Cross-validation tests for cryo-electron microscopy using an independent set of images

by

Sebastián Ortiz Girón

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Physical Sciences

Advisor: Pilar Cossio, Ph.D.
co-advisor: Boris A. Rodríguez,Ph.D

Biophysics of Tropical Diseases - Max Planck Tandem Group
Faculty of Exact and Natural Sciences
December, 2019

Acknowledgments

To the University of Antioquia, Colciencias and the Max Planck Society for funding and supporting this work which is part of the FP44842-292-2017 project.

To members of the Biophysical of Tropical Diseases group, where this work was developed, for their advice, support and friendship.

I would like to thank to my advisors, Dr. Pilar Cossio and Prof. Boris Rodríguez for their constant feedback and the invaluable guidance.

I would like to remark the labor of Dr. Pilar Cossio. I will be in infinity gratitude for her sincere advice, her motivation and exceptional accompaniment.

I want to thank my family for their unconditional support, Mrs. Myriam for her affectionate accommodation and especially Laura Martínez for following this path by my side.

To my parents, brothers and Laura

Abstract

In addition to the chemical composition, information about the three-dimensional structure of a biomolecule is vital for understanding its biological function. For many years, resolving structures of biomolecules was exclusive of X-ray crystallography and nuclear magnetic resonance (NMR) techniques. However, due to technological and software improvements, cryo-electron microscopy (cryo-EM) has emerged as an alternative for resolving complexes that were infeasible for crystallization or too large for NMR. Currently, cryo-EM is able to provide near-atomic resolution and close-to-native structures . Moreover, it enables extracting dynamical information, such as free-energy landscapes, from thermal states in the micrographs. The "resolution revolution" in cryo-EM has provoked an avalanche of reported cryo-EM maps. Recent statistics show an exponentially-growing number of reported maps spatially resolved by cryo-EM with their mean resolution decreasing from $\sim$ 10 Å (in 2013) to 4 Å (for 2018).

The resolution revolution brings with it the need of creating robust and reliable methodologies to validate the increasingly large number of maps. Some advances have been done along these lines: the tilt-pair analysis , the gold-standard procedure and the high-frequency randomization have shown to be reliable validation tools. However, it has recently been shown that these methods remain sensitive to overfitting (treating noise as true signal) and subjective criteria.

In this work, I will present a novel methodology for validating cryo-EM maps. The method is based on cross-validation criteria where the reconstructed maps are compared against a set of experimental images (raw data) not used in the reconstruction procedure. Such comparison is carried out by calculating the probability that an image is the projection of a given map. The information from these probabilities led us to propose two validation criteria, which are tested over three well-behaved systems and two systems that present overfitting. The results prove that our methodology is able to identify overfitted maps.

Contents

List of figures

List of Tables

List of Abbreviations

cryo-EM	cryo-elecron microscopy
SSNR	Spectral signal-to-noise ratio
NJSD	Normalized Jansen-Shannon divergence
FSC	Fourier shell correlation
KLD	Kullback-Leibler divergence
SNR	Signal-to-noise ratio
FT	Fourier transform
DEDs	Direct-electron detectors

1 Introduction

Cryo-electron microscopy (cryo-EM) has become a mainstream technique for resolving biological structures in a close-to-native environment and at near-atomic resolution. Moreover, cryo-EM has the potential for monitoring the dynamics of structural changes induced by thermal fluctuations or chemical interactions [2, 3]. Cryo-EM resolves a biomolecule structure by analyzing hundreds of thousands of two-dimensional projections of the biomolecule's electron density. Such projections are obtained by passing a coherent electron beam through a frozen sample, which contains multiples copies of the biomolecule oriented randomly. With advanced algorithms a 3D density map is reconstructed from the projections, and ultimately an atomic model is fitted into the map.

Most of the reconstruction algorithms are based on the Fourier slice theorem, which states that a central slice (a plane that passes through the origin) of the Fourier transform (FT) of a 3D function corresponds to the FT of the 2D projection of the function over a plane parallel to the slice. The basic idea is to assign an orientation to each experimental cryo-EM projection and to calculate its FT. If there is enough sampling of the orientation space, and it is assumed that all projections correspond to the same conformation, it is possible to obtain a 3D object in Fourier space by combining all the FT-projections as central slices in the corresponding orientations. The inverse FT of this object will be the 3D electron density map. Of course, behind this simple description there are important tasks to solve such as the accuracy of the orientation assignment, reference bias, heterogeneity and others. Moreover, the difficulty is enhanced by the very low signal-to-noise ratios (SNR) present in cryo-EM data sets. These

typically range from 0.1 to 0.001.

In fact, the low SNRs make it challenging for the reconstruction algorithms to distinguish noise from true signal. This has serious consequences because noise can be aligned as true signal (termed as *overfitting*), and hence artificial features can build-up the map. In extreme cases, a reference object can be generated from pure-noise images [4, 5]. For this reason, validation and map quality assessment play a vital role for cryo-EM [6, 7].

Currently, there are some standard methodologies to reduce the risk of overfitting. The most common is the 'gold-standard' procedure [8]. In this method, the image set is divided into two subsets and two independent reconstructions are generated from each half set, then the two maps are correlated for assessing the resolution. Additional strategies to validate the reconstructions are the tilt-pair analysis [9] and the high-frequency substitution [10].

Despite this progress, in the recent Map Challenge [11] it has been shown that the protocols remain user-dependent and there can be biases due to processing workflows, which can lead to overfitted cryo-EM reconstructions. For example, the reported values of the resolution in the atomic model (from the Protein Data Bank or PDB) and in the map (from the Electron Map Deposition Bank or EMDB) are different for about 30% of the deposited data [12]. Moreover, it has been found that more than 70% of the maps in the EMDB have moderate to low agreement with the model, mostly because of the limited resolvable features of the maps [7].

Cross-validation methods are widespread statistical tools for measuring the prediction accuracy of a model over a control dataset not used in the model training. Following this idea, we here propose an unbiased strategy that validates cryo-EM reconstructions using a small control set of images that are omitted from the refinement process. The prediction accuracy will be measured by calculating the probability that control images are projections from the refined maps [13, 14]. Based on such probabilities, we propose two criteria for monitoring the quality and correctness of the refinement procedure. We test the method on different systems and asses its effectiveness to discriminate overfitted maps.

1.1 Problem statement

Validation tests are fundamental to recognize correct maps from those contaminated by noise alignment or template bias in cryo-EM. Moreover, the high rate of reported cryo-EM structures has forced the community to propose more robust validation criteria [11]. Although some advances have been done in this direction, cryo-EM map validation is still an open issue, since it was recognized that current reconstruction strategies remain prone to overfitting.

1.1.1 Objectives

General objective

Propose a robust methodology for validating cryo-EM maps using an independent set of images not used in the refinement procedure.

Specific objectives

- Implement a strategy to compare the refined maps against the validation set.
- Propose robust criteria to identify overfitted maps.
- Create a public code to apply the methodology.

2
Cryo-electron microscopy

There are two main stages to overcome for solving a biological 3D structure by cryo-EM: the experimental part, including sample preparation, data collection and data processing (Fig. 2.1). In this chapter, we will describe each stage briefly, emphasizing the processing part. Basic concepts used in cryo-EM are explained in ref. [15].

2.1 Experimental stage

The sample preparation initializes with the purification, using biochemical methods, of an aqueous solution containing multiples copies of a biomolecule. Since the high-vacuum of the electron microscopes causes dehydration, the biomolecules must be fixed in a close-to-native state: negative staining and vitrification are the most commonly used methods [16].

Jacques Dubochet and his group firstly proposed vitrification [17]. They vitrified a biological sample into a thin film of amorphous ice by freezing it at cryogenic temperatures ($\sim -196°$C) using liquid nitrogen. This procedure was done quickly enough to prevent the formation of ice-cubic cells, which would diffract the electron beam [17]. The single-particle approach of imaging of cryo-fixed samples using an electron microscope became known as cryo-electron microscopy (cryo-EM). In 2017, J. Dubochet was awarded, together with Richard Henderson and Joachim Frank, the Nobel prize in chemistry because of their contribution to the development of the cryo-EM field.

After vitrification, the specimen is loaded onto a thin support film and it is imaged by a

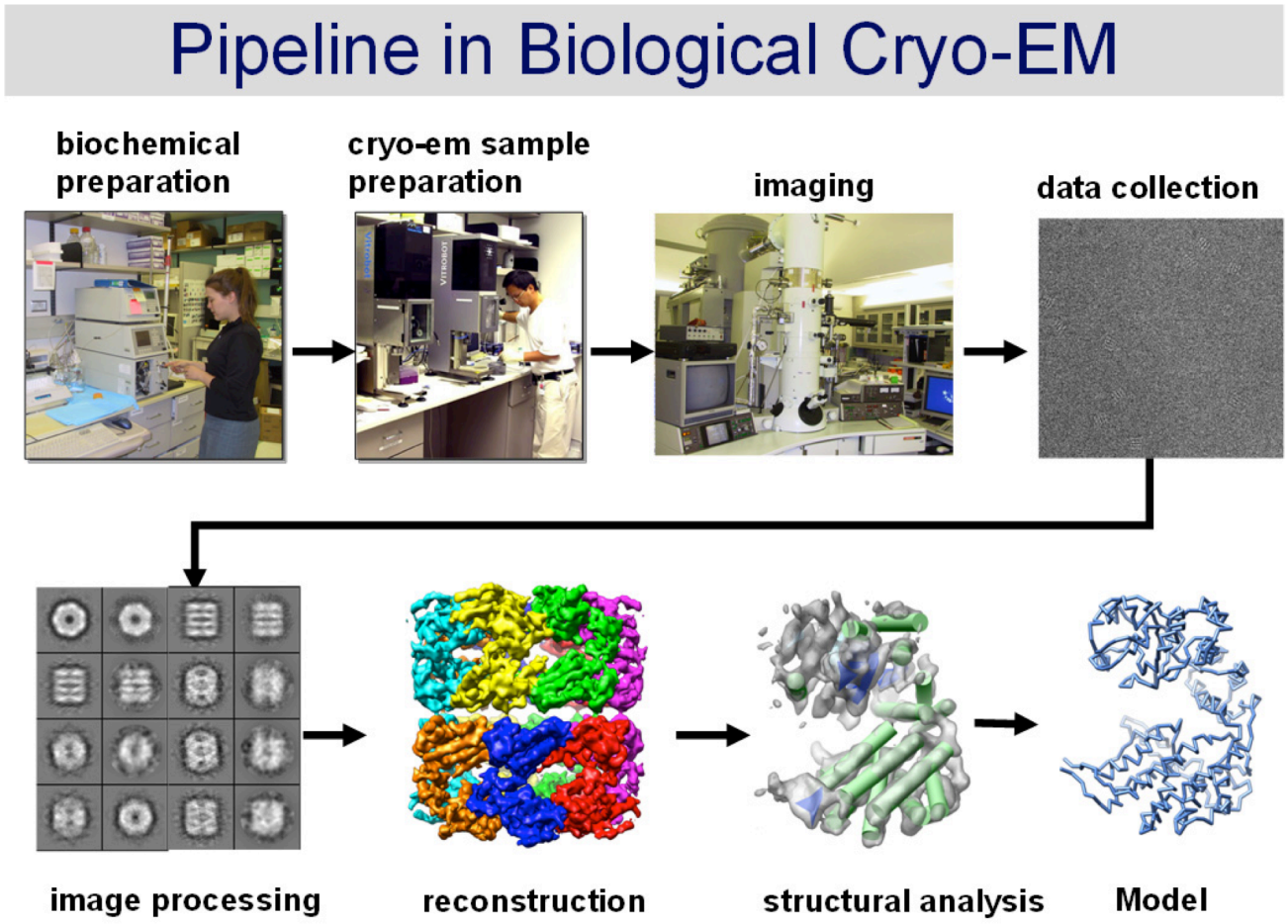


Figure 2.1: Graphical summary for a standard pipeline to solve a biomolecular structure using cryo-EM. The experimental stage includes the specimen purification, vitrification and imaging by a TEM. The series of micrographs (top-right) undergo data processing until a 3D map is obtained. Finally, an atomic model is fitted from the map.

conventional transmission electron microscope (TEM) working from 200 to 300eV. A 2D projection of the sample (called micrograph) is generated (*e.g.*, top-right in Figure 2.1).

Since imaging biological samples requires low radiation doses to avoid chemical bond breaking, the micrographs contain a strong noise component (typical SNRs are in the order of 10^{-3}). This imposed a substantial limitation to the cryo-EM technique. However, the introduction in 2012 of direct-electron detectors (DEDs), [18, 19] with relatively high efficiency (which expresses the degradation over the incoming SNR due to the recording error), led cryo-EM to achieve near-atomic resolution of cryo-EM maps for the first time. DEDs can detect individual electrons allowing a pixel-resolution recording. Moreover, due to the high frame rate of the DEDs, the radiation dose can be spread among different frames creating an exposure movie. These movies are used as input for the computational processing part.

For more details about experimental procedures, such as sample preparation, apparatus description, and data recording, see [16].

2.2 Data processing

In a typical cryo-EM experiment, 2D images of the vitrified sample containing multiples projections of the biological system are recorded in a series of frames (or movies). The analysis of such data is typically divided into three stages. The first stage is the pre-processing of the micrographs, which includes picking, cleaning, and clustering of the individual projections of the biomolecule (henceforth referred to as particles). The second stage is the reconstruction, refinement, and validation of the 3D electron density map generated from the particles. The final stage consists of fitting an atomic model into the map (see Figure 2.1).

Since the interaction between the electron beam and the biomolecules induces mechanical movements, averaging frames of a movie can not be done directly. Instead, motion correction is performed [20, 21], and after a correct alignment, the frames are averaged, producing a micrograph with a higher signal. Using automatic [22], semi-automatic [23] or even manual strategies, the particles are picked from the micrograph. Picking algorithms have to be very careful to avoid template bias because, in extreme cases, pure noise images can be detected as a particles [4]. This can furtheron lead to overfitting (*i.e*, treating noise as true signal) in cryo-EM workflows.

Once the particles are picked, most algorithms perform a 2D classification, where the particles are clustered according to their similarity over rotations in a plane [1, 24]. Particles from the same cluster are averaged. Bad particles (with artifacts because of radiation damage, overlaped projections, etc) can be detected because they produce blurred averaged-projections. The particles chosen from the good clusters are used to generated a low-resolution *de-novo* map [25, 26] (see below how to go from 2D images to a 3D map). Afterwards, a 3D classification that clusters the particles according to certain underlying states present in the experi-

mental sample is performed [27]. The number of 3D states will be limited by the total number of particles. If it is excessively large, the conformational states will have low populations and hence the information will not be enough to reconstruct the corresponding maps. For most cases, 3D classification is used to clean even more the data. The final subset of particles belonging to a unique state is used to obtain a final reconstruction and perform a refinement of the map.

For a complete introduction to processing techniques in cryo-EM reconstructions, we recommend the books [28, 29, 30]. A practical pipeline is described in [31]. Additionally, refs. [32, 33, 34] offer a historical review of cryo-EM development. Below I will focus on explaining in detail the 3D refinement procedure and map validation.

2.2.1 3D reconstruction

Most reconstruction algorithms are based on the *weak-phase approximation* [35], which describes the scattered electron wave function as a phase shift of the incident electron wavefunction, *i.e* $\Psi_{exit} = e^{-i\varphi}\Psi_{incident}$. Such shift φ is proportional to the effective atomic potential of the sample, $\varphi \propto \int V(z)dz$ (assuming $\hat{e}_z$ as the propagation direction). If a thin enough sample (see Appendix) and small angles scattering are considered (this latter assumption is fulfilled by typical cryo-EM experiments, where electron beams are generated at 200-300 keV energies), then the exit wave function is reduced to $\Psi_{exit} = (1 - i\varphi)\Psi_{incident}$.

Taking into account the optical aberrations effects induced by the microscope, the image formation can be modelled by the linear model

$$\mathfrak{F}\{X\} = \text{CTF} * \mathfrak{F}\{\mathbf{P}_{\vec{\mathbf{d}}} V\} + N, \tag{2.1}$$

where the CTF is the FT of the point-spread function of the microscope (see the Appendix), X represents the intensities of a particle (*i.e* a projection of the biomolecule obtained from experimental data), V is the underlying 3D electron density map, $\mathbf{P}_{\vec{\mathbf{d}}}$ is the projection operator along the direction $\vec{\mathbf{d}}$ and N represents the noise (see chapter 9th of ref. [36] for a comprehensive explanation about how such model arises).

Most reconstruction algorithms use an iterative approach to generate better maps at each step. From an initial map V, the best orientations for each particle are calculated using the projection-matching method [37]. Once the orientations are assigned to all particles, a new map is reconstructed using the Fourier-slice theorem [37] which states that a central slice normal to $\vec{\mathbf{d}}$ of a 3D map corresponds to the FT of a projection along $\vec{\mathbf{d}}$. The algorithm is halted when the map quality stops improving.

Algorithms based on maximum-likelihood and Bayesian inference [1, 25, 38, 39], instead of projection-matching, are currently the most popular. They have showed huge power to

resolve the underlying map with the minimal user intervention at accessible computational costs [40].

2.2.2 Validation & Resolution

The gold-standard procedure

Most of the reconstruction algorithms are based on iterative schemes, which initialize from a low-resolution map and improvements are done using, for example, Bayesian or likelihood-based methods. For 'perfect' images, these algorithms should convert to a unique and unambiguous structure. However, if there are high noise levels, differentiating between noise and true signal is difficult. In this case, orientation assessment may be incorrect due to spurious correlations between the noise components and the features from the reference map. Refinement iterations can yielded a template-biased ('overfitted') final map. A famous example is the Einstein-from-noise model, where Einstein's face is reconstructed from pure-noise images [5, 4].

Moreover, even if one uses a low-resolution map, matching noise at higher frequencies is still possible, which can induce the generation of artificial features. Overfitting can be difficult to detect since there is no exact ("true") map for comparison. In several cases, a visual analysis of an expert is required. However, objective tools for overfitting detection such as the tilt-pair analysis and the Fourier Shell Correlation (FSC) have been proposed. We will make a brief introduction to the FSC since it is, by far, the most common tool currently used.

The FSC is 1D curve which measures the correlation between two objects as a function of the spatial frequency. It was proposed by Saxton and Baumeister for image comparison [41] and it was generalized for 3D maps [42].

Let F_1, F_2 be the 3D FT of the maps M_1 and M_2, respectively. The FSC is defined as

$$\mathrm{FSC}(k, \Delta k) = \frac{\sum_{q \in S_{k,\Delta}} F_1(q) * F_2^*(q)}{\sqrt{\sum_{q \in S_{k,\Delta k}} |F_1(q)|^2 \sum_{q \in S_{k,\Delta k}} |F_1(q)|^2}} \tag{2.2}$$

where all the sums run over the Fourier components belonging to the shell $S_{k,\Delta k}$ of radius k and width Δk. By definition, the FSC is bounded between 1 and -1. These points correspond to a perfect correlated or anti-correlated shells, respectively. FSC $= 0$ indicates uncorrelated shells.

Both maps (from set 1 and 2) must have the same statistical significance, *i.e.*, they must correspond to the same structural state. This is guaranteed by performing subsequent non-supervised 3D-classification procedures) for the selecting the particles to run the final gold-standard refinement). For the systems tested in this work, 3D classification was performed

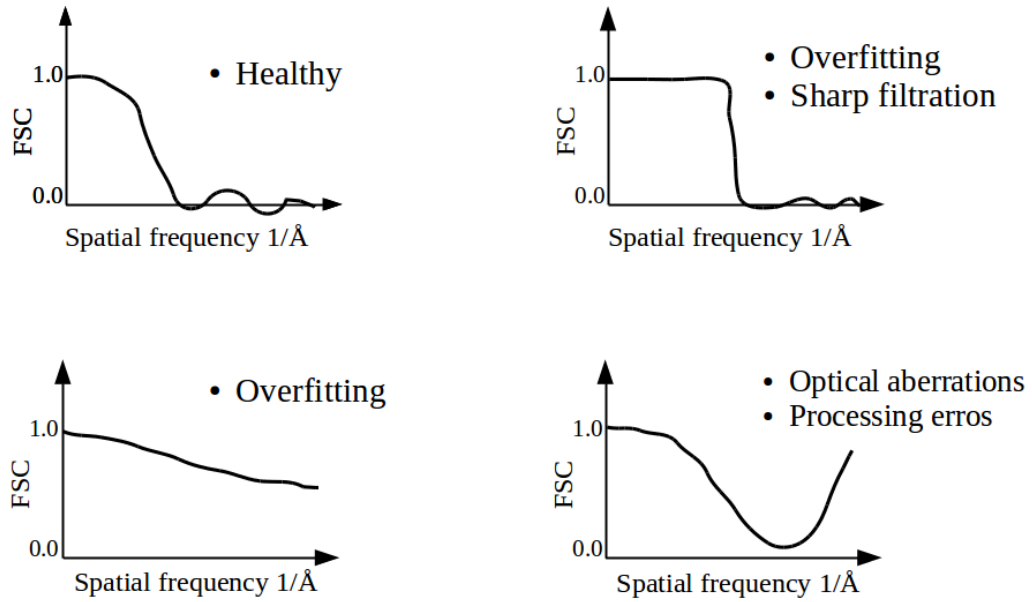


Figure 2.2: FSC curves encountered for cryo-EM maps. Some FSC curves show signs of overfitting. This figure is based on the figure 3.4 from ref. [45].

previous to the refinement.

Besides, in order to avoid artificial correlations, both maps should be independent. Following this idea, the so-called 'gold-standard' procedure consists on dividing randomly the particle set in two halves, and generating independent reconstructions from each half. It is important that reconstructions are performed using independent alignment processes.

The majority of the validation and quality assessment criteria of cryo-EM data is based on the FSC. The popularity of this method is motivated by its relation to the spectral signal-to-noise ratio (SSNR) (the SNR in Fourier shells) [43, 44, 45],

$$\mathrm{FSC} = \frac{SSNR}{1 + SSNR} \,. \tag{2.3}$$

The FSC can be interpreted as a measure of the noise level added at each frequency shell. Therefore, the FSC curve is used as validation criteria [6, 7, 8]. If the reconstruction was performed adequately, the FSC will begin near 1 for the lowest frequencies, decay smoothly and oscillate around zero after a certain frequency less or equal than the Nyquist frequency (see figure 2.2). Any behavior differing from the one mentioned above indicates that some error/overfitting was performed in the reconstruction procedure[45].

Chen *et. al.* [10] propose a method to obtain a more robust validation criteria from FSC curves. They create an image stack identical to the original one, except that, beyond a certain

frequency, the Fourier components of the original images are replaced by random complex numbers which follow a similar distribution to the background noise. The FSC curve between the two maps generated from the manipulated and the original images give an additional validation criteria.

However, some authors [46, 44] have criticized the FSC arguing that it is based on the assumption that both noise components of M_1 and M_2 are orthonormal, which might not be true for all cases.

Resolution and map quality assessment

Quantifying the quality of cryo-EM maps is still under debated in the field, mainly due to the statistical nature of the cryo-EM resolution concept. Unlike with X-ray crystallography where the resolution has physical sense related to Bragg's law, in cryo-EM the resolution measures the statistical consistency of the data. Although several approaches have been proposed, FSC-based methods are the most widespread. For a review of the resolution measure see refs. [47, 45]. The debate has centered on the convenience of the use of a single number for resolution assessment, and how this number is obtained [46, 44, 48, 12].

In the practice, most studies use a fixed-threshold approach. Based on the relation between the FSC and the SSNR (eq. 2.3) authors have proposed the resolution as the frequency where the FSC curve reaches the 0.143 or 0.5 value [49, 41]. The 0.5 threshold indicates that there is as much signal as noise after comparing both maps generated with the gold-standard procedure. The 0.143 threshold has the same significance but comparing a half-map to the hypothetical perfect map generated from the entire dataset.

Better resolution estimates are obtained with reference-free pipelines using the 1/2 bit non-fixed FSC threshold [46, 50]. The local resolution in a map can be evaluated using the background noise of the reconstruction [51] or by masking different regions with the FSC [52, 53]. Predictability of the particle alignment provides quality indicators of the reconstruction [48, 54]. Moreover, several metrics that monitor cross-correlations in real or Fourier space between the maps and models indicate the reliability of the resolution [12, 7, 55]. Recently, deep learning algorithms have been introduced to automatically classify maps into high, medium, and low resolution [56]. However, all these methods have the limitation that they do not use the raw data, which ultimately comes from the individual particles, but they only use the maps or models that are product of processing and averaging. For instance, in cryo-EM there is no cross-validation method, such as the R-free in X-ray crystallography [57], which uses an independent control set from the pure experimental data. This has therefore motivated us to develop cross-validation tests for cryo-EM.

3
Theory and methods

Cross validation is a widespread statistical tool for measuring the prediction accuracy of a model over a control dataset not used in the model fitting. It is commonly used in machine learning for expressing the agreement between a model and the observed data. In structural biology, cross-validation tools have been used for addressing the quality of the atomic models fitted to density maps. A prominent example is the free R-factor used to validate models resolved by X-ray crystallography [57]. In cryo-EM, there are also some proposed methodologies for validating atomic models fitted into 3D maps [58, 59]. However, we are interested in the map validation stage where there is no a cross-validation approach available.

In this work, inspired by the main ideas of cross validation, we propose to validate cryo-EM maps (generated from the gold-standard procedure) by comparing them against a set of particles (Ω) not used in the reconstruction procedure. Such comparison is done by calculating the probability $P_{M\omega}$ that a particle $\omega \in \Omega$ is the projection of a given map M [13]. Based on these probabilities we propose two validation criteria.

The chapter is organized as follows. First, we describe the general theory for the two validation tests. Then, in the Methods section, we describe the details about the benchmarks systems and the reconstruction procedure.

3.1 Theory: Novel cross-validation tests for cryo-EM

We propose a statistical framework for the cross-validation of cryo-EM reconstructions. First, and foremost, the validation analysis is done over a small control set of particle images not used in the refinement process. This independent set should give an unbiased estimate of the quality of the reconstructions.

From a given pre-processed particle set, a small number N_ω of particles are selected randomly as the test set Ω. The rest of the particles will belong to the reconstruction set. The particles from reconstruction set are submitted to a refinement procedure following the gold-standard guidelines. Therefore, the particles from the refinement are split in two halves. Starting from a low-resolution map used as reference (to reduce overfitting and bias template [1]), the refinement is done iteratively, and independently for each half. A series of maps M_i with $i = 1, 2$ for each half are generated (see the Methods section below for details on the refinement procedure).

Fig. 3.1 shows the work-flow of our novel methodology. The refinement is done following the gold-standard procedure (Fig. 3.1–left), where two reconstructions are generated at each iteration step. These two reconstructions are validated using the control particle set (Fig. 3.1–right). At each iteration, the two maps are low-pass filtered to different frequency cutoffs, k_c. Then, the BioEM probabilities are calculated for each low-pass filtered map. The main idea is to measure map probability with the cumulative log-posterior (**Test 1**), and compare these probability distributions using the normalized Jensen-Shannon metric (**Test 2**). Below we will explain each test in detail.

3.1.1 **Test 1:** The BioEM posterior probability

The first direct validation criteria is the cumulative log-probability of each map M given the test set of particles. This measures how probable a 3D map is given the unbiased experimental data. The cumulative evidence should increase or remain constant as a function of the iteration. Moreover, it should also increase as a function of a low-pass filter frequency cutoff. Failing this test is a prime indicative that there is a problem in the refinement process.

The BioEM posterior probability

The probabilities involved in our methodology are calculated using the BioEM (Bayesian Inference of electron microscopy images) algorithm developed by Cossio and Hummer [13]. The BioEM method [13] uses a Bayesian framework to quantify the consistency between an experimental image ω and a given map M (or atomic model) by calculating the posterior probability $P_{M\omega}$.

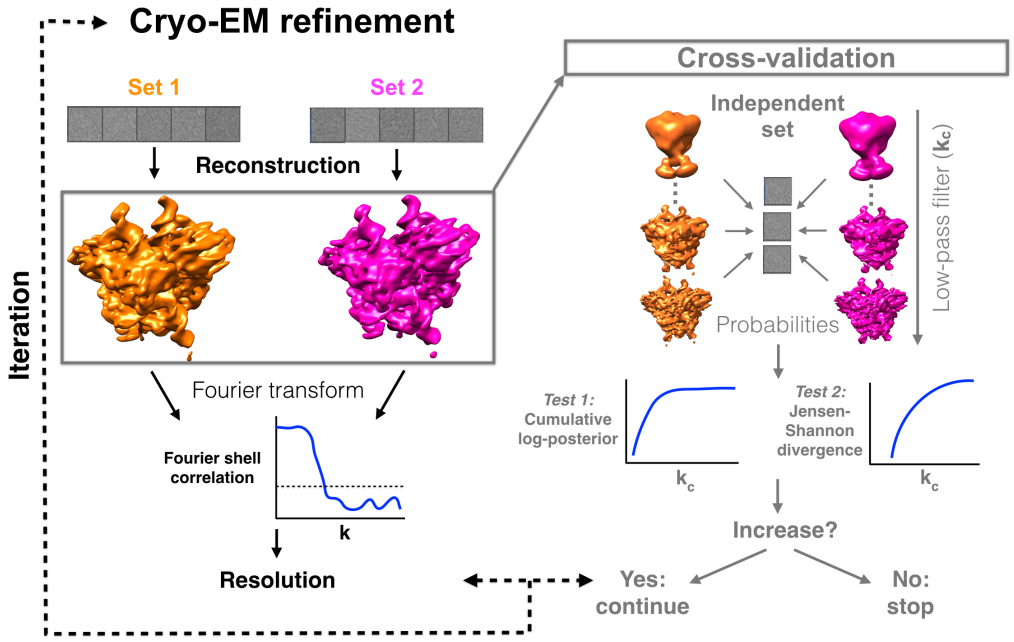


Figure 3.1: Cross-validation protocol for unbiased map validation in cryo-EM. (**left**) Gold-standard refinement procedure in cryo-EM. Two particle sets are used to generate two independent reconstructions. These reconstructions are compared using the Fourier shell correlation (FSC). A fixed FSC threshold is used to extract the resolution of the reconstructions. The process is iterated until the resolution stops improving. (**right**) Novel cross-validation protocol using a small control particle set. At each iteration of the refinement, the reconstructions are low-pass filtered to different frequency cutoffs k_c. The BioEM probabilities [13, 14], over the independent control set, are calculated as a function of k_c. Two tests validate the quality of the reconstructions: 1) the cumulative log-posterior and 2) the statistical similarity between the probability distributions (measured with a normalized Jensen-Shannon divergence). The results from both tests should increase as a function of the frequency cutoff. The maps represented correspond to the RAG1-RAG2 complex (see the Methods).

Unlike Cossio and Hummer's original work [13], we are not interested in ranking atomic models (or model ensembles). Instead, our aims are (i) to extent the BioEM utility for analyzing electron densities -maps and (ii) to evaluate how the consistency of the maps generated in a typical 3D-refinement procedure.These aims are challenging because the changes between the maps from different refinement iterations are, in general, not distinguishable with no noticeable structural changes, but rather by the information contained in the higher frequencies measured with the individual cryo-EM particles.

To calculate posterior probability, BioEM takes into account the relevant physical parameters (Θ) for the image formation: center displacement (x-shift and y-shift), normalization, offset, noise, orientation and CTF parameters (defocus, amplitude, and B-factor). $P_{m\omega}$ is cal-

culated by integrating out all these parameters

$$P_{m\omega} \propto \int L(\omega|\Theta, m)p(\Theta)p(m)d\Theta \,, \tag{3.1}$$

where $p(m)$ and $p(\Theta)$ are the prior probabilities of the map and parameters, respectively, and $L(\omega|\Theta, m)$ is the likelihood function.

Recalling the weak-phase approximation discussed in chapter 2 and assuming that the image is blurred by independent and white noise, distributed normally with mean zero and variance σ^2, then BioEM method postulates the likelihood function as

$$L(\omega|\Theta, m) = \prod_{x,y}^{N_{\text{pix}}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left[I_{\omega}^{(obs)}(x,y) - I^{(cal)}(x,y|m,\Theta)\right]^2/2\sigma^2} \tag{3.2}$$

where $I_{\omega}^{(obs)}$ is the intensity of the experimental image ω and $I^{(cal)}(x, y|m, \Theta)$ is the intensity of the image projected from the map m under the parameters Θ.

Prior probabilities of maps might play an important role in applications such as model-ensemble refinement [13], however, in our case we will consider $p(m)$ constant. Prior probabilities of CTF parameters are Gaussian functions, their means and variances are chosen from previous information about the particles. For several cases, (for example, the CTF parameters), these can be extracted during the experimental stage of cryo-EM. Therefore, the variances and means of the CTF priors are chosen to sample around these reference values in order of reduce the integration space. The prior probabilities of all others parameters are uniform over the integration intervals. In Eq. 3.1, the integrals over the offset, noise and normalization are performed analytically [13], and that over the center displacement is described in ref. [14]. The integral over the orientations and CTF defocus is done using a double-round algorithm, which is described in the BioEM algorithm section.

Similarly as in ref. [13], we define a noise model $P_{\text{Noise}} = (2\pi\lambda^2 e)^{-N_{\text{pix}}/2}$ where N_{pix} is the number of pixels and λ is the image variance (by default $\lambda = 1$). P_{Noise} is used as a reference to compare the posterior probabilities.

In summary, $P_{m\omega}$ measures how probable a 3D map (m) is given an experimental image ω. So we calculate the cumulative log-probability relative to noise $\sum_{\omega} \ln(P_{m\omega})/N_{\omega} - \ln(P_{\text{Noise}})$, over the control set with N_{ω} images, to determine how the probability of maps changes as a function of the low-pass filtering frequency and the refinement iteration step.

3.1.2 **Test 2:** The similarity between the posterior distributions

The second cross-validation test consists on measuring the similarity between the probability distributions of the two reconstructions (generated from the two halves of the gold-standard

procedure) as a function of a low-pass frequency cutoff and refinement iteration. We expect that as more frequencies are added to the reconstructions, more noise is added, and the probability distributions are less similar.

Normalized Jensen-Shannon divergence

Measuring a distance among probability distributions is a common task in statistics which it is of key importance in fields such as applied statistics, machine learning and Bayesian inference. The most popular measure is the Kullback-Leibler divergence (KLD) [60, 61] defined as

$$D_{KL}(P|Q) = \sum_x P(x) \log[P(x)/Q(x)]$$

where the sum runs over all the realizations of the distributions $P(x)$ and $Q(x)$. Thomas' book [62] presents a detailed study about KLD and other statistical measures. However, KLD has several limitations: it is not symmetric and it is not normalized. To overcome this, we define a metric that is the Jensen-Shannon divergence [62, 61] normalized by the individual Shannon entropies

$$\text{NJSD} = \frac{\sum_\omega [P_{1\omega} \ln(P_{1\omega}/M_\omega) + P_{2\omega} \ln(P_{2\omega}/M_\omega)]}{2(\sum_\omega P_{1\omega} \ln(P_{1\omega}) \sum_\omega P_{2\omega} \ln(P_{2\omega}))^{1/2}}, \tag{3.3}$$

where $P_{1\omega}$ and $P_{2\omega}$ are the probabilities of the reconstructions from set 1 and 2, respectively, over particles $\omega \in \Omega$, and $M_\omega = (P_{1\omega} + P_{2\omega})/2$. For simplicity of notation, we have omitted the dependency of the probabilities on the frequency cutoff k_c. To calculate Eq. 3.3, we normalize the posterior probabilities such that $P_{1\omega} + P_{2\omega} = 1$ for each particle ω, frequency cutoff and iteration.

In Eq. 3.3, the numerator measures the correlation between the probability distributions, and the Shannon entropies in the denominator play the role of a normalization factor. The NJSD metric is always positive, symmetric and its lower bound is 0 if and only if $P_{1\omega} = P_{2\omega}$ for all particles $\omega \in \Omega$.

In summary, the NJSD measures how similar are the probability distributions generated from the BioEM probabilities for the two maps from the refinement procedure. These distributions should become less similar as a function of the low-pass frequency cutoff and the refinement iteration step. In chapter 4, we will present in detail the results from these two validation tests over several benchmark systems. In the following, we describe the Methods for performing the calculations.

3.2 Methods

In this section, we describe in detail the BioEM algorithm and benchmark systems used for calculating the cross-validation tests.

3.2.1 The BioEM algorithm

The BioEM posterior calculation is time consuming due to the multidimensional integral that has to be calculated numerically. We found that the orientation integral has a critical influence on the BioEM posterior probability. Then, to ensure an accurate orientation assessment we divided the BioEM calculation into two rounds (Figure 3.2).

In the first round an all-orientations to all-particles algorithm is performed [14], carrying out an uniform sampling of the orientation space. For this propose a grid of 36864 quaternions is generated following the scheme described in ref. [63]. Moreover, to reduce the errors associated to the integration over CTF parameters, the particles were grouped into sets with similar experimental defocus with $0.4\mu m$ range, and an independent orientation search was performed for each group.

As the BioEM input map, we used the final reconstruction from the refinement with a broad mask and without low-pass filtering. The objective of the first round is to obtain the best orientations for each particle. An example of the BioEM input for the first round is presented in the Appendix.

In the second round, a local search around the best 10 orientations from the first round is performed. The zoom around each best orientation is done using 125 quaternions with approximately 0.01 grid spacing, resulting in 1250 zoomed-orientations for each particle. This procedure is described in detail in ref. [64]. The defocus of each particle is fixed to its experimental value. For this round, we select a representative subset of maps from the refinement iterations, each one of those is low-pass filtered using 8 frequencies cutoffs distributed uniformly from $1/(p_s\sqrt{N_{\text{pix}}})$ to $1/(3p_s)$, where p_s is the pixel size. All reconstructions were masked using the same broad mask as for round 1. An example of the BioEM input file for round 2 is presented in the Appendix.

Figure 3.2 shows a graphical summary of the workflow described here.

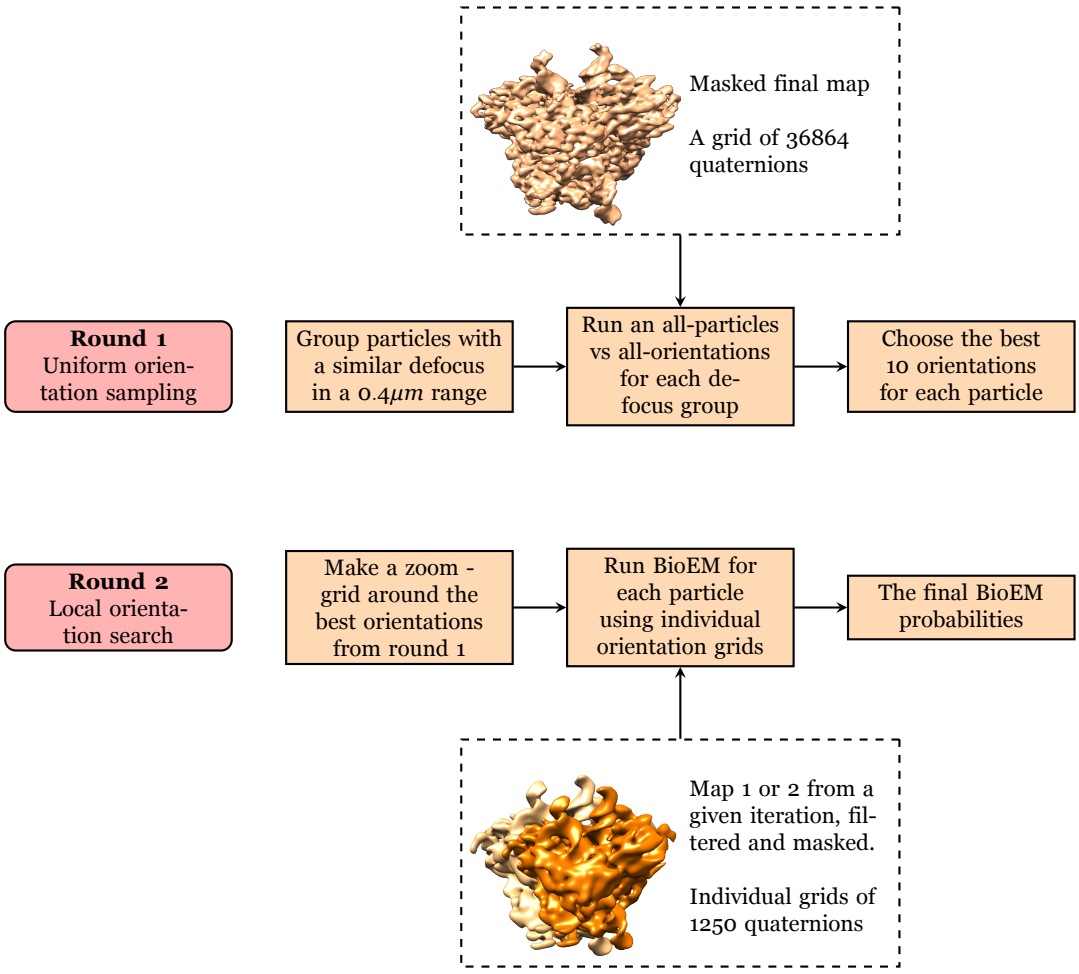


Figure 3.2: Summary of methodology employed for calculating the BioEM probabilities using a double round of orientational search.

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

3.2.2 Cryo-EM benchmark systems

It is worth remembering that our protocol is focused on validating maps and not on monitoring the quality of all the steps involved in the pre-processing stage. Thus, the results presented here start from the refinement procedure itself. This is done using pre-processed particles available in Electron Microscopy Public Image Archive (EMPIAR)[65]

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

- *The human hyperpolarization-activated cyclic nucleotide-gated channel* (HCN1) is a voltage-dependent ion channel, which was resolved to high resolution using cryo-EM [66]. The system was resolved in two conformational states, an *apo* state and a cAMP-bound state, to $\sim$ 3.5 Å using RELION 3D-refinement [1]. 55870 particles images belonging to the *apo* state are available in the EMPIAR with code 10081.

- *The recombination-activating genes RAG1-RAG2* form a complex (RAG1-RAG2) that plays an essential role in the generation of antibodies and antigen-receptor genes in a process called V(D)J recombination. Two main structures of the RAG1-RAG2 complex can be distinguished during the V(D)J recombination, a synaptic paired complex and the signal end complex (SEC). These states were resolved to 3.7 and 3.4 Å, respectively, using cryo-EM [67]. 81946 processed picked particles from the SEC state are deposited in the EMPIAR data bank with code 10049.

- *The mammalian transient receptor potential TRPV1* ion channel (TRPV1) is the receptor for capsaicin. Its structure was determined to 3.4 Å using cryo-EM [68]. A set of 35645 processed particles for this system are found in the EMPIAR data bank with code 10005.

- *The human immunodeficiency virus type 1 envelope glycoprotein trimer* (HIV-ET) is a membrane-fusing machine which mediates virus entry into host cells. The structure of the apo HIV-1 envelope glycoprotein in the trimer-conformation was determined to 6 Å using the 0.5 FSC threshold with cryo-EM [69]. A set of 124478 particles used in the refinement process is available in EMPIAR with code 10008.

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

The defocus information is also available for all these particles. Furthermore, for all of the above cases, a subset of 5000 particles was randomly selected to be used as the cross-validation set. Specifically, these particles are not used in the refinement processes.

3.2.3 3D refinement

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

*using the 0.143 FSC threshold

Table 3.1: Summary of the results from the 3D-refinement using RELION [1] for the cryo-EM systems.

The refinement for all systems was performed using the RELION [1] software. RELION provides the *relion_refine* utility that joins a Bayesian framework and an expectation- maximization strategy [1, 40] to refine iteratively a reference map. This algorithm requires as input an initial map and the set of particles corresponding to an unique state (it is assumed that a 3D-classification was performed previously). The particles from the refinement data are divided into two halves and two independent reconstruction are carried on following the gold-standard procedure. This is done automatically in RELION. In order to guarantee independence of both generated maps, it is suggested [8] to low-pass filter strongly ($50 \sim 60$ Å) the initial reference map. Upon convergence a final map is generated by joining the information from the two halves.

For all systems, we assume that the deposited particles correspond to the same state. Therefore, the preprocessing steps of 2D or 3D classification are not performed. Hence, only the refinement procedure is carried out. As the initial reference map for the 3D refinement, we use the final map reported by the authors low-pass filtered to 60 Å. This was done also to minimize the risk of overfitting [8]. We note that the number of particles used for these reconstructions was slightly less than those of the original works because the particles from the control set were taken out. In all cases, we used the RELION default parameters, and point-group symmetries reported by the authors. Table 1 summarizes the results obtained from the 3D refinement. The resolutions are in accordance with the reported ones, taking into account that the post-processing steps were not performed, and that the control set of particles was excluded from the refinement.

3.2.4 Summary of the protocol application

For a specific system, a subset of 5000 particles is selected from the original particle set. The refinement procedure is carried out with the remaining particles using the RELION software. Several maps corresponding to both refinement sets and representative iterations are chosen for the cross-validation tests. These maps are low-pass filtered using eight frequency cutoffs

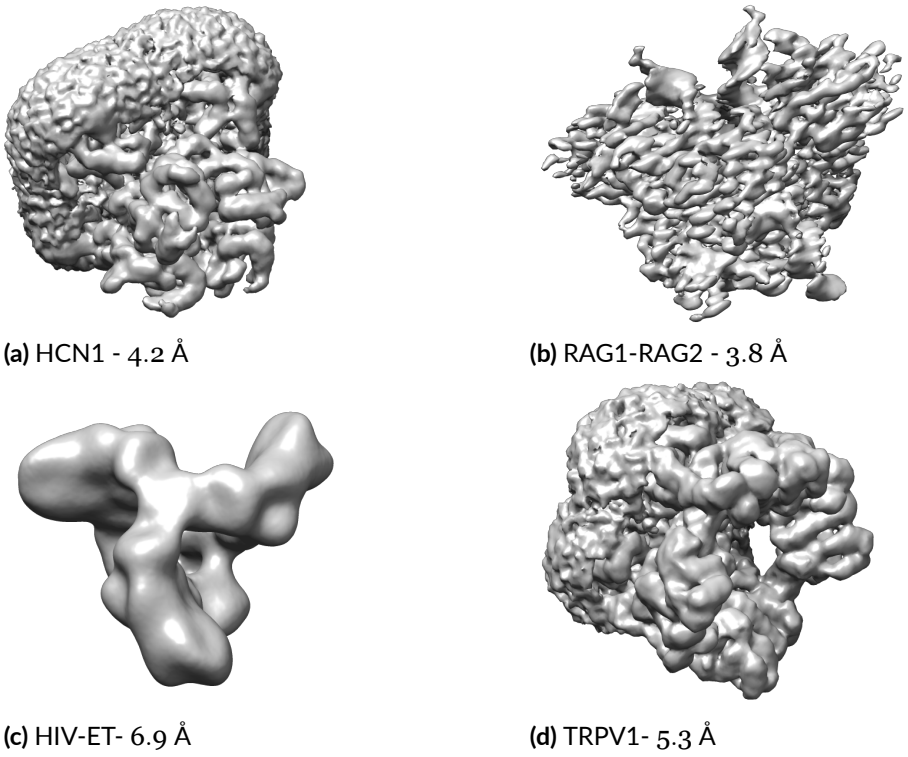
(see details in the Appendix). Each low-pass filtered map is submitted to the first and second rounds of the BioEM algorithm, obtaining a probability for each. These probabilities are used to compute both validation criteria: the cumulative probability and the normalized Jensen-Shannon divergence. In the next chapter, the results over the selected cryo-EM systems will be presented.

4

Cross-validation tests for cryo-EM

In this chapter, we present the main results obtained by applying our cross-validation tests over some systems that represent a diverse set of biomolecular families, including membrane proteins and protein-nucleicacids complexes. We tested the methodology over the cryo-EM datasets: the synaptic RAG1-RAG2 complex (RAG1-RAG2) [67], the human HCN1 channel (HCN1) [66], and the TRPV1 ion channel (TRPV1) [68]. To analyze the impact of overfitting, we studied two additional systems: cryo-EM reconstructions from the HIV-1 envelop trimer (HIV-ET) [69] and a set of synthetic pure-noise images that act as a 'false' control set with the RAG1-RAG2 reconstructions. This was motivated by the fact that some reconstructions might have been generated from pure-noise particles, and their resolution might have been over-estimated [70, 4, 5]. The reconstruction refinement was performed using the gold-standard procedure in RELION [1] yielding resolutions in the 3 to 6 Å range. Figure 4.1 shows the final reconstructions for the four real systems.

We propose two validation tests for monitoring overfitting in cryo-EM. We monitor the cumulative log-posterior and the NJSD increase as function of the refinement iteration and filtering frequency cutoff. With these two observables, we are able to distinguish overfitted maps from non-overfitted ones. All this is discussed more in detail in the next sections.



(a) HCN1 - 4.2 Å

(b) RAG1-RAG2 - 3.8 Å

(c) HIV-ET- 6.9 Å

(d) TRPV1- 5.3 Å

Figure 4.1: Final maps from the RELION refinement for the four systems presented in chapter 3. Resolutions are calculated with the *FSC* with the 0.143 threshold.

4.1 **Test 1**: Map evidence from the BioEM log-posterior.

In Fig. 4.2, we examine the improvement of the maps by monitoring the cumulative log-posterior relative to noise, $\sum_\omega \ln(P_{i\omega}(k_c))/N_\omega - \ln(P_{\text{Noise}})$, over the control set with $N_\omega = 5000$, as a function of the filtering frequency k_c for the reconstructions from sets $i = 1, 2$. The results are shown for different refinement iterations with a gradient color scheme (first iteration: maroon; last iteration: green). These results measure how probable each filtered map is relative to P_{Noise} (see the Methods). For the RAG1-RAG2, HCN1 and TRPV1 systems, we find an increase of the map evidence (given by the cumulative log-posterior) as a function of the frequency cutoff. For very high frequencies, the cumulative evidence plateaus. We only observe minor differences between the results from set $i = 1$ and 2 (solid and dashed lines, respectively, in Fig. 4.2). This is an indication of the similarity between the reconstructions generated from the two sets. Importantly, the results highlight the ability of the BioEM posterior to correctly rank maps of different resolutions. The reconstructions from the last iterations (*i.e.*, the most refined) are the most probable. This is in agreement with what one

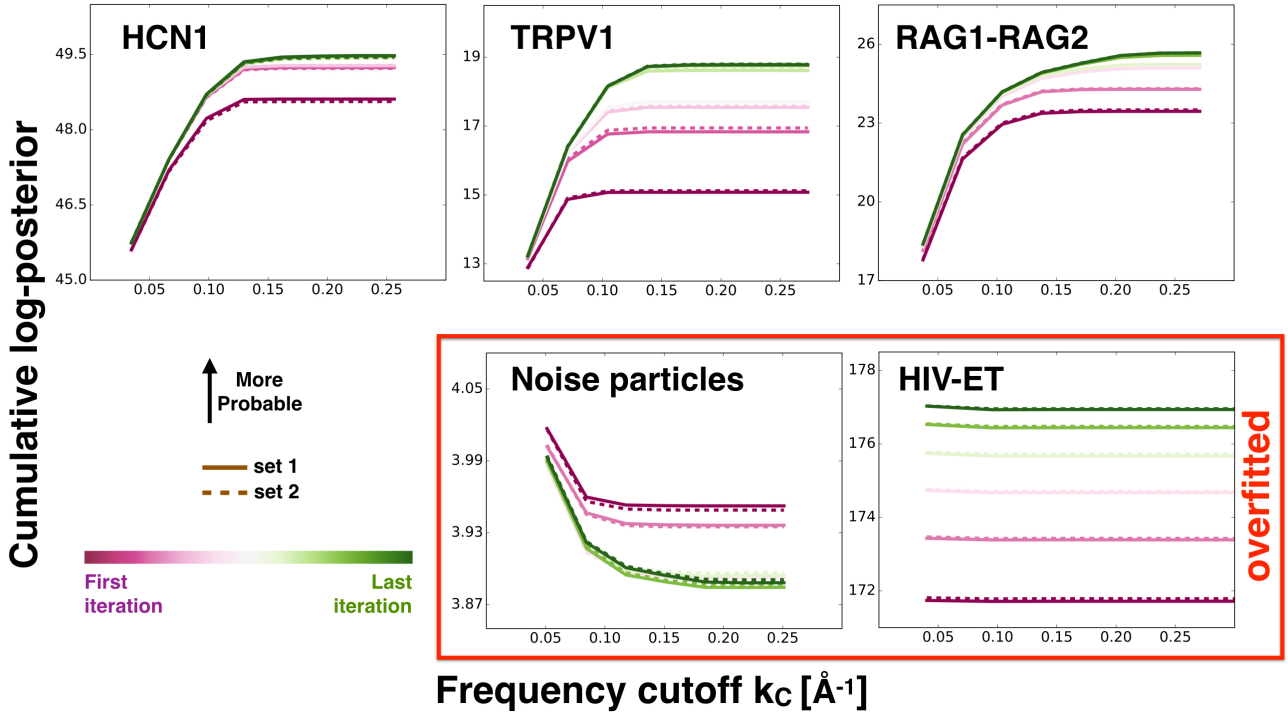


Figure 4.2: The cumulative log-posterior relative to noise $\sum_\omega \ln(P_{i\omega})/N_\omega - \ln(P_{\text{Noise}})$, over the control set with N_ω images, as a function of the frequency cutoff for reconstructions from set $i = 1$ and 2 (solid and dashed lines, respectively). The results are shown for different refinement iteration steps with a gradient color code: the first iteration is maroon and the last iteration is green. On the top row, we show the results for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2 for $N_\omega = 5000$. Systems that exhibit signs of overfitting, *i.e.* a noise-particle control set with $N_\omega = 1000$ and HIV-ET with $N_\omega = 5000$, are shown in the bottom row, highlighted with a red box.

expects from the 3D-refinement algorithms [71].

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

4.2 **Test 2**: Similarity between the probability distributions.

Assuming that homogeneous particle sets were used in the gold-standard procedure, one expects that the BioEM probabilities will be quite similar for both maps from the same iteration. The difference level should depend on the map resolution since overfitting and noise alignment is stronger for higher spatial frequencies [6, 8]. Thus, as a second validation test, we compare the distributions of the posterior probabilities generated by the reconstructions from sets $i = 1, 2$ over the control set. Figure 4.3 shows an example of the probability distributions for the HCN1 system for two frequency cutoffs at a given iteration. We find that the probability distributions, over the independent set, are quite similar for both reconstructions (top panel). However, there are small differences between them, and the higher-frequency maps present larger fluctuations (bottom panel) as expected. It is these difference levels that we seek quantify using the NJSD (see chapter 3 for its definition).

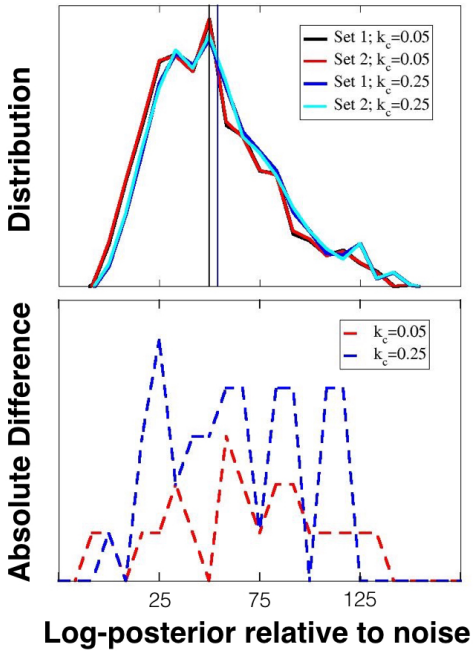


Figure 4.3: *Differences in the log-posterior distributions.* (**top**) Examples of the distributions of the log-posterior relative to noise over the independent particle set. The distributions are calculated for the reconstructions from set 1 and set 2 at two cutoff frequencies $k_c = 0.05$ and 0.25 Å^{-1} for the fifth iteration of refinement of the HCN1 system. The vertical lines are the averages of the distributions. (**bottom**) Absolute value of the difference between the probability distributions from set 1 and set 2 for $k_c = 0.05$ and 0.25 Å^{-1}. The distributions calculated for the maps with higher frequencies are less similar.

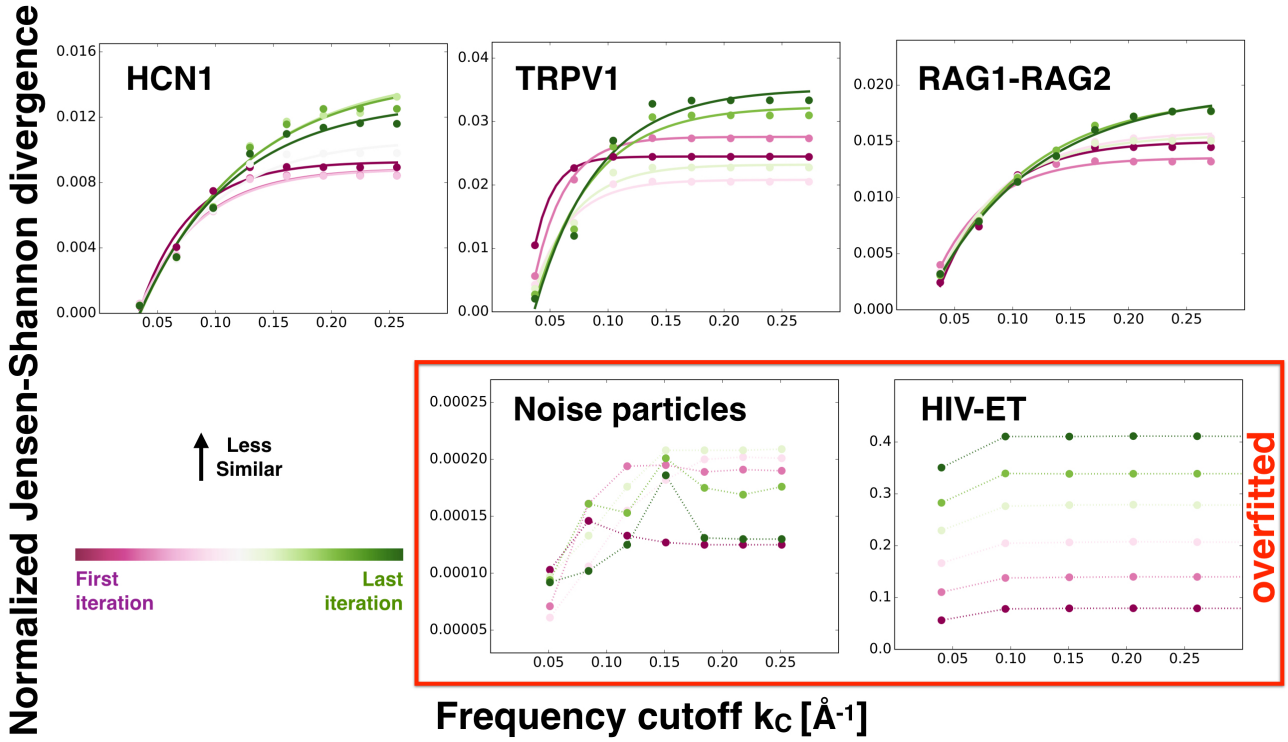


Figure 4.4: Normalized Jensen-Shannon divergence (NJSD) as a function of the frequency cutoff. This metric calculates the similarity between the distributions of the BioEM probabilities computed for the two reconstructions from sets 1 and 2. We use a gradient color code for the refinement iteration steps: the first iteration is maroon and the last iteration is green. On the top row, we show the results for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. For these systems, we fit the data points to an inverse exponential function $-Ae^{-k_c/\gamma}+B$ (solid lines). Systems that present signs of overfitting, a noise-particle control set and HIV-ET, are shown in the bottom row with dashed lines as a guide. The red box highlights the overfitted systems. The number of images in the control sets are the same as for the data in Fig. 4.2.

In Fig. 4.4, we plot the NJSD as a function of the frequency cutoff k_c for all the four systems and the synthetic noise particles set. Interestingly, for the RAG1-RAG2, HCN1 and TRPV1 systems, we observe that as the filtered maps contain higher frequencies, the larger the value of the NJSD. This implies that the probability distributions between maps with higher frequencies are less similar, possibly because they are more uncorrelated due to the high-frequency noise. For these standard systems, we also find that as the iteration increases the NJSD reaches at higher frequencies a plateau value. This behavior can be fit with an inverse exponential function $-Ae^{-k_c/\gamma}+B$ (see below and solid lines in Fig. 4.4). On the contrary, for the HIV-ET and noise-particle set, we find that the NJSD remains constant or has random behavior, suggesting that distributions do not consistently change when higher frequencies

are added to the maps.

Based on the results of both validation criteria for the HIV-ET system, we can conclude that this system shows signs of overfitting. The nature of such overfitting was discussed previously in [4], where the author argue that experimental particles are in most cases, pure noise data with spurious correlations introduced by a biased particle picking. The similarity between the HIV-ET and the *pure noise-images* curves, shown in figures 4.2 and 4.4 support this conclusion.

4.3 Cross-validation tests versus resolution.

An interesting feature of the NJSD curves shown in Fig. 4.4 is that the saturation rate depends on the iteration index (first iterations reach the plateau faster than last iterations), which in turn are a resolution indicative. This motivated us to investigate the correlation between the NJSD curves and the map resolution.

For the HCN1, TRPV1 and RAG1-RAG2 systems, we find that the NJSD curves can be fitted to an inverse exponential function, $-Ae^{-k_c/\gamma} + B$ (solid lines shown in Fig. 4.4). Intuitively, the frequency γ describes how fast the plateau of the NJSD is reached: a larger γ indicates a slower saturation of the NJSD.

In Fig. 4.5, we plot the frequency γ as a function of the inverse of the resolution (calculated using the FSC at the threshold 0.143). Interestingly, we find that the frequency γ is highly correlated to the inverse of the resolution with correlation coefficient $r^2 = 0.93$, 0.91, and 0.85, for HCN1, TRPV1 and RAG1-RAG2, respectively. These results show that even from a small independent control set, it is possible to extract unbiased information about the map resolution. We note the correlation between γ and the FSC resolution is a fortuitous result, signature of the data. In future work, we intend to search for the analytical explanation of this correlation.

We note that for the HIV-ET and noise-particle sets it is not possible to fit the NJSD data to an inverse exponential function. Therefore, we can only estimate the correlation between γ and the inverse of the resolution for the standard cryo-EM systems.

4.4 Convergence over a small cross-validation set.

We assessed how the results depend on the number of particles in the control set. In Fig. 4.6, we show an example of the cumulative log-posterior and NJSD as a function of the number of particles. We find that after approximately 1000 particles these observables converge, suggesting that only a small set is needed to perform the cross-validation analysis. This is confirmed in Fig. 4.7, where we plot the cumulative log-posterior and NJSD as a function of the frequency cutoff for a validation set of 1000 images. For the same set, in Fig. 4.8, we plot

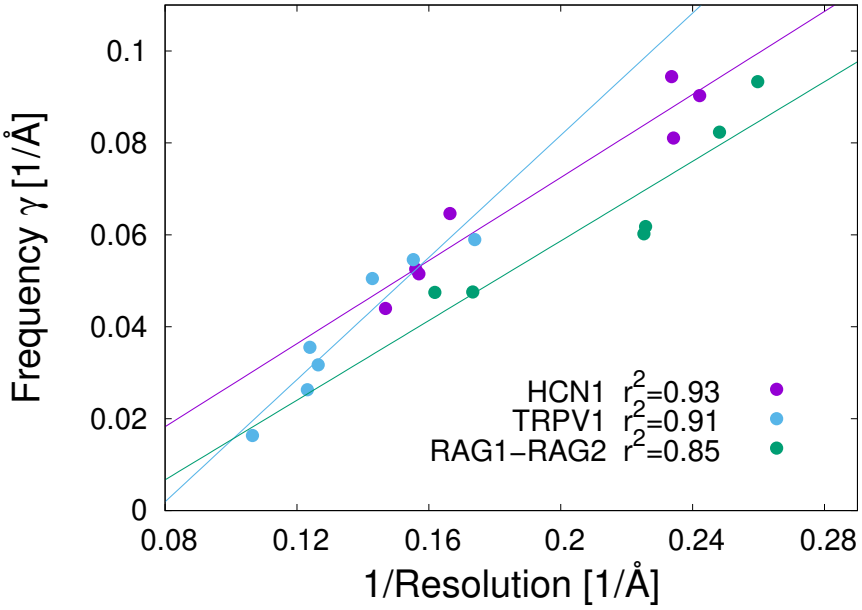


Figure 4.5: Frequency γ versus the inverse of the resolution for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. The NJSD curves for these systems were fitted to an inverse exponential function $-Ae^{-k_c/\gamma} + B$. We find large correlations between γ and the inverse of the resolution (calculated using the 0.143 criteria). The correlation coefficients are $r^2 = 0.93$, 0.91, and 0.85, for HCN1, TRPV1 and RAG1-RAG2, respectively. Solid lines show the linear fits.

the frequency γ as a function of the inverse of the map resolution, showing high correlations for the standard cryo-EM systems. These results are very similar to those obtained for the cross-validation set with 5000 particles.

The convergence of the log-posterior and NJSD, for different sizes of the validation set and for all the systems, can be associated to the fact that the particles sets sample near-uniformly the orientation space. It would be interesting for future work to analyze systems with orientational preference.

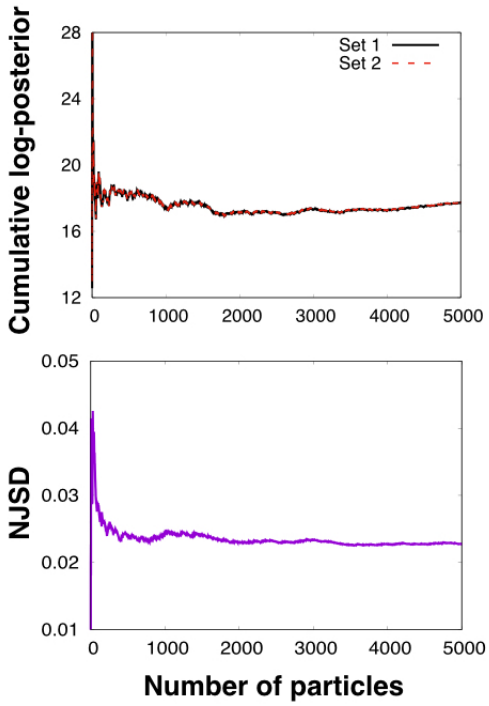


Figure 4.6: *Convergence of the observables.* (**top**) The cumulative log-posterior relative to noise $\sum_\omega \ln(P_{i\omega})/N_\omega - \ln(P_{\text{Noise}})$ for set $i = 1$ and 2 (solid and dashed lines, respectively), and (**bottom**) the normalized Jensen-Shannon divergence as a function of the number of particles in the control set. The results are shown for the TRPV1 system for iteration 12 and cutoff frequency $k_c = 0.21$ Å^{-1}. The observables converge if more than approximately 1000 particles are used.

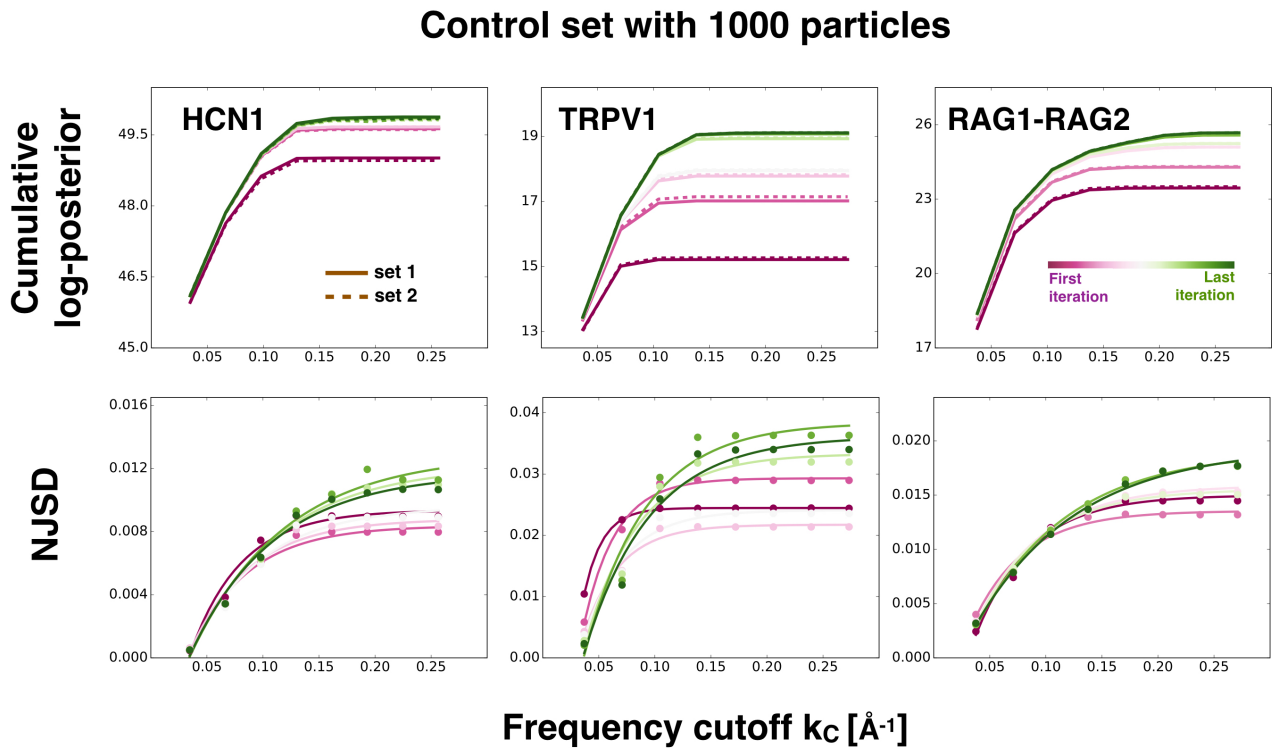


Figure 4.7: *Cumulative log-posterior and NJSD for a control set with 1000 particles.* (**top**) The cumulative log-posterior relative to noise and (**bottom**) the normalized Jensen-Shannon divergence as a function of the frequency cutoff. We use a gradient color code for the refinement iteration steps: the first iteration is maroon and the last iteration is green. The results are shown for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. The cumulative log-posterior is shown for the reconstructions from set 1 as solid lines and set 2 as dashed lines. NSJD data is fit to an inverse exponential function $-Ae^{-k_c/\gamma} + B$ (solid lines; bottom).

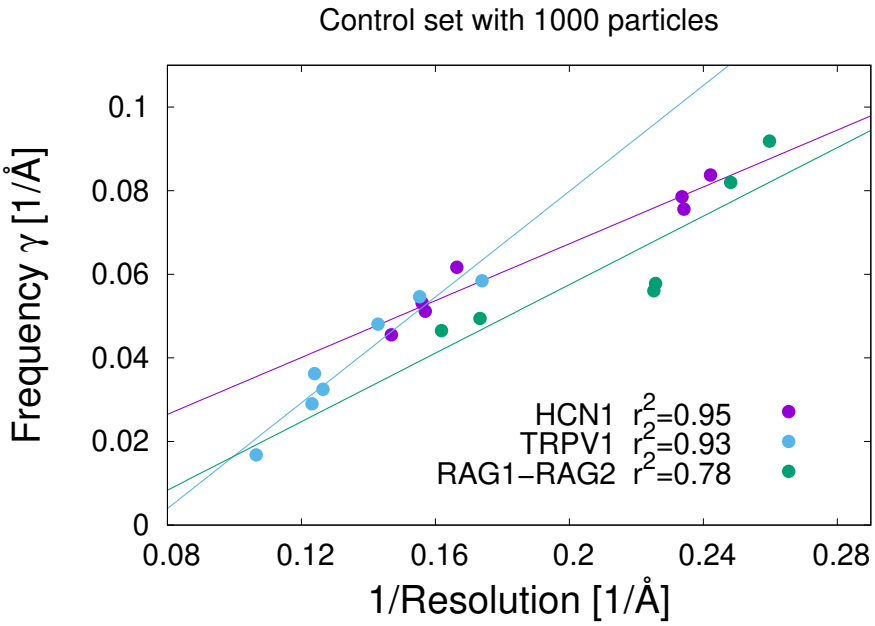


Figure 4.8: *Frequency (γ) versus the inverse of the resolution for a control set with 1000 particles.* The results are shown for the standard cryo-EM systems: HCN1, TRPV1 and RAG1-RAG2. The correlation coefficients are $r^2 = 0.95$, 0.93, and 0.78, respectively. Solid lines show the linear fits.

5 Conclusions

In this work, we have developed a novel methodology to validate cryo-EM maps. The main characteristic of our validation methodology is the employment of control images, which are not used in the reconstruction procedure. This protocol defines clear criteria to detect overfitted reconstructions in cryo-EM. Public-friendly codes and a detailed tutorial are available in https://github.com/bio-phys/BioEM-tutorials.

In summary, we propose monitoring the map probability and the similarity between the two probability distributions, associated to both maps generated during the gold-standard procedure, as function of the filtering frequency cutoff and the refinement iteration. As the similarity measure between distributions, we proposed the NJSD, which is a positive metric and is zero only for identical distributions. The increase of the map probability and the NJSD as a function of the frequency cutoff and the refinement iteration is a reliable validation test, since one expects that higher resolution maps, or less filtered maps, will have a greater correlation with the control particles.

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

Our methodology is general and robust. The mathematical framework is not only valid for

the BioEM posterior but also for any posterior probability that measures the likelihood of a 3D density given a particle set. The tests converge over a small particle set, typically only 1000 particles. Moreover, the methodology has the potential to be applicable for directly refining atomic models (instead of 3D maps) using an independent control set.

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

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

6
Perspectives

As future goals, we seek the optimization of the BioEM algorithm to speed-up round 2 of the analysis. Dr. Luka Stanisic has already implemented some optimization improvements to the BioEM code which are available in the latest BioEM version 2.1. However, we need to investigate code improvements or new strategies to reduce the computing resources and performance time of our protocol.

Furthermore, we are very interested in a deeper analysis of the relation between the frequency γ and the standard resolution. The high correlations between these two quantities motivates us to try to define a resolution measure based on the cross-validation test. This resolution would have a similar meaning to the free R-factor widely used in X-ray crystallography. We are currently researching this topic.

A Appendix

A.1 Image formation and contrast transfer function

In chapters 2 and 3, we discussed several aspects of the image formation process in cryo-EM. Here, we will describe briefly this process and introduce the contrast transfer function (CTF).

The scattering of the incoming electron beam produced by the interaction with the sample can be approximately modelled by a phase shift. If the sample is thin enough and weak scattering is assumed, the phase shift can be expanded in a Taylor series to the first order (this is known as the *weak-phase approximation*) [35]. Thus, all information about the biomolecule structure is encoded in the complex part of the exit wave function. However, to obtain an ideal recording the phase angle must be multiple of $\pi/2$. Unfortunately, in cryo-EM there is no experimental setup able to achieve this yet (but phase-plates are a promise alternative [72]).

To overcome this, the common strategy in cryo-EM is recording data in defocus conditions [36, 15]. The microscope optical aberrations generate a dependency of the phase shift on the spatial frequency components k which induces a phase-contrast. This can be modelled by the *contrast transfer function* [36, 73] in Fourier space

$$\mathrm{CTF}(k) = e^{-k^2/2b^2}(\sqrt{1-A^2}\sin(\xi(k)) + A^2\cos(\xi(k)), \tag{A.1}$$

which generates a phase shift induced by the optical aberrations. The factors A is the *amplitude contrast ratio* and b is the B-factor. $\xi(k)$ retains all the aberrations effects, mainly due to

the defocus, the spherical aberration and the astigmatism. The envelope $e^{-k^2/2b^2}$ is a general function to describe all other incoherent aberrations which tend to suppress high-frequency components [36, 15, 73]. It has been shown that large biomolecules (typically greater than 30nm) requires CTF corrections due to the size effects[74]. However, such effect is negligible for the systems studied here, that have size in the 180-200 Å.

A.2 Map low-pass filtering

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

$$\mathcal{F}_m^{k_c}(\mathbf{k}) = \begin{cases} \mathcal{F}_m(\mathbf{k}) & k \leq k_c \\ 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

We use the code *lowpassmap_fftw* available from the Rubinstein lab webpage [75] to perform this calculation. We then convert the map into real space by applying the inverse Fourier transform of $\mathcal{F}_m^{k_c}(\mathbf{k})$. The real-space filtered map is masked and then used as input for the BioEM computation.

A.3 Pure-noise particles

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

A.4 BioEM input file examples

See the BioEM manual for more in detail information.

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

```
PIXEL_SIZE 1.22
NUMBER_PIXELS 256
```

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

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

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

A.5 Compute performing

With the current version of BioEM, it is recommended to run the validation methodology in a computing cluster -if it is possible-. Since the manipulation of the BioEM code is not in the scope of the present work, we had the collaboration of Ph.D..s Luka Stanisic and Markus Rampp from the Max Planck Computing and Data Facility, who improved the CUDA and MPI parallelization for the Round 2.

The results shown here were computed over a single node with 32 real cores and 2 GPUs. Round 1, takes approximately 4 hours. For Round2, BioEM takes 0.3 seconds per image, so analyzing the 64 maps (8 iterations and 8 frequencies) and 5000 particles takes 1 day with same configuration. .

Minimize this time computing is a important future goal.

References

[1] Sjors H W Scheres. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

[2] Ali Dashti, Peter Schwander, Robert Langlois, Russell Fung, Wen Li, Ahmad Hosseinizadeh, Hstau Y Liao, Jesper Pallesen, Gyanesh Sharma, Vera A Stupina, et al. Trajectories of the ribosome as a brownian nanomachine. *Proceedings of the National Academy of Sciences*, 111(49):17492–17497, 2014.

[3] Joachim Frank. New opportunities created by single-particle cryo-em: The mapping of conformational space, 2018.

[4] Richard Henderson. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences*, 110(45):18037–18041, 2013.

[5] Maxim Shatsky, Richard J Hall, Steven E Brenner, and Robert M Glaeser. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of structural biology*, 166(1):67–78, apr 2009.

[6] Peter B Rosenthal and John L Rubinstein. Validating maps from single particle electron cryomicroscopy. *Current Opinion in Structural Biology*, 34:135–144, oct 2015.

[7] Piotr Neumann, Achim Dickmanns, and Ralf Ficner. Validating Resolution Revolution. *Structure*, 26(5):785–795.e4, may 2018.

[8] Sjors H W Scheres and Shaoxia Chen. Prevention of overfitting in cryo-EM structure determination. *Nature methods*, 9(9):853, 2012.

[9] Richard Henderson, Shaoxia Chen, James Z Chen, Nikolaus Grigorieff, Lori A Passmore, Luciano Ciccarelli, John L Rubinstein, R Anthony Crowther, Phoebe L Stewart, and Peter B Rosenthal. Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. *Journal of molecular biology*, 413(5):1028–1046, 2011.

[10] Shaoxia Chen, Greg McMullan, Abdul R Faruqi, Garib N Murshudov, Judith M Short, Sjors HW Scheres, and Richard Henderson. High-resolution noise substitution to measure overfitting and validate resolution in 3d structure determination by single particle electron cryomicroscopy. *Ultramicroscopy*, 135:24–35, 2013.

[11] J. Bernard Heymann, Roberto Marabini, Mohsen Kazemi, Carlos Oscar S. Sorzano, Maya Holmdahl, Joshua H. Mendez, Scott M. Stagg, Slavica Jonic, Eugene Palovcak, Jean-Paul Armache, Jianhua Zhao, Yifan Cheng, Grigore Pintilie, Wah Chiu, Ardan Patwardhan, and Jose-Maria Carazo. The first single particle analysis Map Challenge: A summary of the assessments. *Journal of structural biology*, 204(2):291–300, NOV 2018.

[12] Pavel V. Afonine, Bruno P. Klaholz, Nigel W. Moriarty, Billy K. Poon, Oleg V. Sobolev, Thomas C. Terwilliger, Paul D. Adams, Alexandre Urzhumtsev, and IUCr. New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallographica Section D Structural Biology*, 74(9):814–840, sep 2018.

[13] Pilar Cossio and Gerhard Hummer. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of structural biology*, 184(3):427–437, 2013.

[14] Pilar Cossio, David Rohr, Fabio Baruffa, Markus Rampp, Volker Lindenstruth, and Gerhard Hummer. BioEM: GPU-accelerated computing of Bayesian inference of electron microscopy images. *Computer Physics Communications*, 210:163–171, 2017.

[15] Fred J Sigworth. Principles of cryo-em single-particle image processing. *Microscopy*, 65(1):57–67, 2016.

[16] Grant Jensen. *Cryo-EM Part A: sample preparation and data collection*, volume 481. Academic Press, 2010.

[17] Marc Adrian, Jacques Dubochet, Jean Lepault, and Alasdair W McDowall. Cryo-electron microscopy of viruses. *Nature*, 308(5954):32, 1984.

[18] Shenping Wu, Jean-Paul Armache, and Yifan Cheng. Single-particle cryo-EM data acquisition by using direct electron detection camera. *Microscopy*, 65(1):35–41, feb 2016.

[19] G. McMullan, A.R. Faruqi, and R. Henderson. Direct Electron Detectors. *Methods in Enzymology*, 579:1–17, jan 2016.

[20] Xueming Li, Paul Mooney, Shawn Zheng, Christopher R Booth, Michael B Braunfeld, Sander Gubbens, David A Agard, and Yifan Cheng. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em. *Nature methods*, 10(6):584, 2013.

[21] Shawn Q Zheng, Eugene Palovcak, Jean-Paul Armache, Kliment A Verba, Yifan Cheng, and David A Agard. Motioncor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature methods*, 14(4):331, 2017.

[22] Feng Wang, Huichao Gong, Gaochao Liu, Meijing Li, Chuangye Yan, Tian Xia, Xueming Li, and Jianyang Zeng. Deeppicker: a deep learning approach for fully automated particle picking in cryo-em. *Journal of structural biology*, 195(3):325–336, 2016.

[23] Sjors HW Scheres. Semi-automated selection of cryo-em particles in relion-1.3. *Journal of structural biology*, 189(2):114–122, 2015.

[24] Dmitry Lyumkis, Axel F Brilot, Douglas L Theobald, and Nikolaus Grigorieff. Likelihood-based classification of cryo-em images using frealign. *Journal of structural biology*, 183(3):377–388, 2013.

[25] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, mar 2017.

[26] Tanvir R Shaikh, Haixiao Gao, William T Baxter, Francisco J Asturias, Nicolas Boisset, Ardean Leith, and Joachim Frank. Spider image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature protocols*, 3(12):1941, 2008.

[27] Sjors HW Scheres. Processing of structurally heterogeneous cryo-em data in relion. In *Methods in enzymology*, volume 579, pages 125–157. Elsevier, 2016.

[28] Joachim Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, 2006.

[29] Grant Jensen. *Cryo-EM Part B: 3-D Reconstruction*, volume 482. Academic Press, 2010.

[30] Grant Jensen. *Cryo-EM, Part C: analyses, interpretation, and case studies*, volume 483. Academic Press, 2010.

[31] Rafael Fernandez-Leiro and Sjors HW Scheres. A pipeline approach to single-particle processing in relion. *Acta Crystallographica Section D: Structural Biology*, 73(6):496–502, 2017.

[32] Eva Nogales. The development of cryo-em into a mainstream structural biology technique. *Nature methods*, 13(1):24, 2015.

[33] Yifan Cheng. Single-particle cryo-em—how did it get here and where will it go. *Science*, 361(6405):876–880, 2018.

[34] Joachim Frank. Generalized single-particle cryo-em–a historical perspective. *Microscopy*, 65(1):3–8, 2016.

[35] Miloš Vulović, Lenard M Voortman, Lucas J van Vliet, and Bernd Rieger. When to use the projection assumption and the weak-phase object approximation in phase contrast cryo-em. *Ultramicroscopy*, 136:61–66, 2014.

[36] C Barry Carter and David B Williams. *Transmission electron microscopy: Diffraction, imaging, and spectrometry*. Springer, 2016.

[37] Eva Nogales and Sjors HW Scheres. Cryo-em: a unique tool for the visualization of macromolecular complexity. *Molecular cell*, 58(4):677–689, 2015.

[38] N. Grigorieff. Frealign: An Exploratory Tool for Single-Particle Cryo-EM. *Methods in Enzymology*, 579:191–226, jan 2016.

[39] Guang Tang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, jan 2007.

[40] Sjors HW Scheres. A bayesian view on cryo-em structure determination. *Journal of molecular biology*, 415(2):406–418, 2012.

[41] W. O. Saxton and W. Baumeister. The correlation averaging of a regularly arranged bacterial cell envelope protein. *Journal of Microscopy*, 127(2):127–138, aug 1982.

[42] George Harauz and Martin van Heel. Exact Filters for General Geometry Three Dimensional Reconstruction. *Optik*, 78(4):6–30, feb 1986.

[43] J Frank and L Al-Ali. Signal-to-noise ratio of electron micrographs obtained by cross correlation. *Nature*, 256(5516):376, 1975.

[44] C O S Sorzano, J Vargas, J Otón, V Abrishami, J M de la Rosa-Trev\'\in, J Gómez-Blanco, J L Vilas, R Marabini, and J M Carazo. A review of resolution measures and related aspects in 3D Electron Microscopy. *Progress in biophysics and molecular biology*, 124:1–30, 2017.

[45] Pawel A. Penczek. Resolution Measures in Molecular Electron Microscopy. *Methods in Enzymology*, 482:73–100, jan 2010.

[46] Marin Van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of structural biology*, 151(3):250–262, 2005.

[47] Hstau Y Liao and Joachim Frank. Definition and estimation of resolution in single-particle reconstructions. *Structure*, 18(7):768–775, 2010.

[48] J. Vargas, R. Melero, J. Gómez-Blanco, J. M. Carazo, and C. O. S. Sorzano. Quantitative analysis of 3D alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis. *Scientific Reports*, 7(1):6307, dec 2017.

[49] Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology*, 333(4):721–745, 2003.

[50] Pavel Afanasyev, Charlotte Seer-Linnemayr, Raimond B. G. Ravelli, Rishi Matadeen, Sacha De Carlo, Bart Alewijnse, Rodrigo V. Portugal, Navraj S. Pannu, Michael Schatz, and Marin van Heel. Single-particle cryo-EM using alignment by classification (ABC): the structure of Lumbricus terrestris haemoglobin. *IUCrJ*, 4(5):678–694, sep 2017.

[51] Alp Kucukelbir, Fred J Sigworth, and Hemant D Tagare. Quantifying the local resolution of cryo-EM density maps. *Nature methods*, 11(1):63–5, jan 2014.

[52] Giovanni Cardone, J Bernard Heymann, and Alasdair C Steven. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *Journal of structural biology*, 184(2):226–36, nov 2013.

[53] Grigore Pintilie, Dong-Hua Chen, Cameron A. Haase-Pettingell, Jonathan A. King, and Wah Chiu. Resolution and Probabilistic Models of Components in CryoEM Maps of Mature P22 Bacteriophage. *Biophysical Journal*, 110(4):827–839, feb 2016.

[54] J. Vargas, J. Otón, R. Marabini, J. M. Carazo, and C. O. S. Sorzano. Particle alignment reliability in single particle electron cryomicroscopy: a general approach. *Scientific Reports*, 6(1):21626, apr 2016.

[55] Alan Brown, Fei Long, Robert A. Nicholls, Jaan Toots, Paul Emsley, Garib Murshudov, and IUCr. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallographica Section D Biological Crystallography*, 71(1):136–153, jan 2015.

[56] Todor Kirilov Avramov, Dan Vyenielo, Josue Gomez-Blanco, Swathi Adinarayanan, Javier Vargas, and Dong Si. Deep Learning for Validating and Estimating Resolution of Cryo-Electron Microscopy Density Maps. *Molecules*, 24(6), MAR 2 2019.

[57] Axel T. Brünger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, jan 1992.

[58] Benjamin Falkner and Gunnar F Schröder. Cross-validation in cryo-em–based structural modeling. *Proceedings of the National Academy of Sciences*, 110(22):8930–8935, 2013.

[59] Frank DiMaio, Junjie Zhang, Wah Chiu, and David Baker. Cryo-em model validation using independent map reconstructions. *Protein Science*, 22(6):865–868, 2013.

[60] Solomon. Kullback. *Information theory and statistics*. Dover Publications, 1968.

[61] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[62] Thomas M Cover and Joy A Thomas. Elements of information theory 2nd edition. 2006.

[63] Anna Yershova, Swati Jain, Steven M. LaValle, and Julie C. Mitchell. Generating uniform incremental grids on SO(3) using the Hopf fibration. *Int. J. Robot. Res.*, 29(7):801–812, JUN 2010.

[64] Pilar Cossio, Matteo Allegretti, Florian Mayer, Volker Mueller, Janet Vonck, and Gerhard Hummer. Bayesian inference of rotor ring stoichiometry from electron microscopy images of archaeal ATP synthase. *Microscopy*, 67(5):266–273, OCT 2018.

[65] Andrii Iudin, Paul K Korir, José Salavert-Torres, Gerard J Kleywegt, and Ardan Patwardhan. EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods*, 13(5):387–388, may 2016.

[66] Chia-Hsueh Lee and Roderick MacKinnon. Structures of the human HCN1 hyperpolarization-activated channel. *Cell*, 168(1-2):111–120, 2017.

[67] Heng Ru and Others. Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. *Cell*, 163(5):1138–1152, 2015.

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

[69] Youdong Mao, Liping Wang, Christopher Gu, Alon Herschhorn, Anik Desormeaux, Andres Finzi, Shi-Hua Xiang, and Joseph G. Sodroski. Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. *Proceedings of the National Academy of Sciences*, 110(30):12438–12443, JUL 23 2013.

[70] Sriram Subramaniam. Structure of trimeric HIV-1 envelope glycoproteins. *Proceedings of the National Academy of Sciences*, 110(45):E4172–E4174, NOV 5 2013.

[71] Pilar Cossio and Gerhard Hummer. Likelihood-based structural analysis of electron microscopy images. *Current Opinion in Structural Biology*, 49:162–168, apr 2018.

[72] Radostin Danev, Bart Buijsse, Maryam Khoshouei, Jürgen M Plitzko, and Wolfgang Baumeister. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proceedings of the National Academy of Sciences*, 111(44):15635–15640, 2014.

[73] Alexis Rohou and Nikolaus Grigorieff. Ctffind4: Fast and accurate defocus estimation from electron micrographs. *Journal of structural biology*, 192(2):216–221, 2015.

[74] Dongjie Zhu, Xiangxi Wang, Qianglin Fang, James L Van Etten, Michael G Rossmann, Zihe Rao, and Xinzheng Zhang. Pushing the resolution limit by correcting the ewald sphere effect in single-particle cryo-em reconstructions. *Nature communications*, 9(1):1–7, 2018.

[75] Rubinstein Lab., https://sites.google.com/site/rubinsteingroup/home.