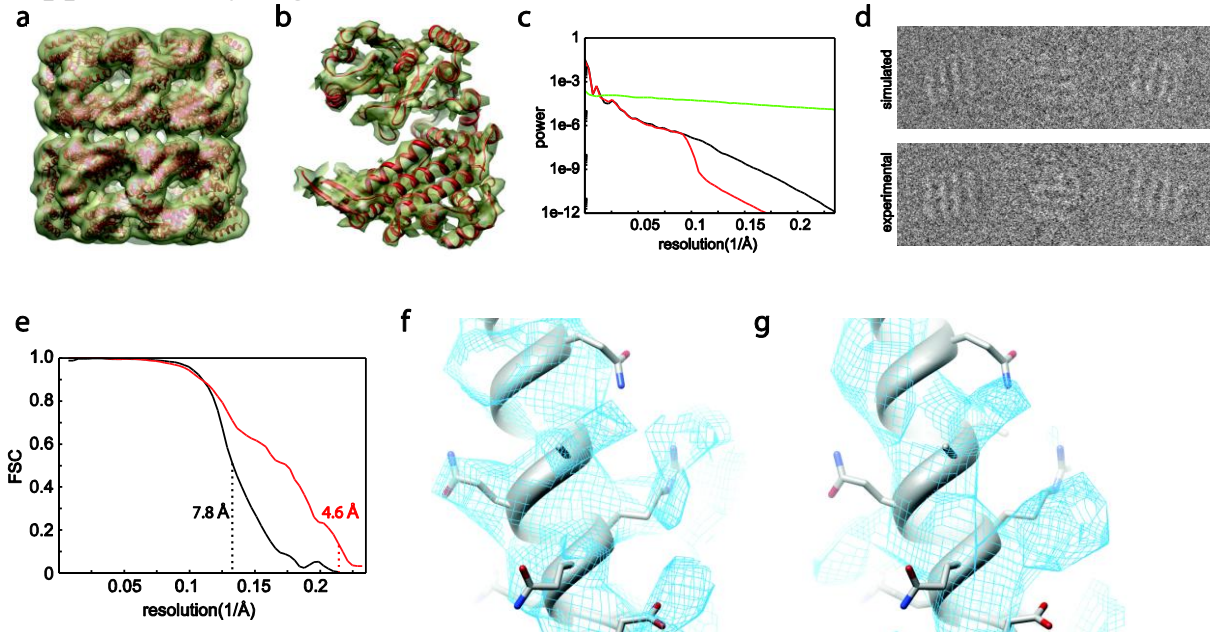
Supplementary information

The prevention of overfitting in cryo-EM structure determination

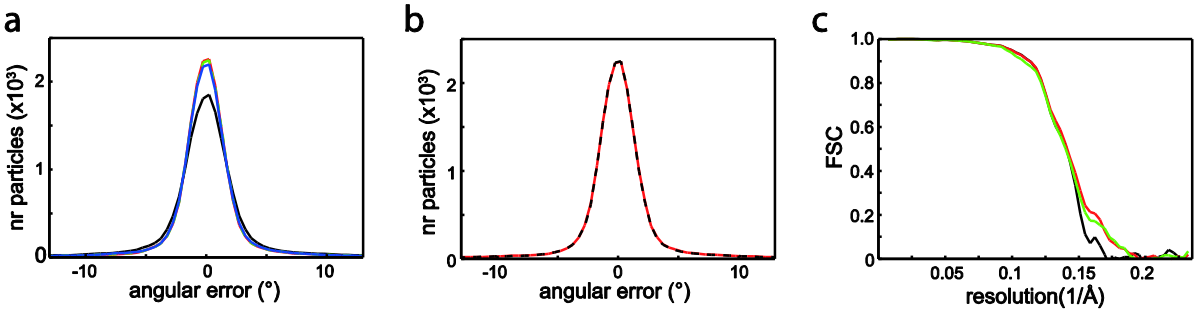
Sjors H.W. Scheres & Shaoxia Chen

Supplementary Figure 1:



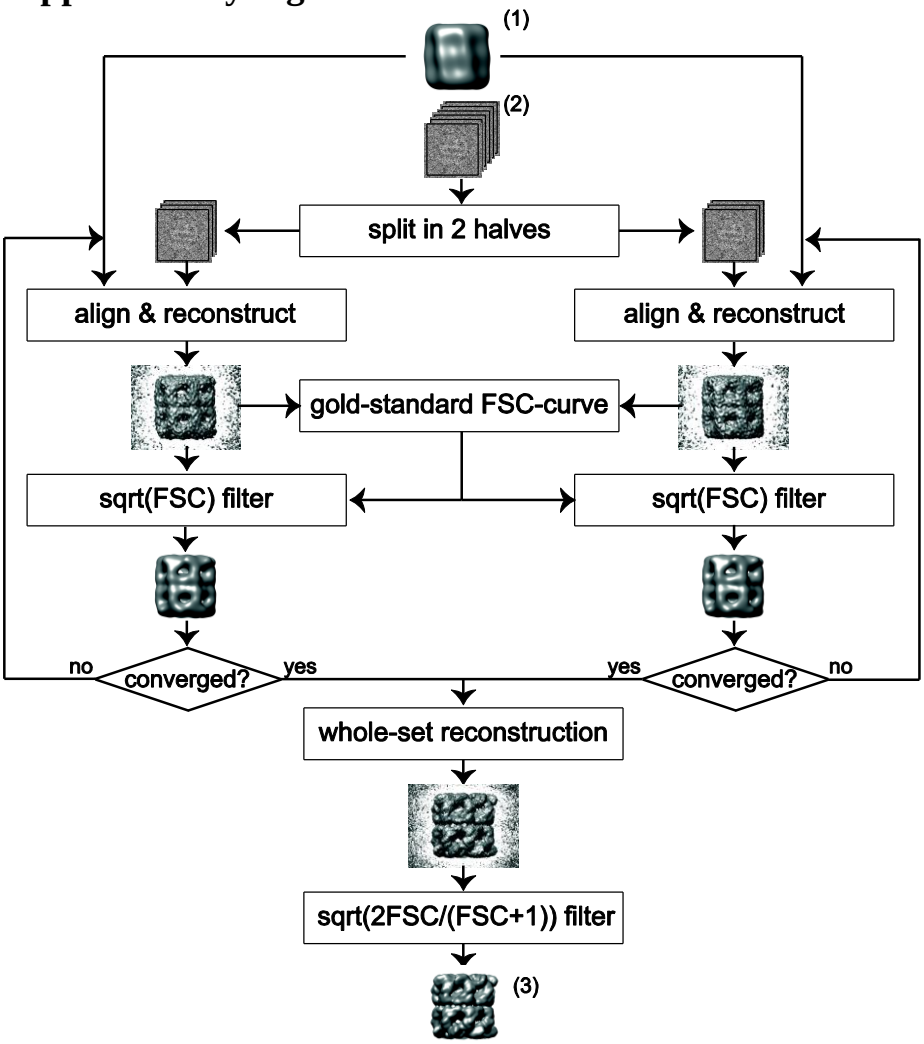
Simulations illustrate the pitfalls of undetected overfitting. (**a**) A simulated data set was designed to resemble an experimental cryo-EM data set of 5,168 GroEL particles that is distributed by the NCMI as part of a workshop on the EMAN2 software package[1]. Using standard procedures in XMIPP[2], all experimental particles were normalized, 115 particles were discarded after initial sorting, and the remaining 5,053 particles were windowed to images of 128×128 pixels, with a pixel size of 2.12 Å. A preliminary refinement with the experimental particles in RELION[3] yielded a 10 Å resolution map (transparent yellow), in which a published GroEL crystal structure[4] (Protein Data Bank ID: 1XCK) was fitted (red). This crystal structure contains 14 unique monomers in its asymmetric unit, and each of these was fitted separately into the reconstruction using UCSF Chimera[5], while for each monomer the equatorial, intermediate and apical domains were allowed to move independently as rigid bodies. (**b**) The fitted atomic model (red) was converted in XMIPP to a density map with a pixel size of 2.12 Å, and D7 symmetry was applied. The resulting map (transparent yellow) shows good density for α-helices and some bulky side chains, but the symmetrisation of 14 different chains made the visualization of details beyond approximately 7 Å cumbersome. (**c**) The phantom map was brought onto the same scale as the experimental map, and applying a B-factor of 350Å^2 yielded a power spectrum (black) that matched the power spectrum of the reconstruction from the experimental particles (red) up to the resolution of the reconstruction. The power spectrum of the noise in one of the micrographs as estimated for the experimental particles in RELION (green) is shown for comparison. (**d**) Projections of the phantom map were made in the orientations as determined for the experimental images (with small random perturbations), and the same CTFs were simulated as those estimated for the experimental particles. Independent Gaussian noise was added to the simulated particles in the Fourier domain using the same power spectra as estimated for the experimental data (e.g. the green line in **c**), which resulted in simulated particles with similar SNRs compared to the experimental ones. Examples of simulated particles are shown in the top row, and their experimental counterparts in the bottom row. To increase the size of the simulated data set, for each experimental particle four simulated particles were generated, resulting in a data set of 20,212 particles. Finally, prior to refinement a B-factor sharpening of −120Å^2 was applied to the images. A similar amount of sharpening was applied in the refinement of the original GroEL data set from which the experimental data set used here is a subset[6]. (**e**) Refinement using the conventional projection matching protocol in XMIPP[7] yielded a reported resolution of 4.6 Å (red), while the FSC with the original phantom map indicated a true resolution of 7.8 Å. (Note that the frequency where the FSC curve between two noisy reconstructions from half of the particles drops below 0.143 indicates where the signal-to-noise ratio in the reconstruction from all particles drops below 1, which is equivalent to the frequency where the FSC curve between the reconstruction from all particles and the noiseless phantom drops below 0.5, see ref. 8 for details). (**f**) The overfitted 4.6 Å map was interpreted in terms of an atomic model, which was obtained by a small rigid-body displacement of an α-helix from the GroEL crystal structure that was used to generate the phantom. Apparent density for side chains and the pitch in the α-helix in the overfitted map may look convincing support for the 4.6 Å resolution claim. (**g**) However, comparison of the same atomic model with the true density map reveals that the high-resolution features are merely due to noise.

Supplementary Figure 2:



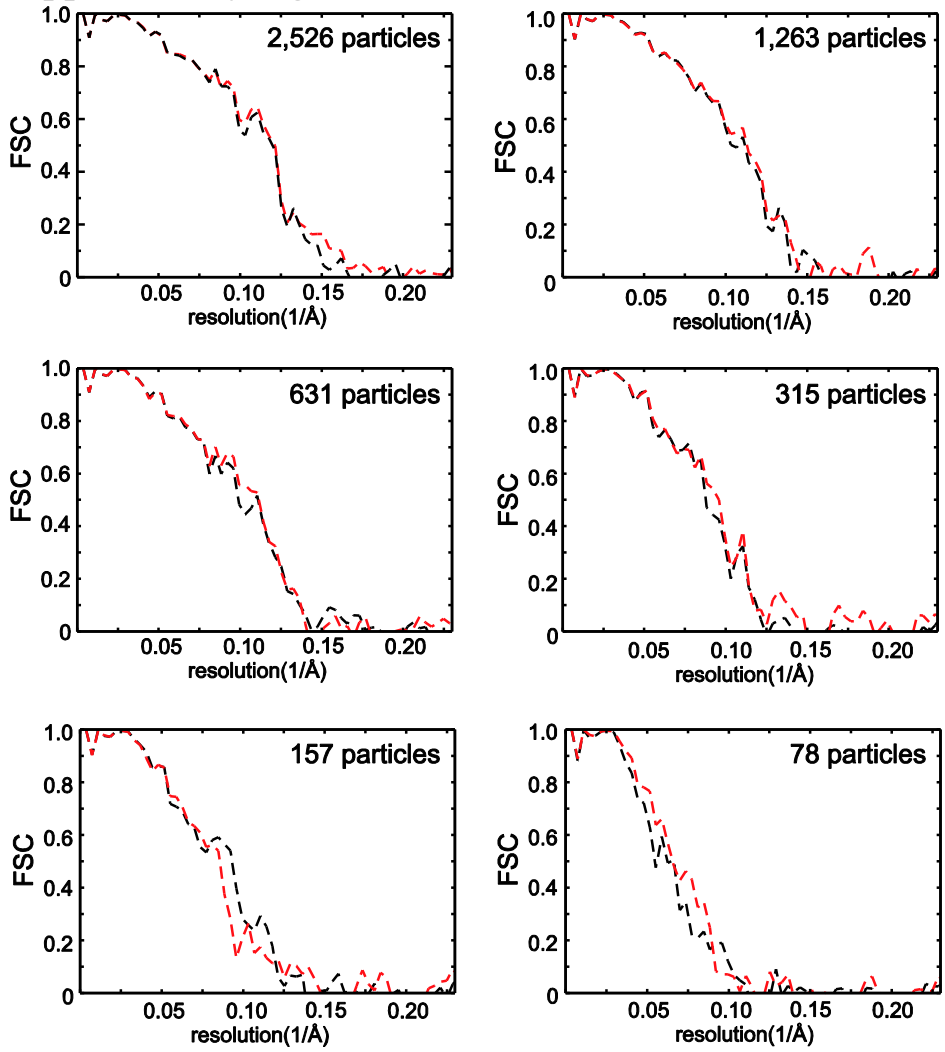
Refinement of two independent models or frequency-limited refinement do not lead to worse orientations or resolutions. (**a**) The simulated particles described in Supplementary Figure 1 were aligned against the true phantom map and the differences between the determined orientations and the true orientations (i.e. the angular errors) were plotted as histograms. Four different calculations were performed where the data that was included in the alignment was limited to 20 Å (black), 10 Å (blue), 8 Å(green) and Nyquist (red) frequencies. Although limiting the data to 20 Å has a notable effect on the quality of the orientations, data beyond 10 Å frequencies seem to contribute only marginally to the accuracy of the orientations. (**b**) A similar experiment was then performed for the alignment against a reconstruction from all particles (red) or reconstructions from only half of the particles (black) in their correct orientations. The two experiments yielded orientations of indistinguishable quality. (**c**) To validate that frequency-limited refinements or refinements against only half of the particles indeed do not result in worse orientations and therefore worse resolutions, three iterative refinements were started from the same 60 Å low-pass filtered phantom: a conventional procedure of a single model that used all data out to Nyquist (red); a conventional procedure that used only data out to 10 Å (green); and a procedure based on gold-standard FSCs (see Supplementary Figure 3) where two models were refined independently against two halves of the data using all data out to Nyquist (black). All three refinements yielded a map that correlated up to 7.1 Å with the phantom.

Supplementary Figure 3:



A generally applicable refinement scheme based on gold-standard FSCs. At the outset of refinement, the whole data set (2) is split into two separate half-sets, and both half-sets are used in independent refinements that start from the same initial reference structure (1). At every iteration, a (gold-standard) FSC curve is calculated between the reconstructions from both half-sets, and the resulting FSC curve is used to apply a $\sqrt{FSC}$-filter to both reconstructions. Upon some convergence criterion (both refinements converge simultaneously), the two half-reconstructions are summed (or a reconstruction from the whole data set is performed). The FSC curve from the final iteration is used to apply a $\sqrt{2FSC/(FSC+1)}$-filter to reflect that all particles contribute to the final reconstruction. Alternatively, more complex weighting schemes could be employed to take into account that the signal occupies only a fraction of the reconstructed volume[9]. To avoid overfitting, the final map should no longer be used in refinement. Note that in the calculations presented in this paper no real-space masking operations were performed on the reconstructions prior to the gold-standard FSC calculations, which may lead to a slight under-estimation of the true signal (also see Supplementary Table 1). However, care should be taken when masking half-reconstructions used for FSC-calculations, as real-space masking introduces correlations in the frequency-domain, which could again lead to spurious FSC curves. Also note that the two refinements start from the same model, which makes them not entirely independent. However, by using a strongly (and strictly) low-pass filtered initial model, and provided that this model lies within the radius of convergence of both refinements, inflated resolution estimates at higher resolutions may be avoided.

Supplementary Figure 4:



Even for relatively small data sets the gold-standard procedure yields similar resolutions as the conventional procedure. Subsets of the cryo-EM GroEL data set described in the main text were subjected to refinements using the conventional or the gold-standard procedures. The number of particles was decreased two-fold in six consecutive steps. Shown are the FSC curves between the resulting reconstructions and the GroEL crystal structure[4] for the gold-standard (black) and the conventional (red) procedure.

Supplementary Table 1:

True and reported resolutions (in Å) for the refinements shown in Figure 1 of the main text. The resolution where the dashed lines in Figure 1 pass through FSC=0.5 indicate the true resolution of the reconstructions from all particles, whereas the resolution where the solid lines pass through FSC=0.143 indicate the reported resolution (see suppl. ref. 8 for more details).

	Gold-standard procedure		**Conventional procedure**	
	Reported	True	Reported	True
GroEL	9.3	8.4	6.8	8.2
β-galactosidase	13.9	12.7	8.6	16.2
hepatitis B	7.6	7.3	7.0	7.3

Supplementary methods:

Electron microscopy on β-galactosidase

Solutions of *E. coli* β-galactosidase (obtained from Sigma; catalog no. G3153) at a concentration of 1 mg/ml were applied to glow-discharged Quantifoil grids (Agar Scientific), blotted, and plunge frozen in liquid ethane. Grids were transferred to an FEI Polara G2 microscope that was operated at 80 kV. Images were recorded on Kodak SO163 film at a calibrated magnification of 40, 956× with defocus between 1.2 μm and 2.7 μm using a dose of approximately 10 electrons/Å^2. Digitization with a KZA scanner and a step size of 6 μm, followed by an additional 2-fold down-sampling yielded a final pixel size of 2.93Å. From a total of 32 micrographs, 50,330 particles were selected manually.

Image processing

Prior to refinement, all particles were normalized using standard XMIPP procedures[7]. Refinements were started from 60 Å low-pass filtered models that were obtained in previous studies on these data sets. The only difference between the conventional refinements and the refinements based on gold-standard FSCs lied in the nature of the FSC calculations, all other parameters and algorithms were kept identical. For all three data sets, twelve iterations were performed with gradually increasing angular sampling rates (down to 1 degree for the GroEL and β-galactosidase data; and down to 0.25 degree for the hepatitis B data). Upon convergence of the gold-standard FSC refinements, the two independent models for each data set were added together, and this model was used to calculate the FSC with the crystal structure. The following crystal structures were used: wild-type GroEL from *E.coli* (PDB-ID: 1XCK)[4], β-galactosidase from *E.coli* (PDB-ID: 3I3E)[10], and human hepatitis B capsid (PDB-ID: 1QGT)[11].

Supplementary Software:

The following shell script was used to implement the gold-standard FSC-based refinement procedure in the projection matching protocol of the XMIPP package[7].

```
#!/usr/bin/env csh

# Number of iterations to perform
set nr_iter = 12
# Selection file with the input particles
set inselfile="betaGal.sel"
# Output rootname
set outroot="ProjMatchGold/run1"
# Python script with the rest of the parameters
set pyfile="xmipp_protocol_projmatch_goldstandard.py"
# Pixel size (in Angstroms)
set angpix="2.93"

##### Do not change anything below here.
# Split the selfile into 2  random halves
xmipp_selfile_split -i ${inselfile} -n 2 -o xmipp_projmatch_goldstandard_split

# Iterate
set iter = 0
while ($iter < $nr_iter)
 @ iter++

 # Run a single iteration of the projmatch protocol for each half of the data
 foreach h (1 2)
  set inselfile=`echo "xmipp_projmatch_goldstandard_split_"${h}".sel"`
  set workdir=`echo ${outroot}"_half"${h}`
  set nextiter=`echo \( ${iter} + 1 \) | bc`
  cat ${pyfile} | sed "s|XXXselfileXXX|${inselfile}|" | sed "s|XXXworkingdirXXX|${workdir}|" \
   | sed "s|XXXnr_itersXXX|${iter}|" | sed "s|XXXcontinueatXXX|${iter}|" > ${pyfile}_half${h}
  python ${pyfile}_half${h}
 end

 # Now calculate FSC between the two unfiltered maps
 xmipp_resolution_fsc -sam $angpix  \
  -i ${outroot}_half1/Iter_${iter}/Iter_${iter}_reconstruction.vol \
  -ref ${outroot}_half2/Iter_${iter}/Iter_${iter}_reconstruction.vol

 # Calculate sqrt(FSC)-filter
 # because correct_bfactor applies sqrt(2FSC/(1+FSC)) calculate FSC/2-FSC first
 awk 'BEGIN{a=0} {if ($1=="#"){print } else { if ($2<0 || a==1) {a=1; print $1,0} else \
  {print $1, $2/(2-$2)}}}'  \
  < ${outroot}_half1/Iter_${iter}/Iter_${iter}_reconstruction.vol.frc \
  > ${outroot}_half1/Iter_${iter}/Iter_${iter}_reconstruction.vol.tmpfsc

 # Filter both maps with sqrt(FSC) filter and OVERWRITE the filtered map from the protocol!
 xmipp_correct_bfactor -i ${outroot}_half1/Iter_${iter}/Iter_${iter}_reconstruction.vol \
  -o ${outroot}_half1/Iter_${iter}/Iter_${iter}_filtered_reconstruction.vol -sampling ${angpix} \
  -maxres 0.1 -adhoc 0 -fsc ${outroot}_half1/Iter_${iter}/Iter_${iter}_reconstruction.vol.tmpfsc
 xmipp_correct_bfactor -i ${outroot}_half2/Iter_${iter}/Iter_${iter}_reconstruction.vol \
  -o ${outroot}_half2/Iter_${iter}/Iter_${iter}_filtered_reconstruction.vol -sampling ${angpix} \
  -maxres 0.1 -adhoc 0 -fsc  ${outroot}_half1/Iter_${iter}/Iter_${iter}_reconstruction.vol.tmpfsc
 xmipp_fourier_filter -low_pass 0.48 \
  -i ${outroot}_half1/Iter_${iter}/Iter_${iter}_filtered_reconstruction.vol
 xmipp_fourier_filter -low_pass 0.48 \
  -i ${outroot}_half2/Iter_${iter}/Iter_${iter}_filtered_reconstruction.vol

end
```

Where, `xmipp_protocol_projmatch_goldstandard.py` is identical to the `xmipp_protocol_projmatch.py` file one would use for a conventional refinement (see ref. 7 for more details), except for the following fields:

```
SelFileName='XXXselfileXXX'
WorkingDir='XXXworkingdirXXX'
NumberofIterations=XXXnr_itersXXX
ContinueAtIteration=XXXcontinueatXXX
```

Installation instructions and further documentation on the XMIPP package and it projection matching protocol can be found at http://xmipp.cnb.csic.es.

Supplementary References:

[1] Tang, G. et al. *J. Struct. Biol.* **157**, 38–46 (2007).
[2] Scheres, S. H. W. *Meth. Enzym.* **482**, 295–320 (2010).
[3] Scheres, S. H. W. *J. Mol. Biol.* **415**, 406-418 (2012).
[4] Bartolucci, C. et al. *J. Mol. Biol.* **354**, 940–951 (2005).
[5] Pettersen, E. F. et al. *J. Comp. Chem.* **25**, 1605–1612 (2004).
[6] Ludtke, S. J. et al. *Structure* **16**, 441–448 (2008).
[7] Scheres, S. H. W. et al., *Nat. Protoc.* **3**, 977-990 (2008).
[8] Rosenthal, P. B. and Henderson, R. *J. Mol. Biol.* **333**, 721–745 (2003).
[9] Sindelar, C.V. and Grigorieff, N. *J. Struct. Biol.* in press.
[10] Dugdale, M. L., Dymianiw, D. L., Minhas, B. K., D'Angelo, I., and Huber, R. E. *Biochem. Cell Biol.* **88**, 861‑869 (2010).
[11] Wynne, S. A., Crowther, R. A., Leslie, A. G. *Mol Cell.* **3**, 771-780 (1999)