Cross-validation tests for cryo-EM maps using an independent particle set

Sebastian Ortiz[1], Luka Stanisic[2], Boris A Rodriguez[3], Markus Rampp[2], Gerhard Hummer[4,5], and Pilar Cossio[1,4,*]

[1] *Biophysics of Tropical Diseases, Max Planck Tandem Group, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.*

[2] *Max Planck Computing and Data Facility, 85748 Garching, Germany.*

[3] *Grupo de Física Atómica y Molecular, Instituto de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.*

[4] *Department of Theoretical Biophysics, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany.*

[5] *Institute of Biophysics, Goethe University, 60438 Frankfurt am Main, Germany.*

[*] *email:* pilar.cossio@biophys.mpg.de; grupotandem.biotd@udea.edu.co

Abstract

Cryo-electron microscopy is a revolutionary technique that can provide 3D density maps at near-atomic resolution. However, map validation is still an open issue in the field. Despite several efforts from the community, it is possible to overfit the reconstructions to noisy data. Here, inspired by modern statistics, we develop a novel methodology that uses a small independent particle set to validate the 3D maps. The main idea is to monitor how the map probability evolves over the control set during the refinement. The method is complementary to the gold-standard procedure, which generates two reconstructions at each iteration. We low-pass filter the two reconstructions for different frequency cutoffs, and we calculate the probability of each filtered map given the control set. For high-quality maps, the probability should increase as a function of the frequency cutoff and of the refinement iteration. We also compute the similarity between the probability distributions of the two reconstructions. As higher frequencies are added to the maps, more dissimilar are the distributions. We optimized the BioEM software package to perform these calculations, and tested the method on several systems, some which were overfitted. Our results show that our method is able to discriminate the overfitted sets from the non-overfitted ones. We conclude that having a control particle set, not used for the refinement, is essential for cross-validating cryo-EM maps.

Cryo-electron microscopy (cryo-EM) has revolutionized structural biology by providing electron density maps of biomolecules that were difficult to resolve with X-ray crystallography or nuclear magnetic resonance [1–3]. The introduction of direct electron detection cameras [4, 5] and novel computational algorithms [6, 7] has enabled the reconstruction of density maps with near-atomic details. To date, thousands of maps, and their corresponding atomic models, have been deposited in the electron microscopy [8] and protein data banks [9] (EMDB and PDB, respectively).

Typically, cryo-EM maps are reconstructed using the gold-standard procedure [10, 11]. The particle images are divided into two sets, and two independent reconstructions are generated. The reconstructions are refined iteratively using maximum-likelihood [12, 13] or Bayesian techniques [14, 15]. At each iteration the Fourier Shell Correlation (FSC) [16, 17] between the two independent reconstructions is computed. Fixed FSC threshold criteria at 0.143 [18] or 0.5 [17] are used to determine the resolution of the reconstructions (*i.e.*, the size of the smallest reliable detail). The refinement process is halted when the resolution of the reconstructions stops improving. In the end, the maps are masked and a final resolution is determined.

However, in spite of several efforts from the cryo-EM community, map validation is still problematic. In the recent Map Challenge it has been shown that there is no absolute 'gold standard' [19]. The protocols are user-dependent and there can be biases due to processing workflows. For instance, in the FSC calculation, the resolution estimate is dependent on the radius of the shell in Fourier space, and on the point symmetry of the molecule [20, 21]. The use of a fixed threshold for the FSC is restricted by the assumption that the noise and the signal are orthogonal [20]. In addition, the mask can be a source for overestimating the resolution [18, 22, 23]. Therefore, the best criteria to estimate the map resolution are still debated in the cryo-EM community [20, 21]. These issues can lead to overfitted cryo-EM reconstructions. For example, the reported values of the resolution in the model (from the PDB) and in the map (from the EMDB) are different for about 30% of the deposited data [24]. Moreover, it has been found that more than 70% of the maps in the EMDB have moderate to low agreement with the model, mostly because of the limited resolvable features of the maps [25]. In extreme cases, maps can be reconstructed from pure-noise images [26, 27].

Therefore, methods that validate the quality of the maps and models are fundamental for cryo-EM. Randomization of the phases beyond a frequency threshold can give signatures of overfitting in the FSC curve [11, 28]. Bet-

ter resolution estimates are obtained with reference-free pipelines using the 1/2 bit non-fixed FSC threshold [20, 29]. The local resolution in a map can be evaluated using the background noise of the reconstruction [30] or by masking different regions with the FSC [23, 31]. Predictability of the particle alignment provides quality indicators of the reconstruction [32, 33]. Moreover, several metrics that monitor cross-correlations in real or Fourier space between the maps and models indicate the reliability of the resolution [24, 25, 34]. Recently, deep learning algorithms have been introduced to automatically classify maps into high, medium, and low resolution [35]. However, all these methods have the limitation that they do not use the raw data, which ultimately comes from the individual particles, but they only use the maps or models that are product of processing and averaging. For instance, in cryo-EM there is no cross-validation method, such as the R-free in X-ray crystallography [36], which uses an independent control set from the pure experimental data.

Inspired by modern statistical methods, we here propose an unbiased strategy that validates cryo-EM reconstructions using a small control set of particle images that are omitted from the refinement process. We do not focus on determining a specific value for the resolution but we develop a simple cross-validation technique that monitors how the quality of the reconstructions evolves during the refinement procedure. We first calculate the BioEM [37, 38] probability of the maps, given the control set, as a function of a low-pass frequency cutoff of the reconstructions. High-quality maps should increase in probability for higher frequency cutoffs and higher refinement iterations. We then show that the similarity between the probability distributions of the two reconstructions from the gold-standard procedure is an additional quality indicator. Finally, we test the method on different systems and asses its effectiveness to discriminate overfitted maps.

Results

Cross-validation protocol.

We propose a statistical framework for the cross-validation of cryo-EM reconstructions. First, and foremost, the validation analysis is done over a small control set of particle images not used in the refinement process. Analogously to the R-free in X-ray crystallography [36], this independent set should give

an unbiased estimate of the quality of the reconstructions.

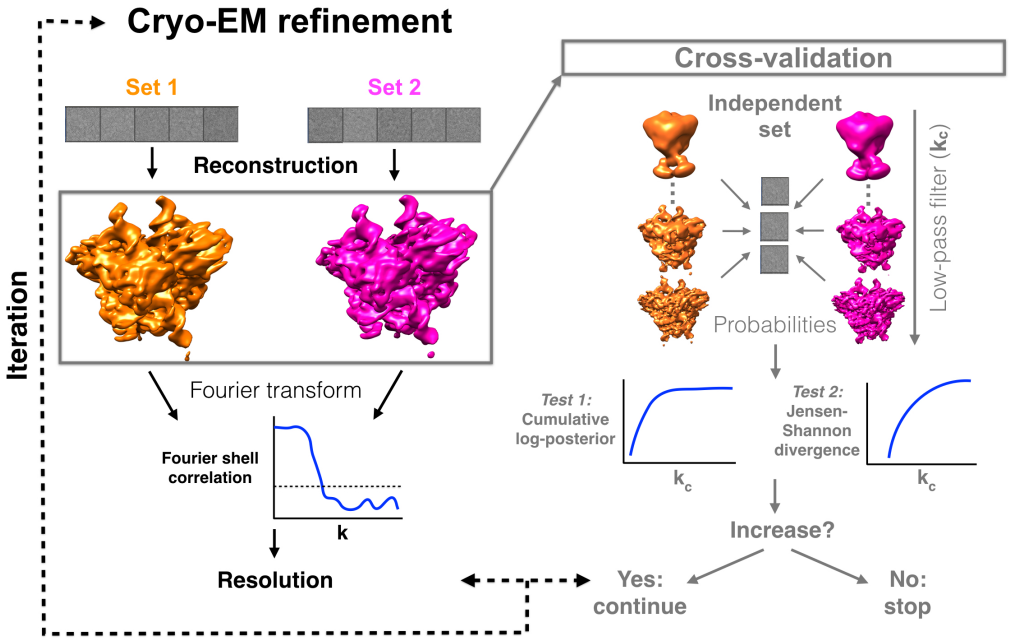


Figure 1: Cross-validation protocol for unbiased map validation in cryo-EM. (**left**) Gold-standard refinement procedure in cryo-EM. Two particle sets are used to generate two independent reconstructions. These reconstructions are compared using the Fourier shell correlation (FSC). A fixed FSC threshold is used to extract the resolution of the reconstructions. The process is iterated until the resolution stops improving. (**right**) Novel cross-validation protocol using a small control particle set. At each iteration of the refinement, the reconstructions are low-pass filtered to different frequency cutoffs k_c. The BioEM probabilities [37, 38], over the independent control set, are calculated as a function of k_c. Two tests validate the quality of the reconstructions: 1) the cumulative log-posterior and 2) the statistical similarity between the probability distributions (measured with a normalized Jensen-Shannon divergence). The results from both tests should increase as a function of the frequency cutoff. The maps represented correspond to the RAG1-RAG2 complex (see the Methods).

Fig. 1 shows the work-flow of the methodology. The refinement is done following the gold-standard procedure (Fig. 1–left), where two reconstruc-

tions are generated at each iteration step. These two reconstructions are validated using the control particle set (Fig. 1–right). At each iteration, the two maps are low-pass filtered to different frequency cutoffs, k_c (see the Methods). The BioEM [37] probability, $P_{i\omega}(k_c)$, for each set $i = 1, 2$ is calculated over the control set, $\omega \in \Omega$, with N_ω particles. As a first cross-validation test, we monitor the cumulative log-posterior, $\sum_\omega \ln(P_{i\omega}(k_c))/N_\omega$, as a function of k_c for each set i. This cumulative evidence should increase or remain constant as higher frequencies are added to the maps. Failing this test is a prime indicative that there is a problem in the refinement process.

The second cross-validation test consists on measuring the similarity between the probability distributions of the two reconstructions, also as a function of the frequency cutoff. For this purpose, we calculate a normalized Jensen-Shannon divergence (NJSD) (see the Methods). The NJSD is a positive, symmetric and bound metric that measures how distinguishable are the probability distributions from the reconstructions sets 1 and 2. We expect that as more frequencies are added to the reconstructions, more noise is added, and the probability distributions are more uncorrelated (*i.e.*, less similar).

In the following, we describe in detail the two cross-validation tests.

Map evidence from the cumulative log-posterior.

We tested the methodology over several cryo-EM datasets: the synaptic RAG1-RAG2 complex (RAG1-RAG2) [39], the human HCN1 channel (HCN1) [40], and the TRPV1 ion channel (TRPV1) [41]. These systems represent a diverse set of biomolecular families, with membrane proteins and protein-nucleicacids complexes. The reconstruction refinement was performed using the gold-standard procedure in RELION [14]. The final resolution of these systems ranges from approximately 3 to 6 Å(see the Methods). To analyze the impact of overfitting, we studied two additional systems: cryo-EM reconstructions from the HIV-1 envelop trimer (HIV-ET) [42] and a set of synthetic pure-noise images that act as a 'false' control set with the RAG1-RAG2 reconstructions (see the Methods). This was motivated by the fact that some reconstructions might have been generated from pure-noise particles, and their resolution might have been over-estimated [26, 27, 43].

In Fig. 2, we examine the improvement of the maps by monitoring the cumulative log-posterior relative to noise, $\sum_\omega \ln(P_{i\omega}(k_c))/N_\omega - \ln(P_{\text{Noise}})$, over the control set with $N_\omega = 5000$, as a function of k_c for the reconstructions

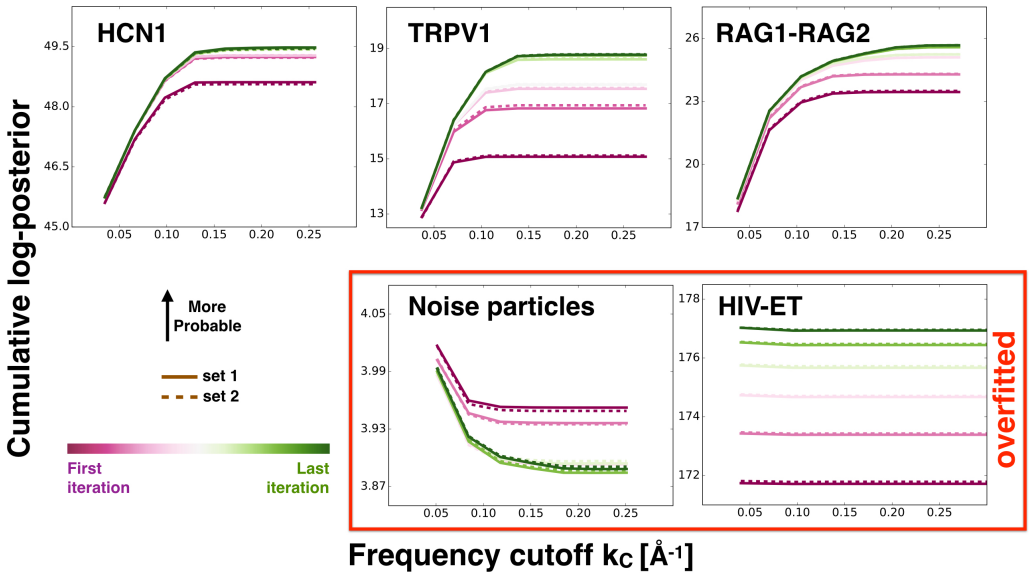


Figure 2: The cumulative log-posterior relative to noise $\sum_\omega \ln(P_{i\omega})/N_\omega - \ln(P_{\text{Noise}})$, over the control set with N_ω images, as a function of the frequency cutoff for reconstructions from set $i = 1$ and 2 (solid and dashed lines, respectively). The results are shown for different refinement iteration steps with a gradient color code: the first iteration is maroon and the last iteration is green. On the top row, we show the results for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2 for $N_\omega = 5000$. Systems that exhibit signs of overfitting, *i.e.* a noise-particle control set with $N_\omega = 1000$ and HIV-ET with $N_\omega = 5000$, are shown in the bottom row, highlighted with a red box.

from sets $i = 1, 2$. The results are shown for different refinement iterations with a gradient color scheme (first iteration: maroon; last iteration: green). These results measure how probable each filtered map is relative to P_{Noise} (see the Methods). For the RAG1-RAG2, HCN1 and TRPV1 systems, we find an increase of the map evidence (given by the cumulative log-posterior) as a function of the frequency cutoff. For very high frequencies, the cumulative evidence plateaus. We only observe minor differences between the results from set $i = 1$ and 2 (solid and dashed lines, respectively, in Fig. 2). This is an indication of the similarity between the reconstructions generated from the two sets. Importantly, the results highlight the ability of the BioEM

posterior to correctly rank maps of different resolutions. The reconstructions from the last iterations (*i.e.*, the most refined) are the most probable. This is in agreement with what one expects from the 3D-refinement algorithms [7].

In contrast, for the HIV-ET and noise-particle set, we find a different behavior of the map evidence. We find that the cumulative log-posterior does not increase as a function of the frequency cutoff but decreases or remains constant. For the noise-particle set, the map evidence relative to P_{Noise} is small, and the differences between iterations are almost two orders of magnitude smaller than for the non-overfitted sets. Moreover, for this case, as the refinement iterations increase, the maps are slightly less probable. This analysis monitors overfitting in cryo-EM: if the map evidence does not increase as a function of the frequency cutoff or the refinement iteration, then there are signs of overfitting in the data.

Similarity between the probability distributions.

As a second validation test, we compare the distributions of the posterior probabilities generated by the reconstructions from sets $i = 1, 2$ over the control set. In the Supplementary Information, we show an example of the probability distributions for the HCN1 system for two frequency cutoffs at a given iteration (Supplementary Fig. 1-top). We find that the probability distributions, over the independent set, are quite similar for both reconstructions. However, there are small differences between them, and the higher-frequency maps present larger fluctuations (Supplementary Fig. 1-bottom). These differences can be quantified using a normalized Jensen-Shannon divergence (NJSD; see the Methods).

In Fig. 3, we plot the NJSD as a function of the frequency cutoff k_c. Interestingly, for the RAG1-RAG2, HCN1 and TRPV1 systems, we observe that as the filtered maps contain higher frequencies, the larger the value of the NJSD. This implies that the probability distributions between maps with higher frequencies are less similar, possibly because they are more uncorrelated due to the high-frequency noise. For these standard systems, we also find that as the iteration increases the NJSD reaches at higher frequencies a plateau value. This behavior can be fit with an inverse exponential function $-Ae^{-k_c/\gamma} + B$ (see below and solid lines in Fig. 3). On the contrary, for the HIV-ET and noise-particle set, we find that the NJSD remains constant or has random behavior, suggesting that distributions do not consistently

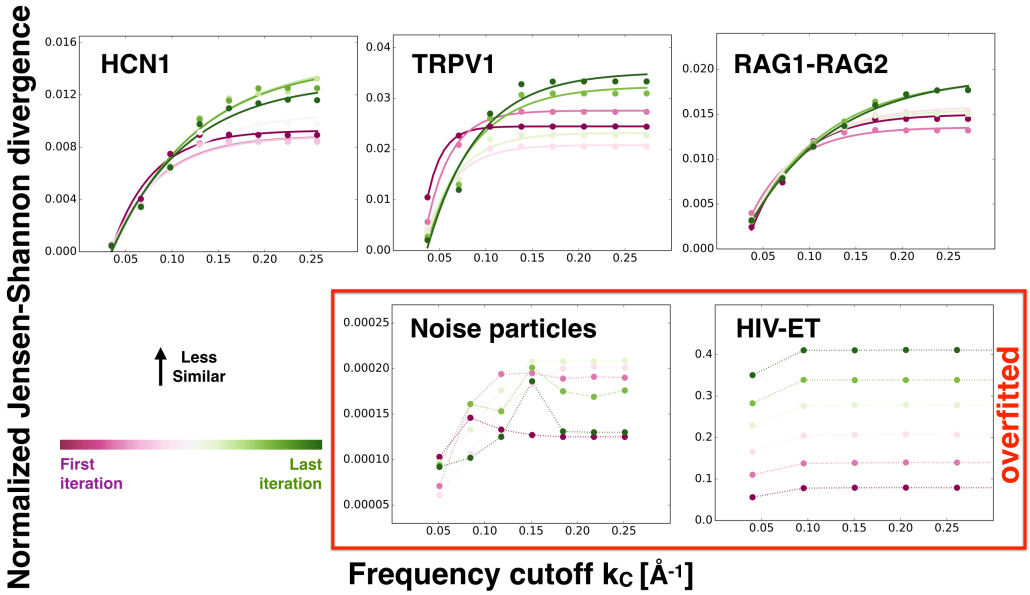


Figure 3: Normalized Jensen-Shannon divergence (NJSD) as a function of the frequency cutoff. This metric calculates the similarity between the distributions of the BioEM probabilities computed for the two reconstructions from sets 1 and 2. We use a gradient color code for the refinement iteration steps: the first iteration is maroon and the last iteration is green. On the top row, we show the results for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. For these systems, we fit the data points to an inverse exponential function $-Ae^{-k_c/\gamma} + B$ (solid lines). Systems that present signs of overfitting, a noise-particle control set and HIV-ET, are shown in the bottom row with dashed lines as a guide. The red box highlights the overfitted systems. The number of images in the control sets are the same as for the data in Fig. 2.

change when higher frequencies are added to the maps.

Cross-validation tests versus resolution.

We explored how the cross-validation results depend on the map resolution. For the HCN1, TRPV1 and RAG1-RAG2 systems, we find that the NJSD curves can be fitted to an inverse exponential function, $-Ae^{-k_c/\gamma} + B$ (solid lines shown in Fig. 3). Intuitively, the frequency γ indicates where the

plateau of the NJSD is reached. In Fig. 3, we can qualitatively see that γ is larger for higher refinement iterations. In Fig. 4, we plot the frequency γ as a function of the inverse of the resolution (calculated using the FSC at the threshold 0.143). Interestingly, we find that the frequency γ is highly correlated to the inverse of the resolution with correlation coefficient $r^2 = 0.93$, 0.91, and 0.85, for HCN1, TRPV1 and RAG1-RAG2, respectively. These results show that even from a small independent control set, it is possible to extract unbiased information about the map resolution. We note that for the HIV-ET and noise-particle sets it is not possible to fit the NJSD data to an inverse exponential function. Therefore, we can only estimate the correlation between γ and the inverse of the resolution for the standard cryo-EM systems.

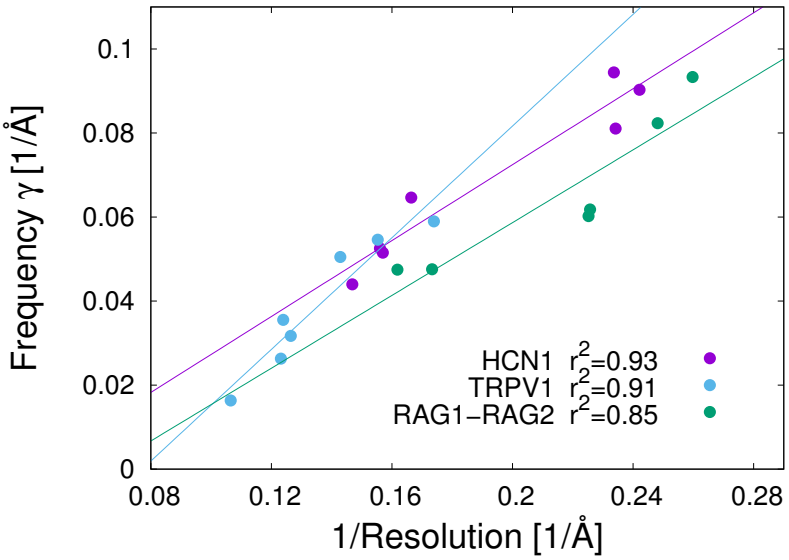


Figure 4: Frequency γ versus the inverse of the resolution for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. The NJSD curves for these systems were fitted to an inverse exponential function $-Ae^{-k_c/\gamma} + B$. We find large correlations between γ and the inverse of the resolution (calculated using the 0.143 criteria). The correlation coefficients are $r^2 = 0.93$, 0.91, and 0.85, for HCN1, TRPV1 and RAG1-RAG2, respectively. Solid lines show the linear fits.

Convergence over a small cross-validation set.

We assessed how the results depend on the number of particles in the control set. In Supplementary Fig. 2, we show an example of the cumulative log-posterior and NJSD as a function of the number of images in the control set. We find that after approximately 1000 particles these observables converge, suggesting that only a small set is needed to perform the cross-validation analysis. This is confirmed in Supplementary Fig. 3, where we plot the cumulative log-posterior and NJSD as a function of the frequency cutoff for a validation set of 1000 images. For the same set, in Supplementary Fig. 4, we plot the frequency γ as a function of the inverse of the map resolution, showing high correlations for the standard cryo-EM systems. These results are very similar to those obtained for the cross-validation set with 5000 particles.

Discussion

In this work, we have developed a novel methodology for cross-validating cryo-EM reconstructions. Importantly, the procedure is performed over an independent particle set that is not used to generate the reconstructions. Two cross-validation tests are proposed. The first consists of monitoring the cumulative log-posterior of the maps as a function of a low-pass filter frequency cutoff. The posterior should increase as a function of the frequency cutoff and the refinement iteration. In the second test, we assess the similarity between the probability distributions generated from the two reconstructions from the gold-standard procedure. The distributions should become less similar as higher frequencies are added to the reconstructions.

We performed the cross-validation tests over several systems: three standard cryo-EM reconstruction sets, and two datasets with noise particles that mimic overfitting. The results show substantial differences. While for the standard cryo-EM sets the results are as expected, the overfitted sets present almost no increment (even sometimes decrease) of the cumulative posterior or the NJSD. Thus, signatures of overfitting can be monitored with the proposed cross-validation tests.

Our methodology is general and robust. The mathematical framework is not only valid for the BioEM posterior but also for any posterior probability that measures the likelihood of a 3D density given a particle set. The tests converge over a small particle set, typically only 1000 particles. Moreover, the

methodology has the potential to be applicable for directly refining atomic models (instead of 3D maps) using an independent control set.

Determining an unbiased estimate of the reconstruction resolution remains an open issue. However, our procedure could shed light on how to tackle this problem with a different perspective. For example, the resolution could be defined as a multiple of γ that determines the frequency at which the information between the probability distributions is governed by noise.

All-in-all, our work provides a novel way to monitor overfitting in cryo-EM. We conclude that having a control particle set which is not used to generate the reconstructions should become a standard for any cryo-EM application.

Methods

Benchmark systems.

We used the following benchmarks that represent diverse biomolecular families and cryo-EM systems:

The human hyperpolarization-activated cyclic nucleotide-gated channel (HCN1) is a voltage-dependent ion channel, which was resolved to high resolution using cryo-EM [40]. The system was resolved in two conformational states, an *apo* state and a cAMP-bound state, to ~ 3.5 Å using RELION 3D-refinement [14]. 55870 particles images belonging to the *apo* state together with the defocus information of each particle are available in the Electron Microscopy Public Image Archive (EMPIAR) [44] with code 10081. Their pixel and image size are also available in that archive.

The recombination-activating genes RAG1-RAG2 form a complex (RAG1-RAG2) that plays an essential role in the generation of antibodies and antigen-receptor genes in a process called V(D)J recombination. Two main structures of the RAG1-RAG2 complex can be distinguished during the V(D)J recombination, a synaptic paired complex and the signal end complex (SEC). These states were resolved to 3.7 and 3.4 Å, respectively, using cryo-EM [39]. 81946 processed picked particles from the SEC state are deposited in the EMPIAR data bank with code 10049. The defocus information is available for these particles.

The mammalian transient receptor potential TRPV1 ion channel (TRPV1) is the receptor for capsaicin. Its structure was determined to 3.4 Å using cryo-EM [41]. A set of 35645 processed particles for this system are found in the EMPIAR data bank with code 10005. The defocus information is also available for these particles.

The human immunodeficiency virus type 1 envelope glycoprotein trimer (HIV-ET) is a membrane-fusing machine which mediates virus entry into host cells. The structure of the apo HIV-1 envelope glycoprotein in the trimer-conformation was determined to 6 Å using the 0.5 FSC threshold with cryo-EM [42]. A set of 124478 particles used in the

refinement process is available in EMPIAR with code 10008. The defocus information is also available for these particles.

For all of the above cases, a subset of 5000 particles was randomly selected to be used as the cross-validation set. Specifically, these particles are not used in the refinement processes.

Pure-noise images: we generated a set of synthetic 1000 pure-noise particles. Each particle contains random intensities following a Gaussian distribution with zero mean and unit variance (for details see the Supplementary Information). These images were used as a "false" control set to assess the RAG1-RAG2 reconstructions.

3D refinement.

System	#Particles	Symmetry	#iterations	Final resolution*
HCN1	50870	C4	17	4.2Å
RAG1-RAG2	79946	C2	26	3.8Å
TRPV1	30645	C4	24	5.3Å
HIV-ET	119478	C3	10	9.9Å

*using the 0.143 FSC threshold

Table 1: Summary of the results from the 3D-refinement using RELION [14] for the cryo-EM systems.

The RELION [14] software was used to reconstruct the cryo-EM maps. For all systems, we assume that the deposited particles correspond to the same state. Therefore, the preprocessing steps of 2D or 3D classification are not performed. As the initial reference map for the 3D refinement, we use the final map reported by the authors low-pass filtered to 60 Å. This was done to minimize the risk of overfitting [11]. The 3D-refinement procedure implements the gold-standard approach by splitting the data into two random halves (sets $i = 1, 2$) and performing two independent reconstructions. We note that the number of particles used for these reconstructions was slightly less than those of the original works because the particles from the control set were taken out. In all cases, we used the RELION default parameters, and point-group symmetries reported by the authors. Table 1 summarizes the results obtained from the 3D refinement. The resolutions are in accordance with the reported ones, taking into account that the post-processing steps were not performed, and that the control set of particles was excluded from the refinement.

Low-pass filter.

Consider a map m generated from an iteration of the 3D refinement. Let $\mathcal{F}_m(\mathbf{k})$ be its 3D-Fourier transform, where $\mathbf{k}$ is the reciprocal vector. We perform a low-pass filter on the map, $\mathcal{F}_m^{k_c}(\mathbf{k})$, up to a frequency cutoff k_c. The resulting filtered map is

$$\mathcal{F}_m^{k_c}(\mathbf{k}) = \begin{cases} \mathcal{F}_m(\mathbf{k}) & k \leq k_c \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We use the code *lowpassmap_fftw* available from the Rubinstein lab webpage [45] to perform this calculation. We then convert the map into real space by applying the inverse Fourier transform of $\mathcal{F}_m^{k_c}(\mathbf{k})$. The real-space filtered map is masked and then used as input for the BioEM computation (see below).

BioEM posterior probabilities.

The BioEM method [37] uses a Bayesian framework to quantify the consistency between an experimental image ω and a given map m (or model) by calculating a posterior probability $P_{m\omega}$. BioEM takes into account the relevant physical parameters (Θ) for the image formation: center displacement, normalization, offset, noise, orientation and CTF parameters (defocus, amplitude, and B-factor). $P_{m\omega}$ is calculated by integrating-out all parameters

$$P_{m\omega} \propto \int L(\omega|\Theta, m)p(\Theta)p(m)d\Theta \ , \quad (2)$$

where $p(m)$ and $p(\Theta)$ are the prior probabilities of the map and parameters, respectively, and $L(\omega|\Theta, m)$ is the likelihood function. We considered the prior probabilities of maps and parameters uniform over the integration intervals. In Eq. 2, the integrals over the offset, noise and normalization are performed analytically [37], and that over the center displacement is described in ref. [38]. The integral over the orientations and CTF defocus is done using a double-round algorithm, which is described in the following subsection.

Similarly as in ref. [37], we define a noise model $P_{\text{Noise}} = (2\pi\lambda^2 e)^{-N_{\text{pix}}/2}$ where N_{pix} is the number of pixels and λ is the image variance (by default $\lambda = 1$). P_{Noise} is used as a reference to compare the posterior probabilities.

BioEM algorithm.

To optimize the computations, we divided the BioEM posterior calculation into two rounds. The objective of the first round is to obtain the best orientations for each particle. In this round, an all-orientations to all-particles algorithm is performed [38]. As the BioEM input map, we used the final reconstruction from the refinement with a broad mask and without low-pass filtering. To sample the orientations, we used 36864 quaternions that sample uniformly orientation space [46]. The particles were grouped into sets with similar experimental defocus with $0.4\mu m$ range, and an independent orientation search was performed for each group. In this round, the best 10 orientations for each particle are obtained. An example of the BioEM input for the first round is presented in the Supplementary Information.

In the second round, a zoom around the best 10 orientations from the first round and experimental defocus is performed for each low-pass filtered reconstruction from the different refinement iterations. The zoom around each best orientation is done using 125 quaternions with approximately 0.01 grid spacing, resulting in 1250 zoomed-orientations

for each particle. This procedure is described in detail in ref. [47]. The defocus of each particle is fixed to its experimental value. We used 8 filtering-frequencies for each reconstruction; these were distributed uniformly from $1/(p_s\sqrt{N_{\rm pix}})$ to $1/(3p_s)$ where p_s is the pixel size. All reconstructions were masked using the same broad mask as for round 1. An example of the BioEM input file for round 2 is presented in the Supplementary Information.

BioEM code.

The BioEM code has been extended with several optimizations, which drastically increase performance for the second round of calculations. Most importantly, the main data structures and algorithm were modified to allow for a parallel comparison of multiple orientations to a single particle image. Initial reading of the input files has been parallelized, and the overall memory consumption decreased. These code changes lead to more efficient utilization of the computing resources, and hence to a faster calculation of posterior probabilities, especially for the workloads specific to the second round. For more information, we refer the reader to the BioEM user manual: https://readthedocs.org/projects/bioem/.

Normalized Jensen-Shannon divergence.

Measuring a distance among probability distributions is a common task in statistics. Most distance measures include concepts from information theory, such as the Kullback-Leibler divergence [48, 49] or the Shannon entropy [50]. In this work, we measure the statistical similarity between the probability distributions from reconstructions from set 1 and set 2 calculated over the control set. We define a metric that is the Jensen-Shannon divergence [49, 50] normalized by the individual Shannon entropies

$$\mathrm{NJSD} = \frac{\sum_\omega [P_{1\omega}\ln(P_{1\omega}/M_\omega) + P_{2\omega}\ln(P_{2\omega}/M_\omega)]}{2(\sum_\omega P_{1\omega}\ln(P_{1\omega})\sum_\omega P_{2\omega}\ln(P_{2\omega}))^{1/2}}, \tag{3}$$

where $P_{1\omega}$ and $P_{2\omega}$ are the probabilities of the reconstructions from set 1 and 2, respectively, over image ω, and $M_\omega = (P_{1\omega}+P_{2\omega})/2$. For simplicity of notation, we have omitted the dependency of the probabilities on the frequency cutoff k_c. To calculate Eq. 3, we normalize the posterior probabilities such that $P_{1\omega}+P_{2\omega}=1$ for each image ω, frequency cutoff and iteration.

In Eq. 3, the numerator measures the correlation between the probability distributions, and the Shannon entropies in the denominator play the role of a normalization factor. Some important properties of the NJSD metric are that it is positive, symmetric and its lower bound is 0 if and only if $P_{1\omega}=P_{2\omega}$ for all particles ω.

Data availability

The BioEM code is available at https://github.com/bio-phys/BioEM. A tutorial to perform the cross-validation protocol is available at:

https://github.com/bio-phys/BioEM-tutorials.

Acknowledgements

The authors thank Dr. Alessandro Laio for insightful discussions, Dr. Jose Maria Carazo for information about the overfitted systems, and Dr. Frank Avila for proof reading. S.O. and P.C. were supported by Colciencias, University of Antioquia and Ruta N, Colombia. G.H., and P.C. acknowledge the support of the Max Planck Society. Some computations were performed on a local server with an NVDIA Titan X GPU. PC gratefully acknowledges the support of NVIDIA Corporation for the donation of this GPU. Other computations were performed at the Max Planck Computing and Data Facility.

Author contributions

G.H. and P.C. conceived the presented idea. S.O., G.H, and P.C. developed the theory. S.O. and P.C. performed the computations. L.S. and M.R. co-developed and optimized the code. All authors discussed the results and contributed to the final manuscript.

References

[1] Kühlbrandt, W. Cryo-EM enters a new era. *eLife* **3**, e03678 (2014).

[2] Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161**, 450–457 (2015).

[3] Murata, K. & Wolf, M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta* **1862**, 324–334 (2018).

[4] Wu, S., Armache, J.-P. & Cheng, Y. Single-particle cryo-EM data acquisition by using direct electron detection camera. *Microscopy* **65**, 35–41 (2016).

[5] McMullan, G., Faruqi, A. & Henderson, R. Direct Electron Detectors. *Methods in Enzymology* **579**, 1–17 (2016).

[6] Kervrann, C., Sanchez Sorzano, C. O., Acton, S. T., Olivo-Marin, J.-C. & Unser, M. A Guided Tour of Selected Image Processing and Analysis Methods for Fluorescence and Electron Microscopy. *IEEE Journal of Selected Topics in Signal Processing* **10**, 6–30 (2016).

[7] Cossio, P. & Hummer, G. Likelihood-based structural analysis of electron microscopy images. *Current Opinion in Structural Biology* **49**, 162–168 (2018).

[8] Lawson, C. L. *et al.* EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* **39**, D456–D464 (2011).

[9] Berman, H. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

[10] Henderson, R. *et al.* Outcome of the First Electron Microscopy Validation Task Force Meeting. *Structure* **20**, 205–214 (2012).

[11] Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature methods* **9**, 853 (2012).

[12] Sorzano, C. *et al.* XMIPP: a new generation of an open-source image processing package for electron microscopy. *Journal of Structural Biology* **148**, 194–204 (2004).

[13] Tang, G. *et al.* EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology* **157**, 38–46 (2007).

[14] Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology* **180**, 519–530 (2012).

[15] Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290–296 (2017).

[16] Saxton, W. O. & Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein. *Journal of Microscopy* **127**, 127–138 (1982).

[17] Harauz, G. & van Heel, M. Exact Filters for General Geometry Three Dimensional Reconstruction. *Optik* **78**, 6–30 (1986).

[18] Rosenthal, P. B. & Rubinstein, J. L. Validating maps from single particle electron cryomicroscopy. *Current Opinion in Structural Biology* **34**, 135–144 (2015).

[19] Heymann, J. B. *et al.* The first single particle analysis Map Challenge: A summary of the assessments. *Journal of structural biology* **204**, 291–300 (2018).

[20] Van Heel, M. & Schatz, M. Fourier shell correlation threshold criteria. *Journal of structural biology* **151**, 250–262 (2005).

[21] Sorzano, C. O. S. *et al.* A review of resolution measures and related aspects in 3D Electron Microscopy. *Progress in biophysics and molecular biology* **124**, 1–30 (2017).

[22] Penczek, P. A. Resolution Measures in Molecular Electron Microscopy. *Methods in Enzymology* **482**, 73–100 (2010).

[23] Pintilie, G., Chen, D.-H., Haase-Pettingell, C., King, J. & Chiu, W. Resolution and Probabilistic Models of Components in CryoEM Maps of Mature P22 Bacteriophage. *Biophysical Journal* **110**, 827–839 (2016).

[24] Afonine, P. V. *et al.* New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallographica Section D Structural Biology* **74**, 814–840 (2018).

[25] Neumann, P., Dickmanns, A. & Ficner, R. Validating Resolution Revolution. *Structure* **26**, 785–795.e4 (2018).

[26] Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences* **110**, 18037–18041 (2013).

[27] Shatsky, M., Hall, R. J., Brenner, S. E. & Glaeser, R. M. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of structural biology* **166**, 67–78 (2009).

[28] Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013).

[29] Afanasyev, P. *et al.* Single-particle cryo-EM using alignment by classification (ABC): the structure of Lumbricus terrestris haemoglobin. *IUCrJ* **4**, 678–694 (2017).

[30] Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature methods* **11**, 63–5 (2014).

[31] Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *Journal of structural biology* **184**, 226–36 (2013).

[32] Vargas, J., Melero, R., Gómez-Blanco, J., Carazo, J. M. & Sorzano, C. O. S. Quantitative analysis of 3D alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis. *Scientific Reports* **7**, 6307 (2017).

[33] Vargas, J., Otón, J., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. Particle alignment reliability in single particle electron cryomicroscopy: a general approach. *Scientific Reports* **6**, 21626 (2016).

[34] Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallographica Section D Biological Crystallography* **71**, 136–153 (2015).

[35] Avramov, T. K. *et al.* Deep Learning for Validating and Estimating Resolution of Cryo-Electron Microscopy Density Maps. *Molecules* **24** (2019).

[36] Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

[37] Cossio, P. & Hummer, G. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of structural biology* **184**, 427–437 (2013).

[38] Cossio, P. *et al.* BioEM: GPU-accelerated computing of Bayesian inference of electron microscopy images. *Computer Physics Communications* **210**, 163–171 (2017).

[39] Ru, H. & Others. Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. *Cell* **163**, 1138–1152 (2015).

[40] Lee, C.-H. & MacKinnon, R. Structures of the human HCN1 hyperpolarization-activated channel. *Cell* **168**, 111–120 (2017).

[41] Liao Maofu, Cao Erhu, Julius David & Cheng Yifan. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107 (2013).

[42] Mao, Y. *et al.* Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. *Proceedings of the National Academy of Sciences* **110**, 12438–12443 (2013).

[43] Subramaniam, S. Structure of trimeric HIV-1 envelope glycoproteins. *Proceedings of the National Academy of Sciences* **110**, E4172–E4174 (2013).

[44] 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).

[45] Rubinstein Lab. URL `https://sites.google.com/site/rubinsteingroup/home`.

[46] Yershova, A., Jain, S., LaValle, S. M. & Mitchell, J. C. Generating uniform incremental grids on SO(3) using the Hopf fibration. *Int. J. Robot. Res.* **29**, 801–812 (2010).

[47] Cossio, P. *et al.* Bayesian inference of rotor ring stoichiometry from electron microscopy images of archaeal ATP synthase. *Microscopy* **67**, 266–273 (2018).

[48] Kullback, S. *Information theory and statistics.* (Dover Publications, 1968).

[49] Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37**, 145–151 (1991).

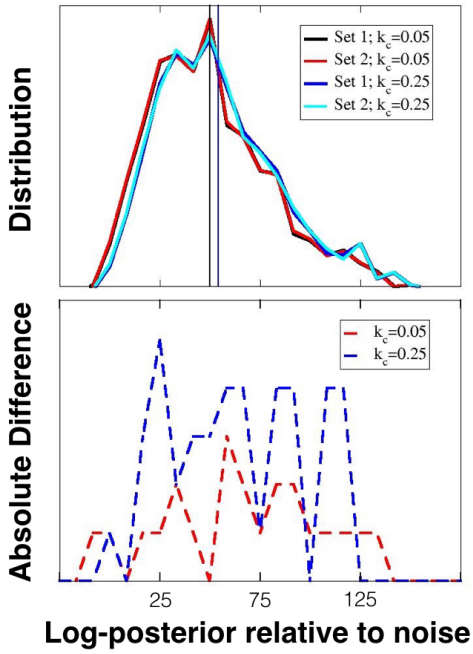
[50] Cover, T. M. & Thomas, J. A. Elements of information theory 2nd edition (2006).

Supplementary Information: Cross-validation tests for cryo-EM maps using an independent particle set

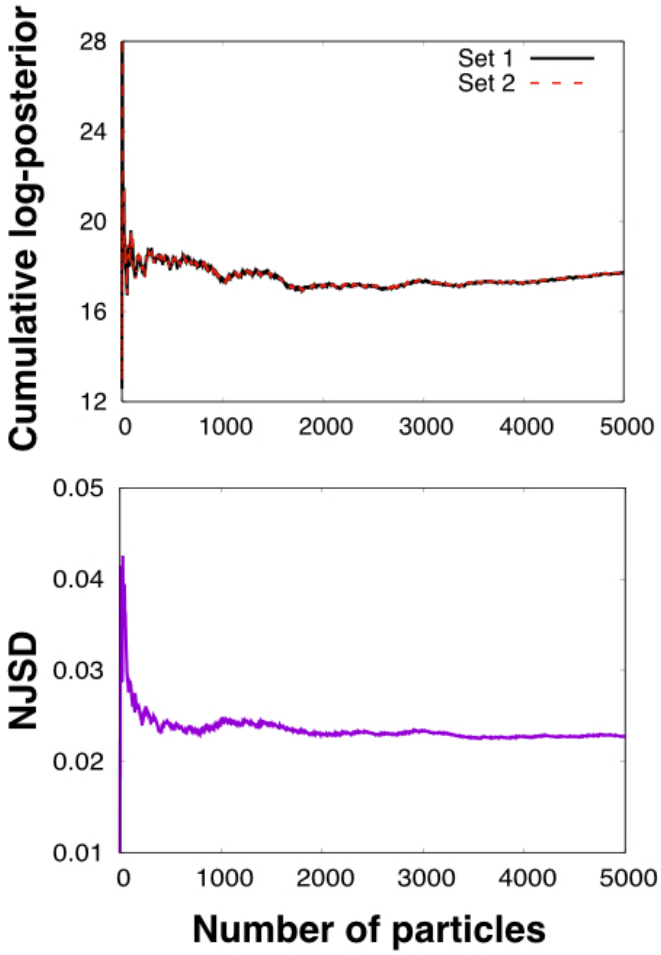
Sebastian Ortiz[1], Luka Stanisic[2], Boris A Rodriguez[3], Markus Rampp[2], Gerhard Hummer[4,5], and Pilar Cossio[1,4,*]

[1] *Biophysics of Tropical Diseases, Max Planck Tandem Group, University of Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.*
[2] *Max Planck Computing and Data Facility, 85748 Garching, Germany.*
[3] *Grupo de Física Atómica y Molecular, Instituto de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.*
[4] *Department of Theoretical Biophysics, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany.*
[5] *Institute of Biophysics, Goethe University, 60438 Frankfurt am Main, Germany.*
[*] *email:* pilar.cossio@biophys.mpg.de; grupotandem.biotd@udea.edu.co

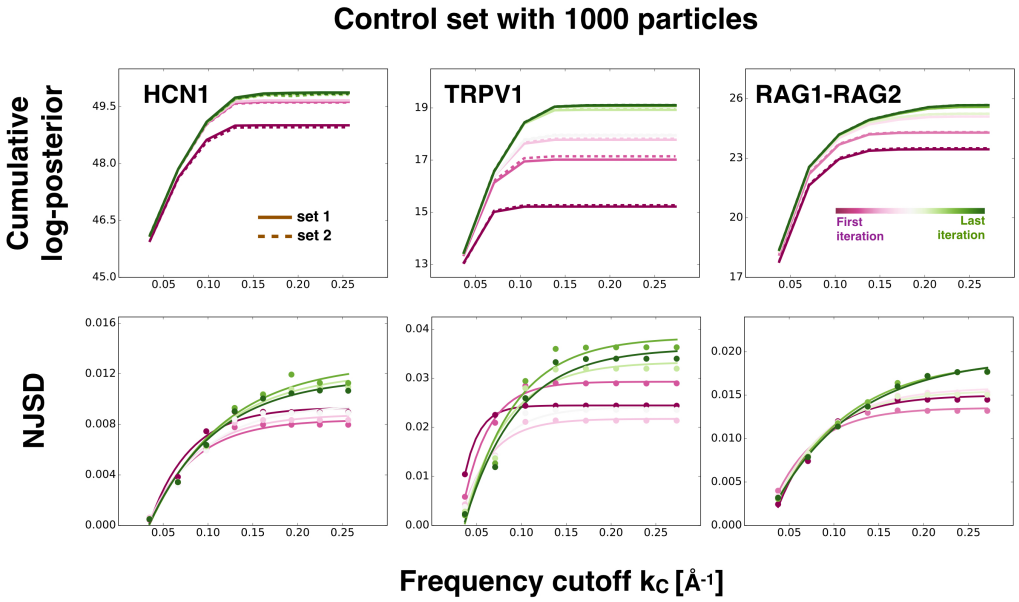
Supplementary Figures



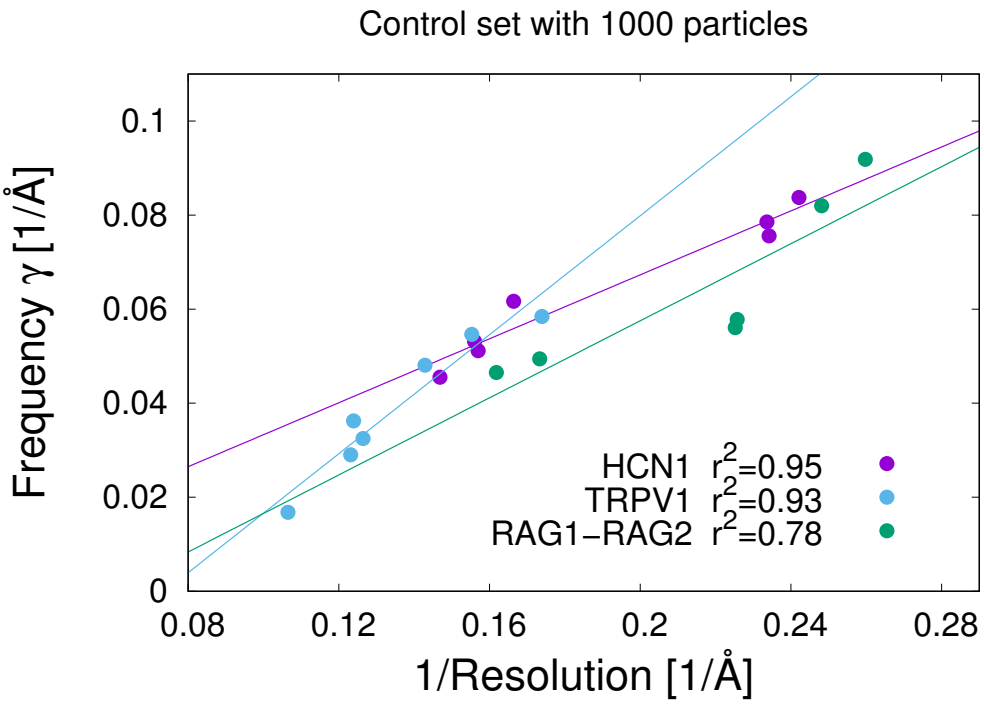
Supplementary Figure 1: *Differences in the log-posterior distributions.* (**top**) Examples of the distributions of the log-posterior relative to noise over the independent particle set. The distributions are calculated for the reconstructions from set 1 and set 2 at two cutoff frequencies $k_c = 0.05$ and 0.25 Å^{-1} for the fifth iteration of refinement of the HCN1 system. The vertical lines are the averages of the distributions. (**bottom**) Absolute value of the difference between the probability distributions from set 1 and set 2 for $k_c = 0.05$ and 0.25 Å^{-1}. The distributions calculated for the maps with higher frequencies are less similar.



Supplementary Figure 2: *Convergence of the observables.* (**top**) The cumulative log-posterior relative to noise $\sum_\omega \ln(P_{i\omega})/N_\omega - \ln(P_{\text{Noise}})$ for set $i = 1$ and 2 (solid and dashed lines, respectively), and (**bottom**) the normalized Jensen-Shannon divergence as a function of the number of particles in the control set. The results are shown for the TRPV1 system for iteration 12 and cutoff frequency $k_c = 0.21$ Å^{-1}. The observables converge if more than approximately 1000 particles are used.



Supplementary Figure 3: *Cumulative log-posterior and NJSD for a control set with 1000 particles.* (**top**) The cumulative log-posterior relative to noise and (**bottom**) the normalized Jensen-Shannon divergence as a function of the frequency cutoff. We use a gradient color code for the refinement iteration steps: the first iteration is maroon and the last iteration is green. The results are shown for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. The cumulative log-posterior is shown for the reconstructions from set 1 as solid lines and set 2 as dashed lines. NSJD data is fit to an inverse exponential function $-Ae^{-k_c/\gamma} + B$ (solid lines; bottom).



Supplementary Figure 4: *Frequency (γ) versus the inverse of the resolution for a control set with 1000 particles.* The results are shown for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. The correlation coefficients are $r^2 = 0.95$, 0.93, and 0.78, respectively. Solid lines show the linear fits.

Supplementary Text

BioEM input file examples

Round 1: Example of the BioEM input file for the TRPV1 system for round 1. The best orientations for each particle are obtained using the final map from the refinement. The following input file is for a subset of particles that have experimental defocus between 1.3 and 1.7 μm. The best 10 orientations for each particle are selected.

```
PIXEL_SIZE 1.22
NUMBER_PIXELS 256
USE_QUATERNIONS
CTF_DEFOCUS 1.3 1.7 10
CTF_B_ENV 0 10 2
CTF_AMPLITUDE 0.1 0.1 1
PRIOR_DEFOCUS_CENTER 1.5
SIGMA_PRIOR_DEFOCUS 0.8
SIGMA_PRIOR_B_CTF 1
DISPLACE_CENTER 30 1
WRITE_PROB_ANGLES 10
```

Round 2: Example of the BioEM input file for the TRPV1 system for round 2. The input file is for a single particle that has an experimental defocus of 1.9 μm.

```
PIXEL_SIZE 1.22
NUMBER_PIXELS 256
USE_QUATERNIONS
CTF_DEFOCUS 1.9 1.9 1
CTF_B_ENV 0 10 2
CTF_AMPLITUDE 0.1 0.1 1
PRIOR_DEFOCUS_CENTER 1.9
SIGMA_PRIOR_DEFOCUS 0.3
SIGMA_PRIOR_B_CTF 1
DISPLACE_CENTER 30 1
```

Pure-noise particles

We generated a set of 1000 synthetic pure-noise particles. Each particle has an image size of 180×180 and a pixel size of 1.23 Å. The particles contain random intensities following a Gaussian distribution with zero mean and unit variance. Because there is no experimental defocus, the BioEM probabilities are computed by performing round 1 with defocus range between 0.5 and 4.5 μm and using 4608 quaternions uniformly distributed in orientation space. This analysis was performed for each of the refined maps of the RAG1-RAG2 system.