Multi-scale Bayesian modeling of cryo-electron microscopy density maps

Samuel Hanot[2,†], Massimiliano Bonomi[1,*†], Charles H. Greenberg[3], Andrej Sali[3], Michael Nilges[2], Michele Vendruscolo[1], Riccardo Pellarin[2,*]

[1]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

[2]*Structural Bioinformatics Unit, Institut Pasteur, CNRS UMR 3528, 75015 Paris, France*

[3]*Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Sciences, and California Institute for Quantitative Biomedical Sciences, University of California, San Francisco, California 94158, United States*

[*]Correspondence to: mb2006@cam.ac.uk; riccardo.pellarin@pasteur.fr

[†]These authors contributed equally to this work.

Summary

Cryo-electron microscopy has become a mainstream structural biology technique by enabling the characterization of biological architectures that for many years have eluded traditional methods like X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. However, the translation of cryo-electron microscopy data into accurate structural models is hampered by the presence of random and systematic errors in the data, sample heterogeneity, data correlation, and noise correlation. As a consequence, in integrative biology approaches, it has been difficult to objectively weigh EM-derived restraints with respect to other sources of information. To address these challenges, here we introduce a Bayesian approach that allows efficient and accurate structural modeling of cryo-electron microscopy density maps at multiple scales, from coarse-grained to atomistic resolution. The accuracy of the method is benchmarked using a set of structures of macromolecular assemblies. The approach is implemented in the open-source Integrative Modeling Platform package (http://integrativemodeling.org) in order to enable structural determination by combining cryo-electron microscopy with other information, such as chemical cross-linking/mass spectrometry, NMR, and small angle X-ray scattering data.

Introduction

Over the last two decades, electron microscopy (EM) has enabled the structural characterization of complex biological systems that were beyond the capabilities of more traditional techniques, such as X-crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy (*1-3*). This tremendous progress has been fuelled by the continuous advances in both instrumentation and software for cryo-EM image processing (*4-6*). As a result, cryo-EM is rapidly approaching the resolution of X-ray crystallography, allowing the structural determination of complex systems of outstanding biological importance, such as the cytoplasmic polyhedrosis virus at 3.88 Å (*7*), the ribosome at 5.4 Å (*8*), the 20S proteasome at 2.8 Å (*9*), human gamma-secretase at 3.4 Å (*10*) and β-galactosidase at 2.2 Å (*11*). Most importantly, cryo-EM does not require to crystallize the system prior to data acquisition, like X-ray crystallography, and it does not need large samples, isotopic enrichment, nor is limited by the size of the system under study, like NMR spectroscopy. Furthermore, cryo-EM has the potential to identify multiple different structural states (*2-5*), provided that they can be disentangled during the class-averaging process.

Despite these recent advances, the translation of cryo-EM three-dimensional (3D) reconstructions (cryo-EM density maps), into structural models still presents several challenges. First, cryo-EM maps are affected by random and systematic errors (*12, 13*), like all experimental data. In particular, radiation damage to the sample upon prolonged exposure to the electron beam often results in regions of the density map at resolution lower than the average. Second, despite the progress in the development of classification software to group single-particle two-dimensional (2D) images into class averages and to translate them into 3D reconstructions, the final density maps might still group different conformations together (*14*). As a result, areas of the map corresponding to regions of high flexibility typically present lower resolution and, therefore, an individual structure might not be able to fit the entire density map with the same accuracy and precision. Finally, density maps are typically constituted by a set of data points (i.e., voxels) with correlated noise. This aspect is particularly relevant when one wants to integrate EM data with other experimental data (*15*), such as chemical cross-linking/mass spectrometry or NMR data. In this case, the information content of each piece of data needs to be accurately weighed to avoid biasing towards a specific dataset during modeling (*12*). Consequently, there is still a pressing need of modeling approaches that can tackle all these challenges and generate accurate and precise structural models from EM density maps.

A number of approaches have been proposed over the years to model EM density maps. We refer to existing reviews for a complete overview of the state of the art (*16, 17*). Generally speaking, these techniques can be classified in five groups: methods for rigid-body fitting, flexible fitting, homology modeling, *de novo* modeling, and integrative approaches. The most popular methods in all these five categories include: Chimera (*18*), EMfit (*19*), Modeller (*20*), SITUS (*21*), MultiFit (*22*), EMFF (*23*), MDFF (*24*), MDFIT (*25*), Fold-EM (*26*), ROSETTA (*27*), EM-fold (*28*), IMP (*29*), RELION (*30*), and Phenix (*31*). The majority of these approaches generate structural models that minimize the deviation between observed and predicted EM density maps, by means of Molecular Dynamics (MD), Monte Carlo (MC) or Normal Modes Analysis techniques (*17*). Typically, this deviation is measured in terms of an electron density-based correlation coefficient (CC) between experimental and predicted maps

calculated over a set of voxels. Additional pseudo-energy terms are routinely applied to enforce correct stereochemistry. Despite the success of these methods, it is still a challenge to tackle all the issues related to the modeling of cryo-EM data outlined above.

Here we introduce a novel approach to model EM density maps based on a Bayesian framework (*32*), which provides an objective way to interpret experimental data and integrate them with prior knowledge. This technique is based on an analytical representation of the input data in terms of a Gaussian Mixture Model (GMM) (*33-35*), rather than the more standard voxel-based approach. This choice has several advantages: a) it circumvents the problem of data-point correlation, by clustering the voxels into independent components (i.e., the Gaussians); b) it is computationally efficient; c) it is compatible with a multi-scale representation of the model, from coarse-grained for initial efficient sampling to atomistic for refinement of high-resolution maps. Furthermore, our approach models the structure of the system and simultaneously quantifies, in an automated way, the level of noise in the data, thus allowing the balanced integration of different experimental data by weighing each piece of information according to its noise content.

In the following, we first outline the modeling protocol, introduce the theory of our Bayesian approach, and then benchmark its accuracy using synthetic low-resolution data of several protein/DNA complexes. This method is implemented in the open-source Integrative Modeling Platform (IMP; http://integrativemodeling.org) (*29*), thus enabling integrative structure determination of biological systems based on a variety of experimental data, including FRET and NMR spectroscopies, chemical cross-linking coupled with mass spectrometry, small angle X-ray scattering (SAXS), and various proteomics data. Furthermore, this approach can be used with a variety of different representations of the system, from coarse-grained to atomistic, for a multi-scale modeling of cryo-EM data.

Materials and Methods

Protocol for low-resolution modeling of EM data.

We implemented a pipeline that enables the low-resolution modeling of cryo-EM data given partial knowledge of subunits structures, and the final atomistic refinement. The procedure is as follows (**Fig. 1**):

1) Gather the data in the form of protein sequences, crystallographic or NMR structures of domains, homology models, or models predicted from evolutionary covariance data, along with the target 3D density map reconstructed form the electron microscopy data.
2) Generate a Gaussian Mixture Model representation of the 3D density map by using the divide-and-conquer algorithm described below (*33*). Assign a representation to the different components of the complex. Subunits are represented by strings of spherical beads and a set of 3D Gaussians. For a given domain, the conformation of the corresponding beads and Gaussians is either constrained into a rigid body or allowed to move flexibly, depending on the availability

of a structure or a model for that domain. The beads represent one or more contiguous residues, depending on the coarse-graining of the model (*34, 36, 37*). The Gaussians describe the electron density of the model and are used to compute the fit to the EM density map.

3) The score, which ranks the models according to how well they fit the input information, is derived from the posterior probability, which includes a likelihood function for the EM data, and prior terms such as the sequence connectivity of macromolecules and excluded volume.
4) The different degrees of freedoms of the model are sampled using Monte Carlo (MC) coupled with Replica Exchange (*38*) and Simulated Annealing (*39*) using IMP (*29*). Each replica outputs models that are stored in files for later analysis. The ensemble of models resulting from the sampling are ranked and analyzed.
5) Sampling exhaustiveness is tested and models are validated.

A Bayesian scoring function for EM data

In general terms, the Bayesian approach (*32*) estimates the probability of a model, given information available about the system, including both prior knowledge and newly acquired experimental data. The posterior probability $p(M|D)$ of model M, which is defined in terms of its structure X and other Bayesian parameters, given data D and prior knowledge is:

$$p(M|D) \propto p(D|M) \cdot p(M) \quad (1)$$

where the *likelihood function* $p(D|M)$ is the probability of observing data D given M and the *prior* $p(M)$ is the probability of model M given the the prior information. To define the likelihood function, one needs a *forward model* $f(X)$ that predicts the data point that would be observed for structure X in the absence of experimental noise, and a *noise model* that specifies the distribution of the deviation between the experimentally observed and predicted data points. The *Bayesian scoring function* is defined as $S(M) = -\log[p(D|M) \cdot p(M)]$, which ranks the models in the same order as the posterior probability $p(M|D)$. The prior $p(M)$ includes the sequence connectivity, the excluded volume and rigid body constraints. To compute these priors, the domains of the proteins are coarse grained using beads of varying size. The sequence connectivity term is a sum of upper harmonic distance restraint that connects all the beads in the sequence, and emulates the covalent structure of the polypeptide/polynucleotide main-chain. The excluded volume is computed from a soft-sphere potential where the radius of a bead is estimated from the sum of the masses of the residues it represents. The structures derived from X-ray data or homology models are coarse-grained using two categories of resolution, where beads represented either individual residues or segments of up to 10 residues. All these beads are constrained into a rigid body, in which relative distances are constrained during sampling. Strings of beads represent parts without structural information, and are kept flexible. In the following, we define the components of the Bayesian scoring function specifically for EM density maps.

Experimental electron density maps. We represent the experimental density map Ψ_D in terms of a Gaussian Mixture Model (GMM) ϕ_D^j with j components:

$$\phi_D^j(\boldsymbol{x}) = \sum_{i=1}^{j} \phi_{D,i}^j(\boldsymbol{x}) = \sum_{i=1}^{j} \omega_{D,i}^j \cdot G\left(\boldsymbol{x} \middle| \boldsymbol{x}_{D,i}^j, \Sigma_{D,i}^j\right) \quad (2)$$

where $\omega_{D,i}^j$ is the (normalized) weight of the i-th component of the GMM and G a normalized Gaussian function with mean $\boldsymbol{x}_{D,i}^j$ and covariance matrix $\Sigma_{D,i}^j$:

$$G\left(\boldsymbol{x} \middle| \boldsymbol{x}_{D,i}^j, \Sigma_{D,i}^j\right) = \frac{1}{(2\pi)^{3/2}\left|\Sigma_{D,i}^j\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}_{D,i}^j\right)^T \left(\Sigma_{D,i}^j\right)^{-1}\left(\boldsymbol{x} - \boldsymbol{x}_{D,i}^j\right)\right] \quad (3)$$

This description presents several advantages. First, it circumvents the problem of dealing with correlations in the data and noise that are typical of voxel-based representations, as each $\phi_{D,i}^j(\boldsymbol{x})$ might be regarded as an independent component of the density map. Second, it provides a computationally-convenient representation of the data in terms of analytical functions. In addition, the fit of the GMM representation can be evaluated by using the correlation coefficient of the GMM with the EM density map.

The posterior probability of model M given the EM density Ψ_D can be written in terms of all possible GMMs that can be used to represent the data:

$$p(M|\Psi_D) = \sum_j p\left(M\middle|\phi_D^j\right) p\left(\phi_D^j\middle|\Psi_D\right) \quad (4)$$

In the following, we will assume that the conditional probability $p\left(\phi_D^j\middle|\Psi_D\right)$ can select a single GMM ϕ_D with N_D components, which optimally represents the data. In this situation:

$$p(M|\Psi_D) \propto p(M|\phi_D) \propto p(\phi_D|M) \cdot p(M) \quad (5)$$

Divide-and-conquer fit of the experimental map. To fit the experimental density map Ψ_D with a GMM ϕ_D, we used an Expectation Maximization algorithm (*33*). This approach determines the parameters of

the GMM (mean, weight, and covariance matrix of each Gaussian) by maximizing the likelihood that the GMM density function generates the density of the voxels in Ψ_D. As the resolution of the map increases, the number of Gaussians required for the GMM to accurately reproduce all the features of the experimental map exponentially increases, also increasing the computational time and memory required to perform the fit. To overcome these problems, we developed a divide-and-conquer approach (**Fig. 2**). First, the map Ψ_D is masked and all voxels with a density lower than the threshold recommended in the EMDB (www.ebi.ac.uk/pdbe/emdb/) are removed. Second, a recursive procedure starts from iteration 1 by fitting the map Ψ_D (**Fig. 2A**) with a GMM consisting of a small number of Gaussians N_D (typically 2 or 4) (**Fig. 2B**). Each of the components $\phi_{D,i}$ of this initial GMM is used to define a partition of the original map into submaps $\Psi_{D,i}$ (**Fig 2C**):

$$\Psi_{D,i}(\boldsymbol{x}) = \Psi_D(\boldsymbol{x}) \cdot \frac{\phi_{D,i}(\boldsymbol{x})}{\sum_{j=1}^{N_D} \phi_{D,j}(\boldsymbol{x})} \quad (6)$$

This partitioning has two properties: *a)* each submap isolates the part of the original map that overlapped with the component ($\phi_{D,i}$); *b)* the sum of all submaps results in the original density map: $\Psi_D(\boldsymbol{x}) = \sum_{i=1}^{N_D} \Psi_{D,i}(\boldsymbol{x})$. The process is repeated, and each submap $\Psi_{D,i}$ is fitted using a GMM with small number of Gaussians N_D (**Fig 2D**). At each iteration, the portion of the original map that is fit by a given GMM is reduced, so that a small number of Gaussians will eventually be sufficient to accurately reproduce high-resolution details. Furthermore, because of property *b)*, the *global* GMM defined by the sum of all the GMMs obtained at any given iteration also fits the original map (**Fig. 2E**). The procedure can be repeated until the global GMM reaches the desired accuracy (**Fig. 2F**), measured for example in terms of correlation coefficient with the original map Ψ_D, which is computed at each iteration (**Fig. 3**). The whole process can be efficiently run on a computer cluster, being highly parallel.

The forward model. We developed a forward model to predict a EM density map from a single structural model. As for the data representation, the forward model ϕ_M is a GMM with N_M components:

$$\phi_M(\boldsymbol{x}) = \sum_{i=1}^{N_M} \phi_{M,i}(\boldsymbol{x}) = \sum_{i=1}^{N_M} \omega_{M,i} \cdot G\left(\boldsymbol{x} \middle| \boldsymbol{x}_{M,i}, \Sigma_{M,i}\right) \quad (7)$$

For high-resolution maps, each atom can be represented by a single Gaussian, whose parameters can be obtained by fitting the electron atomic scattering factors for a given atomic species (*40*). For low-resolution maps or for an efficient initial sampling of high-resolution maps, we use a single Gaussian to represent each coarse-grained bead, with the Gaussian width proportional to the size of the bead. If multiple coarse-grained beads of the model are part of the same rigid body, the parameters of the GMM associated to these beads are computed by applying the Expectation-Maximization algorithm to the positions of the centers of the beads, weighed by their mass. The number of components in the model-GMM is chosen so as to match the number of Gaussians per unit of mass in the data-GMM.

The noise model. The likelihood $p(\phi_D|M)$ is written in terms of the overlap $ov_{MD,k}$ of the k-th component of the data GMM $\phi_{D,k}$ with the entire model ϕ_M, defined as:

$$ov_{MD,k} = \int d\boldsymbol{x}\ \phi_M(\boldsymbol{x})\,\phi_{D,k}(\boldsymbol{x}) \quad (8)$$

Being ϕ_M a GMM, we can write the overlap as the sum overlaps of the individual components:

$$ov_{MD,k} = \sum_j \int d\boldsymbol{x}\ \phi_{M,j}(\boldsymbol{x})\,\phi_{D,k}(\boldsymbol{x}) \quad (9)$$

where the overlap between two Gaussians $\phi_{M,j}$ and $\phi_{D,k}$ is given by:

$$\int d\boldsymbol{x}\ \phi_{M,j}(\boldsymbol{x})\,\phi_{D,k}(\boldsymbol{x}) = \frac{\omega_{M,j}\,\omega_{D,k}}{(2\pi)^{3/2}\left|\Sigma_{M,j}+\Sigma_{D,k}\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}_{M,j}-\boldsymbol{x}_{D,k}\right)^T\left(\Sigma_{M,j}+\Sigma_{D,k}\right)^{-1}\left(\boldsymbol{x}_{M,j}-\boldsymbol{x}_{D,k}\right)\right] \quad (10)$$

We use a log-normal noise model and treat the N_D individual components of ϕ_D as independent pieces of information:

$$p(\phi_D|M) = \prod_{k=1}^{N_D} \frac{1}{\sqrt{2\pi}\,ov_{DD,k}\,\sigma_k} \cdot \exp\left[-0.5\log^2\left(\frac{ov_{DD,k}}{ov_{MD,k}}\right)/\sigma_k^2\right] \quad (11)$$

where σ_k is the unknown tolerance associated with the k-th component of the data GMM and $ov_{DD,k}$ is the overlap of the k-th component with the entire data GMM.

Priors and marginalization. For simplicity, in the following we assume that different parts of the map have the same tolerance σ and we marginalize this variable using an uninformative Jeffreys prior $p(\sigma) = 1/\sigma$. The resulting marginal data likelihood can be written as:

$$p(\phi_D|M) = \frac{2^{-\frac{3}{2}+\frac{N_D}{2}}\,\Gamma\left(\frac{N_D}{2}\right)\left(\sum_{k=1}^{N_D}\log^2\left(\frac{ov_{MD,k}}{ov_{DD,k}}\right)\right)^{-N_D/2}}{\pi^{N_D/2}\prod_{k=1}^{N_D} ov_{DD,k}} \quad (12)$$

Alternatively, one can assume a variable level of noise in the map and marginalize each σ_k using a Jeffreys prior. For the structure X, the prior $p(X)$ depends on the resolution of the model. In case of atomistic representations, we use a molecular mechanics force field, while for more coarse-grained representation a simple excluded volume potential can be used to avoid steric clashes.

Bayesian scoring function. After defining the ingredients of our Bayesian approach and marginalizing the single parameter representing the global uncertainty in the data, the final Bayesian score for cryo-EM maps can be written, omitting constant quantities, as:

$$S(X) = k_B T \cdot \left\{ -\log[p(X)] + \frac{N_D}{2} \log \left[\sum_{k=1}^{N_D} \log^2 \left(\frac{ov_{MD,k}}{ov_{DD,k}} \right) \right] \right\} \quad (13)$$

Benchmark of the divide-and-conquer GMM calculation.

We assessed the accuracy of our divide-and-conquer approach to fitting EM density maps by using experimental density maps of the exosome, a ~400 kDa 10-subunit protein complex (*41-43*) at different resolutions, ranging from 4.2 to 23 Å (**Table 1**). We used the divide-and-conquer approach described above to obtain GMMs of each map with a number of Gaussians varying from 16 to 16384. The accuracy of the fit is defined as the correlation coefficient (*44*) between the EM density map and the map generated by rasterizing the GMM approximation into a 3D grid with the same mesh properties as the original EM density map (i.e., same voxel size, position, and box lengths) (**Fig. 3**). Furthermore, to leverage the effect of the noise, the correlation coefficient is computed using voxels whose density exceeds the recommended threshold value reported in the EMDB. Finally, we heuristically choose the map with the minimal number of Gaussians whose correlation coefficient exceeds 0.95. The simulated GMMs were generated from the reference structures using the program *gmconvert* (*33*). From these data, we determined the stretched-exponential dependence of the resolution of the EM density maps on the number of Gaussians per unit of mass in a GMM with a correlation coefficient of 0.95 (**Fig. 4**).

Benchmark of the modeling protocol.

Data generation. We run the modeling protocol on a benchmark of 12 protein/DNA complexes consisting of 2 to 6 subunits (*45*) (**Table 2**, **Fig. 5**). For each of these complexes, we generated a simulated EM density map, using the coordinates of the PDB structures. We used one Gaussian every 1.090 kDa of assembly mass, which corresponded roughly to the mass of 10 residues and resulted in a resolution of about 11 Å, as obtained by extrapolating from the stretched-exponential regression (**Fig. 4**). For example, the human transferrin receptor complex (PDB 1suv) (*46*) is made of 6 subunits, it has a molecular mass of 290 kDa, and therefore the simulated map was determined using 262 Gaussians.

Subunits representation and forward model. Molecules (protein and DNA chains) are represented by a set of spherical beads, each with the volume of the corresponding residue. When available, the positions of beads are obtained from the PDB structures and are constrained into one or more rigid bodies. Missing regions are constructed as strings of flexible coarse-grained beads (**Figs. S1-S12**). When molecules are intertwined or if a molecule is composed of structurally independent domains, we defined several rigid bodies (ie, 1Z5S, 3LU0, 3PUV, and 3SFD). Furthermore, in some cases, two domains belonging to distinct molecules are merged into the same rigid body, such as the DNA double-strands of 3V6D and 2Y7H or the helical bundle of 3PUV. The model GMM is computed as follows. First, for each rigid body defined above we computed a GMM based on the corresponding atomic coordinates using the implementation of the expectation-maximization algorithm available in the scikit-learn python library (*47*). The number of Gaussians of the GMMs is determined by the molecular weight of the corresponding rigid body, divided by 1090. The center and covariance matrix rotation of each Gaussian are constrained into the corresponding rigid body. Second, each flexible bead is treated as an individual spherical Gaussian.

Model sampling. The positions and the orientations of rigid bodies and flexible beads are initially randomized. The generation of structural models is performed using MC coupled with Replica Exchange (*38*). 40 replicas are used to cover a temperature range between 1 and 2.5 score units (SU). Intermediate temperatures followed a geometrical progression. In addition, we use a simple simulated annealing procedure (*39*) on the weight of the Bayesian EM restraint. We first perform 1000 MC steps with a weight of 0.1, followed by 10000 MC steps with the weight set to 1.0. This alternating pattern is repeated until reaching a total of $3*10^6$ MC steps with a weight of 1.0. We then produce one configuration every 10 MC steps with weight 1.0, for a total of about 300000 models for each system.

Each Monte Carlo step consists of a series of random transformations of the positions of the flexible beads and the rigid bodies. Each flexible bead and rigid body is translated in a random direction by up to 4 Å. In addition, each rigid body is rotated around its center of mass by up to 0.04 radians about a random axis. Furthermore, a similar transformation (random translation of up to 4Å followed by a random rotation of up to 0.04 radian) is applied to the whole system. Each MC step is accepted or rejected according to the Metropolis criterion.

Structural metrics. To compare two models, we used multiple metrics: the root-mean square deviation (rmsd) of residue positions, the rmsd(80), p(10), the positional error and the angular errors of subunits, and the data-model correlation coefficient (CC). The rmsd(80) quantifies the structural difference between two structures as the rms of the 80% smallest pairwise deviations of all residue positions. The rmsd(80) is robust with respect to outliers. For accuracy measurements, the native structure is one of the two structures compared. The deviation of the residue position in two structures is computed between the positions of corresponding centers of the coarse-grained spheres, without structural alignment. When multiple copies of the same protein are present (i.e., pdb codes 1suv, 2wvy, 2y7h, 3lu0, 3nvq, 3pdu, 3puv, 3r5d and 3sfd) ***(46, 48-55)***, one has to determine the copy assignment. The rmsd is initially computed between the two structures by permutation of all possible assignments between the first and the second structure. The rmsd(80) is then computed using the assignment corresponding to the minimal-rmsd permutation. p(10) is the percentage of residues whose deviation between the two

structures is lower than 10 Å. The positional error and the angular errors are respectively the deviation of the positions of the centroids of a subunit in the two structures and the solid angle needed to best-align two subunit structures. The data-model correlation coefficient is defined as $\boldsymbol{CC} = \frac{\boldsymbol{ov_{MD}}}{\sqrt{\boldsymbol{ov_{MM}} \cdot \boldsymbol{ov_{DD}}}}$ and quantifies the agreement of the forward model with the data.

Clustering. For each complex, the 1000 best scoring models are clustered using the following procedure:

1) The best scoring model is assigned to cluster 0.
2) For each model, in order of increasing score, we compute the rmsd(80) with respect to the best scoring model of each cluster. The model is added to the first cluster encountered for which this rmsd(80) is lower than 5 Å. If no cluster is found, then we create a new cluster that initially contains only that model.

For each complex, we compute the cluster dispersion using the Shannon Entropy (SE) measure:

$$SE = \sum_{c \in clusters} -p_c \, log \, (p_c) \quad (14)$$

where p_c is the population of cluster c. A dispersion SE close to 0 indicates very few highly-populated clusters.

Analysis. All models were ranked by the score, and the 1000 best scoring models were considered for further analysis. We clustered all models that had an rmsd(80) from the native structure lower than 5 Å. For each cluster, we computed the precision and the accuracy metrics. The precision is the average rmsd(80) of cluster members with respect to the cluster center. The accuracy of the fit was assessed by computing the rmsd(80), p(10), the overlap between the model- and data- GMMs, as well as the positional and angular errors of the subunits placements for each cluster**.**

Software. All these steps were implemented using IMP.pmi scripts. In particular, the representations and degrees of freedom of each complex were encoded in a standard way using the IMP.pmi topology tables.

Results

Benchmark of the divide-and-conquer GMM calculation.

We initially assessed the accuracy of our divide-and-conquer approach to fitting EM density maps by using experimental density maps of the exosome complex at different resolutions (*41-43*), ranging from 4.2 to 23 Å (**Table 1**). This benchmark revealed that the number of Gaussian components needed to achieve a given accuracy of the final GMM, measured in terms of correlation coefficient with the input map, varies with the resolution of the map (**Fig. 3**). Indeed, for a given number of components, the data-GMM correlation coefficient is lower for higher-resolution maps. In other words, high resolution maps contain more information and therefore require additional components to describe all their features.

Interestingly, our benchmark allows us to classify the maps of this dataset in two clusters: first, the low-information cluster contains all the maps of resolution worse than 10 Å, and second, the high-information cluster that contains all the maps of resolution better than 10 Å. This threshold corresponds to the resolution at which secondary structure elements become visible (*56*).

We calculated the dependence of the resolution of the EM density maps on the number of Gaussians per unit of mass in a GMM with a correlation coefficient of 0.95 (**Fig. 4**). The relationship can be used to: a) estimate the resolution of a GMM generated from a known structure, and b) estimate the number of Gaussians needed to fit a EM density map of a given mass and resolution.

Finally, the divide-and conquer approach allows to overcome the memory and time limitations of the expectation-maximization algorithm when using a large number of Gaussians. In the case of EMD-3366, our approach takes 24 minutes and less than 1GB per process to generate GMMs with 4, 16, 64, 256, 1024 and 4096 components. When performed serially with the *gmconvert* software, a GMM with 4096 requires over 48 hours and 182 GB of memory.

Benchmark of the modeling protocol

We assessed the accuracy of the modeling protocol using a benchmark of 12 protein/DNA complexes consisting of 2 to 6 subunits (**Table 2**) (*45*). The models were fit using a simulated EM density map with a resolution of approximately 11 Å (Materials and Methods). It is worth noting that no additional data (eg, cross-linking experiments) were included, since our specific purpose here is to explore the performance of the EM scoring function alone. Therefore, to some extent, we expect less accurate results than in a real-life application, where multiple datasets would be integrated together to model a complex. The detailed results of the benchmark are reported in **Tables S1-S12**.

Overall, the results of our benchmark suggested that:

1) *The data are explained*. In all cases, the best scoring models fit the data to an overlap higher than 0.7, showing that the sampling explored conformations that very well explain the EM density maps.

2) *Global benchmark accuracy*. The average accuracy p(10) of the benchmark is 85%. p(10) is defined as percentage of residues whose deviation from the native structure is lower than 10 Å.

3) *Classification of the results*. We classified the outcomes of our benchmark into three categories. We consider a *full positive* result when the total rmsd(80) and all the rmsd(80) of the individual subunits are less than 10 Å. A *partial positive* is achieved when the total rmsd(80) is less than 10 Å but some of the subunits are misplaced with an rmsd(80) larger than 10 Å. A *negative* is obtained when the total rmsd(80) is more than 10 Å. Out of the 12 complexes, we obtained 5 full positives (3r5d, 2uzx, 2wvy, 3nvq, 3pdu) (*48, 51, 52, 54, 57*), 3 partial positives (3v6d, 2y7h, 3lu0) (*49, 50, 58*), and 4 negatives (3sfd, 3puv, 1z5s, 1suv) (*46, 53, 55, 59*). In **Fig. 5**, we show one representative example for the cases of full positive (3nvq), partial positive (3lu0), and negative (1suv). Together with the native conformation (**Fig 5A**), we show models from the best scoring cluster (**Fig 5B**), and a color-coded representation of the accuracy of each residue of the best model (**Fig 5C**), along with the plots of the scores of the models as a function of their RMSD with respect to the native state (**Fig 5D**).

The three partial positives had the majority of the residues placed accurately within a rmsd(80)<5.8 and a p(10)>0.89. In each of the three cases we observed a small subunit placed incorrectly. In both 3v6d (*58*) and 2y7h (*49*), a DNA double helix was placed in a wrong orientation with respect to the native structure. For 3lu0 (*50*), the center of mass of subunit E (2.7 % of the total mass) was placed 11.63 Å from the target and rotated by 105°.

The four negatives had a rmsd(80)>17.1 and a p(10)<0.8. In the following paragraphs, we analyze each individual negative case, highlighting the reasons behind the lower accuracy of the reconstructed models.

The 4-subunit **3sfd** (*55*) (rmsd(80)=17.1, p(10)=0.78) failed to reconstruct the helical bundle made by chains C and D (20% of the total mass). The remaining chains A and B (80% of the total mass) were placed with an rmsd(80) of 1.1 and 2.7 Å respectively. The 14 clusters generated from the 1000 best models were very disperse (SE=2.47), and none of the clusters contained a structure close to the native state. In contrast, the PSE (Percentage Score Error) was equal to 0%, meaning that the best-scoring model scores as well as the native state, indicating a sampling problem.

The best-scoring model of the 5-subunit **3puv** (*53*) had the total rmsd(80)=19.3 and p(10)=0.66. The clustering displayed dispersion (SE=1.4) over a total of 5 clusters. In the best scoring cluster, chains A and B (39% of the mass of the complex) were incorrectly placed (rmsd(80) 40Å and 19Å, respectively). In the second best-scoring cluster, subunits A and B are well placed, but all other chains are misplaced. However, the mass of the best scoring model well explains the EM data, having a data-model correlation coefficient of 0.75. PSE is -4%, meaning that the native conformation scores better than the best-scoring model.

The 4-subunit **1z5s** (*59*) had rmsd(80) of 13.1 Å and p(10) of 0.67, and was the smallest complex (m.w. 49kDa). The models were clustered into 19 clusters with a very high dispersion SE=2.92. The best scoring cluster had subunits A and C correctly placed (rmsd(80)=5.3 and 3.8 respectivelly), and subunits B and D were were correctly centered, but mis-rotated resulting in rmsd(80) of 17.2 and 27.0, respectively. The PSE is +8%, meaning that the native conformation scores worse than the best-scoring

model. The reason is that data resolution (11 Å) is insufficient to correctly position the small B and D subunits (78 and 65 residues respectively).

The 6-subunit **1suv** (*46*) (**Fig. 5, Right**) has a rmsd(80) of 27.9 and p(10) of 0.5. This complex has three different proteins A, C and E which have identical copies B, D, and F. The models were grouped into 2 clusters, with little dispersion (SE=0.57). In all clusters subunits A and B (53% of the total mass) are correctly placed with rmsd(80) of about 2 Å. Subunits D and F (26% of the total mass) are correctly centered but mis-rotated. Chain C and E are wrongly centered and rotated. Again, the data-model correlation coefficient is 0.75, and the PSE is +50%. However, it should be noted that proteins C and E (as well as the corresponding copies) are structurally very similar. Therefore, they could be considered indistinguishable upon permutation when computing the rmsd. In fact, permuting the positions of these four subunits results in similar scores. When computing the accuracy of the sample considering C, D, E, and F as interchangeable, we found that all the subunits were correctly centered. The best scoring model has an rmsd(80) of 15.1Å and a p(10) of 0.65. The source of the inaccuracy of the models is attributed to mis-rotations of subunits C, D, E and F. Given that the model-GMMs of these subunits are roughly spherical, it is likely that mis-rotations of these subunits are not penalized by the Bayesian EM scoring function.

Discussion and Conclusions

A major difficulty of integrative modeling, in which data of different nature are combined to model the structure of a biological complex, is to determine the unknown relative weight of each piece of information. Inaccurate weighing results in models that are biased towards a particular type of data. To optimally weigh each piece of information, one should consider two factors: the accuracy or level of noise in the data and the correlation between data points. These two factors determine the overall information content of each source of data. Very noisy experimental data obviously provide lower structural information than highly accurate and precise data. Also, two measurements of the same structural feature by repeating the same experiment cannot be considered as independent observations and therefore they do not provide a substantial amount of additional information on the system and should not be simply counted twice.

In the Bayesian framework, the problem of structural determination from multiple sources of information is formulated in terms of a probability distribution, or posterior, that objectively ranks models by quantifying the corresponding information content. Here we introduced a Bayesian approach to cryo-EM data that can be generally used in integrative approaches across multiple levels of resolution of the input map. This approach addresses the problem of quantifying both the level of noise and the data correlation. Experimental density maps are decomposed into a minimal Gaussian Mixture Model (GMM) that reproduces all the structural features of the data as well as the density resolution. Each Gaussian represents a subset of the voxels in the original experimental map and approximates the unknown correlation, or covariance, between neighboring voxels. Therefore, determining the Data-GMM from the experimental map is equivalent to data-clustering, where correlated data points (voxels) are grouped into a minimal set of independent descriptors (Gaussians).

On the other hand, a molecular model, consisting of particles, either atomic or pseudo-atomic, is naturally described by a Model-GMM, where a Gaussian is centered on the coordinates of each particle, and the width and masses are derived from the particle mass and volume. The likelihood function in Eq. 11 then quantifies the discrepancy between data and model GMM in the following way. First, the model and the data are compared using the metric $\log^2\left(\frac{ov_{DD,k}}{ov_{MD,k}}\right)$. The rationale behind this metric is that a Model-GMM reproduces the *k*-Data-Gaussian when the overlap (**Eq. 10**) between the model and *k*-Data-Gaussian is similar to the overlap of the Data-GMM and the *k*-Data-Gaussian. Second, the metric described above is modulated by an unknown parameter σ_k, that quantifies the level of noise in the region of the experimental map described by the k-Data-Gaussian. Finally, since the components of the GMM can now be considered as independent data points, the global likelihood is written as a product of N_D (number of Gaussians in the Data-GMM) terms, each one assessing the proximity of the model to the experiment using the metric above. When investigating the scoring function, N_D is number of score terms which are driving the model close to the target for each Data-Gaussian, therefore the weight. As desired, decreasing/increasing the resolution of the data (**Fig. 4**), fewer/more Gaussians are needed to reproduce the density map, and as a consequence the restraint have a lower/higher weight.

The extensive benchmark of our modeling protocol based on the Bayesian EM scoring function demonstrated that our approach is, in most cases, capable of correctly positioning and orienting the components of a macromolecular complex. The few cases in which our approach was inaccurate were characterized by the fact that multiple different placements of the subunits could result in a similar overall density. For example, helical bundles are difficult to model at low resolution because they only define a "cylinder" in which two or more helices can be positioned in multiple ways. Similarly, pseudo spherical subunits (such as subunits C-F of 1suv) can be rotated around their center of mass with only minimal penalty. Moreover, the placement of DNA helices is degenerate, because their expected density is symmetric by rotation. However, these issues are not specific to the modeling protocol presented here, but to various extents shared among all techniques aimed at modeling architectures from low-resolution EM data. These results highlight the importance of integrative approaches when dealing with medium- to low-resolution data. In our case, our Bayesian framework allows the correct integration of EM data with other sources of information, such as cross-linking/mass spectrometry data and labelling experiments, which will resolve the inaccuracies in the orientation of the subunits.

EMDB #	Molecular Weight (kDa)	Resolution [Å]	Number of Gaussians at cc=0.95	Reference
3366	420	4.2	9600	(*41*)
3369	420	5.8	4300	(*41*)
3372	350	6.3	4160	(*41*)
3370	350	6.7	3300	(*41*)
3371	350	11	256	(*41*)
3367	350	11.5	150	(*41*)
3368	350	13	360	(*41*)
1708	400	14	16	(*42*)
1438	400	19	41	(*43*)
1439	300	23	16	(*43*)

Table 1: Dataset used to benchmark the divide-and-conquer approach to the GMM creation

pdb id	Reference	Number of subunits	Percentage score error	data-model correlation coefficient	Number of Clusters	Clusters dispersion (SA)	rmsd(80) [Å]	p(10)	Average position error [Å]	Average angular error [°]	Number of misplaced subunits	Result
2uzx	(*57*)	2	0.01	0.79	1	0.00	2.10	0.95	1.49	3.59	0	Full Positive
2wvy	(*48*)	3	0.01	0.96	1	0.00	1.24	0.99	0.89	2.61	0	Full Positive
3nvq	(*51*)	4	0.62	0.96	1	0.00	1.06	1.00	0.75	2.76	0	Full Positive
3pdu	(*52*)	4	0.05	0.84	10	2.04	2.53	1.00	1.91	7.58	0	Full Positive
3r5d	(*54*)	3	-0.03	0.77	6	1.37	5.81	0.99	2.51	7.80	0	Full Positive
2y7h	(*49*)	5	0.49	0.81	2	0.25	6.97	0.98	2.18	6.77	2	Partial Positive
3lu0	(*50*)	5	0.02	0.84	1	0.00	4.59	0.90	1.77	9.56	1	Partial Positive
3v6d	(*58*)	4	0.04	0.87	2	0.55	7.43	0.95	1.42	7.49	2	Partial Positive
1suv	(*46*)	6	0.50	0.75	2	0.57	15.09	0.65	2.74	60.72	3	Negative
3puv	(*53*)	5	-0.04	0.75	5	1.40	27.92	0.64	8.35	36.92	2	Negative
3sfd	(*55*)	4	0.00	0.78	14	2.47	12.81	0.78	5.29	38.65	2	Negative
1z5s	(*59*)	4	0.08	0.58	19	2.92	13.07	0.64	4.64	60.96	2	Negative

Table 2: Results of the modeling protocol benchmark.

References

1. W. Kuhlbrandt, The resolution revolution. *Science* **343**, 1443-1444 (2014).
2. E. Nogales, The development of cryo-EM into a mainstream structural biology technique. *Nat Methods* **13**, 24-27 (2016).
3. E. Callaway, The Revolution Will Not Be Crystallized. *Nature* **525**, 172-174 (2015).
4. X. C. Bai, G. McMullan, S. H. Scheres, How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences* **40**, 49-57 (2015).
5. R. M. Glaeser, How good can cryo-EM become? *Nature Methods* **13**, 28-32 (2016).
6. X. M. Li *et al*., Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584-+ (2013).
7. X. Yu, L. Jin, Z. H. Zhou, 3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature* **453**, 415-419 (2008).
8. X. C. Bai, I. S. Fernandez, G. McMullan, S. H. Scheres, Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2**, e00461 (2013).
9. M. G. Campbell, D. Veesler, A. Cheng, C. S. Potter, B. Carragher, 2.8 A resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *eLife* **4**, (2015).
10. X. C. Bai *et al*., An atomic structure of human gamma-secretase. *Nature* **525**, 212-217 (2015).
11. A. Bartesaghi *et al*., 2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. *Science* **348**, 1147-1151 (2015).
12. D. Schneidman-Duhovny, R. Pellarin, A. Sali, Uncertainty in integrative structural modeling. *Curr Opin Struct Biol* **28**, 96-104 (2014).
13. M. Bonomi, G. T. Heller, C. Camilloni, M. Vendruscolo, Principles of protein structural ensemble determination. *Curr Opin Struct Biol* **42**, 106-116 (2017).
14. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci Adv* **2**, e1501177 (2016).
15. A. B. Ward, A. Sali, I. A. Wilson, Biochemistry. Integrative structural biology. *Science* **339**, 913-915 (2013).
16. G. F. Schroder, Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr Opin Struct Biol* **31**, 20-27 (2015).
17. J. R. Lopez-Blanco, P. Chacon, Structural modeling from electron microscopy data. *Wires Comput Mol Sci* **5**, 62-81 (2015).
18. E. F. Pettersen *et al*., UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605-1612 (2004).
19. M. G. Rossmann, R. Bernal, S. V. Pletnev, Combining electron microscopic with X-ray crystallographic structures. *Journal of Structural Biology* **136**, 190-200 (2001).
20. A. Sali, T. L. Blundell, Comparative Protein Modeling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **234**, 779-815 (1993).
21. W. Wriggers, Conventions and workflows for using Situs. *Acta Crystallogr D* **68**, 344-351 (2012).
22. K. Lasker, M. Topf, A. Sali, H. J. Wolfson, Inferential Optimization for Simultaneous Fitting of Multiple Components into a CryoEM Map of Their Assembly. *Journal of Molecular Biology* **388**, 180-194 (2009).
23. W. S. Zheng, Accurate Flexible Fitting of High-Resolution Protein Structures into Cryo-Electron Microscopy Maps Using Coarse-Grained Pseudo-Energy Minimization. *Biophys J* **100**, 478-488 (2011).

24. L. G. Trabuco, E. Villa, K. Mitra, J. Frank, K. Schulten, Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673-683 (2008).
25. A. H. Ratje *et al.*, Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites. *Nature* **468**, 713-716 (2010).
26. M. Saha, M. C. Morais, FOLD-EM: automated fold recognition in medium- and low-resolution (4-15 A) electron density maps. *Bioinformatics* **28**, 3265-3273 (2012).
27. F. DiMaio, M. D. Tyka, M. L. Baker, W. Chiu, D. Baker, Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta. *Journal of Molecular Biology* **392**, 181-190 (2009).
28. S. Lindert *et al.*, EM-Fold: De Novo Atomic-Detail Protein Structure Determination from Medium-Resolution Density Maps. *Structure* **20**, 464-478 (2012).
29. D. Russel *et al.*, Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* **10**, e1001244 (2012).
30. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180**, 519-530 (2012).
31. P. D. Adams *et al.*, The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94-106 (2011).
32. W. Rieping, M. Habeck, M. Nilges, Inferential structure determination. *Science* **309**, 303-306 (2005).
33. T. Kawabata, Multiple Subunit Fitting into a Low-Resolution Density Map of a Macromolecular Complex Using a Gaussian Mixture Model. *Biophys J* **95**, 4643-4658 (2008).
34. P. J. Robinson *et al.*, Molecular architecture of the yeast Mediator complex. *Elife* **4**, (2015).
35. S. Jonic *et al.*, Denoising of high-resolution single-particle electron-microscopy density maps by their approximation using three-dimensional Gaussian functions. *J Struct Biol* **194**, 423-433 (2016).
36. J. P. Erzberger *et al.*, Molecular architecture of the 40SeIF1eIF3 translation initiation complex. *Cell* **158**, 1123-1135 (2014).
37. J. Fernandez-Martinez *et al.*, Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167**, 1215-+ (2016).
38. R. H. Swendsen, J. S. Wang, Replica Monte-Carlo Simulation of Spin-Glasses. *Physical Review Letters* **57**, 2607-2609 (1986).
39. S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by Simulated Annealing. *Science* **220**, 671-680 (1983).
40. E. Prince, *International Tables for Crystallography Vol. C*. (Wiley, Hoboken, ed. 3rd, 2004), pp. 1032 p.
41. J. J. Liu *et al.*, CryoEM structure of yeast cytoplasmic exosome complex. *Cell Res* **26**, 822-837 (2016).
42. H. Malet *et al.*, RNA channelling by the eukaryotic exosome. *Embo Rep* **11**, 936-942 (2010).
43. H. W. Wang *et al.*, Architecture of the yeast Rrp44-exosome complex suggests routes of RNA recruitment for 3 ' end processing. *P Natl Acad Sci USA* **104**, 16844-16849 (2007).
44. D. Frenkel, B. Smit, *Understanding molecular simulation : from algorithms to applications*. Computational science (Academic, San Diego, Calif. ; London, ed. 2nd, 2002), pp. xxii, 638 p.
45. J. Velazquez-Muriel *et al.*, Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. *P Natl Acad Sci USA* **109**, 18821-18826 (2012).
46. Y. Cheng, O. Zak, P. Alsen, S. C. Harrison, T. Walz, Structure of the human transferrin receptor-transferrin complex. *Cell* **116**, 565-576 (2004).

47. F. Pedregosa *et al*., Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).
48. Y. P. Zhu *et al*., Mechanistic insights into a Ca2+-dependent family of alpha-mannosidases in a human gut symbiont. *Nat Chem Biol* **6**, 125-132 (2010).
49. C. K. Kennaway *et al*., The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein. *Nucleic Acids Res* **37**, 762-770 (2009).
50. N. Opalka *et al*., Complete Structural Model of Escherichia coli RNA Polymerase from a Hybrid Approach. *Plos Biology* **8**, (2010).
51. H. L. Liu *et al*., Structural Basis of Semaphorin-Plexin Recognition and Viral Mimicry from Sema7A and A39R Complexes with PlexinC1. *Cell* **142**, 749-761 (2010).
52. Y. F. Zhang *et al*., Structural characterization of a beta-hydroxyacid dehydrogenase from Geobacter sulfurreducens and Geobacter metallireducens with succinic semialdehyde reductase activity. *Biochimie* **104**, 61-69 (2014).
53. M. L. Oldham, J. Chen, Snapshots of the maltose transporter during ATP hydrolysis. *P Natl Acad Sci USA* **108**, 15152-15156 (2011).
54. R. Schnell *et al*., Tetrahydrodipicolinate N-Succinyltransferase and Dihydrodipicolinate Synthase from Pseudomonas aeruginosa: Structure Analysis and Gene Deletion. *Plos One* **7**, (2012).
55. Q. J. Zhou *et al*., Thiabendazole inhibits ubiquinone reduction activity of mitochondrial respiratory complex II via a water molecule mediated binding feature. *Protein Cell* **2**, 531-542 (2011).
56. M. L. Baker, T. Ju, W. Chiu, Identification of secondary structure elements in intermediate-resolution density maps. *Structure* **15**, 7-19 (2007).
57. H. H. Niemann *et al*., Structure of the human receptor tyrosine kinase met in complex with the Listeria invasion protein InlB. *Cell* **130**, 235-246 (2007).
58. K. Das, S. E. Martinez, J. D. Bauman, E. Arnold, HIV-1 reverse transcriptase complex with DNA and nevirapine reveals non-nucleoside inhibition mechanism. *Nature Structural & Molecular Biology* **19**, 253-259 (2012).
59. D. Reverter, C. D. Lima, Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex. *Nature* **435**, 687-692 (2005).

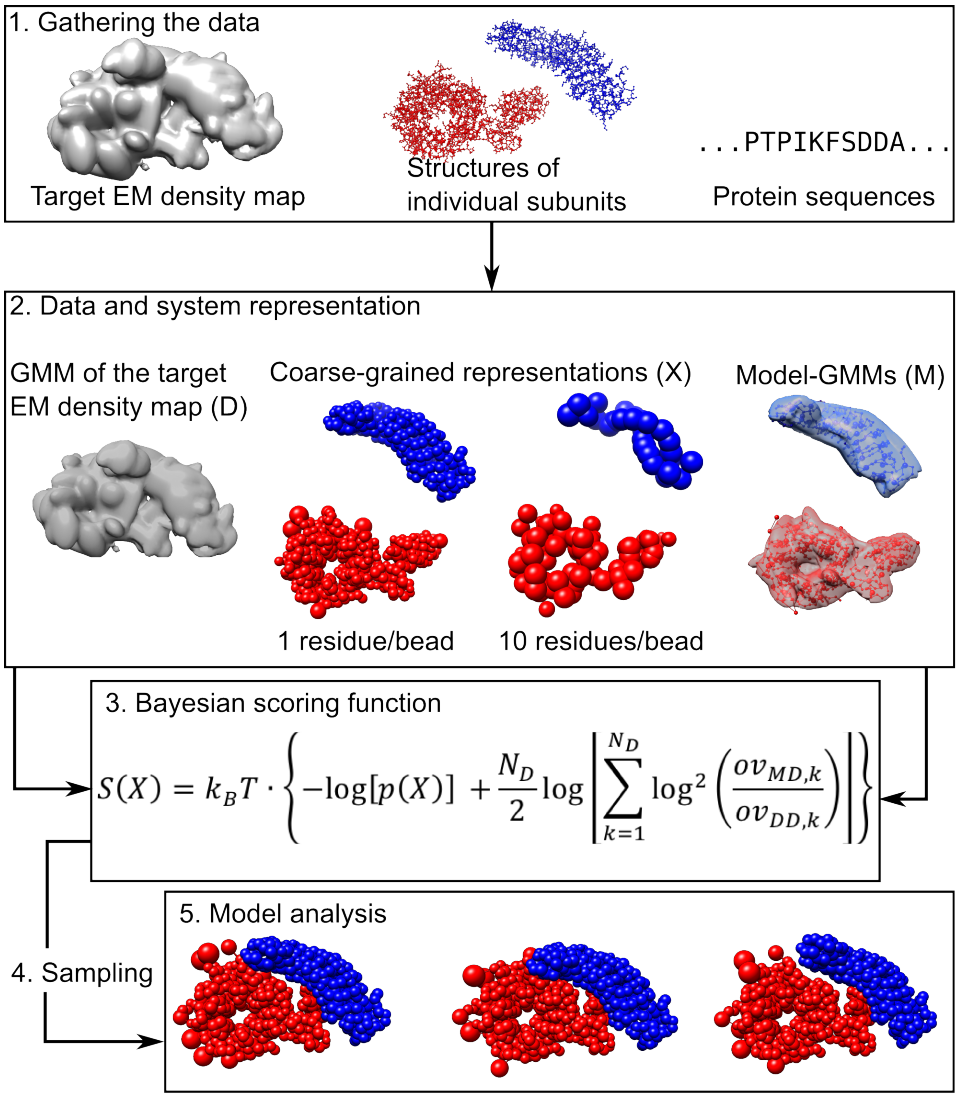

Figure 1: Workflow for low-resolution modeling of EM data. (1) The input information for the modeling protocol is an experimental cryo-EM density map, the (partial) structures of the subunits, and the subunit sequences, which are needed to build missing regions. (2) The density map is fitted with a GMM using our divide-and-conquer approach. In the benchmark, the GMM is obtained directly from the atomic coordinates of the reference assembly structure. The subunits of the complex are represented at coarse-grained level, their model-GMMs are computed and their initial positions randomized. (3) The Bayesian scoring function encodes prior information about the system and measures the agreement between the map predicted from the model and the experimental map. (4) Structural models are sampled by Monte Carlo coupled with Replica Exchange and Simulated Annealing. (5) The generated models are clustered and the precision and accuracy of the ensemble of solutions is assessed.

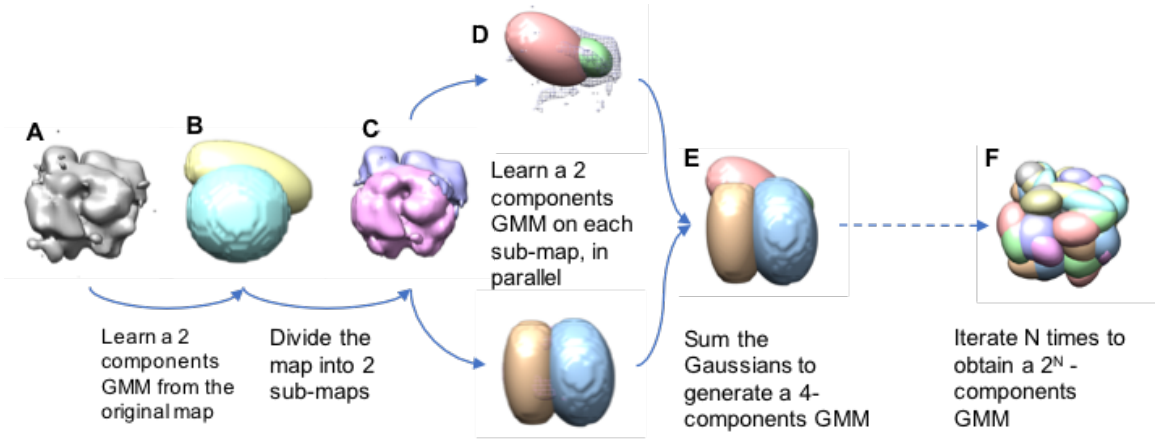

Figure 2: Divide-and-conquer approach for fitting EM density maps with a Gaussian Mixture Model (GMM). The input map (A) is initially fitted using a GMM with a small number of Gaussians (B). Each component of the GMM is used to partition the map into overlapping submaps (C) and each sub-map is then fit using a GMM with a small number of components (D). The sum of the Gaussian of all the GMM for all the submaps result in a GMM that fit the orginal map (E). The fit procedure is reiterated until a global GMM with desired accuracy is obtained (F).

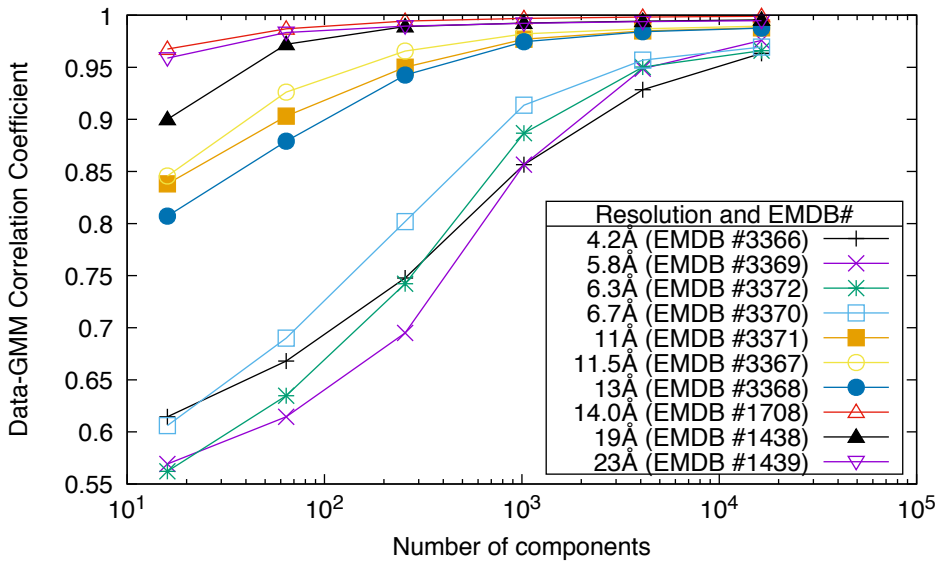

Figure 3: Benchmark of the divide-and-conquer approach to fit EM density maps with a GMM. The accuracy of our divide-and-conquer approach is measured by the correlation coefficient between the input map and the GMM representation. The accuracy increases with the number of components (Gaussians) of the mixture, at a speed that depends on the resolution of the experimental map. The input densities belong to the eukaryotic exosome complex at different resolutions.

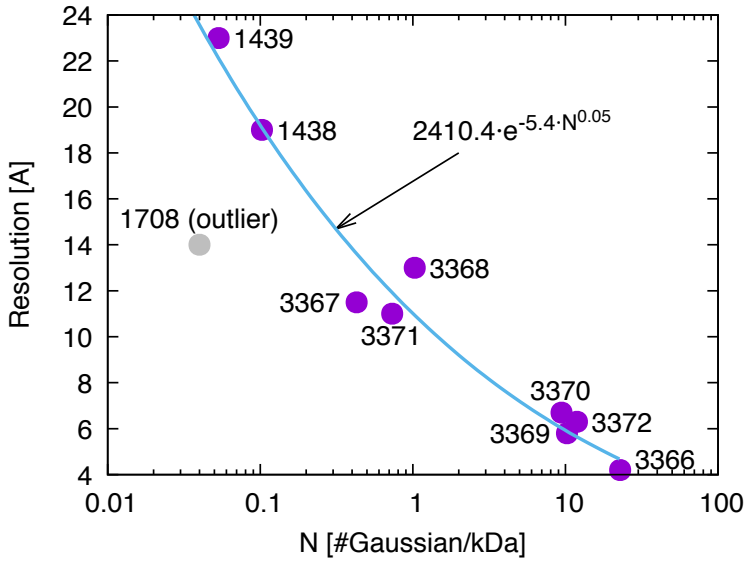

Figure 4: Relationship between map resolution and number of Gaussians in the GMM. For all EMDB maps of Fig. 2, the corresponding resolution is reported as a function of the number of Gaussians needed to achieve a correlation coefficient of 0.95 (purple dots). The number of Gaussians is normalized by the molecular weight of the complex. The points are fit using a stretched exponential regression (blue line). The gray data-point (EMDB 1708) is excluded from the fit as the reported resolution was not determined using the Fourier Shell Correlation 0.143 gold standard.

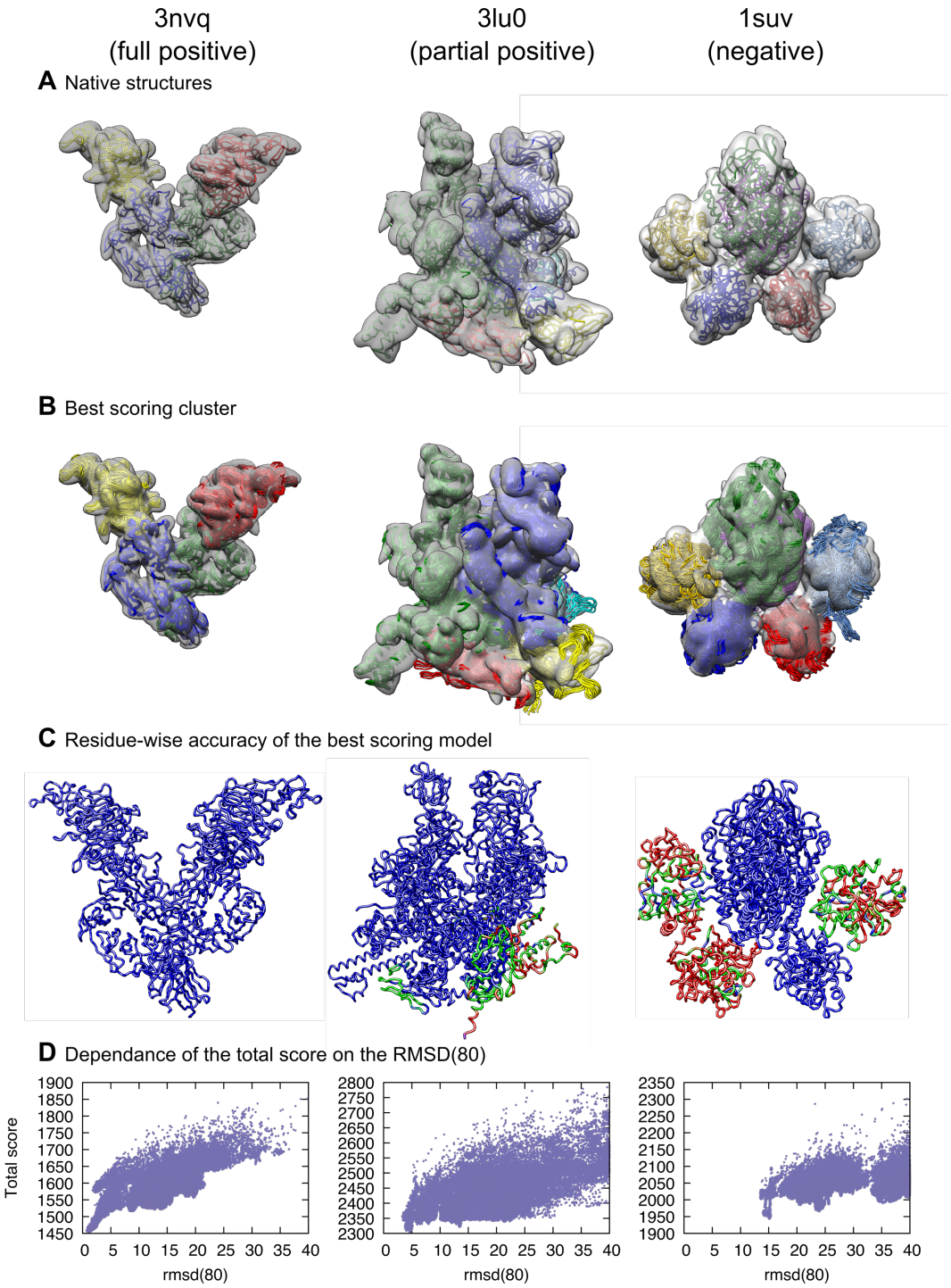

Figure 5: Representative examples of three possible outcomes of the benchmark. (A) Native structures and simulated EM density maps. (B) Best scoring models represented in the simulated EM density maps. Chains of panel B are colored as in panel A. (C) Residue-wise accuracy of the best scoring models: residues with deviation from the native structure less than 10 Å, between 10 and 20 Å, and above 20 Å are colored in blue, green and red, respectively. (D) Total score of all the sampled models as a function of the rmsd(80) from the native structure.