STRUCTURAL BIOLOGY
ISSN 2059-7983

On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy

C. O. S. Sorzano,[a]* A. Jiménez-Moreno,[a] D. Maluenda,[a] M. Martínez,[a] E. Ramírez-Aportela,[a] J. Krieger,[a] R. Melero,[a] A. Cuervo,[a] J. Conesa,[a] J. Filipovic,[b] P. Conesa,[a] L. del Caño,[a] Y. C. Fonseca,[a] J. Jiménez-de la Morena,[a] P. Losana,[a] R. Sánchez-García,[a] D. Strelak,[a,b] E. Fernández-Giménez,[a] F. P. de Isidro-Gómez,[a] D. Herreros,[a] J. L. Vilas,[c] R. Marabini[d] and J. M. Carazo[a]*

[a]Biocomputing Unit, Centro Nacional de Biotecnologia (CNB-CSIC), Calle Darwin 3, 28049 Cantoblanco, Madrid, Spain, [b]Masaryk University, Brno, Czech Republic, [c]School of Engineering and Applied Science, Yale University, New Haven, CT 06520-829, USA, and [d]Escuela Politecnica Superior, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain. *Correspondence e-mail: coss@cnb.csic.es, carazo@cnb.csic.es

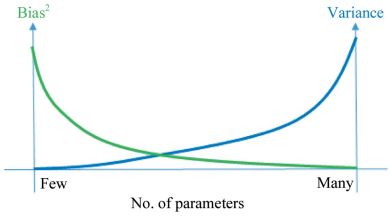
Received 9 August 2021
Accepted 18 February 2022

Edited by T. Burnley, Rutherford Appleton Laboratory, United Kingdom

Keywords: single-particle analysis; cryo-electron microscopy; parameter estimation; image processing; bias; variance; overfitting; gold standard.

Supporting information: this article has supporting information at journals.iucr.org/d

Cryo-electron microscopy (cryoEM) has become a well established technique to elucidate the 3D structures of biological macromolecules. Projection images from thousands of macromolecules that are assumed to be structurally identical are combined into a single 3D map representing the Coulomb potential of the macromolecule under study. This article discusses possible caveats along the image-processing path and how to avoid them to obtain a reliable 3D structure. Some of these problems are very well known in the community. These may be referred to as sample-related (such as specimen denaturation at interfaces or non-uniform projection geometry leading to underrepresented projection directions). The rest are related to the algorithms used. While some have been discussed in depth in the literature, such as the use of an incorrect initial volume, others have received much less attention. However, they are fundamental in any data-analysis approach. Chiefly among them, instabilities in estimating many of the key parameters that are required for a correct 3D reconstruction that occur all along the processing workflow are referred to, which may significantly affect the reliability of the whole process. In the field, the term overfitting has been coined to refer to some particular kinds of artifacts. It is argued that overfitting is a statistical bias in key parameter-estimation steps in the 3D reconstruction process, including intrinsic algorithmic bias. It is also shown that common tools (Fourier shell correlation) and strategies (gold standard) that are normally used to detect or prevent overfitting do not fully protect against it. Alternatively, it is proposed that detecting the bias that leads to overfitting is much easier when addressed at the level of parameter estimation, rather than detecting it once the particle images have been combined into a 3D map. Comparing the results from multiple algorithms (or at least, independent executions of the same algorithm) can detect parameter bias. These multiple executions could then be averaged to give a lower variance estimate of the underlying parameters.



1. Introduction

Single-particle analysis by cryoEM has become a popular technique to elucidate the 3D structure of biological macromolecules. Thousands of projection images from allegedly the same macromolecule are combined into a single density map that is compatible with the acquired measurements. The signal-to-noise ratio of each of the experimental images ranges from 0.1 to 0.01. This reconstruction process requires the estimation of hundreds of thousands of parameters [the

alignment parameters for each experimental image (Sorzano, Marabini *et al.*, 2014) and whether or not they belong to the structural class being reconstructed]. There are six parameters per image (three Euler angles, two in-plane shifts and one parameter for the class the particle belongs to). Additionally, from a mathematical perspective, the reconstructed volume itself is another set of parameters that must be determined (Scheres, 2012*a*). The existence of many iterative reconstruction algorithms attests to this (Sorzano, Vargas *et al.*, 2017), although, to a large extent, once the alignment parameters have been fixed there is very little freedom to choose the reconstructed volume. A different perspective is given by Sharon *et al.* (2020) and all of the previous work by the same group leading to this publication, in which the map is directly reconstructed from the experimental projections without the need to estimate the alignment parameters. This latter family of algorithms is still under development.

For a review of the single-particle analysis technique the reader is referred to Lyumkis (2019), and for a description of the image-processing pipeline the reader is referred to Sorzano, Jiménez-Moreno *et al.* (2021).

In this article, we focus on the structural bias; that is, the difference between our estimated structure, $\hat{V}(\mathbf{r})$, and the true underlying structure, $V(\mathbf{r})$. Obviously, we will never have access to the underlying true structure in a single experiment, if only because the measurement noise will cause some random fluctuation around it. We will model our observation as

$$\hat{V}(\mathbf{r}) = V(\mathbf{r}) + \Delta V(\mathbf{r}) + \varepsilon(\mathbf{r}), \tag{1}$$

where $\mathbf{r} \in \mathbb{R}^3$ is a spatial location in real space, $\Delta V(\mathbf{r})$ is the structural bias and $\varepsilon(\mathbf{r})$ is a random fluctuation with zero mean. The random noise, ε, normally decreases with the number of measurements [for instance, in Unser *et al.* (2005) we explicitly measured how the 3D reconstruction process attenuated white noise], suggesting that it does not pose a major problem in the current era of the automatic acquisition of thousands of micrographs. The problem is with the bias, ΔV, that systematically distorts our structure, preventing us from visualizing the true structure. This bias may be related to missing information, violations of the assumptions of the 3D reconstruction process, incorrect priors about the underlying structure, local minima in the search of the parameters of each image, incorrect use of the programs, software bugs or even the 3D reconstruction workflow itself.

In this article, we have opted for an organization of the work in which all the experiments have been moved to the supporting information. In this way, the main manuscript remains rather narrative and the user is not distracted from the main messages. In Section 2 we set up the statistical framework to analyze bias and variance during the estimation of parameters and to study how they affect the final reconstructed structure. In Sections 3 and 4 we discuss possible experimental and algorithmic sources of bias. In Section 5 we will analyze the currently used tools and recently proposed tools to detect bias. Finally, in Section 6 we draw some conclusions.

2. Bias and variance of parameter estimates

Volume overfitting is a feared feature of electron microscopy, and rightly so because it results in incorrect macromolecular structures (see Fig. 1 in Scheres & Chen, 2012). In the field, it is believed to come from excessive weight on the data, and it is thought to be tackled by providing a suitable weight on a Bayesian prior (Scheres, 2012*a*). Bayesian approaches are handy statistical tools if data are scarce. Interestingly, although the notion of overfitting is generally understood in the structural biology community, to the best of our knowledge there is not a formal, mathematical definition of it in the statistics domain. Overfitting occurs, for example, when the fitted function in a regression problem has too many parameters, so that the function can afford to follow the noise rather than just smoothly following the data trend. Overfitting would be the opposite of the 'principle of parsimony' in which a model should have the smallest number of parameters to represent the data adequately. Even this principle is not formally formulated. Instead, statisticians see overfitting as a trade-off between variance and bias of the parameter estimators (Burnham & Anderson, 1998, chapter 1). Let us assume that we have x and y observations that are related by a functional relationship plus observation noise,

$$y = f(x) + \varepsilon.$$

We will perform a regression with a function parametrized by a set of parameters, Θ, such that our prediction of y is

$$\hat{y} = f_\Theta(x).$$

Then, it can be proved that the mean-squared error (MSE) of our prediction is given by (Section 7.3 of Hastie *et al.*, 2001)

$$\begin{aligned}\mathrm{MSE}_\Theta(x) &= \mathbb{E}_\Theta\{[y - f_\Theta(x)]^2\} \\ &= [y - \mathbb{E}_\Theta\{f_\Theta(x)\}]^2 + \mathbb{E}_\Theta\{[\mathbb{E}_\Theta\{f_\Theta(x)\} - f_\Theta(x)]^2\} + \sigma_\varepsilon^2 \\ &= \mathrm{Bias}_\Theta^2\{f_\Theta(x)\} + \mathrm{Var}_\Theta\{f_\Theta(x)\} + \sigma_\varepsilon^2.\end{aligned}$$

The roles of y, f_Θ and x are played by different elements in each one of the problems addressed below, and this formulation should be taken as a generic framework for understanding some of the important properties of parameter estimation.

Probably the simplest model to illustrate this trade-off is regression by k-nearest neighbors (kNN). In this technique, the predicted value for a given x_0 is the average of the y values of the k nearest neighbors of x_0 (for simplicity of notation, let us illustrate the example for univariate predicted and predictor variables, but the same idea could be extended to multiple dimensions). For the kNN regression, the equation above particularizes to (Section 7.3 of Hastie *et al.*, 2001)

$$\mathrm{MSE}_\Theta(x) = \left[f(x_0) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)})\right]^2 + \frac{\sigma_\varepsilon^2}{k} + \sigma_\varepsilon^2.$$

In kNN regression, the complexity of the regression function (its number of parameters) is inversely proportional to k. That is, a very large k results in very few different predictions (in the limit, eventually all predictions are equal and equal to the input sample mean), with a consequent very low variance of

the predictions [$\mathrm{Var}_{\Theta}\{f_{\Theta}(x)\}$], but a huge bias with respect to the true underlying value $f(x_0)$, while a small k results in many more output possibilities and therefore better adaptation to the local specificities around x_0 (and consequently low bias), but in a much larger variance of the predictions because the second term is divided by a small k.

This example with kNN regression illustrates a much more general principle: as the number of model parameters grows, the bias of the estimated parameters decreases and the variance increases (see Fig. 1). This is known as the bias–variance trade-off (Section 7.3 of Hastie *et al.*, 2001). Models with a low number of parameters cannot explain part of the experimental data. In contrast, models with a large number of parameters do explain the data. Still, they have an unnecessarily large variance with respect to a more parsimonious model that explains the data almost equally well. Model-selection methods such as the Akaike's or Bayesian information criterion (AIC or BIC) try to achieve the minimum of this trade-off between bias and variance (Burnham & Anderson, 1998). To illustrate this idea, let us analyze the formula of the Bayesian information criterion

$$\mathrm{BIC} = 2\log P\{\mathbf{y}|\hat{\boldsymbol{\theta}}\} - k\log N,$$

where $\mathbf{y}$ are the observed data, $\hat{\boldsymbol{\theta}}$ is the set of model parameters, $P\{\mathbf{y}|\hat{\boldsymbol{\theta}}\}$ is the likelihood of observing the data given these parameters, k is the number of parameters of the model and N is the number of observations. The goal is to choose the model that maximizes the BIC. The first term is a data-fidelity term, while the second term is a penalization for the number of parameters in the model.

Although intuitively appealing, having too many parameters is not the explanation for the overfitting observed in EM. Let us consider a set of 100 000 particles of size 200 × 200 pixels. We need to determine a volume of size 200 × 200 × 200 (= 8 000 000 parameters) and 600 000 alignment parameters (five alignment parameters per particle and one additional parameter to decide whether or not the particle belongs to the class that we are reconstructing). This makes a total of 8 600 000 parameters, but there are 4 000 000 000 measurements (pixels). The exact account is not so simple because the quality and the solvability of the map are also related to the angular coverage (Sorzano *et al.*, 2001; Naydenova & Russo, 2017; Sorzano, Vargas, Otón, Abishami *et al.*, 2017; Tan *et al.*, 2017), which mathematically is also related to the null space of the matrix associated with the set of measurements (Sorzano, Vargas, Otón, de la Rosa Trevín *et al.*, 2017), but it brings in the idea that the number of measurements largely surpasses the number of unknowns; correlations among pixels are not considered either. Consequently, the reconstruction artifacts observed in the 3D reconstructions do not come from the freedom of the volume to fit the noise (variance), but from a mismatch between the model and the data (data not originating from this model), incorrect estimates of the particle parameters, or goal functions or algorithms that produce biased results. This statement is experimentally supported in this article by multiple experiments addressing different steps along the image-processing pipeline. The estimation of any parameter can easily fall into local minima, calling for robust image-processing algorithms that are capable of reliably estimating all of these parameters in such a noisy environment. In this article, we argue that there are several sources of bias. Some of them are related to the sample while others are related to the algorithm. Among them, the most important at present is the incorrect estimation of the particle parameters.

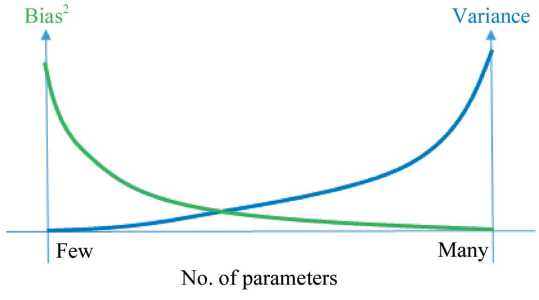


Figure 1
Conceptual trade-off between bias and variance of the parameter estimators depending on the number of parameters.

For example, for the images supposed to correspond to a macromolecule of interest, we must determine whether they all come from a single, homogeneous population of structurally identical particles or exhibit some kind of structural variability. This task is performed by classification of the particles into supposedly homogeneous 3D classes (the cryoEM formulation of this problem is very close to the formulation of mixture models in machine-learning clustering; McLachlan & Basford, 1988). Misclassifying particles results in 3D reconstructions from a mixture of structurally different objects. For instance, let us imagine that we are trying to obtain the structure of class 1 from N_1 images of that structural class, and we have a mixture with N_2 images of class 2. We will have an estimate of the underlying structure that, in a very simple approximation that assumes linearity of the 3D reconstruction process and the same weight for all images, is

$$\begin{aligned}\hat{V}_1(\mathbf{r}) &= \frac{N_1}{N_1+N_2}V_1(\mathbf{r}) + \frac{N_2}{N_1+N_2}V_2(\mathbf{r}) \\ &= V_1(\mathbf{r}) + \frac{N_2}{N_1+N_2}[V_2(\mathbf{r}) - V_1(\mathbf{r})]. \end{aligned} \quad (2)$$

How large the bias that we obtain in our estimate of V_1 is depends on the amount of contamination from class 2, *i.e.* N_2, and the true structural difference between V_2 and V_1. If the 3D reconstruction process is nonlinear (as it is) or the images receive different weights (as they do), then the formula above is not verified in its stated form. However, its simplified version already points out two interesting features that hold in more complicated scenarios: (i) the 3D reconstruction is a mixture of two different structures and (ii) how large our bias in the estimation of V_1 is depends on our number of mistakes, N_2, and the difference between V_2 and V_1. In addition to the difficulties described above, we encounter the additional drawback that the true number of classes is unknown in an experimental setting.

Within a homogeneous population of structurally identical particles, we may encounter the same bias problems if we incorrectly estimate the angular orientation and in-plane shifts of the experimental images or their acquisition parameters (defoci, beam tilt *etc.*). N_1 would now play the role of the particles with correct parameters, N_2 the role of the particles with incorrect parameters and V_2 a 3D reconstruction with the incorrect parameters and their corresponding particles.

One of the major sources of systematic errors when estimating parameters is produced by what we refer to as the 'attraction problem', which was rigorously proved in equation (6) in Sorzano *et al.* (2010). Let us summarize here the main argument. Many algorithms in the field eventually require a comparison between the experimental images and a set of reference images. For example, this comparison is required when assigning an input image to a 2D class, a projection direction or a 3D class. We need to compare the input experimental image with the representative of the 2D class or the reprojection of the current estimate of the map or the 3D class along that direction. The two main tools to perform these comparisons are the Euclidean distance (in real or Fourier space) and the correlation (actually, using relatively mild hypotheses, it can be proved that the reference that maximizes the correlation is the same as the one that minimizes the Euclidean distance). As soon as one of the references starts to get more images, its background will be less noisy because of the higher averaging caused by the larger number of images. However, in the correlation or Euclidean distance calculation the background also contributes, and that with lower noise will contribute less to the Euclidean distance. In this way, it will seem closer to some experimental images, even if these do not correspond to the signal represented by the reference. Consequently, more images are assigned to that reference, and the larger averaging effect is positively reinforced. Euclidean distance and correlation can be considered 'classical' image-similarity tools. With the advent of deep learning, a new approach that has not yet been adopted in cryoEM is the learning of the distance function itself (Wang *et al.*, 2014). This possibility could open new research routes towards more robust image classification and alignment.

Overall, in current practice, we would say that incorrect parameter estimation is the major source of bias in cryoEM. Throughout this article, we discuss several strategies to identify and try to prevent it. One of the most powerful strategies is to analyze the parameter estimates using several independent algorithms, or at least multiple runs of the same algorithm with random initialization. This is in line with the current trend in machine learning of using ensemble approaches such as boosting (an ensemble of many high-bias models that decreases the overall model bias and variance) and bagging (subsampling data with replacement decreasing the overall model variance) (Hastie *et al.*, 2001).

3. Experimental sources of bias

Several sources of bias are primarily related to the sample itself rather than the method of processing the data set, although everything is inter-related. In the following we discuss some algorithmic solutions that we can adopt to prevent structural bias due to these problems.

3.1. Use of incorrect particles

The 3D reconstruction algorithm will assume that a single structure is perfectly compatible with the measurements except for random, zero-mean noise that is superposed on the projection images. This assumption is violated by protein denaturation, conformational heterogeneity and the presence of contaminants, aggregations, disassembled particles, radiation damage, particle superposition within the ice layer *etc.* In a way, deciding whether a patch of a micrograph is a particle or not is the first parameter that we must determine.

All particle pickers have false positives (objects incorrectly identified as particles) and false negatives (missed particles). At present, the trend is to set the picking parameters so that 'all' particles are selected. The idea is to maximally exploit the structural information present in the micrographs. However, this puts pressure on the 2D and 3D image processing to identify particles that do not really correspond to isolated, single particles of the macromolecule under study. Despite the solid image-processing and artificial intelligence background of the most widely used pickers nowadays (Abrishami *et al.*, 2013; Bepler *et al.*, 2019; Wagner *et al.*, 2019), all of them have a false-positive rate that ranges between 10% and 30% depending on the data set (Sanchez-Garcia *et al.*, 2020). Algorithms such as *Deep Consensus* (Sanchez-Garcia *et al.*, 2018) were specifically designed to take all of these candidates to particle centers and apply a deep-learning algorithm to learn their commonalities and decide which of the coordinates really correspond to a particle and which are false positives. There are also algorithms that try to remove the coordinates of protein aggregations, carbon edges, contaminants or any other sample defect (Sanchez-Garcia *et al.*, 2020). In our experience, the combination of *crYOLO* (Wagner *et al.*, 2019), *Xmipp* picking (Abrishami *et al.*, 2013), *Deep Consensus* (Sanchez-Garcia *et al.*, 2018) and *Micrograph Cleaner* (Sanchez-Garcia *et al.*, 2020) produces very few false-positive particles. Any denoising algorithm such as that described by Bepler *et al.* (2020) can also help to produce cleaner micrographs in which particle finding is simplified and, consequently, presumably more accurate. A dangerous practice is to select particles using a reference external to the study, because it may lead to biased reconstruction, as in the famous case of the HIV trimer (Mao, Wang *et al.*, 2013; Mao, Castillo-Menendez *et al.*, 2013; Henderson, 2013; Subramaniam, 2013; van Heel, 2013).

Obviously, an important consequence of not using the correct particles for the macromolecule being reconstructed is that the presence of incorrect particles will contaminate the 3D reconstruction, as described in equation (2).

As we mentioned above, the 3D reconstruction process assumes that the projection of the reconstructed map matches the experimental projection except for the noise. This assumption is violated in a series of cases.

(i) If our particles are not isolated (they can be nearby or superposing; Noble *et al.*, 2018), they do not correspond to the structure under study (but instead to contaminants, images of ice, aggregations, particles on carbon *etc.*) or they correspond to the structure being reconstructed plus some attached flexible matter (antibodies, surrounding membrane, factors that may or may not be bound *etc.*). In these circumstances, the 3D alignment algorithm will try to satisfy the reconstructed particle and its surrounding matter. As shown in Supplementary Experiment 1, the current practice of taking as many particles as possible, disregarding their quality and hoping that the algorithm will manage is very likely to be counterproductive. The reader may note that the induced artifacts are not constrained to the area outside the macromolecule. Inside the macromolecule there are also important structural differences caused by the nearby entities.

The solution to this problem would consist of powerful particle picking, as described above, and 2D image analysis: particle screening (Vargas *et al.*, 2013), 2D classification (Sorzano *et al.*, 2010; Scheres, 2012*b*; Punjani, Rubinstein *et al.*, 2017) and outlier analysis of the classes (Sorzano, Vargas *et al.*, 2014). Choosing those particles from classes in which particles are isolated should be preferred, and if this is not feasible due to the high concentration of particles, then making the 3D alignment with a tight mask (used for alignment, not reconstruction) could help, but in general this is not a solution. Note that the tight mask can be used in two places: (i) to construct the reference mask to apply to the input volume so that it removes the information around the particles and (ii) to mask the reconstructed map so that we can 'hide' the artifacts outside the particle, but not the structural modifications inside the mask. We do not object to its first use, as removing artifacts from the reference volume prevents artifactual features from acting as noise anchors. However, we discourage its second use as we may not see possible biases whose effect is more easily detected outside the macromolecule.

As a final warning, we should be aware that choosing particles only from 'good-looking' 2D classes does not guarantee good particles due to the attraction problem in 2D classes (Sorzano *et al.*, 2010) (depending on the algortihm used, for example *RELION* 2D or *cryoSPARC* 2D) or the scattering of bad particles into the existing classes (Sorzano, Vargas *et al.*, 2014).

(ii) If particles are misclassified during the 3D classification (either because the classification is performed attending to the angular orientation of the particles, several populations are mixed or because differences in the 3D classes are found by incorrect angular assignments; Sorzano *et al.*, 2020). Supplementary Experiment 2 shows examples of the instability of the 3D classification process and 3D attraction problems.

A possible strategy to avoid misclassification could be to repeat the 3D classification process several times and with different algorithms (Scheres *et al.*, 2007; Scheres, 2012*a*; Lyumkis *et al.*, 2013; Punjani, Brubaker *et al.*, 2017), keeping only those images that are consistently classified together. The reason is that performing this classification is extremely challenging for currently existing algorithms, and we cannot just take 'the first classification result', as it will most likely contain important mixtures of different subpopulations.

3.2. Use of incorrect symmetry

If our structure is pseudosymmetric and we reconstruct it as symmetric, we will lose the small differences between subunits. If our structure has some symmetry parameters, such as a helix, and we use different parameters, we will strongly distort our structure. These symmetry-related biases can occur in standard single-particle studies (Ludtke *et al.*, 2004), electron crystallography (Gil *et al.*, 2006; Biyani *et al.*, 2018), helical reconstructions (Egelman, 2014) and studies of icosahedral viruses (Koning *et al.*, 2016). Additionally, macromolecules are intrinsically flexible objects that could be fluctuating around an energetically stable solution (Sorzano *et al.*, 2019), and these fluctuations automatically break all symmetries at high resolution.

Pseudosymmetry is currently one of the most active lines of research. This is useful for analyzing macromolecules with almost equivalent subunits and for analyzing the asymmetric part of particles in which a part is symmetric or has a different symmetry (for instance, nucleic acids inside an icosahedral virus capsid, virus portals *etc.*). One of the solutions is to perform symmetric and asymmetric reconstructions to verify the consistency between the two structures. However, this option is not always feasible due to the 3D attraction problem. Alternatives are symmetry expansion (Scheres, 2016) or symmetry relaxation (Huiskonen, 2018), in which the method tries to separate the particles into structurally homogeneous groups. Another solution would be to analyze the set of images around its symmetric conformation using continuous heterogeneity tools (Dashti *et al.*, 2014; Jin *et al.*, 2014; Haselbach *et al.*, 2018) and focusing on groups of particles according to their deformation parameters (Jin *et al.*, 2014).

3.3. Missing information

If we lack information from some projection directions, this may cause, depending on which directions are missing, empty regions in the Fourier domain for which we simply do not know what the protein looks like. Filling this region with zeroes is usually a bad choice as it results in an elongation of the structure along the missing direction. The absence of measurements in some regions of the Fourier domain is well known in the field because it occurs in some data-collection schemes such as random conical tilt (Radermacher & Hoppe, 1980; Radermacher *et al.*, 1987; Radermacher, 1988; Sorzano, Alcorlo *et al.*, 2015) and orthogonal tilt (Leschziner & Nogales, 2006). An ideal single-particle analysis should not suffer from this problem as in principle particles can be acquired from any possible orientation. An alternative to filling the Fourier space with zeroes is to provide information that guarantees some kind of continuity (Moebel & Kervrann, 2020). However, many projects in SPA face the problem of uneven angular distributions, potentially causing severe artifacts in the 3D reconstructions. Some of these uneven angular distributions are truly caused by a preferential interaction of the

macromolecule with the water–air interface or the sample support (Tan *et al.*, 2017; Noble *et al.*, 2018). In these cases, the lack of experimental data could be complemented with *a priori* volumetric constraints (external surface, total mass, non-negativity *etc.*; Sorzano *et al.*, 2008). This is not an easy task as it involves iterative reconstruction algorithms, which are now in disuse because they are much slower than their Fourier gridding counterparts. At present, it is preferred to tilt the sample (Tan *et al.*, 2017) or to look for different sample-preparation conditions that do not cause preferential views.

4. Algorithmic sources of bias

Structural biologists are very much aware of the problems referred to above and try their best to overcome them. However, algorithmic reasons may also prevent us from achieving an unbiased estimate of the structure under study. Some of them are very well known, such as the dependence of the final structure on the initial guess or the existence of software bugs. Some others are suspected, such as the existence of local minima in the parameters to estimate. Yet others are buried deep in the 3D reconstruction and classification process and are seldom exposed, but are critical.

In this section, we discuss sources of bias that are more related to the image processing itself. We focus on problems that presently remain a bottleneck or that have received less attention from the community. The initial volume problem is of primary importance. As such, it has received all kinds of attention, from descriptions of the problem to algorithmic proposals to tackle it. Although it can cause really poor results if it is not properly selected, our view is that it is now no longer a major bottleneck in most projects as one of the many existing algorithms will be able to find a suitable initial volume. However, our view is that at present the use of incorrect parameters for the particles is the greatest source of structural bias (along with the population mixture that is still observed after 3D classification). The 3D attraction problem causes a major algorithmic problem in some experiments in which the angular assignment is totally biased. Incorrect masking can be a source of structural bias if it truncates part of the structure or leaves extra masses that do not correspond to the macromolecule under study. Still, otherwise, it is not a large challenge except in that it may give us a false sense of good quality by inflating the Fourier shell correlation (FSC). Finally, the image metric and 3D reconstruction algorithm are sources of bias that have never been put forward. Although they do not represent a major problem, it is worth enumerating them in this article and making users aware that the choice of the 3D reconstruction algorithm also introduces its own contribution to the reconstructed structure that might be confounded with true structural features of the macromolecule being reconstructed.

We may identify the following sources of bias induced during the image-processing procedure.

4.1. Initial volume

The dependence of the final structure on the initial volume used to be one of the most severe problems some years ago (Henderson, 2013; Subramaniam, 2013; van Heel, 2013). The 3D alignment and reconstruction process is normally some variant of gradient descent. For this reason, the starting point of the iterations plays a crucial role in the optimization process. This is the reason behind the well known Einstein from noise effect (Shatsky *et al.*, 2009). Several solutions have been proposed in recent years to tackle this problem, such as stochastic optimization algorithms (Ogura & Sato, 2006; Elmlund *et al.*, 2013; Vargas *et al.*, 2014; Punjani, Brubaker *et al.*, 2017), slowly converging algorithms (Scheres, 2012*a*; Sorzano, Vargas *et al.*, 2015) and consensus algorithms (Sorzano, Vargas *et al.*, 2018; Gómez-Blanco *et al.*, 2019). Thanks to all of these new algorithms, the initial volume dependence is no longer a major bottleneck in the image-processing pipeline as long as these algorithms are judiciously used. There are also ways to validate the initial volume through external measurements such as SAXS data (Jiménez *et al.*, 2019). In any case, it should be noted that a bad choice of the initial volume very often leads to erroneous results.

4.2. Incorrect alignment parameters

One of the most important sources of bias is an inaccurate estimation of the alignment parameters. Stewart & Grigorieff (2004) reported important differences in the image alignment depending on the goal function that is being optimized. Consequently, differences in the angular assignment between different programs should be expected. We can think of two different kinds of errors: (1) the alignment parameters found are a small, randomly perturbed version of the true (although unknown) alignment and (2) the alignment parameters found are in a region of the projection sphere or in-plane alignment totally unrelated to the true alignment. In any case, both kinds of mistakes result in an error in the particle orientation, with some errors larger than others depending on whether that specific particle is in case (1) or (2) (see equation 2).

In a previous section, we discussed projects with incomplete angular coverage due to experimental reasons. It is less well known that the angular assignment algorithm itself can also cause this bias in the angular assignment through the previously mentioned 3D angular attraction (Sorzano, Semchonok *et al.*, 2021). As shown in Supplementary Experiments 2 and 3, the 3D attraction effect is a problem that severely affects the validity of the reconstructed map. This behavior has a doubly deleterious effect: firstly, it places experimental projections in incorrect directions, causing structural bias by a mixture of signals and, secondly, it may deplete low-populated directions in favor of nearby directions, causing structural bias by lack of information.

Supplementary Experiments 4 and 5 show that, depending on the data set, uncertainty about the 3D orientation and in-plane alignment of experimental images affects 10–50% of the data set (note that these numbers are based on our experiments and different data sets may yield different limits). Two different algorithms may disagree in the angular assignment of up to 50% of the images (Supplementary Experiment 4). This disagreement may also be found in multiple runs of the same

algorithm (Supplementary Experiment 5). This indicates the variability of the alignment parameters, but also that for any particular execution a fraction of the parameters are significantly biased. These inconsistent parameters can be identified if the outputs of several program runs are compared, but this is seldom performed. The extent of the effect in a particular study is impossible to determine if only one 3D classification or angular assignment is performed for a particular set of images. Its detection necessarily requires multiple independent estimations of the underlying parameters (class membership and/or angular assignment), preferably using algorithms based on different mathematical principles. After comparing their different outputs, one may identify those particles for which the estimates agree. Unfortunately, for those for which they disagree it is difficult at the moment to decide which are the true parameters. Some algorithms are more prone to 3D attraction. Those related to a Euclidean distance between two images (such as *RELION*, *cryoSPARC* and *cisTEM*) are more susceptible to suffering it [see equation (6) of Sorzano *et al.* (2010) for the mathematical explanation]. *Xmipp HighRes* uses a weighting scheme based on the significance of two score functions in its global alignment stage, which might be the reason for its higher immunity to this problem. It should be noted that an absolute consensus algorithm that only keeps the images for which all alignment algorithms agree on their angular assignment would not solve the 3D attraction problem, as the 'attracted' algorithm would prevent the rest from filling the depleted regions. More creative strategies, probably involving three or more independent assignments, should be devised in this case, and this problem is foreseen to be an active research topic in the future.

Another way to identify misaligned particles is through the use of multiple objective functions. Most algorithms optimize a single objective function (log likelihood in the case of *RELION* autorefine or cross-correlation in a maximal circle in *Xmipp HighRes* local alignment). From the point of view of a single numerical observer, it is normally not possible to recognize the presence of misaligned particles. However, the calculation of several similarity measures may help recognize the set of misaligned particles or nonparticles still in the data set. Supplementary Experiment 6 shows how the calculation of the *Xmipp HighRes* local alignment similarity measure can identify two subpopulations where *RELION* autorefine cannot. In general, each different similarity measure 'sees' different features of the same alignment. Using tools such as the different similarity measures shown above or the alignability of the particles shown in Vargas *et al.* (2016, 2017) and Méndez *et al.* (2021), we should also be able to identify those particles for which the angular assignment is in doubt. We have also found it very useful to perform a 3D classification of the particles in two classes without re-estimating the angles. Particles with an incorrect alignment tend to cluster in one of the classes, while the other class retains the particles with good alignment (Sorzano *et al.*, 2020).

A similar situation of alignment bias occurs if the handedness of the images is mixed, as reported in Sanz-García *et al.* (2010) (see Supplementary Experiment 7). Once the angular assignment falls into this situation, it is challenging to disentangle the hand mixture. A possible way is by constructing an initial volume from the particles assigned to a 3D class and verifying that the reconstructed structure resembles it.

4.3. Incorrect CTF correction

Another source of bias may come from inaccuracy in the estimation of the CTF parameters. In its most simplified version, the CTF formula is $\sin(\pi\lambda\Delta f R^2 + \dots)$, where λ is the electron wavelength, Δf is the defocus and R is the frequency at which we evaluate the CTF (Sorzano *et al.*, 2007). The two most important parameters of the CTF are the microscope voltage (which determines the electron wavelength) and the micrograph defoci (Sorzano *et al.*, 2009). Zhang & Zhou (2011) stated that the maximum defocus error to achieve high resolution should be below 100 Å. Larger errors would result in incorrect compensation of the phase shift introduced by the microscope. In the CTF challenge, the discrepancy between different CTF estimation software programs was around 200 and 300 Å (Marabini *et al.*, 2015). As with any other parameter, random fluctuations around the true value must be expected, and these will naturally limit the maximum achievable resolution. However, if these estimation errors are not random (as assumed, for instance, in Penczek *et al.*, 2014) but systematic, we may consistently overcompensate or undercompensate some frequencies (this effect is significant at medium frequencies; at high frequencies the CTF oscillates more rapidly and it is more difficult to make systematic errors). Systematic errors in the CTF normally translate into a dark halo around the macromolecule, as seen in many EMDB entries (see, for instance, EMDB entries EMD-20310 and EMD-20702 as examples of recent releases from October 2019), and a haze on top of the macromolecule. In Supplementary Experiment 8 we show an example in which the dark halo around the particle and the haze on top of it are generated by systematic errors in the estimation of the CTF defocus. Note that in the CTF formula the pixel size participates through the frequency term [$R = i/(NT_s)$, where i is the index of the frequency term in the fast Fourier transform of the input image, N is the image size, which is assumed to be square for simplicity, and T_s is the pixel size or sampling rate]. Consequently, a small error in the pixel size also systematically causes miscorrections of the phase flip. In Supplementary Experiment 8 we also show how systematic errors in the pixel size also translate into dark halos. At present, it is customary to refine the CTF parameters per particle locally (Zhang, 2016; Bartesaghi *et al.*, 2018; Sorzano, Vargas *et al.*, 2018; Zivanov *et al.*, 2018). Although these optimizations are not expected to be particularly biased, the amount of signal available to estimate the CTF parameters per particle is so small that large variances should be anticipated. To the best of our knowledge, there has not been any rigorous work that has tried to estimate the variability of the per-particle parameters. In real practice, dark halos around the reconstructed maps are very often observed. These are probably caused by a mixture of random and systematic errors in the pixel size (which should be small

and can be corrected with a recalibration using an atomic model of the structure) and random and systematic errors in the defocus estimates (which can be minimized by averaging the defocus values reported by several CTF estimation algorithms). The problem is not the halos and hazes themselves, which the isosurface visualization programs can easily ignore. The problem is that we know that the presence of the halo and haze implies the existence of fine structure differences inside the macromolecule, as shown in Supplementary Experiment 8. Note that we cannot show evidence that there are systematic errors in determining the defocus in published experiments. However, we can reproduce the same kind of errors as those in published experiments by forcing a systematic error in the defocus determination.

4.4. Image normalization

The 3D reconstruction process assumes that the acquired images are projections of the macromolecule under study in different poses. The weak phase object approximation gives the relationship between the projection image and the volume to be reconstructed (Koeck & Karshikoff, 2015). In this approximation, a transmitted beam gives rise to a baseline rate of electron arrivals modulated by the matter along their path (the weak phase object approximation states that the modulation is linear). However, this model implies that the raw images acquired by the microscope must be normalized before entering the 3D reconstruction process (for example the ice thickness is not the same for all particles). The normalization normally sets the statistical properties of the ice to some prespecified values (Sorzano *et al.*, 2004). However, this normalization is affected by outlying pixels, nearby particles, contaminations or carbon edges, illumination gradients, inhomogeneous camera gain images (Sorzano, Fernández-Giménez *et al.*, 2018) *etc*. For this reason, the particle normalization must be refined in order to make the projection images maximally consistent with the reconstructed volume (Scheres *et al.*, 2009; Sorzano, Vargas *et al.*, 2018). We may think of the normalization process as a linear transformation of the input images $I' = aI + b$. Systematic errors in b translate into a sphere of density around the reconstructed molecule (the reason is that the backprojection of an additive constant in all possible projection directions is not a constant map, but a sphere whose density increases with the radius up to the maximum radius that can be embedded in a box of the size of the particles). This kind of systematic error is seldom seen in 3D reconstructions of single particles. Instead, random errors in b should be more common. Similarly, it is hard to think up situations in which the image-normalization process systematically biases a. However, images participate in the 3D reconstruction process with some weight (Grigorieff, 2007; Scheres *et al.*, 2007; Sorzano, Vargas *et al.*, 2018), and one can imagine systematically high or low weights depending on the projection direction (for instance, the attraction problem in 3D places more images along specific directions, resulting in a higher weight of that direction with respect to the rest). This situation could not be distinguished from systematic errors in a.

4.5. Incorrect masking in real or Fourier space

Incorrect masking in real space or Fourier space can either cut out valid regions of the map or, on the contrary, leave regions that do not correspond to the structure of interest but may serve as anchors for noise alignment (a related problem can be seen for membrane proteins, where the density of the membrane may drive the angular alignment in undesired ways).

The use of masks during alignment is recommended. They prevent the alignment from being driven by artifacts around the 3D reconstruction that are unrelated to the structure under study (Sorzano, Vargas *et al.*, 2018; see also Supplementary Fig. 2). The same logic applies to masks in the Fourier domain: the FSC can serve as an indicator of the reliability of the different Fourier components. This reliability can be explicitly used during the alignment phase to limit the amount of unreliable content that can serve as noise anchors (Scheres, 2012*a*; Grant *et al.*, 2018; Sorzano, Vargas *et al.*, 2018).

We should distinguish between real-space and Fourier space masks for angular alignment or as post-processing tools. The use of real-space masks after reconstruction should be discouraged because they could hide possible biases. Similarly, modifications of the amplitude spectrum after reconstruction, such as the B-factor correction, normally lead to a biased overboosting of the high-frequency components (Ramírez-Aportela *et al.*, 2020), resulting in publicly deposited maps that do not comply with the expected behavior of the diffraction of macromolecules (Vilas, Vargas *et al.*, 2020). Other modifications that try to match the amplitude spectrum of the reconstructed map to that of its atomic model (Jakobi *et al.*, 2017) explicitly address the minimization of this bias as long as the atomic model is correct (otherwise, this match would induce another bias). Interestingly, current post-processing approaches such as that in *RELION* basically amount to a mask and B-factor correction. After this transformation, the FSC significantly improves, reporting a higher resolution for the reconstructed map (see Supplementary Experiment 9). However, this increase in resolution is merely due to the change of the mask between that used during reconstruction and that used during post-processing because the FSC is invariant to radially symmetric transformations such as the B-factor correction (Sorzano, Vargas, Otón, Abrisham *et al.*, 2017). Masking in real space translates into a convolution in Fourier space [we denote the Fourier transforms of the estimated map and the applied mask $\hat{\mathcal{V}}(\mathbf{R})$ and $\mathcal{M}(\mathbf{R})$, respectively],

$$\begin{aligned}\mathcal{FT}\{\hat{V}(\mathbf{r})M(\mathbf{r})\} &= \hat{\mathcal{V}}(\mathbf{R}) \star \mathcal{M}(\mathbf{R}) \\ &= \hat{\mathcal{V}}(\mathbf{R}) + [\hat{\mathcal{V}}(\mathbf{R}) \star \mathcal{M}(\mathbf{R}) - \hat{\mathcal{V}}(\mathbf{R})];\end{aligned}$$

that is, we are biasing our reconstructed volume by another volume whose Fourier transform is $[\hat{\mathcal{V}}(\mathbf{R}) \star \mathcal{M}(\mathbf{R}) - \hat{\mathcal{V}}(\mathbf{R})]$. The absence of a mask is equivalent to a constant mask of value 1 everywhere. Its Fourier transform would be a delta

function in Fourier space, and the bias term would be equal to zero. However, tight masks are significantly broad in the Fourier domain, resulting in a large bias, and as shown in Section 5.1 this can make the FSC arbitrarily large, as shown in Supplementary Experiment 9. It should also be noted that a bias with respect to the reconstructed map is not necessarily bad, as the bias should be measured with respect to the true underlying structure, not the reconstructed map. In this regard, the masked volume may be closer to the underlying structure than the reconstructed map if the mask removes map artifacts. Unfortunately, the true structure is never known, and introduction of the mask and its effect on the FSC may result in overconfidence in the quality of the map.

4.6. The 3D reconstruction algorithm

As we have seen, some metrics may be better suited than others to identify population mixtures or misaligned particles. These metrics are translated into different weights of the particles in the 3D reconstruction. In turn, this projection-weighting scheme plays an important role in the final reconstruction. For instance, incorrectly aligned particles would not have any effect if their weight is minimal. On the other hand, the possibility of assigning multiple weights to the same image in different orientations will necessarily introduce structural bias in the 3D reconstruction, especially if these weights are similar.

Strictly speaking, the voxel values of the reconstructed map are parameters to determine (8 000 000 in our example in Section 2). Still, once the alignment parameters have correctly been determined (600 000 in the example in Section 2), an unbiased determination of the volume parameters is almost guaranteed with the existing algorithms (Penczek *et al.*, 2004; Scheres, 2012*b*; Abrishami *et al.*, 2015; Sorzano, Vargas, Otón, de la Rosa-Trevín *et al.*, 2017). In this way, most of the effort should be concentrated on determination of the alignment parameters. In the EM community, these parameters have been considered to be nuisance (secondary) parameters (Scheres, 2012*a*; Lyumkis *et al.*, 2013; Punjani, Brubaker *et al.*, 2017) arising from some statistical distribution with a Gaussian or uniform prior. This assumption has proved to be very useful in converging from a wide range of initial volumes, as shown by the success of these methods. However, a consequence of adopting a maximum-likelihood approach is that an image is allowed to occupy multiple orientations with different probabilities. This is a violation of the image-formation model, as an image truly arises from only one, although unknown, orientation. Projection matching does not suffer from this drawback (although it has others, such as a much smaller radius of convergence). An image can have a single set of alignment parameters. As argued in Sorzano, Vargas *et al.* (2018), the probability of making an angular error if a single projection direction is allowed is lower than that of making an angular error if two or more projection directions are allowed (because all except at most one must necessarily be wrong). In Supplementary Experiment 6 we show the distribution of the number of significantly different alignment parameters contributing to each of the experimental images in *RELION*. As can be seen, most of the images have between 1 and 10 significant contributions, with a maximum of about 200. This multiplicity of orientations is translated into a weighting scheme that places the same image at different orientations with different weights. In Supplementary Experiment 10, we show the difference between the 3D reconstruction performed within *RELION* autorefine and *RELION* reconstruction with the same angular distribution. The weighting scheme in *RELION* autorefine results in a low-pass filter of the reconstructed map and a low-frequency white halo superposed on the map. These differences with respect to the true underlying structure are a different kind of bias, in this case caused by the weighting scheme of the reconstruction algorithm.

Maximum-likelihood methods were extended to Bayesian methods by adding a prior on the Fourier coefficients of the reconstructed map (Fourier components that are independent of each other and whose real and imaginary parts are also independent). This prior acts as a low-pass filter (Scheres, 2012*a*) and, as for any prior, it results in a regularization that unavoidably leads to bias (Fessler, 1996; this is the very purpose of Bayesian methods when data are scarce). Additionally, the specific prior used in the community so far has experimentally been shown to be incorrect for macromolecular structures (Sorzano, Vargas, Otón *et al.*, 2015), and consequently as an incorrect prior it systematically biases the reconstructions obtained with this objective function. Still, this prior has proved to be very useful for its EM application, although in the future research on new priors based on the nature of macromolecules could be exploited. These priors would not bias the reconstruction process as they would incorporate prior knowledge matching with the objects being imaged.

Overall, the problem of obtaining an incorrect structure is an open problem in the field. How incorrect it is depends on the countermeasures that we have taken to prevent structural bias. Heymann *et al.* (2018) reached a similar conclusion from the outcome of the map challenge and suggested a set of safe practices that are very much in line with those suggested here. In the following section, we show that our most common strategies to prevent (gold-standard data analysis) and detect (Fourier shell correlation) biased reconstructions can be fooled by systematic errors so that additional measures are necessary.

5. Detection and avoidance of bias

In this section, we discuss the tools that are currently in use to detect biased reconstructions. As explained, it is easier to detect biased estimates of the different parameters than their combined effect in the reconstructed map. We start by analyzing the FSC as the most widely used map-quality tool. We show that this tool can easily be fooled by systematic errors (the kind of errors that bias gives rise to). Similarly, it is easier to avoid bias in the various parameter-estimation steps than when splitting the data into two halves from the beginning. We argue that this practice is not common among

data-analysis applications as it underutilizes the available data and cannot guarantee the lack of bias. Additionally, we show that splitting the data into two is not necessary to avoid bias. Practices closer to cross-validation or separation into training and test data sets could instead be adopted. Finally, we discuss the current implementation of phase randomization. The original idea is worth pursuing, but its current implementation does not adhere to the original plan.

5.1. On the use of the FSC to detect overfitting

The most common tool to detect the overfitting is the FSC between the two maps,

$$\mathrm{FSC}(R) = \frac{\sum_{\mathbf{R}\in\mathcal{S}(R)} \hat{\mathcal{V}}_1(\mathbf{R})\hat{\mathcal{V}}_2^*(\mathbf{R})}{\left[\sum_{\mathbf{R}\in\mathcal{S}(R)} |\hat{\mathcal{V}}_1(\mathbf{R})|^2\right]^{1/2}\left[\sum_{\mathbf{R}\in\mathcal{S}(R)} |\hat{\mathcal{V}}_2(\mathbf{R})|^2\right]^{1/2}}, \quad (3)$$

where $\mathbf{R}$ is the 3D frequency vector, R is its magnitude, $\mathcal{S}(R)$ is the Fourier shell whose center has radius R, and $\hat{\mathcal{V}}_1$ and $\hat{\mathcal{V}}_2$ are the Fourier transforms of the two maps reconstructed from the two data halves. [Note that in this formulation we are not analyzing the statistical distribution of the FSC, and this is why we have not further expanded $\hat{\mathcal{V}}_1$ and $\hat{\mathcal{V}}_2$ into their deterministic and random components. For a deep analysis of these distributional properties, the reader is referred to Sorzano, Vargas, Otón, Abrishami *et al.* (2017).] If both reconstructions are biased, $V_1 + \Delta V_1$ and $V_2 + \Delta V_2$, and the FSC becomes

$$\mathrm{FSC}(R) = \frac{\sum_{\mathbf{R}\in\mathcal{S}(R)} [\hat{\mathcal{V}}_1(\mathbf{R}) + \Delta\mathcal{V}_1(\mathbf{R})][\hat{\mathcal{V}}_2(\mathbf{R}) + \Delta\mathcal{V}_2(\mathbf{R})]^*}{\left\{\sum_{\mathbf{R}\in\mathcal{S}(R)} |[\hat{\mathcal{V}}_1(\mathbf{R}) + \Delta\mathcal{V}_1(\mathbf{R})]|^2\right\}^{1/2}\left\{\sum_{\mathbf{R}\in\mathcal{S}(R)} |[\hat{\mathcal{V}}_2(\mathbf{R}) + \Delta\mathcal{V}_2(\mathbf{R})]|^2\right\}^{1/2}}. \quad (4)$$

With this measurement, the FSC can be made arbitrarily close to 1 by making $\Delta\mathcal{V}_1 \simeq \Delta\mathcal{V}_2 \gg \hat{\mathcal{V}}_1, \hat{\mathcal{V}}_2$. This bias of the FSC is at the core of some of its reported failures (Borgnia *et al.*, 2004; Egelman, 2014; Subramaniam *et al.*, 2016; Tan *et al.*, 2017).

We may pose the 3D reconstruction problem as that of estimating a set of parameters $\boldsymbol{\Theta}$ that include the reconstructed volume, the 3D alignment parameters of each of the experimental projections and any per-particle imaging parameter. The problem is estimating $\boldsymbol{\Theta}$ from the observed data $\mathbf{y}$,

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\Theta})f(\boldsymbol{\Theta})}{f(\mathbf{y})} = \arg\max_{\boldsymbol{\Theta}} f(\mathbf{y}|\boldsymbol{\Theta})f(\boldsymbol{\Theta}).$$

This is a Bayesian regression problem, as stated in Scheres (2012*a*). The first term aims to look for a solution that is consistent with the acquired data. The second term looks for a solution that is consistent with what is known, in general, about biological macromolecules. The drawback of the Bayesian approach is that, for the moment, we do not have a realistic prior for the set of macromolecules being reconstructed. In any case, as with any other regression problem in statistics, the validity of the result should be include the residuals of the regression; that is, comparing the observed $\mathbf{y}$ with the predicted $\hat{\mathbf{y}}$. Different strategies, such as regressing with a large subset of the data and evaluating with a small subset, could be devised (Ortiz *et al.*, 2019), as is the standard practice in statistics and X-ray crystallography (free *R* value; Brünger, 1992). This is also the spirit of measures based on the spectral signal-to-noise ratio (SSNR; Penczek, 2002; Unser *et al.*, 2005). However, it is not at the core of the FSC. The FSC compares two sets of regression parameters $\hat{\boldsymbol{\Theta}}_1$ and $\hat{\boldsymbol{\Theta}}_2$. This has the drawback of being heavily affected by bias: systematic errors are rewarded by the FSC. There is a connection between the FSC and the SSNR when the errors are supposed to be random. However, the functional nature of the relationship is unknown (Sorzano, Vargas, Otón, Abrishami *et al.*, 2017), and its sensitivity to bias should be considered before adopting it as a universal descriptor of the map quality.

At this point, we would like to highlight that the 0.143 threshold normally used in the field is derived under the assumption of no bias and linearity of the 3D reconstruction process. This latter assumption is broken by some algorithms, for example *Xmipp HighRes* (Sorzano, Vargas *et al.*, 2018).

As a consequence, the FSC between two halves for these algorithms may sometimes not cross the 0.143 threshold (in Supplementary Experiment 10, we show the impact of the nonlinear processing of *Xmipp HighRes* on the FSC). For this class of algorithms, we have heuristically found that a threshold of 0.5 is often a better estimate of the resolution (despite this concept being ill-defined). This tends to be true not only for *Xmipp HighRes* but in general for most 3D reconstruction algorithms that we have used (*RELION*, *cryoSPARC* and *Xmipp*). We have normally observed that the FSC = 0.143 resolution, in most reconstruction algorithms, is usually the best resolution in the best voxel of the local resolution map (Vilas *et al.*, 2018; Ramírez-Aportela *et al.*, 2019; Vilas, Tagare *et al.*, 2020) and that the FSC = 0.5 resolution is more representative of the most common local resolution value.

We have also observed that the FSC typically presents a change of decaying regime at a frequency that is better related to the frequency at which the map is reconstructed (Supplementary Experiment 12). However, it is difficult to determine these regime changes automatically, and a straightforward, objective criterion cannot be given at this moment.

5.2. The gold standard and cross-validation

Throughout this article, we have analyzed the most common sources of bias in cryoEM single-particle analysis. For each of the different sources, we have suggested ways to detect and avoid these biases. The most common way to avoid overfitting in cryoEM is the so-called gold-standard data processing, which divides the data into two independently processed halves (Scheres & Chen, 2012). The idea is based on previous work (Grigorieff, 2000) and, as stated in Chen, McMullan *et al.* (2013),

> Grigorieff showed that when signal-to-noise ratio in the images becomes low enough, it is impossible to avoid overfitting when

the two half sets are aligned against the same reference structure, regardless of how the procedure is initiated. He was the first to conclude that a reliable estimation of resolution using FSC can be obtained only when the two half datasets are independently aligned against two independent reference structures.

However, a careful reading of Grigorieff (2000) reveals that what he showed was an experiment in which, with low SNR and a low number of images, the FSC of two halves aligned against the same reference was not representative of the FSC of the full set against the known ground truth. From this experiment, we cannot take as a general result that the data set needs to be split into two halves aligned to independent references. One of the main differences from this result, 20 years later, is that the number of particles used nowadays largely outpaces the number of particles with which this experiment was performed (1000 images).

Actually, this procedure of splitting the data into two halves goes against the most advanced practices in statistics. The standard recommended approach would be cross-validation (dividing the data set into K pieces, typically $K = 10$, processing $K - 1$ to produce the map and using the remaining piece to evaluate the quality of the map; the process is repeated K times, with each one of the pieces playing the role of the validation subset; Picard & Cook, 1984). This approach is computationally expensive since the full 3D alignment and reconstruction process must be repeated K times, and it may have problems in the case of very imbalanced classes or with the attraction problem (Sorzano, Semchonok *et al.*, 2021). In many domains, the procedure has been simplified to separating the data set into training (80–90% of the data) and validation (20–10%) subsets. This is, for example, the case in deep learning, where it is a well accepted practice. This is also the approach suggested by Ortiz *et al.* (2019).

This article argues that the gold-standard approach is neither necessary nor sufficient to guarantee a lack of bias. It is not sufficient because the two halves can be easily led into the same kind of bias (biased initial volume, missing information induced by the alignment and reconstruction algorithm, incorrect symmetry, the use of incorrect particles, biased objective function, incorrect masking and Fourier filtering *etc*.). If this is the case, both data halves will have the same (or similar) bias. This nonsufficiency argument is well known in the field, in which incorrectly reconstructed structures are reconstructed despite following the gold standard. What is not generally considered in the field is that the gold standard is not necessary either in the sense that processing workflows that systematically interchange images between the two halves do not necessarily show signs of overfitting. This is exemplified in Supplementary Experiment 13. In this example, the set of input images is randomly split at every iteration and assigned to one of the two halves. This strategy is similar to the approach of stochastic gradient descent (Punjani, Brubaker *et al.*, 2017) with two current solutions instead of one, and it was generalized to multiple solutions in Sorzano, Vargas *et al.* (2018). In this way, the two reconstructed volumes will surely share common images along the reconstruction history at the end of the processing. Even though this strategy goes against the current belief that total independence of input data is an absolute requirement, we show that there is no obvious sign of overfitting. This experiment is justified by our claim that the overfitting observed in cryoEM is more related to systematic bias than to variance associated with an excessive number of parameters or the lack of independence between the data. In this way, the emphasis in data processing should be more on removing biased parameters rather than the use of half the data for each reconstruction. Still, this practice of constructing two half volumes is useful for calculating the FSC at a particular iteration. This example shows that strategies other than the gold standard are also possible and may appear in the future.

5.3. Randomized phases to detect overfitting

Chen, McMullan *et al.* (2013) suggested the randomization of phases in the experimental images beyond a given frequency as a way to detect overfitting. The calculation of the true SNR in Chen, McMullan *et al.* (2013) is affected by a problem of zeroth-order Taylor expansion (Sorzano, Vargas, Otón *et al.*, 2017). This invalidates a faithful calculation of the true SNR based on the FSC of the two halves and the FSC of the two halves after randomizing the phases. In any case, the suggestion makes sense as a characterization of the ability of the 3D reconstruction process to identify overfitting, rather than as a true detector of the overfitting present in the reconstruction without any randomization. A problem with the most used implementation of phase randomization, *RELION*, is that it works at the volume level and not at the level of experimental images as originally suggested. This annuls the whole validation idea. Supplementary Experiment 14 presents the differences between the analysis when the phase randomization is performed at the level of images or volumes. Performing it at the level of volumes does not confirm the validity of the reconstructed volume, and the reported randomized phase FSC is the logical result (a large degree of agreement up to the frequency of randomization and decay from this frequency) of the operation performed.

6. Conclusions

The goal of cryoEM is to elucidate the 3D structures of biological macromolecules. This task can be hindered by many pitfalls leading to an incorrect structure determination. The difference between the true (unknown) structure and the obtained structure is our bias. The importance of bias depends on the specific violation and the amount of violation of our data set and parameter estimates. It may affect only small details in the reconstruction or lead to a completely wrong reconstruction. Bias can be induced by the following.

(i) Sample-related sources, such as using particles that do not correspond to our structure, imposing an incorrect symmetry or lacking projection directions. For each of these

problems, we have suggested tools that are capable of detecting them and avoiding them if possible.

(ii) Algorithmic related sources such as those related to the initial volume, the particle parameters (angular assignment and in-plane alignment, defocus and acquisition parameters, normalization *etc.*) or the image processing itself (its objective functions or steps inducing bias, in particular masking in real or Fourier space *etc.*). Image-processing biases are hard to fight as they are at the core of the tools available. Still, we should be aware that the image-processing workflow is in itself another source of bias and should do our best to identify incorrectly estimated particle parameters.

Generally speaking, we can consider two different kinds of errors when estimating parameters: (1) random errors around the true solution and (2) parameter estimates significantly away from the true solution. Given a single parameter estimation, it is impossible to know which situation we are in. Even if we are given multiple estimates of the alignment and imaging parameters for a single projection, it is impossible to know which situation, (1) or (2), each estimate is in. However, with multiple estimations (at least two), assuming that the algorithms producing them are reasonably correct, we could adopt the following strategy: comparing the estimates and deciding whether most of them agree in some particular region of the parameter space. If this is the case, we may assume that we are in error case (1), and then averaging the parameters would reduce the variability due to each of the estimation processes. If they disagree, we would not know which of the two, or more than two, clusters of parameters is the correct one. We might choose the most populated cluster (if we have more than two estimates of the parameters for the same image), hoping that, since the algorithms estimating them are reasonable, the most populated cluster is close to the (unknown) ground truth. Then, we could average the parameters in that region. We could also ignore those particles for which not all algorithms agree in the parameter region.

Experimentalists are very much aware of sample-related errors, and they try their best to avoid them. Algorithmic errors have been overlooked in the community, and most structures are reconstructed using a single estimate of the particle parameters (a single run of the 3D classification and alignment algorithm), trusting the underlying algorithm always to find the 'right' answer and, if not, being capable of dealing with incorrect estimates. Moreover, as a community, we have largely adopted tools (FSC) and strategies (gold standard) that we think protect us from overfitting, that is, structure bias. However, they do not. Unfortunately, there are no statistical means to identify bias without prior knowledge about the reconstructed structure (but this is unknown, as it is the whole purpose of cryoEM). This is a general problem in statistics: bias cannot be estimated from a set of samples.

In this article, we have shown that there might be significant differences between the particle-parameter estimates from different algorithms or even different runs of the same algorithm. Whichever strategy we choose to deal with incorrect estimates at this level, identifying possible bias and reducing the variance invariably requires multiple independent estimations of the alignment and class-membership parameters. Although not entirely protected against bias (two coincident estimates could be simultaneously biased), this approach could help to produce better, more robust and reliable 3D reconstructions of cryoEM data.

7. Related literature

The following references are cited in the supporting information for this article: Charbonnier *et al.* (1992), Chen, Pfeffer *et al.* (2013), Jaume *et al.* (2001), Shen *et al.* (2014) and Thévenaz & Unser (2000).

Acknowledgements

The authors acknowledge the support and the use of resources of Instruct, a Landmark ESFRI project.

Funding information

The authors would like to acknowledge economic support from The Spanish Ministry of Economy and Competitiveness through grants PID2019-104757RB-I00 (AEI/FEDER, UE) and SEV 2017-0712, the 'Comunidad Autónoma de Madrid' through grant S2017/BMD-3817, Instituto de Salud Carlos III through grant PT17/0009/0010 (ISCIII-SGEFI/ERDF), European Union (EU) and Horizon 2020 through grants CORBEL (INFRADEV-1-2014-1, Proposal 654248), INSTRUCT-ULTRA (Proposal 731005), EOSC Life (Proposal 824087), HighResCells (Proposal 810057), IMpaCT (Proposal 857203), EOSC-Synergy (Proposal 857647), iNEXT-Discovery (Proposal 871037) and European Regional Development Fund-Project 'CERIT Scientific Cloud' (No. CZ.02.1.01/0.0/0.0/16_013/0001802). The project that gave rise to these results received the support of a fellowship from the 'la Caixa' Foundation (ID 100010434). The fellowship code is LCF/BQ/DI18/11660021. This project received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 713673.

References

Abrishami, V., Bilbao-Castro, J. R., Vargas, J., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. (2015). *Ultramicroscopy*, **157**, 79–87.
Abrishami, V., Zaldívar-Peraza, A., de la Rosa-Trevín, J. M., Vargas, J., Otón, J., Marabini, R., Shkolnisky, Y., Carazo, J. M. & Sorzano, C. O. S. (2013). *Bioinformatics*, **29**, 2460–2468.
Bartesaghi, A., Aguerrebere, C., Falconieri, V., Banerjee, S., Earl, L. A., Zhu, X., Grigorieff, N., Milne, J. L. S., Sapiro, G., Wu, X. & Subramaniam, S. (2018). *Structure*, **26**, 848–856.
Bepler, T., Kelley, K., Noble, A. J. & Berger, B. (2020). *Nat. Commun.* **11**, 5208.
Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A. J. & Berger, B. (2019). *Microsc. Microanal.* **25**, 986–987.
Biyani, N., Scherer, S., Righetto, R. D., Kowal, J., Chami, M. & Stahlberg, H. (2018). *J. Struct. Biol.* **203**, 120–134.
Borgnia, M. J., Shi, D., Zhang, P. & Milne, J. L. S. (2004). *J. Struct. Biol.* **147**, 136–145.
Brünger, A. (1992). *Nature*, **355**, 472–475.
Burnham, K. P. & Anderson, D. R. (1998). *Model Selection and Inference*. New York: Springer-Verlag.

Charbonnier, P., Blanc-féraud, L. & Barlaud, M. (1992). *J. Vis. Commun. Image Represent.* **3**, 338–346.

Chen, S., McMullan, G., Faruqi, A. R., Murshudov, G. N., Short, J. M., Scheres, S. H. W. & Henderson, R. (2013). *Ultramicroscopy*, **135**, 24–35.

Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M. & Förster, F. (2013). *J. Struct. Biol.* **182**, 235–245.

Dashti, A., Schwander, P., Langlois, R., Fung, R., Li, W., Hosseinizadeh, A., Liao, H. Y., Pallesen, J., Sharma, G., Stupina, V. A., Simon, A. E., Dinman, J. D., Frank, J. & Ourmazd, A. (2014). *Proc. Natl Acad. Sci. USA*, **111**, 17492–17497.

Egelman, E. H. (2014). *eLife*, **3**, e04969.

Elmlund, H., Elmlund, D. & Bengio, S. (2013). *Structure*, **21**, 1299–1306.

Fessler, J. A. (1996). *IEEE Trans. Image Process.* **5**, 493–506.

Gil, D., Carazo, J. M. & Marabini, R. (2006). *J. Struct. Biol.* **156**, 546–555.

Gómez-Blanco, J., Kaur, S., Ortega, J. & Vargas, J. (2019). *J. Struct. Biol.* **208**, 107397.

Grant, T., Rohou, A. & Grigorieff, N. (2018). *eLife*, **7**, e35383.

Grigorieff, N. (2000). *Acta Cryst.* D**56**, 1270–1277.

Grigorieff, N. (2007). *J. Struct. Biol.* **157**, 117–125.

Haselbach, D., Komarov, I., Agafonov, D. E., Hartmuth, K., Graf, B., Dybkov, O., Urlaub, H., Kastner, B., Lührmann, R. & Stark, H. (2018). *Cell*, **172**, 454–464.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Heel, M. van (2013). *Proc. Natl Acad. Sci. USA*, **110**, E4175–E4177.

Henderson, R. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 18037–18041.

Heymann, J. B., Marabini, R., Kazemi, M., Sorzano, C. O. S., Holmdahl, M., Mendez, J. H., Stagg, S. M., Jonic, S., Palovcak, E., Armache, J.-P., Zhao, J., Cheng, Y., Pintilie, G., Chiu, W., Patwardhan, A. & Carazo, J. M. (2018). *J. Struct. Biol.* **204**, 291–300.

Huiskonen, J. T. (2018). *Biosci. Rep.* **38**, BSR20170203.

Jakobi, A. J., Wilmanns, M. & Sachse, C. (2017). *eLife*, **6**, e27131.

Jaume, S., Ferrant, M., Warfield, S. K. & Macq, B. M. M. (2001). *Proc. SPIE*, **4322**, 633–642.

Jiménez, A., Jonic, S., Majtner, T., Otón, J., Vilas, J. L., Maluenda, D., Mota, J., Ramírez-Aportela, E., Martínez, M., Rancel, Y., Segura, J., Sánchez-García, R., Melero, R., del Cano, L., Conesa, P., Skjaerven, L., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. (2019). *Bioinformatics*, **35**, 2427–2433.

Jin, Q., Sorzano, C. O. S., de la Rosa-Trevín, J. M., Bilbao-Castro, J. R., Núñez-Ramírez, R., Llorca, O., Tama, F. & Jonić, S. (2014). *Structure*, **22**, 496–506.

Koeck, P. J. B. & Karshikoff, A. (2015). *J. Microsc.* **259**, 197–209.

Koning, R. I., Gomez-Blanco, J., Akopjana, I., Vargas, J., Kazaks, A., Tars, K., Carazo, J. M. & Koster, A. J. (2016). *Nat. Commun.* **7**, 12524.

Leschziner, A. E. & Nogales, E. (2006). *J. Struct. Biol.* **153**, 284–299.

Ludtke, S. J., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. (2004). *Structure*, **12**, 1129–1136.

Lyumkis, D. (2019). *J. Biol. Chem.* **294**, 5181–5197.

Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. (2013). *J. Struct. Biol.* **183**, 377–388.

Mao, Y., Castillo-Menendez, L. R. & Sodroski, J. G. (2013). *Proc. Natl Acad. Sci. USA*, **110**, E4178–E4182.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Désormeaux, A., Finzi, A., Xiang, S. H. & Sodroski, J. G. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 12438–12443.

Marabini, R., Carragher, B., Chen, S., Chen, J., Cheng, A., Downing, K. H., Frank, J., Grassucci, R. A., Heymann, J. B., Jiang, W., Jonic, S., Liao, H. Y., Ludtke, S. J., Patwari, S., Piotrowski, A. L., Quintana, A., Sorzano, C. O. S., Stahlberg, H., Vargas, J., Voss, N. R., Chiu, W. & Carazo, J. M. (2015). *J. Struct. Biol.* **190**, 348–359.

McLachlan, G. J. & Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Méndez, J., Garduño, E., Carazo, J. M. & Sorzano, C. O. S. (2021). *J. Struct. Biol.* **213**, 107771.

Moebel, E. & Kervrann, C. (2020). *J. Struct. Biol. X*, **4**, 100013.

Naydenova, K. & Russo, C. J. (2017). *Nat. Commun.* **8**, 629.

Noble, A. J., Dandey, V. P., Wei, H., Brasch, J., Chase, J., Acharya, P., Tan, Y. Z., Zhang, Z., Kim, L. Y., Scapin, G., Rapp, M., Eng, E. T., Rice, W. J., Cheng, A., Negro, C. J., Shapiro, L., Kwong, P. D., Jeruzalmi, D., des Georges, A., Potter, C. S. & Carragher, B. (2018). *eLife*, **7**, e34257.

Ogura, T. & Sato, C. (2006). *J. Struct. Biol.* **156**, 371–386.

Ortiz, S., Stanisic, L., Rodriguez, B. A., Rampp, M., Hummer, G. & Cossio, P. (2019). *arXiv*:1908.01054.

Penczek, P. (2002). *J. Struct. Biol.* **138**, 34–46.

Penczek, P. A., Fang, J., Li, X., Cheng, Y., Loerke, J. & Spahn, C. M. T. (2014). *Ultramicroscopy*, **140**, 9–19.

Penczek, P., Renka, R. & Schomberg, H. (2004). *J. Opt. Soc. Am. A*, **21**, 499–509.

Picard, R. R. & Cook, R. D. (1984). *J. Am. Stat. Assoc.* **79**, 575–583.

Punjani, A., Brubaker, M. A. & Fleet, D. J. (2017). *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 706–718.

Punjani, A., Rubinstein, J., Fleet, D. J. & Brubaker, M. A. (2017). *Nat. Methods*, **14**, 290–296.

Radermacher, M. (1988). *J. Elec. Microsc. Tech.* **9**, 359–394.

Radermacher, M. & Hoppe, W. (1980). *Proceedings of the Seventh European Congress on Electron Microscopy*, edited by P. Brederoo & G. Boom, Vol. I, pp. 132–133. Leiden: Seventh European Congress on Electron Microscopy Foundation.

Radermacher, M., Wagenknecht, T., Verschoor, A. & Frank, J. (1987). *J. Microsc.* **146**, 113–136.

Ramírez-Aportela, E., Mota, J., Conesa, P., Carazo, J. M. & Sorzano, C. O. S. (2019). *IUCrJ*, **6**, 1054–1063.

Ramírez-Aportela, E., Vilas, J. L., Glukhova, A., Melero, R., Conesa, P., Martínez, M., Maluenda, D., Mota, J., Jiménez, A., Vargas, J., Marabini, R., Sexton, P. M., Carazo, J. M. & Sorzano, C. O. S. (2020). *Bioinformatics*, **36**, 765–772.

Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J. M. & Sorzano, C. O. S. (2018). *IUCrJ*, **5**, 854–865.

Sanchez-Garcia, R., Segura, J., Maluenda, D., Sorzano, C. O. S. & Carazo, J. M. (2020). *J. Struct. Biol.* **210**, 107498.

Sanz-García, E., Stewart, A. B. & Belnap, D. M. (2010). *J. Struct. Biol.* **171**, 216–222.

Scheres, S. H. W. (2012*a*). *J. Mol. Biol.* **415**, 406–418.

Scheres, S. H. W. (2012*b*). *J. Struct. Biol.* **180**, 519–530.

Scheres, S. H. W. (2016). *Methods Enzymol.* **579**, 125–157.

Scheres, S. H. W. & Chen, S. (2012). *Nat. Methods*, **9**, 853–854.

Scheres, S. H. W., Núñez-Ramírez, R., Gómez-Llorente, Y., San Martín, C., Eggermont, P. P. B. & Carazo, J. M. (2007). *Structure*, **15**, 1167–1177.

Scheres, S. H. W., Valle, M., Grob, P., Nogales, E. & Carazo, J. M. (2009). *J. Struct. Biol.* **166**, 234–240.

Sharon, N., Kileel, J., Khoo, Y., Landa, B. & Singer, A. (2020). *Inverse Probl.* **36**, 044003.

Shatsky, M., Hall, R. J., Brenner, S. E. & Glaeser, R. M. (2009). *J. Struct. Biol.* **166**, 67–78.

Shen, B., Chen, B., Liao, H. & Frank, J. (2014). *Computational Methods for Three-Dimensional Microscopy Reconstruction*, edited by G. T. Herman & J. Frank, pp. 67–95. New York: Springer.

Sorzano, C. O. S., Alcorlo, M., de la Rosa-Trevín, J. M., Melero, R., Foche, I., Zaldívar-Peraza, A., del Cano, L., Vargas, J., Abrishami, V., Otón, J., Marabini, R. & Carazo, J. M. (2015). *Sci. Rep.* **5**, 14290.

Sorzano, C. O. S., Bilbao-Castro, J. R., Shkolnisky, Y., Alcorlo, M., Melero, R., Caffarena-Fernández, G., Li, M., Xu, G., Marabini, R. & Carazo, J. M. (2010). *J. Struct. Biol.* **171**, 197–206.

Sorzano, C. O. S., de Isidro-Gómez, F., Fernández-Giménez, E., Herreros, D., Marco, S., Carazo, J. M. & Messaoudi, C. (2020). *J. Struct. Biol. X*, **4**, 100037.

Sorzano, C. O. S., de la Fraga, L. G., Clackdoyle, R. & Carazo, J. M. (2004). *Ultramicroscopy*, **101**, 129–138.

Sorzano, C. O. S., Fernández-Giménez, E., Peredo-Robinson, V., Vargas, J., Majtner, T., Caffarena, G., Otón, J., Vilas, J. L., de la Rosa-Trevín, J. M., Melero, R., Gómez-Blanco, J., Cuenca, J., del Cano, L., Conesa, P., Marabini, R. & Carazo, J. M. (2018). *J. Struct. Biol.* **203**, 90–93.
Sorzano, C. O. S., Jiménez, A., Mota, J., Vilas, J. L., Maluenda, D., Martínez, M., Ramírez-Aportela, E., Majtner, T., Segura, J., Sánchez-García, R., Rancel, Y., del Caño, L., Conesa, P., Melero, R., Jonic, S., Vargas, J., Cazals, F., Freyberg, Z., Krieger, J., Bahar, I., Marabini, R. & Carazo, J. M. (2019). *Acta Cryst.* F**75**, 19–32.
Sorzano, C. O. S., Jiménez-Moreno, A., Maluenda, D., Ramírez-Aportela, E., Martínez, M., Cuervo, A., Melero, R., Conesa, J. J., Sánchez-García, R., Strelak, D., Filipovic, J., Fernández-Giménez, E., de Isidro-Gómez, F., Herreros, D., Conesa, P., Del Caño, L., Fonseca, Y., de la Morena, J. J., Macías, J. R., Losana, P., Marabini, R. & Carazo, J. M. (2021). *Methods Mol. Biol.* **2305**, 257–289.
Sorzano, C. O. S., Jonic, S., Núñez-Ramírez, R., Boisset, N. & Carazo, J. M. (2007). *J. Struct. Biol.* **160**, 249–262.
Sorzano, C. O. S., Marabini, R., Boisset, N., Rietzel, E., Schröder, R., Herman, G. T. & Carazo, J. M. (2001). *J. Struct. Biol.* **133**, 108–118.
Sorzano, C. O. S., Marabini, R., Vargas, J., Otón, J., Cuenca-Alba, J., Quintana, A., de la Rosa-Trevín, J. M. & Carazo, J. M. (2014). *Computational Methods for Three-Dimensional Microscopy Reconstruction*, edited by G. T. Herman & J. Frank, pp. 7–42. New York: Springer.
Sorzano, C. O. S., Otero, A., Olmos, E. M. & Carazo, J. M. (2009). *BMC Struct. Biol.* **9**, 18.
Sorzano, C. O. S., Semchonok, D., Lin, S.-C., Lo, Y.-C., Vilas, J. L., Jiménez-Moreno, A., Gragera, M., Vacca, S., Maluenda, D., Martínez, M., Ramírez-Aportela, E., Melero, R., Cuervo, A., Conesa, J. J., Conesa, P., Losana, P., Caño, L., de la Morena, J. J., Fonseca, Y. C., Sánchez-García, R., Strelak, D., Fernández-Giménez, E., de Isidro, F., Herreros, D., Kastritis, P. L., Marabini, R., Bruce, B. D. & Carazo, J. M. (2021). *J. Struct. Biol.* **213**, 107695.
Sorzano, C. O. S., Vargas, J., de la Rosa-Trevín, J. M., Jiménez, A., Maluenda, D., Melero, R., Martínez, M., Ramírez-Aportela, E., Conesa, P., Vilas, J. L., Marabini, R. & Carazo, J. M. (2018). *J. Struct. Biol.* **204**, 329–337.
Sorzano, C. O. S., Vargas, J., de la Rosa-Trevín, J. M., Otón, J., Álvarez-Cabrera, A. L., Abrishami, V., Sesmero, E., Marabini, R. & Carazo, J. M. (2015). *J. Struct. Biol.* **189**, 213–219.
Sorzano, C. O. S., Vargas, J., de la Rosa-Trevín, J. M., Zaldívar-Peraza, A., Otón, J., Abrishami, V., Foche, I., Marabini, R., Caffarena, G. & Carazo, J. M. (2014). *Proceedings of International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2014)*, p. 950. Granada: Copicentro Granada.
Sorzano, C. O. S., Vargas, J., Otón, J., Abrishami, V., de la Rosa Trevín, J. M., del Riego, S., Fernández-Alderete, A., Martínez-Rey, C., Marabini, R. & Carazo, J. M. (2015). *AIMS Biophys.* **2**, 8–20.
Sorzano, C. O. S., Vargas, J., Otón, J., Abrishami, V., de la Rosa-Trevín, J. M., Gómez-Blanco, J., Vilas, J. L., Marabini, R. & Carazo, J. M. (2017). *Prog. Biophys. Mol. Biol.* **124**, 1–30.
Sorzano, C. O. S., Vargas, J., Otón, J., de la Rosa-Trevín, J. M., Vilas, J. L., Kazemi, M., Melero, R., Del Caño, L., Cuenca, J., Conesa, P., Gómez-Blanco, J., Marabini, R. & Carazo, J. M. (2017). *Biomed. Res. Int.* **2017**, 6482567.
Sorzano, C. O. S., Vargas, J., Vilas, J. L., Jiménez-Moreno, A., Mota, J., Majtner, T., Maluenda, D., Martínez, M., Sánchez-García, R., Segura, J., Otón, J., Melero, R., del Cano, L., Conesa, P., Gómez-Blanco, J., Rancel, Y., Marabini, R. & Carazo, J. M. (2018). *Appl. Anal. Optim.* **2**, 299–313.
Sorzano, C. O. S., Velázquez-Muriel, J. A., Marabini, R., Herman, G. T. & Carazo, J. M. (2008). *Pattern Recognit.* **41**, 616–626.
Stewart, A. & Grigorieff, N. (2004). *Ultramicroscopy*, **102**, 67–84.
Subramaniam, S. (2013). *Proc. Natl Acad. Sci. USA*, **110**, E4172–E4174.
Subramaniam, S., Earl, L. A., Falconieri, V., Milne, J. L. & Egelman, E. H. (2016). *Curr. Opin. Struct. Biol.* **41**, 194–202.
Tan, Y. Z., Baldwin, P. R., Davis, J. H., Williamson, J. R., Potter, C. S., Carragher, B. & Lyumkis, D. (2017). *Nat. Methods*, **14**, 793–796.
Thévenaz, P. & Unser, M. (2000). *IEEE Trans. Image Process.* **9**, 2083–2099.
Unser, M., Sorzano, C. O. S., Thévenaz, P., Jonić, S., El-Bez, C., De Carlo, S., Conway, J. & Trus, B. L. (2005). *J. Struct. Biol.* **149**, 243–255.
Vargas, J., Abrishami, V., Marabini, R., de la Rosa-Trevín, J. M., Zaldivar, A., Carazo, J. M. & Sorzano, C. O. S. (2013). *J. Struct. Biol.* **183**, 342–353.
Vargas, J., Álvarez-Cabrera, A. L., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. (2014). *Bioinformatics*, **30**, 2891–2898.
Vargas, J., Melero, R., Gómez-Blanco, J., Carazo, J. M. & Sorzano, C. O. S. (2017). *Sci. Rep.* **7**, 6307.
Vargas, J., Otón, J., Marabini, R., Carazo, J. M. & Sorzano, C. O. S. (2016). *Sci. Rep.* **6**, 21626.
Vilas, J. L., Gómez-Blanco, J., Conesa, P., Melero, R., de la Rosa-Trevín, J. M., Otón, J., Cuenca, J., Marabini, R., Carazo, J. M., Vargas, J. & Sorzano, C. O. S. (2018). *Structure*, **26**, 337–344.e4.
Vilas, J. L., Tagare, H. D., Vargas, J., Carazo, J. M. & Sorzano, C. O. S. (2020). *Nat. Commun.* **11**, 55.
Vilas, J. L., Vargas, J., Martínez, M., Ramírez-Aportela, E., Melero, R., Jiménez-Moreno, A., Garduño, E., Conesa, P., Marabini, R., Maluenda, D., Carazo, J. M. & Sorzano, C. O. S. (2020). *J. Struct. Biol.* **209**, 107447.
Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., Quentin, D., Roderer, D., Tacke, S., Siebolds, B., Schubert, E., Shaikh, T. R., Lill, P., Gatsogiannis, C. & Raunser, S. (2019). *Commun. Biol.* **2**, 218.
Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. & Wu, Y. (2014). *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1386–1393. Piscataway: IEEE.
Zhang, K. (2016). *J. Struct. Biol.* **193**, 1–12.
Zhang, X. & Zhou, Z. H. (2011). *J. Struct. Biol.* **175**, 253–263.
Zivanov, J., Nakane, T., Forsberg, B. O., Kimanius, D., Hagen, W. J., Lindahl, E. & Scheres, S. H. W. (2018). *eLife*, **7**, e42166.

STRUCTURAL
BIOLOGY

Volume 78 (2022)

Supporting information for article:

On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy

C. O. S. Sorzano, A. Jiménez-Moreno, D. Maluenda, M. Martínez, E. Ramírez-Aportela, J. Krieger, R. Melero, A. Cuervo, J. Conesa, J. Filipovic, P. Conesa, L. del Caño, Y. C. Fonseca, J. Jiménez-de la Morena, P. Losana, R. Sánchez-García, D. Strelak, E. Fernández-Giménez, F. P. de Isidro-Gómez, D. Herrero, J. L. Vilas, R. Marabini and J. M. Carazo

Supplementary Material

Experiment 1. Non-isolated particles.

We selected 32,589 particles from the T20S proteasome of the dataset EMPIAR 10025. These particles were purposefully selected from those having a nearby particle (see Fig. 2). We used as starting volume the correct structure of the proteasome lowpass filtered at 30Å and we imposed D7 symmetry. Relion autorefine reported a resolution of 3.9Å (before post-processing), the same resolution as the one reported for 26,945 isolated particles. In this example, the effect of incorrectly chosen particles has been exaggerated due to the absence of isolated particles ($N_1 = 0$ in Eq. 2) in the dataset. However, it illustrates the general principle that our reconstruction will be a mixture of a correct and incorrect structure. Depending on the proportion of incorrect particles, the incorrect structure will be more or less visible on top of the correct structure. In any case, it will be biasing our result.

Experiment 2. 3D classification stability.

In the following, we describe the results of three different classification experiments:

1. Absence of heterogeneity. We selected 24,199 particles from the Brome mosaic virus sample of EMPIAR 10010. We used Relion to separate these images into two classes (the two classes were internally initialized by Relion by randomly assigning the input images to one of the two classes, which is the most common way of executing Relion 3D classification). We used the 3D reconstruction of all these particles as initial volume, which yielded a reconstruction of 4.1Å resolution. During the classification, we imposed the icosahedral symmetry. We repeated this step three times, and the number of particles yielding the virus structure ranged from 68.4% to 82.8%. In all cases, the remaining particles gathered into a class in which there was a mixture of several kinds of misalignment (including mirroring of the particles). Only 55.8% of the images were always assigned to the correct virus structure in all three classifications, meaning that the class assignment was unclear for 45% of the particles. This result is very interesting because it shows the difficulty of determining the class parameter, even for an easy example in which there is a single class.
2. Continuous heterogeneity. We selected 34,268 particles from the ribosome sample of EMPIAR 10028. The head of this ribosome is moving and we submitted the set of particles to a classification with Relion in three classes. We used as initial volume the reconstruction of all the particles at 4.4Å and imposed no symmetry. We repeated this classification three times obtaining the following class distributions:

Run	Class 1	Class 2	Class 3
Run 1	46.3%	46.0%	7.7%
Run 2	37.2%	31.8%	31.0%
Run 3	40.1%	31.1%	28.8%

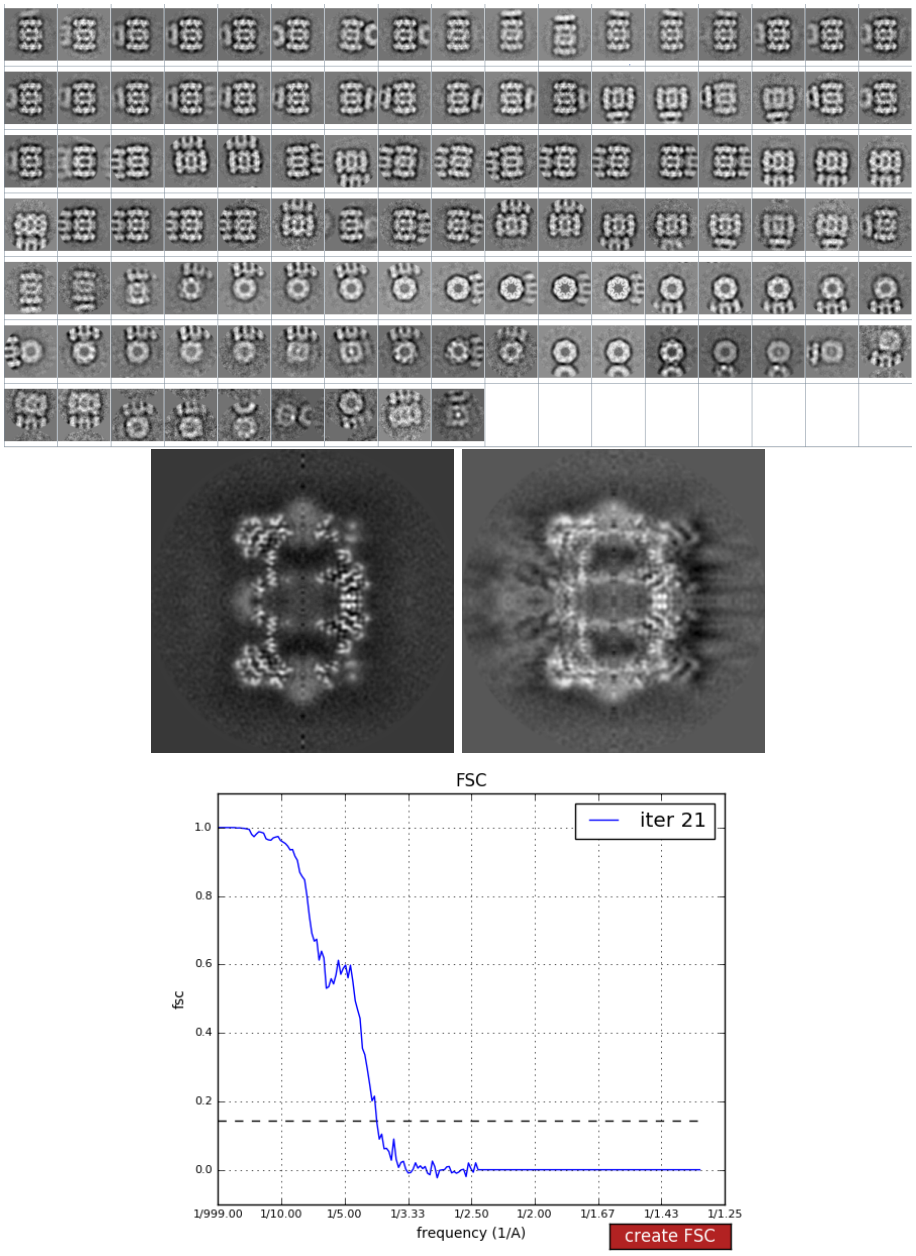


Figure 2: (This Figure corresponds to Experiment 1.) Top: Gallery of the 2D class averages of the particles selected for 3D reconstruction. Middle: Central slice of the 3D reconstruction with isolated (left) and joint (right) particles. Bottom: FSC of the reconstruction with joint particles.

If we compute the subsets of particles that were always classified together, one might expect that in a stable classification, most of the particles in a class stay together independently of the classification run. There are up to 27 ($= 3^3$) possible subsets, of which 19 have sizes of 8.0%, 6.8%, 6.0%, 5.8%, 5.8%, 5.6%, 5.6%, 5.2%, 5.2%, 5.1%, 4.9%, 4.7%, 4.3%, 4.3%, 4.0%, 3.9%, 3.9%, 3.7%, 3.6%, and 8 small classes have less than 1% each. This means that the three classifications are not really consistent with each other. The input particles are almost "uniformly" scattered among the three classes even though true classes could have been formed due to the head movement.

3. Discrete heterogeneity.
    - Case 1. 10,000 images of the Ribosome 70S subunit has served for years as a benchmark for 3D classification [69]. The first 5,000 images of this dataset are supposed to belong to one class and the last 5,000 to another class. We used as initial volume the reconstruction using all particles (with a resolution of 13.5Å). We performed a 3D classification of this dataset into two classes using Relion. Most of the images were assigned to a large class (containing 87-89% of the particles), while the remaining images were assigned to a small class. 85% of the original particles were always classified together. This may indicate two possibilities: a) these 8,500 images really belong to a single class (instead of the 5,000 originally assumed); b) the larger class attracts particles from the smaller class by an effect of 3D attraction. With experimental data, it is not possible to distinguish the true underlying state of the class labels. Still, the dataset has been used in the field several times as a 3D classification benchmark.
    - Case 2. The EMPIAR entry 10333 contains 23,307 images of Human FACT in Complex with Partially Assembled Sub-nucleosomes. The original authors classified these images into two classes of 16,317 (70% of the dataset) and 6,990 (30%) images, respectively. We repeated the classification in two classes five times using Relion. We used the reconstruction performed by the original group as the initial volume (EMDB entry 20840, 4.9Å). In each one of the runs, the largest class obtained: 58%, 71%, 87%, 92%, and 93% of the dataset. This means that the two classes were not separated in the last three of the five runs (another manifestation of the 3D attraction). Once a class starts increasing its SNR, it attracts all particles, disregarding whether they truly belong to that class or not). In the first two runs, the presence of the original classification in the recalculated classification is only between 60-65% (a little bit lower than the 70% that one would expect from a random classification; and much lower than the 100% for a classification that matches the original one). If we analyze the set of images that were consistently put together into the same class, the decomposition of the original dataset is: 43%, 19%, 15%, 7%, and many small groups with less than 2% of the images. This experiment

shows how difficult 3D classification is in such noisy environments, and it suggests that just taking "the first classification result" is not a safe strategy.

Experiment 3. 3D Angular attraction caused by overrepresented directions.

We simulated 36,000 projections of the EMDB entry 10077 (Erythromycin Resistant *Staphylococcus aureus* 50S ribosome (delta R88 A89 uL22) in complex with erythromycin). The structure was solved to 2.3Å. 10,000 of these images were randomly distributed over the projection sphere, while the rest were concentrated in a cone of 30 degrees. We added noise to an SNR of 0.1. We ran Relion autorefine starting from the already solved structure. We used C1 symmetry, and Relion automatically determined the number of iterations. The isosurface of the resulting reconstruction by Relion autorefine is shown in Fig. 3. It can be seen that there is a strong elongation along the overloaded (vertical in the figure) direction caused by the attraction of that direction of nearby particles (the non-overloaded particles had a probability of being found in this region that was 50% larger than expected by a uniform angular distribution). An angular distribution analysis revealed that the 3D attraction did not cause any missing region in this case. However, it attracted to the over-represented region many images that did not belong there (see the histogram in Fig. 3).

Experiment 4. Different angular assignments.

We selected 15,396 images of the β-galactosidase sample of EMPIAR 10013. We used an initial volume calculated by Xmipp `reconstruct_significant` [83] exploiting the D2 symmetry of the macromolecule. We reconstructed it with Xmipp highres and Relion autorefine. The resolution calculated by the FSC at 0.143 was 2.9 Å and 3.9Å, respectively (the reported resolution at EMDB was 3.2 Å; we report the resolution calculated without postprocessing). When we compare the angular assignment of both algorithms, we observe that there is a wider empty region in the case of Relion and that 48% of the particles obtain an angular assignment significantly different between Xmipp highres and Relion autorefine (see Fig. 4), meaning that at least one of the two angular estimates has to be wrong for the particles in which both algorithms do not coincide.

Experiment 5. Angular assignment stability.

We selected 34,268 particles from the ribosomal dataset at EMPIAR 10028. We used as initial volume the reconstruction of all the particles at 4.4Å and imposed no symmetry. We performed five independent angular assignments using Relion autorefine (Relion automatically determined the number of iterations). Between any two angular assignments, about 7% of the particles obtained a significantly different angle (larger than 1.5 degrees in a structure whose size is 400 voxels wide, that is an uncertainty larger than 5 pixels in the border of the image). Only 88% of the particles consistently obtained an angular assignment whose angular difference was smaller than 1.5 degrees.

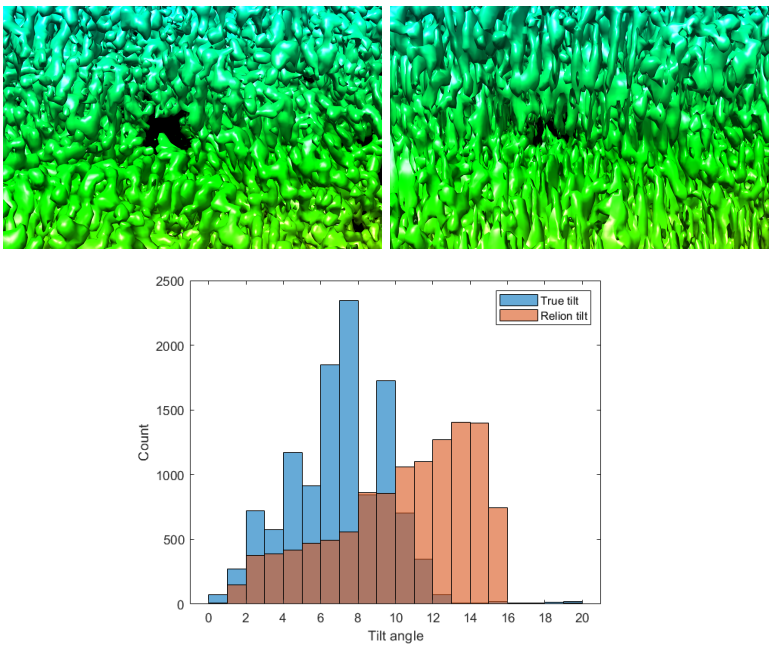


Figure 3: (This Figure corresponds to Experiment 3.) Top: Isosurface of the 50S ribosome in complex with erythromycin reconstruction when a particular direction is overloaded reconstructed by Xmipp highres (left) and Relion 3D autorefine (right). The surface has been colored as a way to help visualization (according to the volume height). Bottom: Ground-truth tilt angle distribution and its estimated distribution by Relion 3D Autorefine.

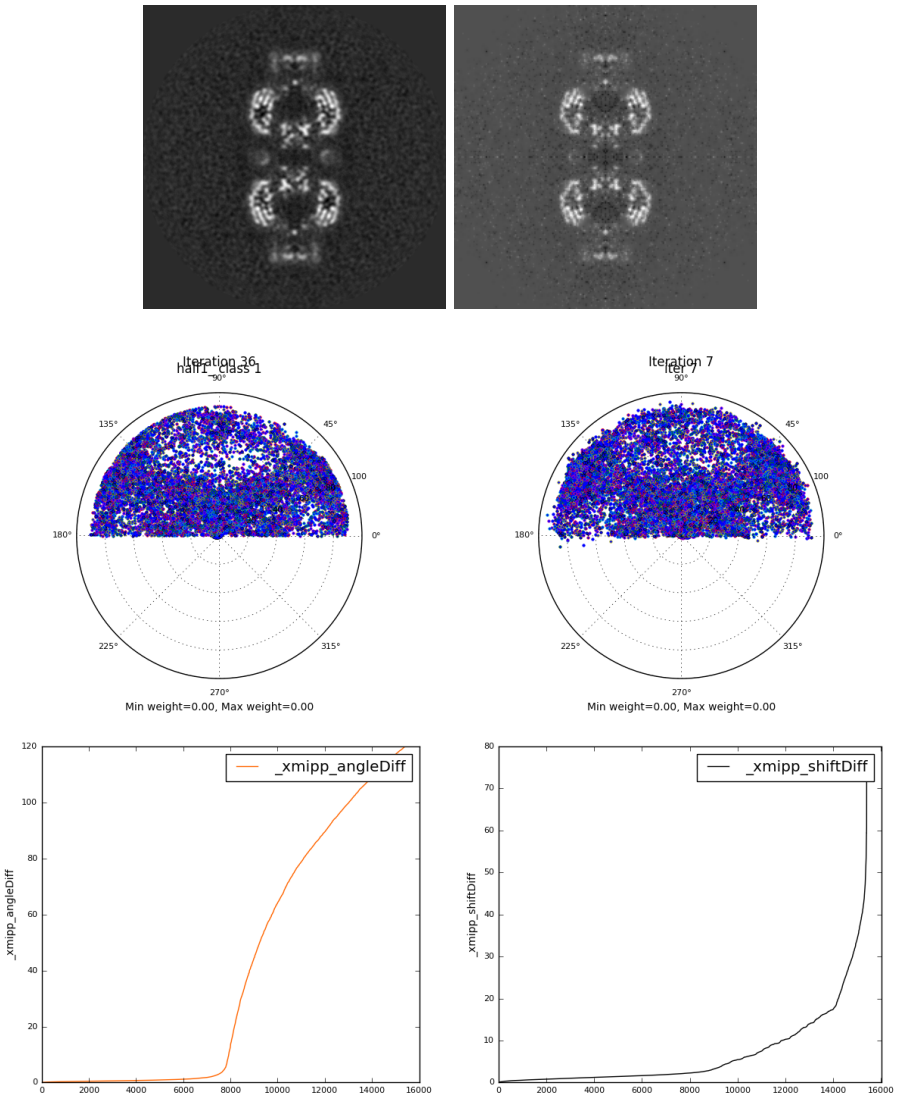


Figure 4: (This Figure corresponds to Experiment 4.) Top row: Central reconstruction slice by Relion autorefine (left) and Xmipp highres (right). Middle row: Angular distribution of Relion (left) and Xmipp highres (right). Bottom: Comparison of the angular difference (left) and shift difference (right) between both algorithms (note that these figures are not histograms but a plot of the sorted values).

The same experiment performed with 26,945 projections of isolated the T20S dataset of EMPIAR 10025 revealed that only 62% of the particles obtained an angular alignment within 1.5 degrees when two independent angular assignments by Relion autorefine were compared. For that experiment, we used an initial volume obtained by Xmipp `reconstruct_significant` [83], lowpass filtered at 30Å. We imposed D7 symmetry.

These experiments show that the uncertainty of the angular assignment is not only due to the use of different assignment algorithms (see Experiment 4) but also due to an uncertainty intrinsic to each algorithm and dataset.

Experiment 6. Different similarity measures.

We analyzed the 26,945 T20S particles of the previous experiment. We plotted the histogram of the Log-Likelihood and the maximum probability as calculated by Relion autorefine. Neither of the two plots reveals the presence of two subpopulations (a population of correctly assigned particles and a population of incorrectly assigned particles or non-particles, for instance), see Fig. 5. However, when we compute the Xmipp highres local similarity (the cross-correlation within a maximal circle), the existence of the two populations can be easily recognized (see Fig. 5). This example illustrates that a single metric can measure only one feature of the data being analyzed. Different metrics allow us to see the dataset from different perspectives. In some of these perspectives, data pathologies like the presence of multiple populations of particles may be more visible than in others (i.e., in other data sets, the difference of behavior in the metric that in this case favors Xmipp highres could have played the other way around and favor Relion). Additionally, we split the input dataset into two groups depending on the low or high local cross-correlation. Performing two reconstructions with these two groups shows that the FSC of one of them is much better than the other one, indicating that one of the groups, the one with low local correlation, had been incorrectly aligned by Relion 3D autorefine (see Fig. 5).

In Maximum-likelihood or Maximum a Posteriori methods, particles can have different angular assignments, each one with a different probability. In the best case, it is expected that the probability distribution for a particle converges to a delta function so that the particle has a unique angular assignment. However, this is not the general case, and a particle usually has several orientations with a probability significantly different from zero. This is a direct consequence of the metric being used and its smoothness. To illustrate this effect, Fig. 6 shows the distribution of the number of significantly contributing alignment parameters in Relion autorefine for these experimental images.

Experiment 7. Handedness mixture.

We selected 26,945 particles of the proteasome dataset EMPIAR 10025. We generated an initial volume that had mirrored and unmirrored 2D classes mixed (we did so by mirroring the initial volume of Experiment 5 and adding it to itself). We then refined this initial model with the particles using Xmipp highres.

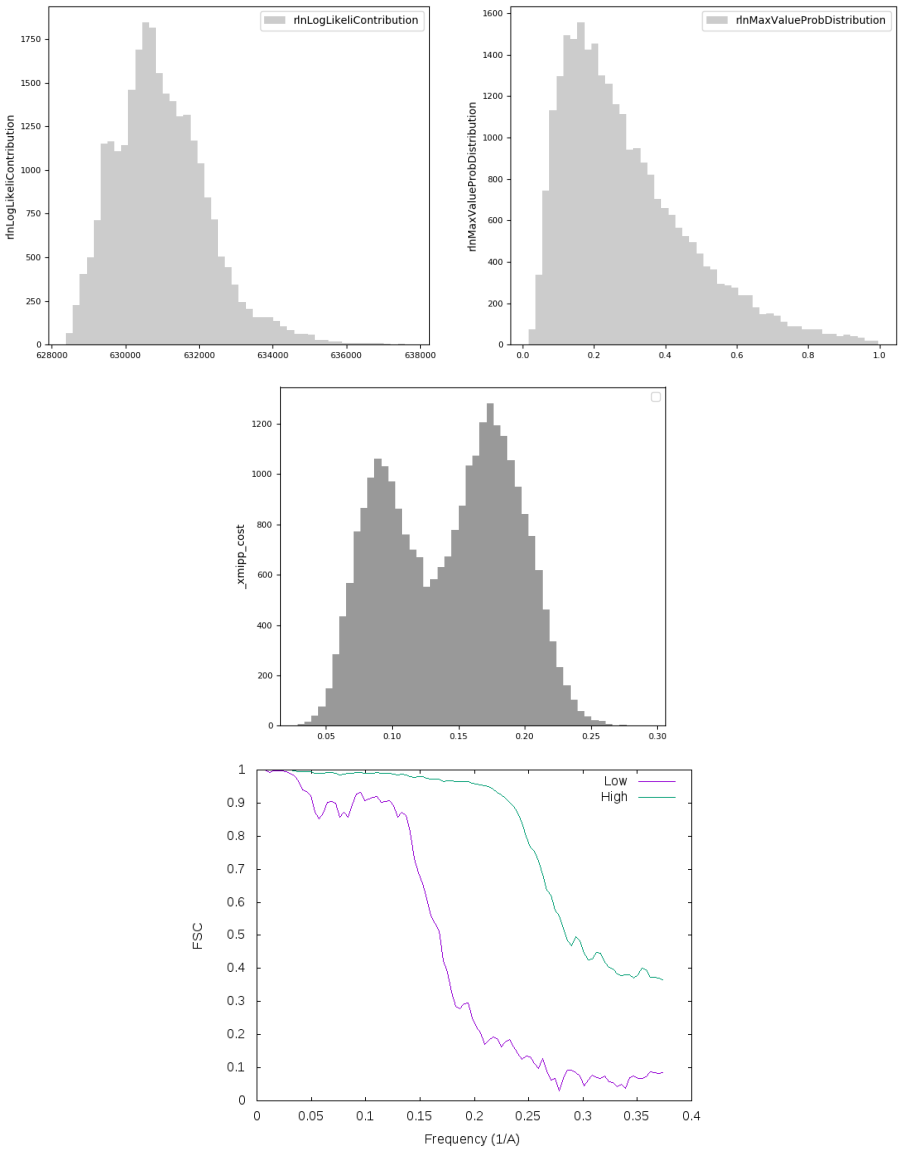


Figure 5: (This Figure corresponds to Experiment 6.) Top: Histogram of the log-likelihood (left) and maximum probability (right) of Relion autorefine. Middle: Histogram of the cross-correlation index within a maximal circle. Bottom: FSCs of two reconstructions performed with the particles in the low and high local correlation groups.

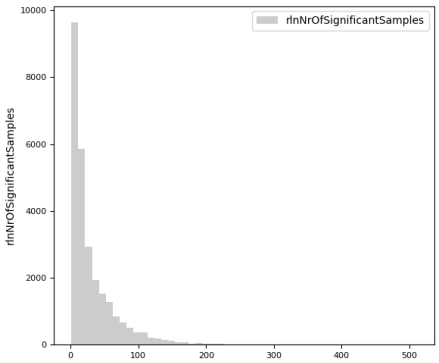


Figure 6: Histogram of the number of significantly contributing alignment parameters for Relion autorefine.

We performed a global angular assignment, but projection matching starting from an initial model (even with all the safeguards built in highres) is, by nature, a local optimization in the volume space. As shown in Fig. 7, the refined volume keeps the hand mixture resulting in an incorrect map. Since the hand mixture is present in both halves, the FSC reports a resolution of 4Å at a threshold of 0.143. If the multiresolution reconstruction scheme of highres is employed, then the algorithm converges to the right structure. Solving the problem at low resolution helps to smooth the landscape of solutions in the volume space and helps the algorithm find a better minimum. This smoothing of the goal function is an effect that has already been reported in other image processing domains like image registration [94], image segmentation [28], or image restoration [10]. However, note that there is no guarantee that by smoothing the goal function, the algorithm will be able to find the global minimum starting from any point.

Experiment 8. Incorrect CTF correction

We selected 22,824 particles from the proteosome dataset EMPIAR 10013. We constructed the initial volume in the same way as in Experiment 4. We reconstructed it with Relion autorefine obtaining the reconstruction shown in Fig. 8a. We then reduced all the defoci of the particles by -0.5μm (this corresponds approximately to 20% of the average nominal defocus). The corresponding reconstruction (Fig. 8b) has a dark halo surrounding the particle (this halo is often observed in public EM maps). The problem is not only the halo itself but the fact that there are also structural differences between reconstructing using the correct defoci and the incorrect defoci, as shown in Fig. 8c. The difference between the reconstruction with the supposedly correct defoci and the ones with the systematically wrong defoci is a structural bias that is superposed to our final 3D reconstruction. However, the FSC reported the same resolution, 3.8Å, for the incorrect and correct defocus experiment.

Errors in the CTF correction can also be induced by incorrect pixel size since the CTF formula depends on it, too, as shown in the main text. In

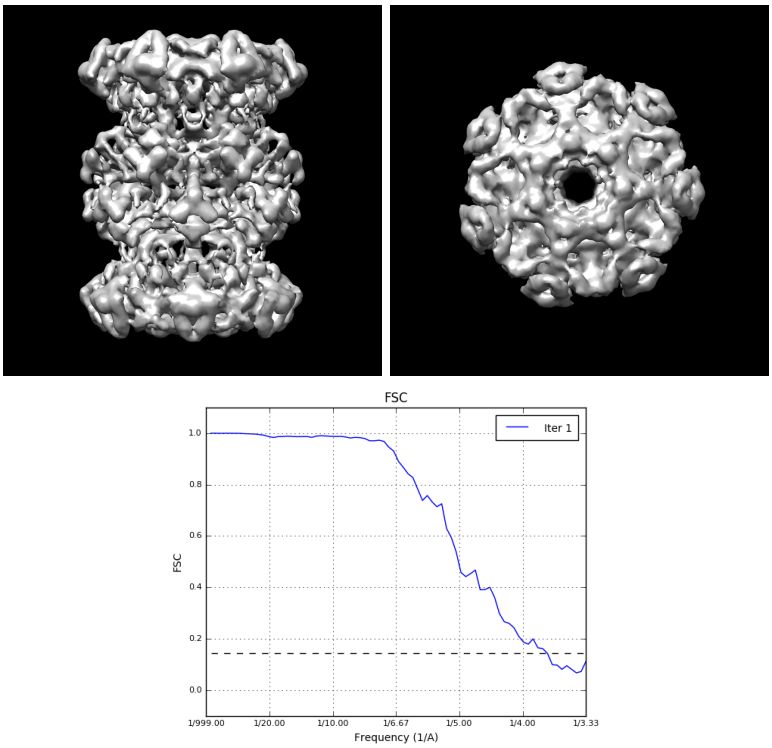


Figure 7: (This Figure corresponds to Experiment 7.) Top: Side and top views of a proteasome dataset when the initial volume has a hand mixture. Bottom: FSC of the reconstruction according to Xmipp highres. Despite the seemingly correct FSC, the reconstruction is totally artifactual as half of the images have been mirrored. Consequently, we have the addition of two volumes: one with the correct hand and another one with the incorrect hand.

our next experiment, we systematically reduced the pixel size by a factor of 20%. Although these large errors are not expected in a real experiment, it illustrates the problem and, in practice, several algorithms like Xmipp highres [82] or cisTEM [20] can optimize for the magnification at the level of particle and, if not properly constrained or monitored, pixel size deviations as large as 20% can be observed for some images. The FSC reported in this case a resolution of 3.1Å (actually, the FSC frequency axis is itself heavily affected by a biased pixel size; the reason is that the Fourier index (k_x, k_y, k_z) of a volume of size $N_x \times N_y \times N_z$ voxels is translated into a continuous frequency given by $(f_x, f_y, f_z) = \left(\frac{k_x}{N_x T}, \frac{k_y}{N_y T} \frac{k_z}{N_z T}\right)$, where T is the pixel size in Å/pixel, in this way, errors in the pixel size are translated into errors of the frequency axis of the FSC and, because of the inverse relationship, errors in which the pixel size is underestimated are amplified by the inverse relationship). Fig. 8 shows how the FSC is, in this case, misleading in order to distinguish a correct from an incorrect reconstruction.

Experiment 9. Volume postprocessing

We selected 34,268 from the ribosomal dataset EMPIAR 10028 (in a similar way to Experiment 5). We reconstructed it with Relion autorefine and post-processed it with Relion. The FSC is shown in Fig. 9. As argued in the text, the change in FSC is solely due to the change in the mask between the 3D reconstruction and the post-processed one. Masking is another source of bias, this time with a more visible effect in Fourier space. In Sec. 5.1 we discuss that a bias in the volume results in arbitrarily high FSCs.

Relion's implementation of the postprocessing also reports the Phase Randomized FSC and its correction. However, the randomized maps FSC is performed at the level of volumes (the phases of the volumes are randomized beyond a given frequency), giving a false impression of reliability on the result obtained from the 3D reconstruction process (the corrected FSC is also affected by the bias induced by the masking). The randomization should be performed at the level of images, as originally suggested by [11], and then the whole 3D reconstruction and alignment procedure should be repeated (see Suppl. Material Experiment 14). In this example, the resolution estimated by the corrected FSC was 3.53Å, while the resolution estimated by the unmasked maps (as reported by Relion autorefine) was 4.54Å. To ascertain which of the two numbers was closer to the underlying resolution, we independently measured local resolution with MonoRes [101]. The histogram of local resolutions is also represented in Fig. 9. We can see that the peak of local resolution is around 4Å and that between 4-5Å there is a large fraction of the voxels.

Experiment 10. Algorithm weight

We selected 10,908 particles from the ribosomal dataset EMPIAR 10028 (the initial steps of the processing are the same as in Experiments 5 and 9). We reconstructed it with Relion autorefine obtaining the reconstruction shown in Fig. 10 left. Using the same angular assignment within Relion reconstruct,

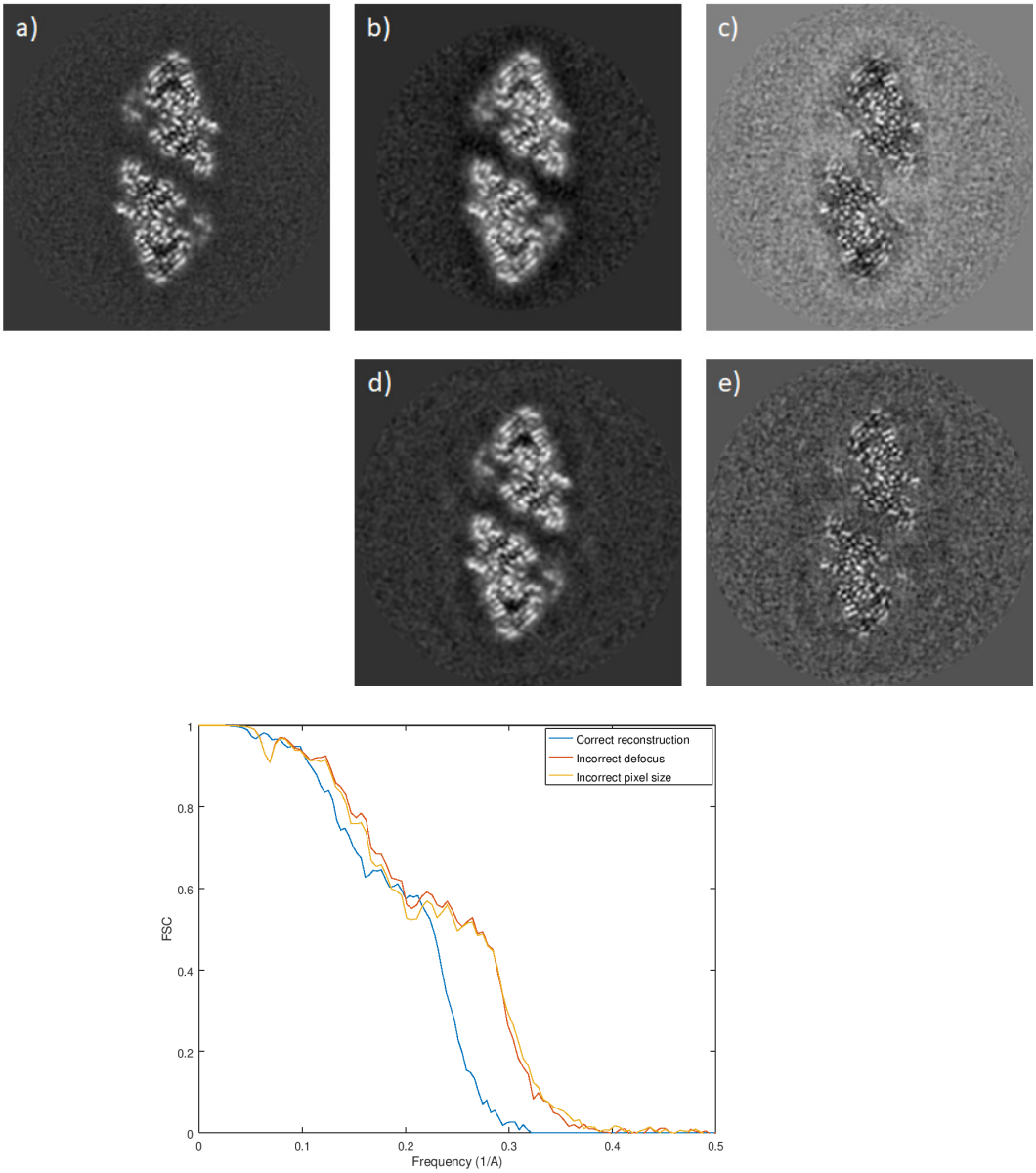


Figure 8: (This Figure corresponds to Experiment 8.) Representative slices of a reconstruction with the correct CTF (a), a reconstruction with biased defocus (b), the difference between a and b (c), a reconstruction with biased pixel size (d), and the difference between a and d (e). Bottom: FSC of the correct reconstruction and the reconstruction with incorrect defoci and incorrect pixel size.

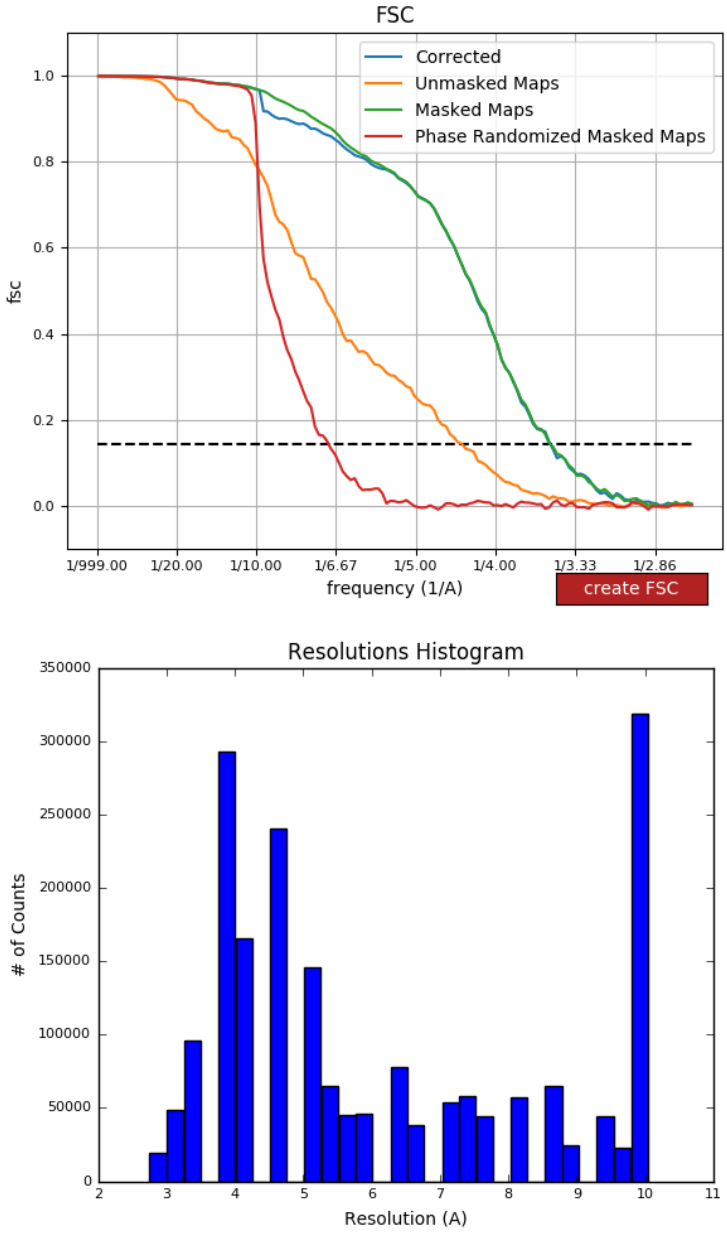


Figure 9: (This Figure corresponds to Experiment 9.) Top: Different FSCs reported by Relion: Unmasked (just coming out from the reconstruction process), Masked (after changing the mask), Corrected (after changing the mask and B-factor corrected), Phase Randomized at the volume level (after randomizing the two half maps from 10Å). Bottom: Histogram of the local resolution as calculated by MonoRes.

we obtain the reconstruction shown in Fig. 10 right. The difference between the two volumes is a bias between both reconstructions. The only difference between these two reconstructions is the internal weight assigned to each of the projections, each of the frequencies, and the fact that in autorefine, the same particle can occupy multiple locations (these weights are the ones at Eq. (9) of Scheres [62]). This example illustrates the fact that the 3D reconstruction algorithm itself is another source of bias.

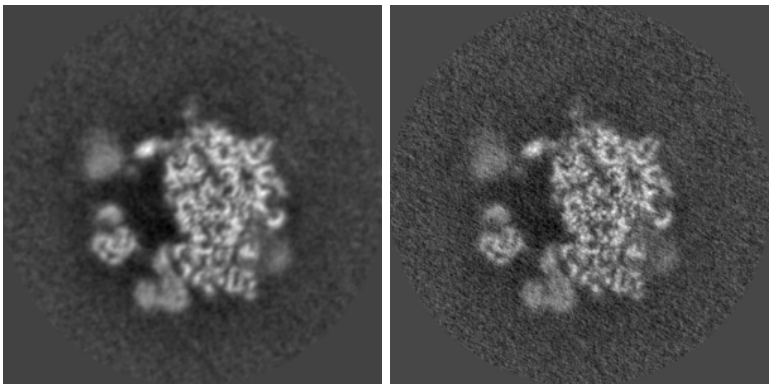


Figure 10: (This Figure corresponds to Experiment 10 .) Central slice of the ribosome reconstructed with Relion autorefine (left) and Relion reconstruct (right).

Experiment 11. Effect of non-linear processing on the FSC

We converted the atomic model of the Peripheral domain of open complex I during turnover (PDB entry: 6ZK9) into an electron density map using electron atomic scattering factors [85]. We then added two realizations of Gaussian noise with zero mean and standard deviation 0.3 (see Fig. 11). These two volumes would be the equivalent of the two half maps from the gold standard approach. Then we calculated the average of the two half maps and the non-linear processing of Xmipp highres (it includes denoising in real space, filterbanks, denoising after applying a Laplacian transformation, deconvolving, applying a soft-negative undershooting remover, and a weight based on the pair differences, see the Suppl. Material of [82] for the algorithmic details). Note that the non-linear processing does not involve any masking (although a mask is used to define a region in which noise statistics can be measured) and that the concept of statistical significance is very much involved in all steps.

We then compared the FSCs coming from linear and non-linear processing. As expected, the 0.143 threshold is a good estimate of the resolution of the two gold-standard maps (see Fig. 12 top). It effectively represents the resolution when the average of the two maps is compared to the atomic model. Interestingly, if we remove the surrounding noise with a relatively generous mask (see Fig. 11), the FSC is largely boosted. However, the map has not been touched, and it could be rightly argued that the map's resolution is the one after masking

(as we have only removed noise that is known not to belong to the structure). At this point, we do not enter into the issue of the right resolution value and only highlight the dependence of the FSC on two reasonable approaches to its calculation.

We next applied the series of non-linear processing steps described above and used by Xmipp highres. We compare the gold standard FSC to the FSC of the highres map to the atomic model (see Fig. 12 middle). For comparison purposes, we also included the comparison of the average of the two half maps (linear processing) to the atomic model. The FSC of the highres gold standard shows that the non-linear processing of the two half maps is extremely consistent (the FSC does not even cross the 0.5 threshold). By comparing the highres result (defined as the average of the two non-linearly processed half maps) to the atomic model, we can estimate whether these restored maps have introduced consistent but incorrect features. From this comparison and the corresponding one of the linear processing, we see that the non-linear processing outperforms the linear one. We wondered then if the improvement was due to an implicit masking effect of the non-linear processing, and for checking that, we applied the same mask to the linearly and non-linearly processed maps. We then compared the masked volumes to the atomic model (see Fig. 12 bottom). From the FSC, we can see that the improvement is not only due to masking, but there are improvements inside the mask too.

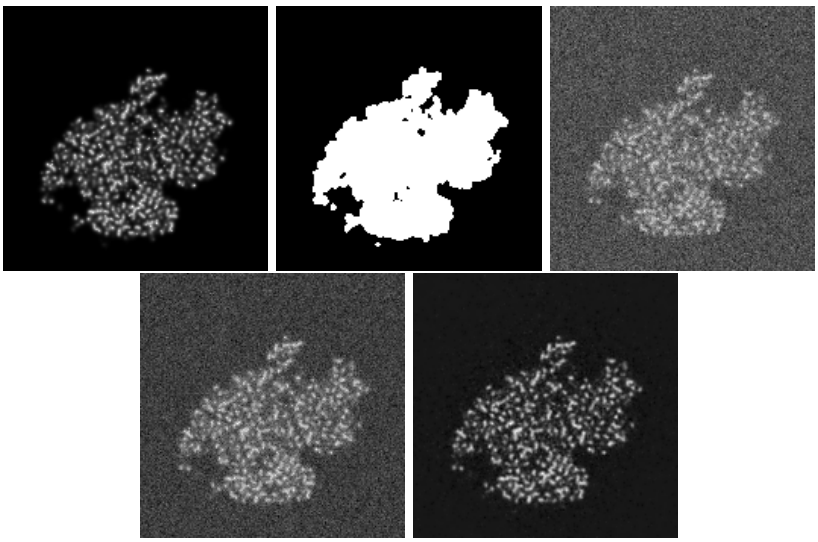


Figure 11: (This Figure corresponds to Experiment 11.) From top to bottom, left to right, slice 84 of: the atomic model, the mask, one of the half maps, the average of the two half maps, and the non-linearly processed map.

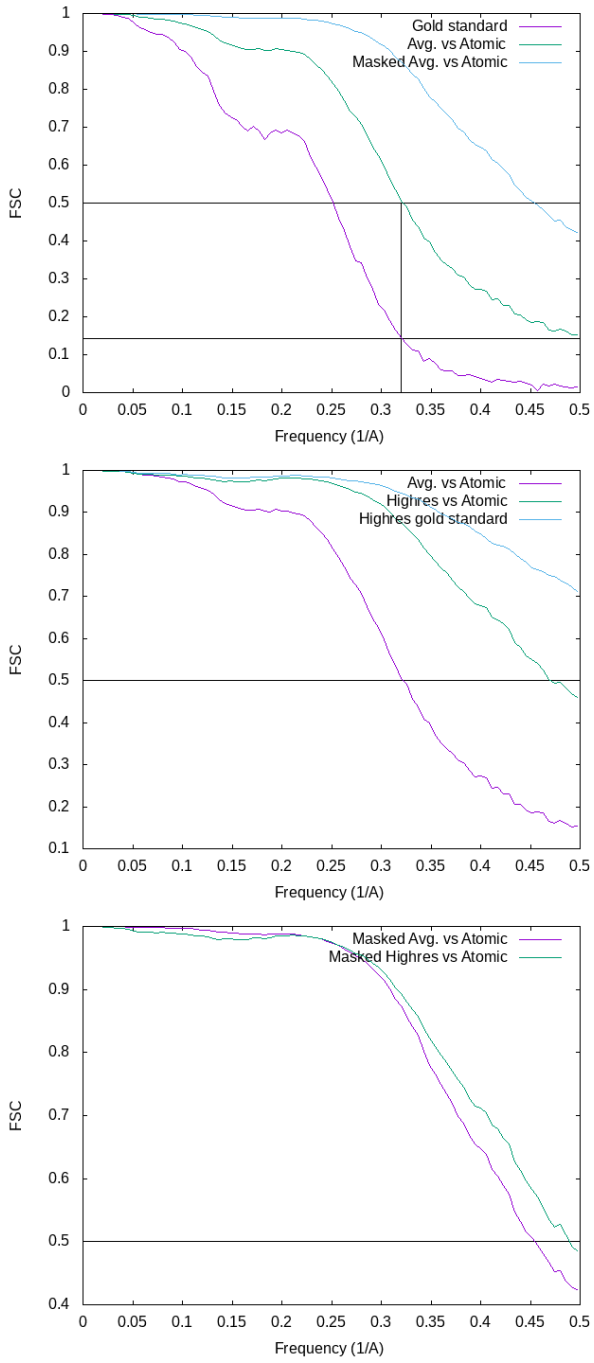


Figure 12: (This Figure corresponds to Experiment 11.) Top: FSC corresponding to linear processing. Middle: FSC corresponding to non-linear processing. Bottom: Comparison of the FSCs between linear and non-linear processing after masking.

Experiment 12. Change of regime of the FSC

We selected 34,268 particles of the ribosome dataset EMPIAR 10028 (the initial volume was constructed as in Experiment 5). We filtered them with a lowpass filter starting to decay at 5Å and totally vanishing at 4.6Å with a raised cosine decay in between. We applied 6 iterations of Xmipp highres with an increasing target resolution (this algorithm is multiresolution, and it downsamples the input images to a pixel size such that the target resolution is at 2/3 of the Fourier frequency range; this is the reason why the FSCs of the different iterations, shown in Fig. 13, finish at different frequencies). In this plot, we see that at 4.6Å there is a clear regime change in the FSC for highres. The same change is observed in Relion's autorefine (this time at 4.8Å). The comparison of the two maps (highres and Relion; see slices in Fig. 13) reveals that there are two fundamental differences: 1) the noise outside the molecule is practically inexistent in highres (there is no hard masking, but there is an automatic detection of the statistical distribution of the noise under several transformations) [82], and a dampening by significance in the whole volume); 2) the details of highres inside the molecule are finer than in Relion. Remember also that the FSC at a given frequency can be understood as the correlation in real-space of two bandpass filtered volumes at that frequency [86]. The fact that the FSC on the left does not go to 0 simply indicates that the two volumes have a medium correlation to each other, disregarding their absolute amount of energy. Actually, if part of the postprocessing dampens the noise at high-frequency, then the two half-maps at high frequency may correlate better despite having much less energy due to the dampening.

This experiment illustrates the fact that the 0.143 FSC threshold is based on an assumption of linearity of the 3D reconstruction process that the 3D reconstruction algorithm may not fulfill (in particular, 3D Fourier gridding is not linear, nor are the noise removal steps of highres; additionally, although there is no explicit masking, the noise reduction of highres results in a smoothing of the solvent outside the particle, see the slice of Fig. 13, that is a soft mask that induces correlation at high frequency). Rather than a specific threshold, the resolution of a map is better determined by a change in the FSC behavior. However, it is more difficult to give a formal criterion to detect this point.

Experiment 13. The gold standard is not necessary

We selected 16,703 particles of the brome mosaic virus dataset EMPIAR 10010. The initial volume calculation was performed as in Experiment 3. We ran Xmipp highres following the gold standard strategy of splitting the dataset into two independent halves that never share information over the iterations. The results of this execution are shown in Fig. 14. We also ran Xmipp highres with a stochastic split at each iteration (9 iterations were performed). At each split, each experimental image may go to either half so that the halves are not independent anymore over iterations. As shown in Fig. 14, there is no significant difference between the two strategies. Icosahedral symmetry was used during the reconstruction. Similar results have been obtained with other macromolecules with lower symmetry (even C1).

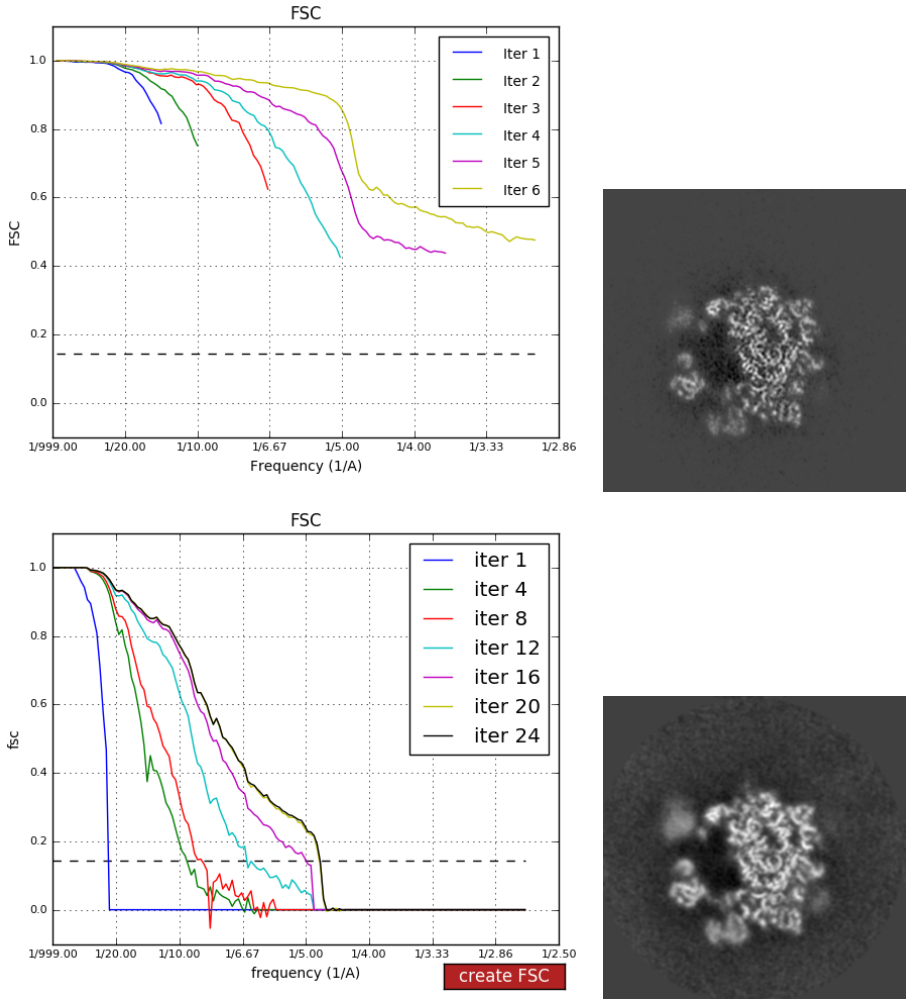


Figure 13: (This Figure corresponds to Experiment 12.) Top row: FSCs and a central slice of Xmipp highres when the input data is filtered at 5Å, the filter fully vanishes at 4.6Å. Bottom row: Same information for Relion autorefine.

This experiment shows that there is not a strict need to separate the dataset into two independent halves. Overfitting can also be avoided if specific measures are taken during the 3D reconstruction process (in the case of highres, the projection weighting scheme depending on the confidence on its angular assignment and the noise attenuation steps performed after reconstructing).

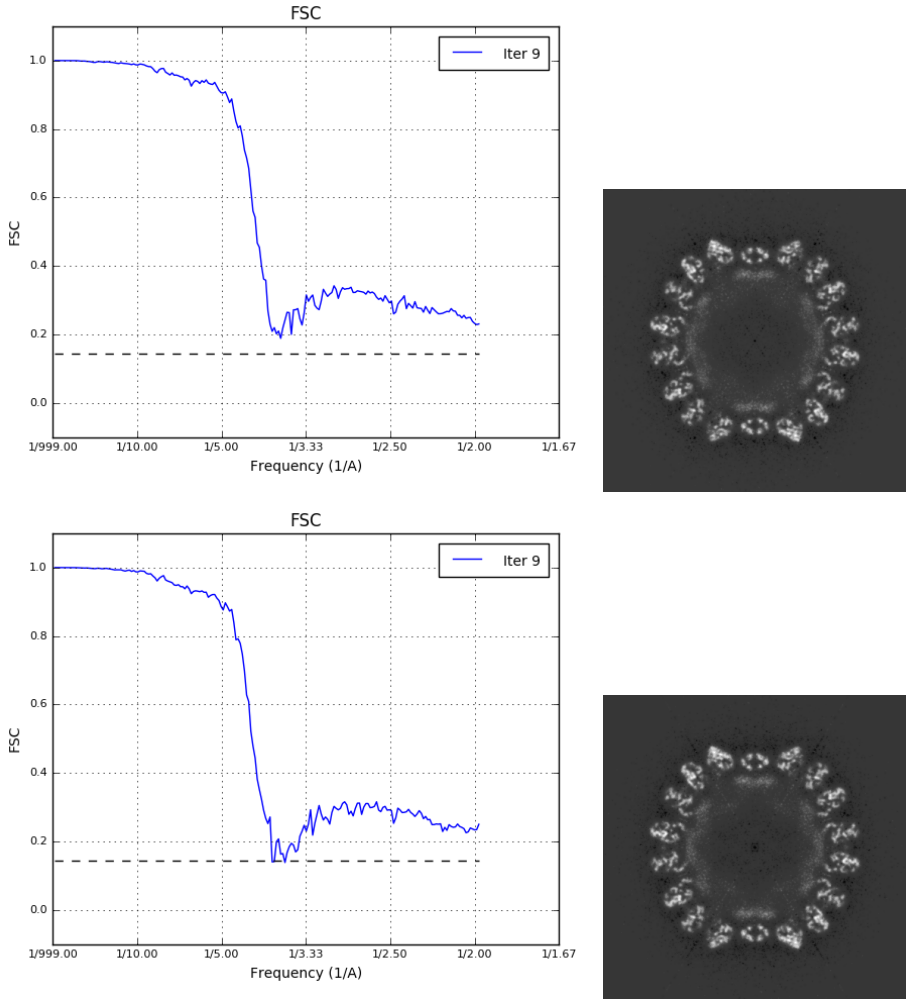


Figure 14: (This Figure corresponds to Experiment 13.) Top row: FSC and a central slice of Xmipp highres running the gold standard. Bottom row: FSC and a central slice of Xmipp highres running in stochastic split mode.

Experiment 14. Phase randomization

We selected 34,268 particles of the ribosome dataset EMPIAR 10028 (the initial steps of the processing were like in Experiment 3). We randomized the

phases of these particles beyond 5Å and performed a 3D angular assignment and reconstruction using Xmipp highres and Relion autorefine. In both cases, we observe the change of FSC regime exactly at 5Å as expected, meaning that both algorithms do not overfit the Fourier components beyond 5Å. You may compare the shape of these FSC curves obtained when images are phase randomized and those obtained when the randomization is performed at the level of volumes (see Suppl. Material Experiment 9). The alignment of the phase randomized images using Relion autorefine took more than 7h using 2 GPUs. In contrast, the phase randomization curve calculated by Relion postprocessing took 17s (this time difference shows why practitioners prefer the result from the postprocessing).

This experiment shows that the FSC obtained with phase randomization at the level of images (Fig. 15) is not the same as the FSC obtained at the level of volumes (Fig. 9). Note that if the phase randomization is performed on the images, the high frequencies of the images cannot contribute to determining the alignment of the particles. However, if the phase randomization is performed only on the volumes, the high-frequency components of the images have contributed to determining better estimates of the angular orientation of the images, and the phase randomization loses its validation power.

Additionally, at the level of volumes, there is an overestimation of the FSC randomized (see that in Fig. 9, the phase randomized FSC crosses the standard 0.143 threshold at 6.6Å, instead of 10Å that is where the phase randomization in the images really started, note that this number 10 was arbitrarily chosen for the purpose of this particular example). This overestimation of the phase randomized FSC is later translated into an overestimation of the corrected FSC through Eq. (4) of Chen et al. [12].

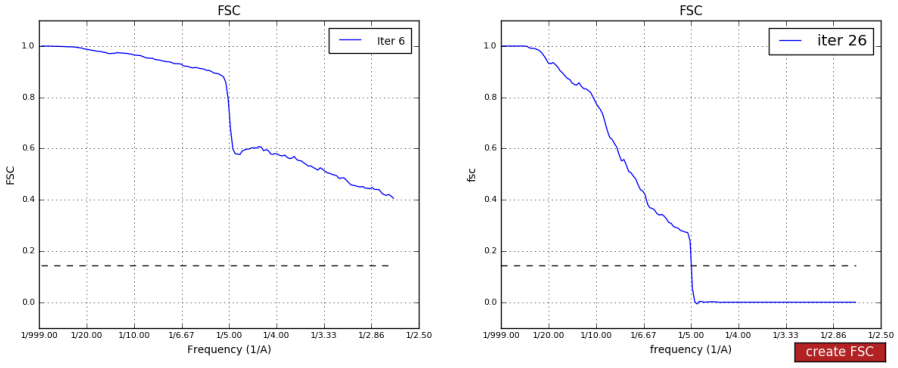


Figure 15: (This figure corresponds to Experiment 14.) FSC of Xmipp highres (left) and Relion (right) after randomizing the phases of the experimental images beyond 5 Å.