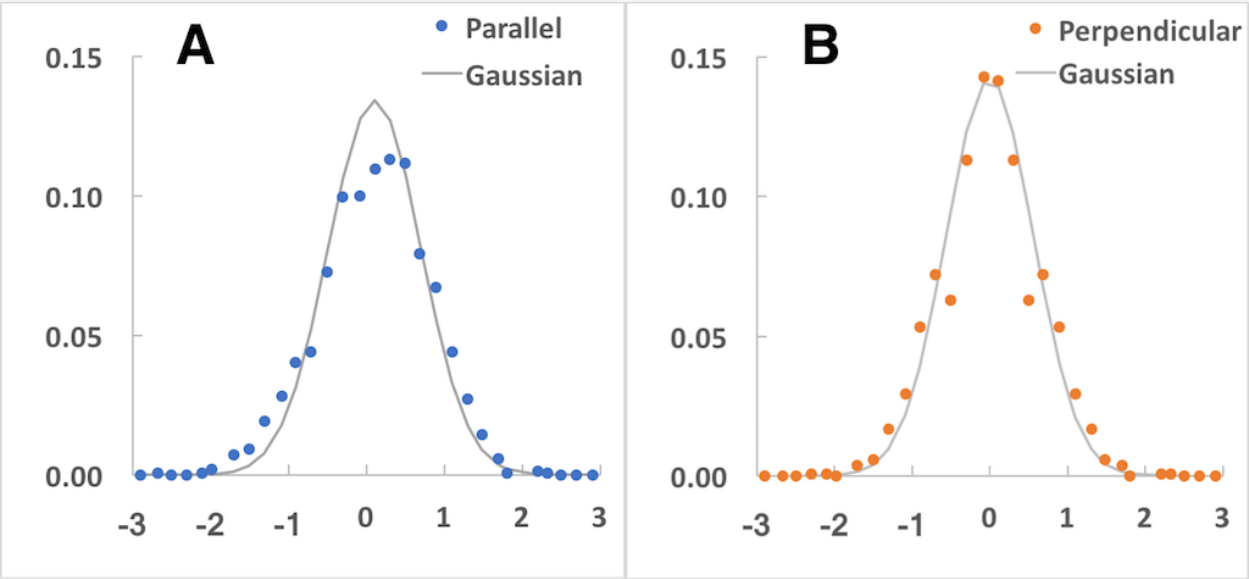

Fig. 4. Analysis of distribution of errors in Fourier coefficients for apoferritin 3.1 Å map. A. Estimated errors in Fourier coefficients parallel to coefficients for 3.1 Å map. B. Errors perpendicular to coefficients for 3.1 Å map. Solid lines in each case correspond to a Gaussian fitted to the values shown (see text).

Procedure for density modification of cryo-EM maps

Density modification of a cryo-EM map is based on the maximum-likelihood formalism that we developed previously for crystallographic density modification[7]. There are two important differences, however. One is that the starting probability distributions for Fourier coefficients (those available before density modification) are very different in the two cases. The other is that typically in a cryo-EM experiment, two independent half-maps are available (two maps with errors that are uncorrelated).

Maximum-likelihood density modification has two overall steps. In the first step a probability distribution (called the "map-phasing" probability distribution) is obtained for each Fourier coefficient. This map-phasing distribution has errors which, in an optimal situation, are independent of the errors in the corresponding Fourier coefficient in the starting map. In the second step the map-phasing probability distribution for each Fourier coefficient is recombined

with the starting information about that Fourier coefficient to yield a new "density modified" estimate of that Fourier coefficient.

The first stage is the same in crystallographic and cryo-EM cases. It starts with a map represented by Fourier coefficients. It requires a function that describes how the likelihood (believability) of that map would change if the values in the map change[7]. This likelihood function might for example say that if the values in the map in the region outside the macromolecule all move towards a common value, the believability increases. It might also say that if the distribution of values in the region of the macromolecule becomes closer to an idealized distribution, that map's believability improves. A specific example of a likelihood function that has both these properties has been described[7] (Eq. 17 in this reference). Given such a map and likelihood function, it is possible to calculate a "map-phasing" probability distribution and its maximum or weighted mean for each Fourier coefficient[33]. This yields a "map-phasing" map.

The map-phasing map has the important property that the new estimate of a Fourier coefficient does not depend at all on the value of that Fourier coefficient in the starting map[33]. This rather non-intuitive situation is possible because the map-phasing probability distribution for a particular Fourier coefficient comes only from all the other Fourier coefficients and the characteristics of the map as reflected in the likelihood function. In other methods of density modification such as solvent flipping a similar effect is obtained by specifically removing the information corresponding to the original map[6,30]. It should be noted however that if the other Fourier coefficients have information about the errors in the Fourier coefficient in question (for example through previous density modification or by masking of the map around the molecule), that information can indeed affect the map-phasing estimate of the Fourier coefficient of interest.

The key differences in implementation between crystallographic and cryo-EM cases arise in the second step. First, the information about the Fourier coefficient that is available before density

modification is very different in the two cases. In the crystallographic case, the amplitude of each Fourier coefficient is typically known quite accurately (often in the range of 5-30% uncertainty) and there may be some information about the phase (this might range from no information to phase uncertainties in the range of 45 degrees). The resulting distribution of likely values for a particular Fourier coefficient might essentially be a ring of relatively constant amplitude or a "boomerang" with partially-defined phase and relatively constant amplitude. In contrast, in the cryo-EM case, phase and amplitude are both uncertain, and the distribution of likely values before density modification can be represented by a two-dimensional Gaussian in the complex plane[10].

This qualitative and very substantial difference in the form of the distribution of likely values for a Fourier coefficient prior to density modification means that when recombining information between the starting map and map-phasing distributions, different approaches are best suited to the two situations. For crystallographic applications, recombination essentially amounts to testing possible values of the phase of a Fourier coefficient at constant amplitude for consistency with prior and map-phasing distributions. In contrast, for cryo-EM applications as described here, recombination consists of calculating the product of two 2-dimensional distributions and finding the maximum of the resulting distribution. If the distributions are Gaussian, this amounts to a simple weighted average of the prior and map-phasing Fourier coefficients.

The second key difference is that cryo-EM analyses are typically carried out in a way that yields two half-maps with largely independent errors. This means that overall mean-square values of errors can be estimated in a straightforward way in bins or shells of resolution by comparison of Fourier coefficients from the two half-maps and from the two map-phasing half maps (see below).

The overall procedure for density modification of two half-maps is then: (1) average the two half maps and sharpen/blur the resulting map based on resolution-dependent half-map

correlation ([4,17]) to obtain an optimized starting full map, (2) calculate target histograms for the macromolecule and non-macromolecule regions of this full map, (3) use the histograms and Fourier coefficients representing each half-map in the first step of density modification to obtain a map-phasing probability distribution for each Fourier coefficient for that half-map, and (4) calculate a weighted average of values of each Fourier coefficient obtained from the two starting half-maps and the map-phasing maps obtained from them to yield a "density-modified" map along with corresponding weighted half-maps and resolution-dependent estimates of the accuracies and resolution of each map. The optimal weighting is discussed below in terms of a simple error model. Finally (5) the entire process can be repeated, using the density modified half maps from one cycle in step (3) of the next cycle. It is also possible (but not done by default in our current procedure) to use the full density modified map from one cycle to obtain histograms for the next cycle. The process is concluded when the estimated improvement in resolution is small (typically less than 0.005 Å).

Target histograms of density distributions

A key element of the maximum-likelihood density modification procedure is the use of target histograms representing the expected distribution of map values for the “true” (desired) map in the region containing the macromolecule and in the region outside it[7]. These histograms can be obtained in any of several ways. One is to use a map or maps corresponding to high-quality structures that are already determined. Another is to use histograms used previously for crystallographic analyses. The method used here is to use histograms derived from the full map obtained by averaging the two current half-maps. These histograms have the advantage that they are automatically at the correct resolution and represent macromolecule and surrounding region in just the same way as the half map to be analyzed.

Error model for analysis of FSC curves and its use in optimizing weights and estimating correlations to true maps.

We use a simple error model with the following assumptions:

A. Starting half-maps have errors that are uncorrelated between half-maps. This assumption is based on the construction of half-maps, in which they are derived from independent subsets of the data. There are however some aspects of half-map construction that could lead to correlation of errors, including the use of the same reference in some stages of analysis and masking of the maps ([4]).

B. Density-modified half-maps have errors that are uncorrelated between half-maps and they may also have errors that are correlated with the corresponding starting half-maps and errors that are correlated with the each other. Errors correlated with the corresponding starting half-maps could come from the density modification procedure not yielding fully independent information. Errors correlated with the other density-modified half-map could come from masking effects introduced from solvent flattening procedures in density modification.

C. Mean square values of errors are resolution-dependent. This assumption simplifies the analysis of errors by describing the errors in terms of resolution and allowing them to be estimated in shells of resolution.

D. Mean square errors for each member of a pair of half maps are the same. This assumption comes from the construction of half-maps, where they typically come from equal numbers of images.

E. Errors have two-dimensional Gaussian distributions with mean expected values of zero. This simplifies the analysis.

F. Fourier coefficients representing individual half-maps are equal to the true Fourier coefficients plus uncorrelated errors unique to that map and correlated errors shared

among two or more maps. This yields a simple form for the Fourier coefficients that is amenable to estimation of errors from correlations of Fourier coefficients.

These assumptions yield a simple error model where Fourier coefficients for the original and density-modified half-maps can be represented as:

Original half-maps:

$$F_{1a} = F + \gamma_a + \sigma_a \qquad (1a)$$

$$F_{1b} = F + \gamma_b + \sigma_b \qquad (1b)$$

Density-modified half-maps:

$$F_{2a} = F + \gamma_a + \alpha + \beta_a \qquad (2a)$$

$$F_{2b} = F + \gamma_b + \alpha + \beta_b \qquad (2b)$$

In this description, F represents the true value of one Fourier coefficient. There are estimates of F that come from each half-map and each density-modified half-map. F_{1a} and F_{1b} represent Fourier coefficients for half-maps a and b, and F_{2a} and F_{2b} represent Fourier coefficients for density-modified half-maps a and b. The terms σ_a and σ_b represent uncorrelated errors in half maps F_{1a} and F_{1b}. The mean square values of each are S: $<\sigma_a^2> = <\sigma_b^2> = S$, and the mean values of all errors in this analysis are zero. The term γ_a represents errors that are correlated between half map F_{1a} and its corresponding density-modified half map F_{2a} (present in half map F_{1a} and not corrected by density modification), and the term γ_b represents errors correlated between F_{1b} and F_{2b}. The mean square values of γ_a and γ_b are $<\gamma_a^2> = <\gamma_b^2> = C$. The terms β_a and β_b represent uncorrelated errors in half maps F_{2a} and F_{2b}, where $<\beta_a^2> = <\beta_a^2> = B$. The term α represents errors correlated between half maps F_{2a} and F_{2b}, where $<\alpha^2> = A$. As it is assumed that errors are resolution-dependent, the estimates of mean square errors (A, B, C, S) are in

turn assumed to be resolution-dependent and in our procedure they are estimated in shells of resolution.

For simplicity in notation, we assume below that the Fourier coefficients for each half-map are normalized in such a way that the mean square value of F is unity. As the following calculations only involve correlation coefficients, the overall scale on Fourier coefficients has no effect on the values obtained, so this simplification does not affect the outcome of the analysis.

Using this error model and normalization, the expected values of correlations between half-maps can be calculated. These are as follows, where the brackets represent expected values, and the notation *CC(x,y)* represents the correlation coefficient relating values of *x* and *y*.

The expected correlation between half-maps *a* and *b* is given by,

$$<CC(F_{1a},F_{1b})> = [<(F + \gamma_a + \sigma_a)(F + \gamma_b + \sigma_b)>/(<(F + \gamma_a + \sigma_a)^2><(F + \gamma_b + \sigma_b)^2>]^{½} \qquad (3)$$

which reduces to,

$$<CC(F_{1a},F_{1b})> = 1/(1 + C + S) \qquad (4a)$$

Similarly, correlation between density-modified half-maps *a* and *b* is given by:

$$<CC(F_{2a},F_{2b})> = (1 + A)/(1 + A + B) \qquad (4b)$$

Cross-correlation between half map *a* and density-modified half map *a* (and also between corresponding maps *b*) is given by:

$$<CC(F_{1a},F_{2a})> = (1 +C)/[(1 +C + S)(1 + C + A + B)]^{½} \qquad (4c)$$

Cross-correlation between half map *a* and density-modified half map *b* (and also between half map *b* and density-modified half map *a*) is given by:

$$<CC(F_{1a},F_{2b})> = 1/[(1 + C + S)(1 +C + A + B)]^{½} \qquad (4d)$$

As there are four relationships and four parameters describing errors, the relationships can be used to estimate the values of the errors *A*, *B*, *C*, and *S*, leading to the formulas:

$$A = [CC(F_{1a},F_{1b}) CC(F_{2a},F_{2b}) + CC(F_{1a},F_{2b})^2 CC(F_{2a},F_{2b}) - CC(F_{1a},F_{2b})^2 - CC(F_{1a},F_{2b}) CC(F_{1a},F_{2a}) CC(F_{2a},F_{2b})] /CC(F_{1a},F_{2b})^2 \qquad (5a)$$

$$B = ((CC(F_{2a},F_{2b}) - 1) [-CC(F_{1a},F_{1b}) - CC(F_{1a},F_{2b})^2 + CC(F_{1a},F_{2b}) CC(F_{1a},F_{2a})]) /CC(F_{1a},F_{2b})^2 \qquad (5b)$$

$$C = [CC(F_{1a},F_{2a})/CC(F_{1a},F_{2b})] – 1 \qquad (5c)$$

$$D = [1/CC(F_{1a},F_{1b})] – [CC(F_{1a},F_{2a})/CC(F_{1a},F_{2b})] \qquad (5d)$$

As noted in more detail below, for shells at high resolution the uncertainties in the correlations such as $CC(F_{1a},F_{1b})$ can be large compared to the correlations themselves. In these situations the values of correlations are smoothed and additional assumptions are made about the relationships among the error estimates in order reduce the number of parameters that need to be obtained from the data at that resolution.

After estimation of errors, all estimates of *F* can be averaged with resolution-dependent weighting factors w. Based on the assumption of equal mean square errors in members of a pair of half-maps, the weights on each half map in a pair are always equal. The recombined (density-modified) estimate (*G*) of *F* is then given by,

$$G = wF_1 + (1-w)F_2 \qquad (6a)$$

Where w is the weight on the original averaged half maps (F_1) and $(1-w)$ is the weight on averaged density-modified averaged half-maps (F_2) and the averaged maps are given by:

$$F_1 = ½ (F_{1a} + F_{1b}) \qquad (6b)$$

$$F_2 = ½ (F_{2a} + F_{2b}) \qquad (6c)$$

The weight w that maximizes the expected correlation of the estimated Fourier coefficient, G with the true one, F, is:

$$w = (2 A + B)/(2 A + B + S) \qquad (7)$$

and the estimated correlation of G with F, the correlation of the final estimate of the Fourier coefficient with the true one, represented[4] by C_{ref} is:

$$C_{ref} = [2 (A + ½B + ½S)/[A(C + S + 2) + B(½C + ½S + 1) + (½C + 1)S]]^{½} \qquad (8)$$

As mentioned above, assumptions are made to allow estimation of correlated and uncorrelated errors from FSC plots for resolution shells where uncertainties in correlation estimates are large. These additional assumptions are:

A. In the highest resolution shell considered there is no remaining signal and all correlations of Fourier coefficients are due to correlated errors. The limiting resolutions of FSC plots in this analysis are set in such a way that there is little signal at those resolutions and correlations are largely due to noise and correlated errors. FSC plots are calculated in shells of resolution (d). The highest resolution in error analyses considered (d_{min}) is the resolution used for density modification multiplied by a fixed ratio (typically

5/6). As the resolution used for density modification is normally about 0.5-1 Å finer than the nominal resolution of the overall map, for a 4 Å map this high-resolution d_{min} would typically be in the range of 2.5 Å to 3 Å.

B. In high-resolution shells where there is substantial uncertainty in the estimates of errors (typically where half-map correlations are less than about 0.05), ratios of correlated to uncorrelated errors are assumed to be the same as those estimated in lower-resolution shells.

C. For shells of resolution (d) where the values of FSC are below a fixed minimum FSC, (typically FSC_{min}=0.2), smoothed FSC values are calculated by fitting the observed values to a simple exponential function with one free parameter. The function used is $FSC=(FSC_{d1} - FSC_{d_min})exp(-H/d^2)+ FSC_{d_min}$. The free parameter is H, the fall-off with $1/d^2$). FSC_{d1} is the value of FSC at resolution $d1$, the highest resolution where FSC is higher than a fixed minimum value (typically 0.2). FSC_{d_min} is the estimated FSC at the highest resolution in the analysis. As noted above it is assumed that any non-zero FSC found at this resolution is due to correlated errors in the analysis.

Real-space weighting and weighting of individual Fourier coefficients in calculation of the final map

An option available at the end of a cycle of density modification is to apply a local weighting scheme to the final combination of original and density-modified maps. The idea is to identify local accuracy in the original map from local similarity between original half-maps, and also local accuracy in the density modified maps from local similarity between density-modified half maps. The procedure for one pair of half-maps is to subtract the maps, square the resulting map, and smooth the squared map with a smoothing radius typically given by twice the resolution of the map to give a local variance for those half-maps. Then a local weight for each set of half-maps obtained as the inverse of the local variance of those half-maps. These local weights are then scaled to yield an average local weight of unity and then are applied to the individual half-maps before they are averaged.

A second option for recombination of maps is to weight individual Fourier coefficients based on the estimated variance for these coefficients. The variance for an individual coefficient is estimated from the four Fourier coefficients representing the four half-maps available at the end of the procedure (the two original half-maps and the two density-modified half-maps).

Boxing of cryo-EM maps

In our procedure, a rectangular solid portion of a cryo-EM map that contains the macromolecule of interest is cut out from the map and is used in the analysis. This "boxing" of the map is carried out with a "soft" (Gaussian) mask with a smoothing radius typically equal to the resolution of the map to reduce the introduction of correlations in Fourier coefficients between different maps boxed in the same way[4]. The edges of the box are chosen using bounds in each direction identified using a low-resolution (typically 20 Å) mask calculated from the full map with a volume based on the expected molecular volume of the macromolecule. Typically a buffer of 5 Å is added to the bounds in each direction to yield a box that has dimensions 10 Å bigger than the size in each direction of the macromolecule.

There are two important effects of boxing. One is to reduce the variation of noise in the map in the region outside the macromolecule. In a typical cryo-EM map there is substantial noise (fluctuation in map values not representing the macromolecule) near the macromolecule, and progressively less further from the macromolecule (the variation in noise levels may also be more complicated). This variation in noise levels is largely due to the use of procedures that smooth Fourier coefficients in reciprocal space with the effect of masking around the macromolecule[25]. In our procedure it is assumed that the distribution of map values in the region outside the macromolecule can be represented by a simple histogram. As there is a distance-dependent variation in the level of noise in unboxed cryo-EM maps, our procedure can be made more applicable by boxing the maps.

The second effect of boxing the map is to reduce the correlation of Fourier coefficients in the map. If a small object is placed in a large box and Fourier coefficients are calculated representing the object in the box, coefficients with similar indices (neighboring Fourier coefficients) will be highly correlated[34,35]. The significance of correlations between Fourier coefficients is that errors may be correlated as well, resulting in map-phasing Fourier coefficients that are not fully independent from the original Fourier coefficients. Boxing reduces the empty volume of the map and reduces this correlation.

Resolution cutoff used for density modification.

In order to include information at high resolution, the resolution ($d_{_dm}$) of Fourier coefficients used in the density modification procedure is typically finer than the resolution of the full map. The relationship between the resolution d of a map and the optimal resolution $d_{_dm}$ for density modification is not obvious, so we used an analysis of 51 half-maps from the EMDB and associated models from the PDB to develop an empirical relationship. The empirical function was obtained by carrying out the entire density modification procedure for each dataset, each with a range of values of $d_{_dm}$. Then the average FSC between map and model-based map was calculated for each analysis and a simple function was developed for choosing the resolutions where density modification was optimal. This function, valid for resolutions between 2.4 Å and 5 Å, was:

$$d_{_dm} = 2.4 + 0.99\,(d-3) - 0.2\,(d-3)^2 \qquad (9)$$

At a resolution $d=2.4$ Å, this yields $d_{_dm} = 1.9$. For resolutions finer than 2.4 Å, we simply subtract 0.5 Å from the resolution to yield $d_{_dm}$, except that $d_{_dm}$ is never allowed to be less than $\frac{1}{2}d$.

Three options for choice of resolution cutoff for density modification $d_{_dm}$ are available in the current implementation of density modification. One is directly specifying $d_{_dm}$, one is using Eq. 9 to estimate $d_{_dm}$, and the last is to try various values of $d_{_dm}$ and choose the one that leads to the most favorable estimated improvement in the resolution where C_{ref} is ½ based on Eq. 8.

Adjustable parameters

There are many adjustable parameters in our procedure but by default all are set to the values used in the examples given here. Some of the parameters that can substantially affect the results and that a user might vary if the initial results are not optimal are listed here. The resolution used for density modification is not fully optimized and can affect the outcome substantially. The number of shells of resolution used in the calculation of correlations between Fourier coefficients has a default of 20; more shells can potentially improve the accuracy by not grouping coefficients that have very different values simply due to resolution-dependent variation but could reduce it due to fewer coefficients in a calculation. The optional use of real-space weighting or individual weighting of Fourier coefficients at the end of the procedure can sometimes affect the resulting map.

Software availability

All the procedures described in this work are available in versions 3689 and later of the *Phenix* software suite[36].

Acknowledgements

This work was supported by the NIH (grant GM063210 to PDA, RJR and TT), the Wellcome Trust (grant 20947/Z/17/Z to RJR), and the Phenix Industrial Consortium. This work was supported in

part by the US Department of Energy under Contract No. DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory.

Author contributions

SL carried out image processing of test datasets to evaluate varying reconstruction procedures, RJR and TCT contributed ideas on the form of errors in cryo-EM, PA developed tools for the testing infrastructure, TT developed the software for error analysis, and PDA and TCT supervised the work.

Author information

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to:
tterwilliger@newmexicoconsortium.org.

References

1 Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nat Methods* **13**, 24-27, (2016).
2 Marques, M. A., Purdy, M. D. & Yeager, M. CryoEM maps are full of potential. *Current Opinion in Structural Biology* **58**, 214-223, (2019).
3 Terwilliger, T. C., Adams, P. D., Afonine, P. V. & Sobolev, O. V. Cryo-EM map interpretation and protein model-building using iterative map segmentation. *Protein Sci*, (2019).
4 Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol* **333**, 721-745, (2003).
5 Rosenthal, P. B. & Rubinstein, J. L. Validating maps from single particle electron cryomicroscopy. *Current Opinion in Structural Biology* **34**, 135-144, (2015).
6 Cowtan, K. Recent developments in classical density modification. *Acta Crystallographica Section D* **66**, 470-478, (2010).
7 Terwilliger, T. Maximum-likelihood density modification. *Acta Crystallographica Section D* **56**, 965-972, (2000).

8 Wang, B. C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol* **115**, 90-112, (1985).
9 Podjarny, A. D., Rees, B. & Urzhumtsev, A. G. in *Crystallographic Methods and Protocols* (eds Christopher Jones, Barbara Mulloy, & Mark R. Sanderson) 205-226 (Humana Press, 1996).
10 Scheres, S. H. A Bayesian view on cryo-EM structure determination. *J Mol Biol* **415**, 406-418, (2012).
11 Sindelar, C. V. & Grigorieff, N. Optimal noise reduction in 3D reconstructions of single particles using a volume-normalized filter. *J Struct Biol* **180**, 26-38, (2012).
12 Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A primer to single-particle cryo-electron microscopy. *Cell* **161**, 438-449, (2015).
13 Murshudov, G. N. in *Methods in Enzymology* Vol. 579 (ed R. A. Crowther) 277-305 (Academic Press, 2016).
14 Lawson, C. L. *et al.* EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* **39**, D456-D464, (2011).
15 Masuda, T., Goto, F., Yoshihara, T. & Mikami, B. The universal mechanism for iron translocation to the ferroxidase site in ferritin, which is mediated by the well conserved transit site. *Biochemical and Biophysical Research Communications* **400**, 94-99, (2010).
16 Urzhumtsev, A., Afonine, P. V., Lunin, V. Y., Terwilliger, T. C. & Adams, P. D. Metrics for comparison of crystallographic maps. *Acta Crystallographica Section D* **70**, 2593-2606, (2014).
17 Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. Automated map sharpening by maximization of detail and connectivity. *Acta Crystallographica Section D* **74**, 545-559, (2018).
18 Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallographica Section D* **74**, 531-544, (2018).
19 Afonine, P. V. *et al.* FEM: feature-enhanced map. *Acta Crystallogr D Biol Crystallogr* **71**, 646-666, (2015).
20 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242, (2000).
21 Bartesaghi, A. *et al.* 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science* **348**, 1147, (2015).
22 Gao, Y., Cao, E., Julius, D. & Cheng, Y. TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature* **534**, 347, (2016).
23 Yin, Y. *et al.* Structure of the cold- and menthol-sensing ion channel TRPM8. *Science* **359**, 237, (2018).
24 Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods* **13**, 387-388, (2016).
25 Tang, G. *et al.* EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology* **157**, 38-46, (2007).
26 Jakobi, A. J., Wilmanns, M. & Sachse, C. Model-based local density sharpening of cryo-EM maps. *Elife* **6**, (2017).
27 Skubak, P. *et al.* A new MR-SAD algorithm for the automatic building of protein models from low-resolution X-ray data and a poor starting model. *IUCrJ* **5**, 166-171, (2018).

28 Terwilliger, T. Improving macromolecular atomic models at moderate resolution by automated iterative model building, statistical density modification and refinement. *Acta Crystallographica Section D* **59**, 1174-1182, (2003).
29 Bricogne, G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallographica Section A* **30**, 395-405, (1974).
30 Abrahams, J. P. & Leslie, A. G. Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr D Biol Crystallogr* **52**, 30-42, (1996).
31 Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr D Biol Crystallogr* **71**, 136-153, (2015).
32 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, (2004).
33 Terwilliger, T. Map-likelihood phasing. *Acta Crystallographica Section D* **57**, 1763-1775, (2001).
34 Sousa, D. & Grigorieff, N. Ab initio resolution measurement for single particle structures. *J Struct Biol* **157**, 201-210, (2007).
35 Shaikh, T. R., Hegerl, R. & Frank, J. An approach to examining model dependence in EM reconstructions using cross-validation. *Journal of Structural Biology* **142**, 301-310, (2003).
36 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D* **75**, 861-877, (2019).