Prevention of overfitting in cryo-EM structure determination

To the Editor: In the field of single-particle analysis of electron cryomicroscopy (cryo-EM) data, a growing concern that some resolution claims might not be substantiated by the data has helped instigate community-wide efforts to develop new validation tools[1]. A known issue with commonly used cryo-EM structure determination procedures is their tendency to overfit the data. Most procedures counter overfitting by low-pass filtering, but the effective frequencies for these filters are often based on suboptimal Fourier shell correlation[2] (FSC) procedures. In the suboptimal procedure, FSC curves are calculated between reconstructions from two halves of the data, and a single model is used to determine the relative orientations of all particles. It is well known that bias toward noise in the single model may inflate the resulting resolution estimates. To illustrate this, we applied the suboptimal procedure to a simulated cryo-EM data set of 20,212 GroEL particles. Whereas the reported resolution was 4.6 Å, the true resolution of the map was only 7.8 Å. Also, the presence of expected density features in the map does not necessarily provide sufficient evidence for a resolution claim: we made convincing figures of apparent side-chain density that in reality corresponded to overfitted noise (**Supplementary Fig. 1**). Consequently, overfitting may remain undetected, and interpretation of cryo-EM maps may be subject to errors.

The dangers of overfitting have been recognized, and refinement procedures with resolution-dependent weighting schemes to reduce overfitting have been proposed[3,4]. However, two known solutions to prevent overfitting are not in common use. By refining two models independently (one for each half of the data), so-called gold-standard[1] FSC curves may be calculated that are free from spurious correlations. Alternatively, the data used for the orientation determination may be limited to a user-specified frequency so that model bias beyond that frequency may be avoided. However, the argument that withholding part of the data from the refinement would substantially deteriorate the orientations and thereby the quality of the structure has prevented the widespread use of either of these solutions. In what follows, we prove this thesis to be false.

Analysis of simulated data with realistic signal-to-noise ratios (SNRs) indicates that the accuracy of the orientation determination is not affected either by the exclusion of high-frequency terms or by the use of a model that is reconstructed from only half of the particles (**Supplementary Fig. 2**). These simulations illustrate that only the low- to medium-frequency terms in the individual particles contain sufficiently high SNRs to contribute significantly to the orientation determination, which is in agreement with experimental evidence that cryo-EM particles may be aligned accurately using only low-frequency data[5]. Because in most cryo-EM studies, the low-medium frequencies of reconstructions from half of the particles are not expected to be significantly worse than those of reconstructions from all particles, we hypothesize that overfitting may be prevented without a notable loss of resolution using either frequency-limited refinement or

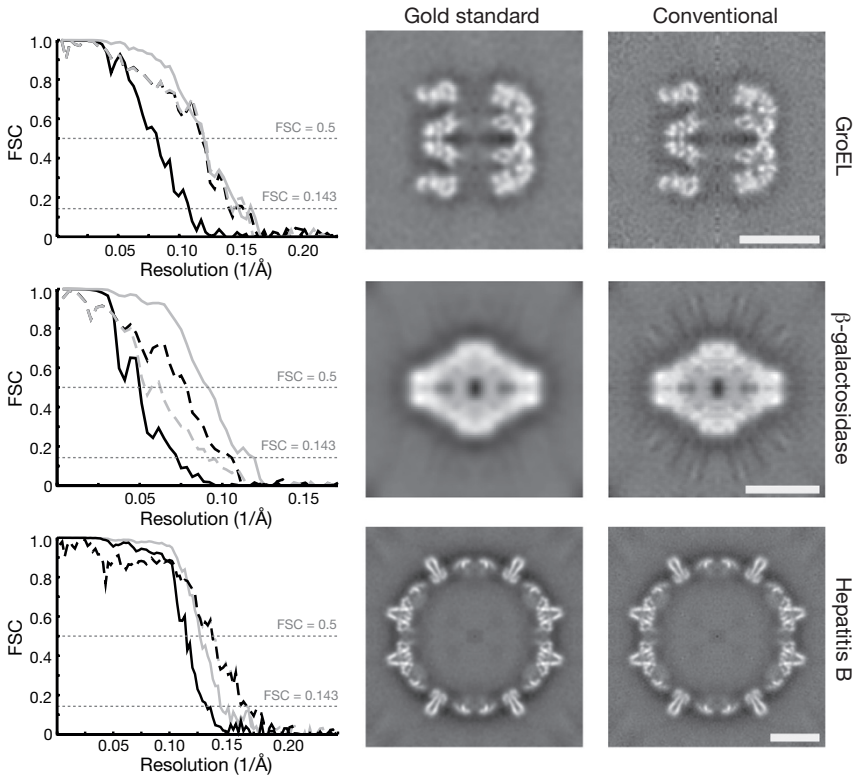

Figure 1 | The prevention of overfitting. Tests with three cryo-EM data sets (GroEL, β-galactosidase and hepatitis B) illustrate that overfitting may be avoided without compromising resolution. FSCs between reconstructions from random halves of the data (solid lines) and FSCs between the reconstruction from all particles and the crystal structures (dashed lines) show that the gold-standard procedure (black) does not yield lower resolutions, whereas the conventional procedure (gray) overestimates resolution (left column). The frequency at which the dashed lines pass through FSC = 0.5 indicate the true resolution of the reconstructions from all particles, and the frequency at which the solid lines pass through FSC = 0.143 indicate the reported resolution[8]. Corresponding values are given in **Supplementary Table 1**. Central slices through the reconstructions for the gold-standard procedure show less noisy maps with fewer artifacts (middle column) than those for the conventional procedure (right column). Scale bars, 100 Å.

refinement based on gold-standard FSCs. Because the former involves a decision by the user—that is, choosing the frequency at which to limit the refinement—we favor gold-standard FSCs and implemented a procedure to independently refine two models as a script on top of the conventional projection-matching protocol in the XMIPP package[6] (**Supplementary Fig. 3** and **Supplementary Software**).

We tested our hypothesis using three cryo-EM data sets: 5,053 GroEL particles that are distributed by the National Center for Macromolecular Imaging, an in house–collected data set of 50,330 β-galactosidase particles (**Supplementary Methods**) and 5,403 hepatitis B capsid particles from a previously published study[7]. High-resolution crystal structures are available for all three data sets, and these were used to assess the 'true' resolution obtained using refinements based on either gold-standard or conventional FSC procedures (**Fig. 1**). For all three cases, the conventional procedure reported apparently better FSC curves than did the gold-standard procedure, but in no case did the gold-standard procedure actually result in a lower-resolution map as measured against the crystal structure. On the contrary, for the β-galactosidase data, the gold-standard procedure yielded a map that correlated up to higher frequencies with the crystal structure than did the map obtained with the conventional procedure. The latter suffered from severe overfitting and gave rise to strong artifacts in the map. We also note that, in the absence of overfitting, the frequency at which the gold-standard FSC drops below 0.143 is a good indicator of the true resolution of the map (**Supplementary Table 1**), which is as expected from theory[8]. Finally, in small data sets, division of the data into two halves might affect resolution. However, calculations with subsets of the GroEL particles suggest that this only becomes an issue for data sets that are much smaller than those typically used in cryo-EM reconstructions (**Supplementary Fig. 4**).

The principal conclusion is therefore that overfitting of noise using suboptimal FSCs causes worse orientations and leads to a worse structure. In contrast, the use of gold-standard FSCs provides a realistic estimate of the true signal, which ultimately leads to a better map. The procedures proposed here are straightforward to implement in existing programs, and their application will eradicate the hazards of overfitting from cryo-EM structure determination procedures.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2115.

ACKNOWLEDGMENTS

We are grateful to T. Crowther and R. Henderson for helpful discussions and to J. Grimmett for help with computing. T. Crowther provided hepatitis B data, and the National Center for Macromolecular Imaging, which is funded by US National Institutes of Health grant P41RR02250, provided GroEL data. This work was funded by the UK Medical Research Council through grant MC_UP_A025_1013 to S.H.W.S.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Sjors H W Scheres & Shaoxia Chen

Medical Research Council Laboratory of Molecular Biology, Cambridge, UK.
e-mail: scheres@mrc-lmb.cam.ac.uk

PUBLISHED ONLINE 29 JULY 2012; DOI:10.1038/NMETH.2115

1. Henderson, R. *et al. Structure* **20**, 205–214 (2012).
2. Saxton, W.O. & Baumeister, W. *J. Microsc.* **127**, 127–138 (1982).
3. Stewart, A. & Grigorieff, N. *Ultramicroscopy* **102**, 67–84 (2004).
4. Scheres, S.H. *J. Mol. Biol.* **415**, 406–418 (2012).
5. Henderson, R. *et al. J. Mol. Biol.* **413**, 1028–1046 (2011).
6. Scheres, S.H. *et al. Nat. Protoc.* **3**, 977–990 (2008).
7. Böttcher, B., Wynne, S.A. & Crowther, R.A. *Nature* **386**, 88–91 (1997).
8. Rosenthal, P.B. & Henderson, R. *J. Mol. Biol.* **333**, 721–745 (2003).

The Coherent X-ray Imaging Data Bank

X-ray lasers produce pulses of X-rays 1 billion times brighter than those of synchrotrons; this capability creates extraordinary research opportunities in physics, chemistry and biology[1,2]. The high repetition rate of these machines leads to large volumes of data. For example, the Coherent X-ray Imaging beamline at the Linac Coherent Light Source (LCLS) produces three times more data per second than the ATLAS detector of the Large Hadron Collider. This data volume increase has created a need for tools to analyze and make efficient use of the new observations.

To address this issue, I created the Coherent X-ray Imaging Data Bank (CXIDB, http://cxidb.org/) as a permanent public repository of data from coherent X-ray sources (**Fig. 1a–c**), modeled after the Protein Data Bank[3] and the Electron Microscopy Data Bank[4]. CXIDB allows the community to share and organize their data as well as distribute the burden of data analysis, thereby effectively speeding up the rate of scientific discovery. It also increases the availability of data from X-ray lasers and the impact of these facilities by allowing groups that were not awarded beamtime to test hypotheses and develop analysis methods using data from others. Finally, it enables the reproduction of calculations from observations, and the verification of conclusions, a process that is one of the cornerstones of science.

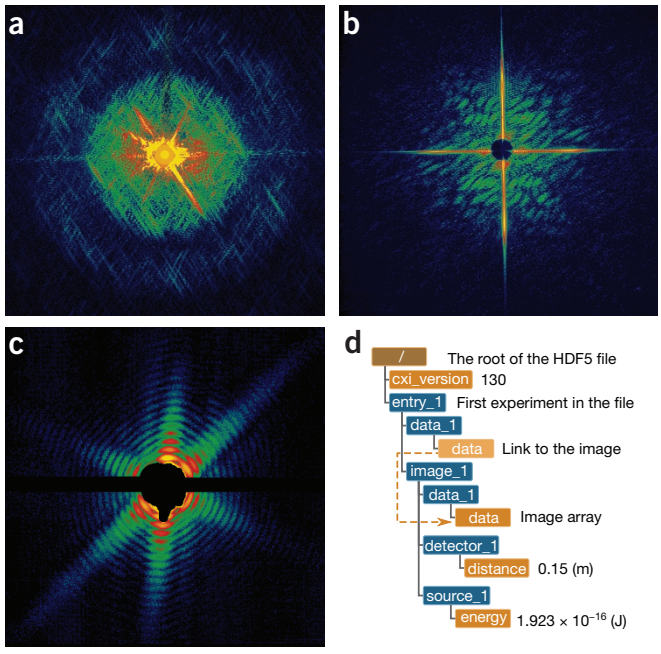

Figure 1 | Sample images from the CXIDB, covering different light sources and experimental techniques, and their file structure. (**a**) Diffraction pattern of a gold ball pyramid, part of a three-dimensional data set obtained at the Advanced Light Source[5]. (**b**) Diffraction pattern of a nanofabricated structure recorded with a single shot at the FLASH free-electron laser in Hamburg, Germany. (**c**) Diffraction pattern of a mimivirus particle intercepted in flight by the LCLS hard X-ray laser in Stanford, California[2]. (**d**) Simplified schematic representation of the hierarchic group structure of a CXI file.