A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps

Thomas C. Terwilliger*, Paul D. Adams3,4, Pavel V. Afonine3,5 and Oleg V. Sobolev3

We report a fully automated procedure for the optimization and interpretation of reconstructions from cryo-electron microscopy (cryo-EM) data, available in Phenix as phenix.map_to_model. We applied our approach to 476 datasets with resolution of 4.5 Å or better, including reconstructions of 47 ribosomes and 32 other protein–RNA complexes. The median fraction of residues in the deposited structures reproduced automatically was 71% for reconstructions determined at resolutions of 3 Å or better and 47% for those at resolutions worse than 3 Å.

Improvements to methods for cryo-EM data collection and image reconstruction have made it possible to obtain 3D images of macromolecules at resolutions that allow ready visualization of structural details such as the locations of side chains1,2. A key limiting factor in structure determination by cryo-EM is now the effort required for interpretation of these reconstructed images in terms of an atomic model. Algorithms for de novo model building using cryo-EM maps have been reported recently3–8 and are capable of building complete models in some cases, but they are not fully automated. None, for example, can carry out map segmentation; automatically sharpen a map; or identify parts of the map that correspond to protein, RNA, or DNA when more than one type of chain is present in the same map; or take reconstruction symmetry into account.

Here we report an integrated procedure for map interpretation without user intervention. The most important new algorithms include automatic map segmentation, automatic map sharpening, automatic interpretation of multiple chain types, and automatic application of reconstruction symmetry. The procedure requires as input only a cryo-EM map, the nominal resolution of the reconstruction, the sequences of the molecules present, and any symmetry used in the reconstruction. Map interpretation begins with automatic image sharpening using an algorithm that maximizes connectivity and detail in the map9. The unique parts of the structure in the sharpened map are then identified with a segmentation algorithm that extends previous methods10 by taking into account reconstruction symmetry and choosing a contour level based on the expected contents of the structure (details in Methods and Supplementary Fig. 1; examples in ref. 11). For each part of the structure and for each type of macromolecule present, atomic models are generated via several independent methods, and then extensive real-space refinement12 is applied with restraints based on automatically identified secondary structure elements. Finally, our automated procedure carries out assembly and refinement of the entire structure, including any symmetry that is present. The procedure is carried out by the Phenix tool phenix.map_to_model, available at https://www.phenix-online.org.

One of the strengths of our automated model-building procedure is that a novice user will generally obtain the same result as an expert. Whereas most automatic tools have many parameters that the user can adjust to improve results in difficult cases, our approach, generally speaking, does not. Each time we identify a parameter that needs to be adjusted depending on the situation, that parameter is automatically varied and optimized in our procedure. For example, the sharpening of a map is a very important adjustable parameter that is set by the user in most approaches. We developed a metric that we could use to evaluate map sharpening and automatically optimize the sharpening of the map during the analysis. The result of this strategy is that our procedure has just one overall parameter (“quick” or “not quick”) that a user might normally adjust for a difficult case. Another strength of automated procedures such as this is that they conduct analyses that would be challenging to carry out with tools that require manual input. With automated interpretation of a map, for example, one can obtain error estimates by repeating the interpretation using different algorithms or different random seeds in appropriate stages of the analysis (Supplementary Results).

The steps in map interpretation are illustrated in Fig. 1. Part of a deposited cryo-EM reconstruction of lactate dehydrogenase (EMDB entry 819114) obtained at a resolution of 2.8 Å is illustrated in Fig. 1a, along with an atomic model (PDB entry 5K0Z15) representing the authors’ interpretation of the reconstructed map. Our procedure optimizes map sharpening and yields an atomic model as an interpretation of the map (Fig. 1b). Features such as side chains can be clearly visualized without the need for manual adjustment of the map sharpening. In this case (Fig. 1b) our automated interpretation did not determine the identity of the side chains in this region of the map, but in other cases, such as the reconstruction of β-galactosidase16 at a resolution of 2.2 Å (Fig. 1c), side chains were identified automatically for most of the structure.

Figure 1d illustrates the deposited cryo-EM reconstruction of glutamate dehydrogenase (EMDB–663017) obtained at a resolution of 3.3 Å. Using the automatically sharpened map, we created three models, each with a different Phenix algorithm (tracing polypeptide backbones by following tubes of density, feature-based identification of secondary-structure elements, and template matching followed by extension with fragments from known structures; Methods) so as to...
Fig. 1 | Automated interpretation of cryo-EM maps. **a**, Section through the deposited cryo-EM reconstruction and deposited model of lactate dehydrogenase \(^1\) (EMDB-8191, PDB 5K02). **b**, Automatically sharpened version of the map in **a** with the automatically generated model. **c**, Automatic interpretation of β-galactosidase \(^2\) at a resolution of 2.2 Å. **d**, Automatically sharpened version of a map of glutamate dehydrogenase \(^3\) (EMDB-6630) with three independent interpretations automatically generated by Phenix (using chain-tracing (green), feature-based helix and strand identification (blue), and pattern-based secondary structure identification with fragment-based extension (magenta)). **e**, Composite model derived from the three models in **d**. **f**, Automatically interpreted reconstruction of the \(M.\) smegmatis ribosome (yellow tubes; EMDB-6789) compared with the deposited model (green ribbons; PDB 5XYM). **g**, Automatically generated model of the ERAD channel HRD1 (purple ribbons; EMDB-8637) compared with the deposited model (brown ribbons; PDB 5V6P). **h,i**, Detail of the automatically built \(M.\) smegmatis ribosome (**h**) and corresponding portion of the deposited model (**i**). Graphics created with Chimera \(^22\).
to minimize correlated errors. For each model-building algorithm, the main chain was built first, then the sequence was assigned and side chains were added on the basis of the fit to the density. These three models are superimposed on the map in Fig. 1d (after application of reconstruction symmetry). The best-fitting parts of each model were combined and extended to fill gaps, and the resulting composite model was refined to yield an interpretation of the reconstruction (Fig. 1e). For this glutamate dehydrogenase structure, 66% of the residues in the deposited model (PDB 3JCZ) were reproduced in the automatically generated model (C$_\alpha$ atoms in the deposited model matched within 3 Å), with a root-mean-square (r.m.s.) coordinate difference of 0.8 Å for matching C$_\alpha$ atoms. The automatically generated model represents 60% of the RNA and 48% of the protein in the deposited model.

At lower resolution, a smaller fraction of the structures is typically reproduced by our methods, but secondary structural elements such as protein and RNA helices are often identifiable. Figure 1g shows the automatically sharpened map and compares the automatic interpretation for the protein-conducting ERAD channel HRD1 (EMDB-8637) with the deposited model (PDB 5V6P). This reconstruction at a resolution of 4.1 Å was previously interpreted through a combination of helix-fitting into density, manual model building, and Rosetta$^{26}$ modeling with distance restraints from evolutionary analysis.

We evaluated the overall effectiveness of our procedure by applying it to all 476 high-resolution cryo-EM reconstructions in the EMDB that we could extract with simple tools and match to an entry in the PDB. For chains in maps reconstructed at resolutions of 3 Å or better (24 structures containing protein and 6 containing RNA), the median fractions of protein and RNA residues in the deposited models reproduced by our approach were 71% and 45%, respectively (Fig. 2a). At resolutions lower than 3 Å (452 structures containing protein and 73 with RNA), 47% of protein residues and 34% of RNA residues were reproduced. The median r.m.s. differences for matching C$_\alpha$ atoms in protein chains and for matching P atoms in RNA chains were each about one-third the resolution of the reconstructions (Fig. 2b). The median fraction of the sequence of the deposited structure that could be reproduced (Fig. 2c) was 28% for protein chains at higher resolution and 9% at lower resolution (the expected fraction matching based on random sequence assignment was roughly 6% for protein with 20 amino acids with frequencies in eukaryotes and roughly 25% for RNA with 4 bases
and similar frequencies). For RNA chains, the sequence match was 49% at higher resolution and 42% at lower resolution. An analysis of the geometries of the models is presented in Supplementary Figs. 2 and 3. We found that substantial maps could be obtained even at resolutions lower than the resolution of 4.5 Å for which the procedure was designed. Figure 2d,e shows a comparison of our automatic analysis of horse spleen apoferritin at a resolution of 4.7 Å with the deposited model.

We compared our automatic model-building approach with two recently developed methods: MAINMAST8 and de novo Rosetta21 model building. For our analysis we chose maps for two other recently developed methods: MAINMAST8 and de novo Rosetta21 model building. For our analysis we chose maps for them by cutting out just the region surrounding a single chain, other methods are not fully automated, we prepared maps suitable for them by cutting out just the region surrounding a single chain, and we used these artificial maps to test each approach. We applied our automated approach to the 22 unique maps used in the previous work1 with a matching entry in the PDB, segmented to show a single chain as described8, and compared the models (Fig. 2f and Supplementary Table 1) with those obtained previously8. Our method yielded an average coverage higher than that of Rosetta and lower than that of MAINMAST. The models built by our automated procedure had accuracy (r.m.s. difference from deposited models of matching Cα atoms of 1.31 Å) the same as or better than that of the MAINMAST models (r.m.s. deviation of 1.51 Å) and the Rosetta models (r.m.s. deviation of 1.33 Å).

With our automated procedure, essentially any high-resolution reconstruction with suitable metadata describing the reconstruction, its resolution, and reconstruction symmetry can be interpreted and a first atomic model can be generated without any manual intervention. The models produced are not complete at this stage. We anticipate that combination of the integration available in our approach with other algorithms3–8 could lead to both a high degree of automation and high model completeness.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0173-1.

Received: 2 February 2018; Accepted: 6 August 2018; Published online: 30 October 2018

References
1. Kühlbrandt, W. Science 343, 1443–1444 (2014).
2. Henderson, R. Arch. Biochem. Biophys. 581, 19–24 (2015).
3. Wang, R. Y. et al. Nat. Methods 12, 335–338 (2015).
4. Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D. & DiMaio, F. Nat. Methods 14, 797–800 (2017).
5. Chen, M., Baldwin, P. R., Ludtke, S. J. & Baker, M. L. J. Struct. Biol. 196, 289–298 (2016).
6. DiMaio, F. & Chiu, W. Methods Enzymol. 579, 255–276 (2016).
7. Zhou, N., Wang, H. & Wang, J. Sci. Rep. 7, 2664 (2017).
8. Terassi, G. & Kihara, D. Nat. Commun. 9, 1618 (2018).
9. Terwilliger, T. C., Sobolev, O. V., Afonine, P. V. & Adams, P. D. Acta Crystallogr. D Struct. Biol. 74, 545–559 (2018).
10. Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. J. Struct. Biol. 170, 427–438 (2010).
11. Terwilliger, T. C., Adams, P. D., Afonine, P. V. & Sobolev, O. V. J. Struct. Biol. 204, 338–343 (2018).
12. Afonine, P. V. et al. Acta Crystallogr. D Struct. Biol. 74, 531–544 (2018).
13. Lawson, C. L. et al. Nucleic Acids Res. 44, D396–D403 (2016).
14. Merk, A. et al. Cell 165, 1698–1707 (2016).
15. Berman, H. M. et al. Nucleic Acids Res. 28, 235–242 (2000).
16. Bartsch, A. et al. Science 348, 1147–1151 (2015).
17. Borgia, M. et al. Mol. Pharmacol. 89, 645–651 (2016).
18. Terwilliger, T. C. Acta Crystallogr. D Biol. Crystallogr. 59, 45–49 (2003).
19. Li, Z. et al. Protein Cell 9, 384–388 (2018).
20. Schoebel, S. et al. Nature 548, 352–355 (2017).
21. Wang, R. Y. R. et al. Nat. Methods 12, 335–338 (2015).
22. Pettersen, E. F. et al. J. Comput. Chem. 25, 1605–1612 (2004).

Acknowledgements
This work was supported by the NIH (grant GM063210 to P.D.A. and T.C.T.) and the Phenix Industrial Consortium. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-06-SC11311 at Lawrence Berkeley National Laboratory. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the US Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396. We thank G. Terashi and D. Kihara for providing the MAINMAST and Rosetta models built into segmented maps.

Author contributions
P.V.A. and O.V.S. developed core tools for map segmentation and secondary-structure restraints; P.D.A. and P.V.A. developed the real space refinement tools; T.C.T. developed the phenix.map_to_model tool and carried out the analyses; and P.D.A. and T.C.T. supervised the work.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0173-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.C.T.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2018
Methods

Summary of automated modeling procedure. The first step in our automated modeling procedure is automatic map sharpening. This yields an optimized map suitable for interpretation. The second step is segmentation of the map, in which the unique part of the map is identified and further split into parts containing regions of contiguous density. The third step is modeling of each region in the unique part of the map as protein (and RNA/DNA if present in the sequence file) via several independent map-interpretation methods, which yields overlapping interpretations of each part of the map. The fourth step is selection of the best-fitting interpretation of each region of the map and creation of a composite model representing the entire unique part of the map. The final step entails application of the reconstruction symmetry of the map and refinement of the resulting full model.

Map sharpening. Maps are automatically sharpened (or blurred) with phenix.auto_sharpen, which maximizes the level of detail in the map and the connectivity of the map by optimizing an overall sharpening factor $B_{\text{sharpen}}$ applied to Fourier coefficients representing the map up to its effective resolution. Beyond that resolution, a blurring exponential factor $B_{\text{sharpen}}$ with a value of 200 Å is applied. This blurring procedure enables the user to dampen high-resolution information without precise knowledge of the optimal resolution cutoff.

Map segmentation. Map segmentation consists of the identification of all regions with density above an automatically determined threshold, followed by selection of a unique set of density regions that maximizes connectivity and compactness, taking into account the symmetry present.

Contiguous regions above a threshold in a map are identified by a region-finding algorithm. This algorithm chooses all the grid points in a map that are above a given threshold. Then it groups these grid points into regions in which every point is above the threshold and is connected to every other point in that region through adjacent grid points that are also above the threshold. The choice of threshold for defining regions of density is a critical parameter in segmentation. We set this value automatically by finding the threshold that optimizes a target function that is based on three factors. One factor targets a specific volume of the map above the threshold. The second factor targets the expectation that if $n$-fold symmetry is present, then groups of $n$ regions should have approximately equal volumes. The third factor targets regions of density of a specific size. The desired volume above the threshold is chosen on the basis of the molecular volumes of the molecules expected to be present in the structure and the assumption that a fraction $f$ (typically 0.2) of the volume inside a molecule will have high density (the parts very near atoms) and that only these high-density locations should be part of the map above the threshold. The desired size of individual regions of density is about the size occupied by 50 residues of the macromolecule, as this is a factor representing the map up to its effective resolution. Beyond that resolution, a blurring exponential factor $B_{\text{sharpen}}$ with a value of 200 Å is applied. This blurring procedure enables the user to dampen high-resolution information without precise knowledge of the optimal resolution cutoff.

The desired volume ratio of the $n$th region, $v_n$, to that of the first region, $v_1$, is unity, and the value of this ratio is

$$v_n = v_1 / v_1$$

For a specific threshold, the volumes of regions above the threshold and the median volume $v_{\text{median}}$ of the first $m_{\text{target}}$ of these regions, after sorting from largest to smallest, are noted.

The desired median volume $v_{\text{median}}$ is $v_{\text{region, target}}$. We use the target function

$$v_{\text{target}} = \begin{cases} v_{\text{region, target}} & \text{if } v_{\text{region, target}} < v_{\text{median}} \\ {v_{\text{median}}} / v_{\text{region, target}} & \text{otherwise} \end{cases}$$

where $a$ is expected to increase from a value of 1 if all regions are the same size to larger values if regions are of different sizes, as the largest regions will have volumes greater than the median.

The desired volume of the largest region, $v_1$, is also $v_{\text{region, target}}$. The target function

$$v_{\text{target}} = \begin{cases} v_1 / v_{\text{region, target}} & \text{if } v_1 < v_{\text{region, target}} / v_{\text{target}} \\ v_{\text{region, target}} & \text{otherwise} \end{cases}$$

expresses this.

Finally, empirically we find that a threshold $t$ on the order of unity is typically optimal (after scaling of the map as described above). We express this with a final ratio,

$$v_{\text{target, threshold}} = \begin{cases} t & \text{if } t < 1, 1 / t \text{ otherwise} \end{cases}$$

where larger values are again desired.

The total score $Q$ for a threshold $t$ is given by

$$Q = A v_{\text{ratio}} + B \left( v_{\text{ratio, median}} + v_{\text{ratio, target}} \right) + C v_{\text{target, threshold}}$$

where $A$, $B$, and $C$ have default values that we set by limited experimentation using a few test cases. Values of the threshold $t$ are automatically tested, and the value that maximizes the total score is used.

Once a threshold is chosen and the resulting set of regions of connected density above that threshold are found, these regions are assembled into groups with members related by symmetry (if any is present). A unique set of density regions is chosen through selection of one region from each symmetry-related group. The choice of regions is optimized to yield a compact structure and high connectivity. The compactness of the structure is represented by the radius of gyration of randomly sampled points from all chosen regions. The connectivity of a set of regions is calculated on the basis of the r.m.s. of the maximum gaps that would have to be spanned to connect each region to one central region, going through any number of regions in between. For any pair of regions, the gap is defined as the smallest distance between randomly sampled points in the two regions. For a set of regions, the overall gap is the largest of the individual gaps that would have to be crossed to go from one region to another, going through any other regions in the process. The overall connectivity score is the r.m.s. of these gaps for connections between each region and one central region.

The central region and all the other regions are chosen to minimize both the connectivity score and the radius of gyration. The relative weighting of the two scores is determined by a user-definable parameter with a default value of weight on the radius of gyration of weight_rad_gyr = 0.1. The weight on the radius of gyration is then normalized to the size of the molecule by multiplication of this parameter by the largest cell dimension divided by 300 Å. (This dimension is arbitrary; the key relationship is that the radius of gyration scales with the size of the molecule.)

The goal of the segmentation procedure described so far is to yield density corresponding to a single molecular unit. If the segmentation procedure yields only density corresponding to parts of several molecular units, then complete chain tracing will not be possible. To increase the proportion of complete chains, regions neighboring the initial regions that would lead to an increase in the overall radius of gyration of 1 Å or less are added to the segmented region.

Chain types to be examined. The chain types (protein, RNA, DNA) to be tested in model building were automatically deduced from the contents of the sequence file with the Phenix method phenix.guess_chain_types from_sequences.

Protein model building. In our new core method for protein model building, the phenix.trace_chain algorithm is used to build a polypeptide backbone through a map following high contour levels. These preliminary models are then improved.
by automatic iterative identification of secondary structure and refinement of the models including hydrogen-bond restraints representing this secondary structure with the phenix.real_space_refine approach \(^2\). Because the connectivity in a map can sometimes be misleading at lower resolutions, which builds two chains of duplexes separately, frequently resulted in poorly base-paired strands. To build RNA helices directly, we used very similar overall strategies, except that the templates were all base-paired, and base-paired nucleotides were always considered as a single unit. This automatically led to the same favorable base-pairing found in the structures used to derive the templates. The atoms used for Serpentine chains and bases were the C4' and C3' atoms of one nucleotide and the C1' atom of the base-paired nucleotide. As in the previous method for sequence assignment, both bases in each base-paired set were considered together, which led again to a substantial increase in map-based information about the identities of bases in the model. Supplementary Fig. 4 illustrates the models obtained with the two methods for a small region of RNA density from the Leishmania ribosome \(^27\) at a resolution of 2.5 Å (EMDB-7025).

Combining model information from different sources and removing overlapping fragments. Model-building into local regions of density and into the entire asymmetric unit of the map normally yields a set of partially overlapping segments. These segments were used to build a basis of the sharpened map with the Phenix tool phenix.real_space_refine.\(^2\) The refined segments were scored on the basis of map–model correlation multiplied by the square root of the number of atoms in the segment (related to fragment scoring in RESOLVE model building, except that density at the coordinates of atoms was used instead of map–model correlation as in that work).\(^6\) Then a composite model was created from these fragments, starting from the highest-scoring one and working down, and including only nonoverlapping parts of each new fragment considered, as implemented in the Phenix tool phenix.combine_models. When symmetry was used in reconstruction, all the symmetry-related copies of each fragment were considered for evaluation of whether a particular part of a new fragment would overlap with the existing composite model.

Evaluation of model similarity to deposited structures. We developed the Phenix tool phenix.chain_comparison as a way of comparing the overall backbone (C or P atoms only) similarity of two models. The unique feature of this tool is that it considers each segment of each model separately so that it does not matter whether the chain is complete or broken into segments. Additionally, the tool can separately identify segments that have matching C or P atoms that proceed in the same direction and those that are reversed, as well as those that have insertions and deletions, as is common in low-resolution model building. The phenix.chain_comparison tool also identifies whether the sequences of the two models are similar counting the number of matching C or P atoms that are associated with matching residue names. These analyses are carried out with a default criteria that C or P atoms that are within 3 Å are matching and those further apart are not. This distance is arbitrary but was chosen to allow atoms to match in chains that superimpose secondary structure elements such as helices even if the secondary structure elements do not superimpose exactly.

In comparisons of models reconstructed into a reconstruction that has internal symmetry, the appropriate pairs of matching atoms may require application of that symmetry. The phenix.chain_comparison tool allows the inclusion of symmetry in the analysis.

Datasets used. We selected reconstructions for analysis on the basis of the following:

1. Availability of a reconstruction in the EMDB as of November 2017
2. Resolution of the reconstruction of 4.5 Å or better
3. Presence of a unique deposited model in the PDB matching the reconstruction
4. Consistent resolution in PDB and EMDB
5. Ability to use Phenix tools to automatically extract model and map from PDB and EMDB, apply symmetry if present in the metadata, and write the model

This resulted in 502 map–model pairs extracted from a total of 882 single-particle and helical reconstructions in the EMDB in this resolution range. (Note that only 660 of the 882 structures have one or more associated PDB entries.)

After our initial analysis, we further excluded reconstructions that had the following characteristics:

1. Map–model correlation for the deposited map and deposited model of less than 0.3 after extraction of map and model and analysis with phenix.map_model_cc (18 reconstructions)
(2) Deposited model in the PDB represents less than half of the structure (9 reconstructions).

This yielded the 476 map–model pairs described in this work (Supplementary Data 1). We downloaded the maps from the EMDB and used them directly in phenix.map_to_model, with the exception of EMDB-6531. For that map, a pseudo-helical reconstruction, we were able to deduce reconstruction symmetry only for the part of the map corresponding to the deposited model (PDB 3JAR), so we used the phenix.map_box tool to cut out a box of density around the region defined by the deposited model, analyzed this map, and at the conclusion of the process translated the automatically generated models to match the deposited map.

Parameters used when running phenix.map_to_model. All of the reconstructions selected were analyzed with official version 1.13-3015 of Phenix, with all default values of parameters except those specifying file names for the reconstructed map, sequence file, and symmetry information; the resolution of the reconstruction; and any control parameters specific to the computing system and processing approach (for example, the number of processors to use, queue commands, parameters specifying what parts of the calculation to carry out or what parts to combine in a particular job, and the level of verbosity in the output).

The phenix.map_to_model tool allows an analysis to be broken up into smaller tasks, after which all the results are combined to produce essentially the same result as would be obtained if the entire procedure were run in one step. We used this approach for two reasons. First, some of the datasets we analyzed required a very large amount of computer memory in certain stages of the analysis, and in particular in the map-segmentation and final model-construction steps. For other datasets such as ribosomes, we were able to speed up the analysis substantially by running smaller tasks on many processors.

Computation. We carried out most of the analyses in this work on the Grizzly high-performance computing cluster at Los Alamos National Laboratory, typically with one task per node. Some of the analyses were carried out on a dedicated computing cluster at Lawrence Berkeley National Laboratory, and some large analyses in particular were carried out on a machine with 1 TB of memory.

We monitored the CPU use for 175 of our analyses (generally of smaller structures) that were each carried out in a single step on a single machine using a single processor. These analyses took from 15 min to 12 h to complete on the Grizzly cluster. For example, the analysis of EMDB-6630 in Fig. 2c required 3 CPU hours. We also monitored the CPU use for one of the largest structures (EMDB-9565), which required 129 CPU hours to complete.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The phenix.map_to_model tool and all other Phenix software are available along with full documentation in source and binary forms from the Phenix website at https://www.phenix-online.org as part of the Phenix software suite.

Data availability
The datasets generated and/or analyzed during the current study are available in the Phenix data repository at https://phenix-online.org/phenix_data/terwilliger/map_to_model_2018. All the parameters, including resolution and specifications of reconstruction symmetry used in this work, are available on this site, along with all of the models produced.

References
23. DeLaBarre, B. & Brunger, A. T. Acta Crystallogr. D Biol. Crystallogr. 62, 923–932 (2006).
24. Terwilliger, T. C. Acta Crystallogr. D Biol. Crystallogr. 66, 285–294 (2010).
25. Terwilliger, T. C. Acta Crystallogr. D Biol. Crystallogr. 66, 268–275 (2010).
26. Terwilliger, T. C. Acta Crystallogr. D Biol. Crystallogr. 59, 38–44 (2003).
27. Terwilliger, T. C. et al. Acta Crystallogr. D Biol. Crystallogr. 69, 2244–2250 (2013).
28. Shalev-Benami, M. et al. Nat. Commun. 8, 1589 (2017).
29. Zhang, R., Alushin, G. M., Brown, A. & Nogales, E. Cell 162, 849–859 (2015).
30. Adams, P. D. et al. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221 (2010).
Experimental design

1. Sample size
   Describe how sample size was determined.

   We selected reconstructions to analyze based on:
   - availability of a reconstruction in the EMDB as of Nov, 2017
   - resolution of the reconstruction 4.5 Å or better
   - presence of a unique deposited model in the PDB matching the reconstruction
   - consistent resolution in the PDB and EMDB
   - ability to use Phenix tools to automatically extract model and map from PDB and EMDB,
   - apply symmetry if present in the metadata, and write the model

   This resulted in 502 map-model pairs extracted from a total of 882 single-particle and helical
   reconstructions in the EMDB in this resolution range. (Note that only 660 of the 882 have
   one or more associated PDB entries).

   We downloaded the maps from the EMDB16 and used them directly in
   phenix.map_to_model, with the exception of one map (EMDB entry 6351). For EMDB entry
   6351, a pseudo-helical reconstruction33, we could only deduce reconstruction symmetry for
   the part of the map corresponding to deposited model (PDB entry 3jar), so we used the
   phenix.map_box tool to cut out a box of density around the region defined by the deposited
   model, analyzed this map, and at the conclusion of the process translated the automatically-
   generated models to match the deposited map.

2. Data exclusions
   Describe any data exclusions.

   After our initial analysis we further excluded reconstructions that had the following
   characteristics:
   - map-model correlation for the deposited map and deposited model of less than 0.3 after
     extraction of map and model and analysis with phenix.map_model_cc (18 reconstructions)
   - deposited model in the PDB represents less than half of the structure (9 reconstructions)

   This yielded the 476 map-model pairs that are described in this work.

3. Replication
   Describe the measures taken to verify the reproducibility
   of the experimental findings.

   The successful application of our method to all 476 map-model pairs is our principal
   demonstration of reproducibility of the methods. Additionally, during the development of the
   software, many individual datasets were analyzed multiple times.

4. Randomization
   Describe how samples/organisms/participants were
   allocated into experimental groups.

   There was no randomization carried out

5. Blinding
   Describe whether the investigators were blinded to
   group allocation during data collection and/or analysis.

   Investigators were not blinded.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|-----|-----------|
| ☒   | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☒   | A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒   | A statement indicating how many times each experiment was replicated |
| ☐   | The statistical test(s) used and whether they are one- or two-sided Only common tests should be described solely by name; describe more complex techniques in the Methods section. |
| ☒   | A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☒   | Test values indicating whether an effect is present Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted. |
| ☒   | A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☒   | Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation) |

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The Phenix software was used for this study. The source code and documentation is available at www.phenix-online.org

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No restrictions are present

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

Eukaryotic cell lines were not used

b. Describe the method of cell line authentication used.

Eukaryotic cell lines were not used
c. Report whether the cell lines were tested for mycoplasma contamination.

Eukaryotic cell lines were not used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

Eukaryotic cell lines were not used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used
Policy information about studies involving human research participants

12. Description of human research participants
   Describe the covariate-relevant population characteristics of the human research participants.

No human subjects were used