Bayesian multi-scale modeling of macromolecular structures based on 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.

‡Lead contact

Summary

Cryo-electron microscopy (cryo-EM) has become a mainstream technique for the determination of structures of biological architectures. However, accurate structural modeling has been hampered by difficulties with objectively weighing cryo-EM-derived restraints against other sources of information due to random and systematic errors in the data and sample heterogeneity. To address this challenge, we introduce a Bayesian scoring function that efficiently and accurately ranks alternative structural models of a macromolecular system based on its cryo-EM density map, and that is compatible with atomic and coarse-grained model representations. The accuracy of our approach is benchmarked using several complexes of known structure and illustrated in two real applications, in which cryo-EM is combined with chemical cross-linking/mass spectrometry data to model the structure of the RNA polymerase II and of the exosome complexes. The approach is implemented in the open-source *Integrative Modeling Platform* (http://integrativemodeling.org), enabling integrative structure determination by combining cryo-EM data with other information.

Introduction

Over the last two decades, cryo-electron microscopy (cryo-EM) has enabled structural characterization of complex biological systems that were beyond the capabilities of traditional techniques, such as X-crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy [1-3]. This progress has been fuelled by the continuous advances in both instrumentation and software for cryo-EM image processing [4-6]. As a result, the resolution of the structures from cryo-EM is rapidly approaching the atomic resolution of X-ray crystallography, as illustrated by structures of many complex systems, 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, unlike X-ray crystallography, needs a small amount of sample, does not require isotopic labeling, and is applicable to systems larger than approximately 100 kDa. Furthermore, cryo-EM has the potential to identify multiple different structural states in a single experiment [2-5], provided that they can be disentangled during image classification analysis.

Despite these recent advances, the translation of cryo-EM three-dimensional (3D) reconstructions into structural models still presents several challenges. First, cryo-EM density 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 progress in methods for classifying 2D images of single-particles into class averages and for converting them into 3D reconstructions, the final density maps might still depend on images of particles in different conformations [14]. As a result, areas of the map corresponding to regions of high flexibility typically present lower resolutions. Thus, an individual molecular structure might not be able to fit the entire density map with the same accuracy and precision. Finally, cryo-EM maps are typically defined by a set of data points representing the electron density on a grid in real space (i.e., voxels). Neighbouring voxels, which correspond to regions close in space, are not expected to provide independent information about the system, but to be affected by a certain degree of correlation. Accounting for such correlation is particularly relevant when integrating cryo-EM with other experimental data [15], as the information content of each piece of data needs to be accurately weighed to avoid biasing a model [12]. Consequently, there is still a pressing need for modeling approaches that can correctly weigh cryo-EM data taking into account random and systematic error in the data, data-point correlation, conformational and compositional heterogeneity to generate accurate and precise structural models from cryo-EM density maps.

A number of approaches have been proposed to model macromolecular structures based on cryo-EM density maps [16, 17]. Generally speaking, these techniques can use one or more of the following strategies: rigid-body fitting of components of known structure, flexible refinement, use of homology modeling or *de novo* protein structure prediction of the components, and integrative modeling based on multiple types of experimental data. The most popular software packages for cryo-EM-based modeling include Chimera [18], EMfit [19], Modeller [20], SITUS [21], MultiFit [22], EMFF [23], MDFF [24],

Flex-EM [25], γ-TEMPy [26], COAN [27], MDFIT [28], Fold-EM [29], ROSETTA [30], EM-fold [31], IMP [32], RELION [33], ISD [34], and Phenix [35]. The majority of these approaches generate structural models that minimize the deviation between observed and predicted cryo-EM density maps, including by Molecular Dynamics (MD), Monte Carlo (MC) or Normal Modes Analysis techniques [17]. Typically, this deviation is measured by the cross-correlation between experimental and predicted maps calculated over the set of voxels. Additional pseudo-energy terms are routinely applied to enforce correct stereochemistry. Despite the utility of these methods, it is still a challenge to tackle cryo-EM data weighing, the presence of error in the data, data-point correlation, and sample heterogeneity.

Here, we introduce an approach to model a given macromolecular system based on its cryo-EM density map, using a Bayesian scoring function to rank a set of models compatible with the input information [36]. This scoring function objectively interprets experimental data and integrates them with prior knowledge. The method is based on an analytical representation of the input data in terms of a Gaussian Mixture Model (GMM) [37-39], rather than the more standard voxel-based approach. This choice has several advantages: a) it circumvents the problem of data-point correlation, by assigning the original correlated data-points to a set of uncorrelated GMM components (i.e., the Gaussians); b) it is computationally efficient; and c) it enables 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 while simultaneously automatically quantifying the level of noise in the data. Thus, different experimental data are properly integrated by weighing each piece of information according to its noise content.

Next, 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. Finally, we apply our approach to the integrative modeling of the RNA polymerase II and of the exosome complexes, in which we combine an experimental cryo-EM density map with chemical cross-linking/mass spectrometry (XL-MS) data. Our method is implemented in the open-source Integrative Modeling Platform (IMP) package (http://integrativemodeling.org) [32], thus enabling integrative structure determination of biological systems based on a variety of experimental data, including FRET and NMR spectroscopies, XL-MS, small angle X-ray scattering, 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 multi-scale modeling of cryo-EM data.

Materials and Methods

Protocol for low-resolution modeling of cryo-EM density maps

We implemented in IMP [32] a pipeline that enables the multi-scale modeling of macromolecular structures based on cryo-EM data, given partial knowledge of subunits structures. The general 4-stage procedure is as follows (**Fig. 1**):

1) Gather the data, including the sequences of subunits, their structures (*eg*, from X-ray crystallography, NMR spectroscopy, homology modeling, and *ab initio* prediction), and the target 3D density map reconstructed form the electron microscopy data (**Fig. 1.1**).
2) Convert the data into a scoring function for ranking alternative structural models:
    A. Generate a GMM representation of the 3D density map (data-GMM) by using the divide-and-conquer algorithm described below (**Fig. 1.2A**).
    B. Assign a representation to the different components of the complex (**Fig. 1.2B**). Subunits are represented by spherical beads, corresponding to a coarse-grained representation of the atomic degrees of freedom. For a given domain, the beads are either constrained into a rigid body or allowed to move flexibly, depending on the uncertainty about the domain structure (*cf*, crystallographic structure *vs ab initio* model). The beads represent one or more contiguous residues, depending on the level of coarse-graining [38, 40, 41].
    C. The electron density of the model is also described by a GMM (model-GMM) and will be used to compute the fit of the model to the cryo-EM density map (**Fig. 1.2C**).
    D. The final scoring function that 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 cryo-EM data, and prior terms such as the bead sequence connectivity and excluded volume (**Fig. 1.2D**).
3) Sample models using MC and Replica Exchange methods [42], with an iterative approach to maximize sampling exhaustiveness (**Fig. 1.3**).
4) Analyze the sampled models in terms of their variability and clustering, data satisfaction, *etc.* (**Fig. 1.4**).

A Bayesian approach to ranking models based on cryo-EM density maps

In general terms, the Bayesian approach [36] 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 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 restraints that apply to all pairs of consecutive beads in the sequence, implied by 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 its residues. The structures derived from X-ray data or homology models are coarse-grained using two categories of resolution, where beads are represented either individual residues or segments of up to 10 residues. Beads can be constrained into a rigid body, in which relative distances are fixed during sampling. Alternatively, strings of beads representing parts without structural information can be flexible with respect to each other. In the following, we define the components of the Bayesian scoring function specifically for an cryo-EM density map.

Experimental cryo-EM density map. We represent the experimental density map Ψ_D in terms of a Gaussian Mixture Model (GMM) ϕ_D^j with j components (*ie*, data-GMM):

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

where $\omega_{D,i}^j$ is the (normalized) weight of the i-th component of the GMM and G is 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 three 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. Finally, it allows representing the density map at multiple resolutions, which is exploited here to accelerate sampling of structural models compatible with the data (*ie*, see Model Sampling paragraph below).

The posterior probability of model M given the cryo-EM density map 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 $p\left(\phi_D^j\middle|\Psi_D\right)$ selects 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 data-GMM. To fit the experimental density map Ψ_D with a GMM ϕ_D, we used the Expectation Maximization algorithm implemented in the *gmconvert* software [37]. This approach determines the parameters of the GMM (mean, weight, and covariance matrix of each Gaussian component) 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 increases exponentially along with the computational time and memory required to perform the fit. To overcome these challenges, 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 database are removed. Second, a recursive procedure starts from a first iteration in which the map Ψ_D (**Fig. 2A**) is fit 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 partition the original map into sub-maps $\Psi_{D,i}$ (**Fig. 2C**):

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

This partitioning has two properties: *a)* each sub-map selects the part of the original map that overlaps with the component ($\phi_{D,i}$); *b)* the sum of all sub-maps 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 sub-map $\Psi_{D,i}$ is again fitted using a GMM with a small number of Gaussians N_D (**Fig. 2D**), dividing the process into as many branches as the number of Gaussians N_D.

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**). This procedure is repeated until the global GMM reaches the desired accuracy (**Fig. 2F**). The accuracy of the fit was defined as the correlation coefficient CC [43] between the cryo-EM density map Ψ_D and the map generated by rasterizing the

data-GMM into a 3D grid with the same mesh properties as the original density map (*ie*, voxel size, offsets, and box lengths) (**Fig. 3A**). The CC was computed using only those voxels whose density exceeds the recommended threshold value reported in the EMDB. A given branch was stopped when the local CC between a sub-map and its GMM was greater than 0.95. Since the resolution of the original map can vary locally, individual branches will be terminated at different iterations.

This procedure generates at each step a global data-GMM with increasing number of components, thus with increasing resolution. To quantify the resolution of each of these global data-GMMs, we computed their Fourier Shell Correlations (FSCs) with respect to the original map Ψ_D. By analogy with the method of the two half-maps [44], the resolution was defined as the inverse of the frequency at which the FSC crossed the 0.143 threshold (**Fig. S1A**). Finally, we defined the *optimal* data-GMM as the fit with resolution closest to the original map Ψ_D (**Fig. S1B**).

The entire process was parallelized to run efficiently on a computer cluster.

The forward model. We developed a forward model to compute an cryo-EM density map from a single structural model. As for the data representation above, the forward model ϕ_M is a GMM with N_M components (*ie*, model-GMM):

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

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 atom type [45]. 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 model-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 noise model. Most of the refinement techniques use the correlation coefficient (CC) as measure of agreement between predicted and observed density maps [17]. In our approach, we use a different metric and define the global overlap between model and data density maps as:

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

The standard CC can then be expressed in terms of the overlap functions as [38]:

$$\mathrm{CC} = \frac{2ov_{MD}}{ov_{MM} + ov_{DD}} \quad (9)$$

Notably, maximum correlation is obtained with maximum overlap ov_{MD}, since the quantities at the denominator of Eq. 9 do not depend on the coordinates of the particles in the structural model.

The global overlap ov_{MD} can be expressed in terms of local overlaps $ov_{MD,k}$ between model and the k-th component of the data-GMM $\phi_{D,k}$:

$$ov_{MD} = \sum_{k=1}^{N_D} ov_{MD,k} = \sum_{k=1}^{N_D} \int d\boldsymbol{x}\ \phi_M(\boldsymbol{x})\ \phi_{D,k}(\boldsymbol{x}) \quad (10)$$

Each local overlap measures the agreement of the model with the part of the experimental density map represented by a component of the data-GMM. Because ϕ_M is also a GMM, we can write the local overlap as the sum of overlaps for the individual components:

$$ov_{MD,k} = \sum_{j} \int d\boldsymbol{x}\ \phi_{M,j}(\boldsymbol{x})\ \phi_{D,k}(\boldsymbol{x}) \quad (11)$$

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

$$\int d\boldsymbol{x}\ \phi_{M,j}(\boldsymbol{x})\ \phi_{D,k}(\boldsymbol{x}) = \frac{\omega_{M,j}\ \omega_{D,k}}{(2\pi)^{3/2}\left|\Sigma_{M,j} + \Sigma_{D,k}\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}_{M,j} - \boldsymbol{x}_{D,k}\right)^T\left(\Sigma_{M,j} + \Sigma_{D,k}\right)^{-1}\left(\boldsymbol{x}_{M,j} - \boldsymbol{x}_{D,k}\right)\right] \quad (12)$$

We treat the N_D individual components of the data-GMM as independent pieces of information and express the data likelihood in terms of local overlaps $ov_{MD,k}$, using a log-normal noise model:

$$p(\phi_D|M) = \prod_{k=1}^{N_D} \frac{1}{\sqrt{2\pi}\ ov_{DD,k}\ \sigma_k} \cdot \exp\left[-0.5 \log^2\left(\frac{ov_{DD,k}}{ov_{MD,k}}\right) / \sigma_k^2\right] \quad (13)$$

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.

Marginal likelihood. 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:

$$p(\phi_D|M) = \frac{2^{-\frac{3}{2}+\frac{N_D}{2}}\Gamma\left(\frac{N_D}{2}\right)\left(\prod_{k=1}^{N_D}\log^2\left(\frac{ov_{MD,k}}{ov_{DD,k}}\right)\right)^{-N_D/2}}{\pi^{N_D/2}\prod_{k=1}^{N_D} ov_{DD,k}} \quad (14)$$

Alternatively, one can assume a variable level of noise in the map and marginalize each σ_k using a Jeffreys prior. The marginal likelihood in Eq. 14 is maximized when the local overlap $ov_{MD,k}$ reproduces the overlap $ov_{DD,k}$ for all data-GMM components (**Fig. S2**).

Bayesian scoring function. Omitting constant quantities, the final Bayesian scoring function for a fit of a model to a cryo-EM map can be written as:

$$S(X) = k_B T \cdot \left\{-\log[p(X)] + \frac{N_D}{2}\log\left[\sum_{k=1}^{N_D}\log^2\left(\frac{ov_{MD,k}}{ov_{DD,k}}\right)\right]\right\} \quad (15)$$

Where $p(X)$ is the structural prior and depends on the resolution of the model. In our coarse-grained representation, $p(X)$is the sum of a excluded volume potential to avoid steric clashes and a sequence connectivity restraint.

Importantly, the number of components N_D of the data-GMM determines the overall weight of the cryo-EM restraint by increasing the number of log-harmonic functions. On the other hand, the weight is less sensitive to the number of components of the model-GMM. As a consequence, the data-GMM has to be rigorously fit to the experimental density map with the divide-and-conquer approach, while there are no strict guidelines for the maximum number of components of the model-GMM. Here, we followed a parsimonious approach and we empirically chose the number of components in the model-GMM to match the number of Gaussians per unit of mass in the data-GMM. With high-resolution density maps and atomistic models, we expect to use one component of the model-GMM per heavy atom of the system.

Benchmark of the divide-and-conquer fit of the data-GMM

We assessed the accuracy of our divide-and-conquer approach to computing a data-GMM by using 20 experimental density maps of protein complexes at different resolutions, ranging from 3.6 to 25 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. At each step of the divide-and-conquer each sub-map was fit using a GMM with 4 components.

Benchmark of the modeling protocol

Data generation. We benchmarked our modeling protocol using 21 protein/DNA complexes consisting of 2 to 7 subunits [46] (**Table 2** and **Fig. 4**). For each of these complexes, we generated a simulated cryo-EM density map, using the structures extracted from the PDB. We used one Gaussian for every 1.090 kDa of assembly mass, which corresponded approximately to the mass of 10 residues and resulted in a resolution of approximately 10 Å, as obtained by extrapolation from the stretched-exponential regression (**Fig. 3B**). For example, the human transferrin receptor complex (PDB code 1SUV) [47] consists of 6 subunits and has a molecular mass of 290 kDa. Therefore, the simulated map was determined using 262 Gaussians. The simulated GMMs were generated from the reference structures using the program *gmconvert* [37].

Subunits representation and forward model. Molecules (protein and DNA chains) were represented by a set of spherical beads, each with the volume of the corresponding residue. When available, the positions of beads were obtained from the PDB structures and constrained into one or more rigid bodies. Missing regions were constructed as strings of flexible coarse-grained beads**.** When molecules were intertwined or if a molecule was composed of structurally independent domains, we defined several rigid bodies, one for each domain. Furthermore, in some cases, two domains belonging to distinct molecules were merged into the same rigid body, such as the DNA double-strands in 3V6D and 2Y7H or the helical bundle of 3PUV. The model-GMM was 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 [48]. The number of Gaussians of a model-GMM was determined by dividing the molecular weight of the corresponding rigid body by the average weight of a 10-residue peptide (1.09kDa). The center and covariance matrix rotation of each Gaussian were constrained into the corresponding rigid body. Second, each flexible bead was treated as an individual spherical Gaussian.

Model sampling. The initial positions and orientations of rigid bodies and flexible beads were randomized. The generation of structural models was performed using MC coupled with Replica Exchange [42]. 48 replicas were used to cover a temperature range between 1 and 2.5 Score Units (SU). Intermediate temperatures followed a geometrical progression. Each MC step consisted of: A) a series of random transformations of the positions of the flexible beads and the rigid bodies, B) rigid body transformation of the whole system, and C) rigid-body swapping moves. In (A), each individual flexible bead and rigid body was translated in a random direction by up to 4 Å, and each individual rigid body was rotated around its center of mass by up to 0.04 radians about a randomly oriented axis. In (B), a rigid-body transformation was applied to the whole system. In (C) we swapped the position and orientation of two rigid-bodies, randomly chosen among those with similar shape, to allow efficient sampling of alternative conformations equally consistent with the data. The shape similarity

was assessed by computing the root mean square deviation of the inertia moments of the two rigid-bodies. Each MC step was accepted or rejected according to the Metropolis criterion.

Our scoring function becomes more rugged at higher resolutions. In fact, as discussed above, the number of components for the data-GMM (N_D) increases with the resolution, and at the same time their variance decreases to better describe high-frequency features. As a consequence, the log-square score term (Eq. 15) becomes more peaked, thus increasing the frustration of the total score. To alleviate this issue, we implemented an iterative sampling procedure (**Fig. S3**). The idea is to progressively increase N_D using all fits obtained at different stages of the divide-and-conquer procedure, from the minimum (ie, 4) to the N_D of the optimal data-GMM. In each iteration, we 1) sampled the models at a given N_D ; 2) generated a pool of initial models (seeds) for the next iteration; and 3) incremented N_D. During step 1), we produced an ensemble of 12,000 models using the MC and Replica Exchange protocol described above. After extracting the 100 best scoring models of the resulting ensemble, we identified a subset of models as structurally diverse as possible, using the root-mean square deviation (rmsd) criterion. The number of models in the subset was constrained to the number of replicas (ie, 48). In the first iteration (N_D =4), the score-landscape is shallow, which allows the system to explore a large variety of conformations. As N_D increases, the structural variability among seeds is reduced.

Finally, to assess sampling exhaustiveness, at the end of each iterative modeling run we analyzed the agreement of the best scoring models with the cryo-EM map by computing the cross correlation between the model-GMM and data-GMM. If the resulting cross correlations were below 0.7, we started another iteration run. The threshold of 0.7 was chosen by experience, as we noticed that lower cross correlation coefficients usually indicate poor agreement of the model with the cryo-EM data with clearly misplaced subunits. This procedure was applied to all cases for which the simple replica exchange above did not yield satisfying results (PDB codes 1MDA, 1SUV, 1TYQ, and 3PUV in the synthetic benchmark and the application to the RNA polymerase II and exosome complexes).

Analysis. All models produced by the modeling protocol described above were ranked by score, and the 1000 best scoring models were considered for further analysis. The accuracy of the fit was assessed by computing a series of structural metrics, namely the root mean square deviation (rmsd), p(10), the correlation coefficient between the model- and data- GMMs, as well as the average placement score of the best scoring model. These metrics are defined in the next paragraph.

Structural metrics. To compare two models, we used several metrics, including the rmsd of residue positions, p(10), the Average Placement Score (APS), and the data-model correlation coefficient CC. The rmsd of residue positions was defined as the rmsd between the positions of corresponding centers of the coarse-grained spheres in two structures, without structural alignment. When multiple copies of the same protein were present, the rmsd was defined as the minimum rmsd across all possible assignments of the identical components. p(10) was defined as the percentage of residues whose deviation between the two structures is lower than 10 Å. The Placement Score of the model is a two-number metrics that measures the translation and the rotation needed to optimally align each subunit of

the model to a reference. The APS is average of the Placement Score calculated over all subunits and weighted by the number of residues. The data-model correlation coefficient (CC) was defined by Eq. 9 and quantifies the agreement of the model with the data.

Clustering. For each complex, the 1000 best scoring models selected for analysis were clustered using a hierarchical clustering approach [49]. Initially, all 1000 models were placed in a list $\mathcal{L}$ of models not yet clustered. Then, we applied the following iterative procedure:

1. The best scoring model m_0 from the list $\mathcal{L}$ was selected to define a new cluster *Ck* and removed from $\mathcal{L}$.
2. All models m_i from $\mathcal{L}$ with rmsd from m_0 lower than 10 Å were defined as members of the cluster *Ck* and removed from $\mathcal{L}$.
3. We iterated step 1 and 2 until all models were clustered into spheres of radius equal to 10 Å.

At the end of this iterative procedure, we merged all those pairs of clusters that contained at least two elements within 10 Å one from each other. By construction, the first cluster produced by the algorithm (labelled as C1) contained the best scoring model.

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, and the analysis was performed using the IMP.pmi features.

Integrative modeling of the RNA polymerase II complex

We modelled the architecture of the 12-subunit RNA polymerase II using cryo-EM data at 20.9 Å resolution (EMDB code 1883) [50] and two datasets of 108 [51] and 157 [38] cross-links (107 inter- and 158 intra-molecular) (**Fig. 5** and **Fig. S4**). The RNA polymerase II complex consists of 12 subunits, named Rpb1 to Rpb12. The optimal data-GMM contained 16 components (**Fig. 5A**). We followed the modeling protocol and the model representation used for the benchmark with synthetic cryo-EM maps (**Fig. S5**). The coordinates of the beads used to represent our system were obtained from PDB code 1WCM [52]. Based on a prior domain analysis of Rpb1 and Rpb2, we constrained the coordinates of these two large subunits into four rigid-bodies corresponding to: 1) residues 1141-1274 of Rpb1, 2) residues 1275-1733 of Rpb1, 3) residues 1-1102 of Rpb2, and 4) residues 1-1140 of Rpb1 together with residues 1103-1224 of Rpb2. The number of residues per coarse-grained bead was set to 20, and the number of residues per component of the model-GMM was set to 10. The quality of the resulting models was assessed using the same structural metrics as in the benchmark with synthetic data, using the structure of PDB code 1WCM as reference. The XL-MS data was encoded using a previously developed Bayesian scoring function [53].

Integrative modeling of the exosome complex

We modelled the architecture of the 10-subunit yeast exosome complex using cryo-EM data at 11.5 Å resolution (EMDB code 3367) [54] and a dataset of 98 cross-links (26 inter- and 72 intra-molecular) [53] (**Fig. 6** and **Fig. S6**). The exosome complex consists of a core complex of 9 proteins (Csl4, Mtr3, Rrp4, Rrp40, Rrp42, Rrp43, Rrp45, Rrp46, and Ski6), and an RNase protein (Dis3). The top of the core complex recruits RNAs that are then transferred to Dis3 through a central channel in the core complex. The optimal data-GMM contained 256 components (**Fig. 6A** and **Fig. S1B**). We followed the modeling protocol and the model representation (**Fig. S7**) used for the benchmark with synthetic cryo-EM maps, with few variations. We split the largest subunit, Dis3, into three rigid-bodies, corresponding to residues 1-237, 238-471, and 472-1001, which is the domain organization of this subunit based on its structure in PDB code 4IFD (**Fig. S6**) [55]. The quality of the resulting models was assessed using the same structural metrics as in the benchmark with synthetic data. The only difference was that the reference structure used to compute the accuracy (PDB code 5G06) was different from the structure used to initialize the positions of beads in rigid-bodies (PDB code 4IFD) [54]. As for the modeling of RNA polymerase II, XL-MS data was encoded using a Bayesian scoring function [53].

Results

Benchmark of the divide-and-conquer fit of the data-GMM

We assessed the accuracy of our divide-and-conquer approach by determining the data-GMM of 20 experimental density maps at different resolutions [54, 56, 57], ranging from 3.6 Å to 25 Å (**Table 1**). This benchmark revealed that the number of Gaussian components needed to achieve a given accuracy of the optimal data-GMM varies with the resolution of the map and molecular weight of the complex (**Fig. 3**). Indeed, for a given number of components and molecular weight, the data-GMM correlation coefficient is lower for higher-resolution maps. In other words, high resolution maps and maps of high molecular weight complexes contain more information and therefore require additional components to describe the ensemble of their features.

Our analysis allows us to classify the maps of this dataset in two clusters: a low-resolution and a high-resolution cluster that contains all maps with a resolution greater than 10 Å, and the latter is a high-resolution cluster that contains all maps with a resolution lower than 10 Å. Interestingly, this threshold has been reported to be the resolution at which secondary structure elements become visible [58].

From the results of our benchmark, we calculated the dependence of the resolution of the cryo-EM density maps on the number of Gaussians per mass-unit of the optimal data-GMM (**Fig. 3B**). 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 cryo-EM density map of a given mass and resolution.

Finally, the divide-and-conquer approach allowed us to overcome the computational inefficiency of the traditional expectation-maximization algorithm for fitting GMM with a large number of components. For example, in the case of the yeast cytoplasmic exosome at 4.2 Å resolution (EMDB code 3366)

[54], our approach required 24 minutes and less than 1GB to generate GMMs with 4, 16, 64, 256, 1024, and 4096 components. In contrast, a serial implementation on a single computer required over 48 hours and 182 GB of memory for fitting with 4096 components.

Benchmark of the modeling protocol

We assessed the accuracy of the modeling protocol using a benchmark of 21 protein/DNA complexes consisting of 2 to 7 subunits (**Table 2**) [46] and simulated cryo-EM density maps with a resolution of approximately 10 Å (Materials and Methods). No additional experimental data beside the crystal structure of the individual components were included, in order to explore the performance of the cryo-EM scoring function alone. The detailed results of the benchmark are reported in **Table S1**.

The average accuracy p(10) of the whole benchmark was 88%. We classified the outcomes of our benchmark into three categories. We defined a *full positive* result when the global rmsd with respect to the reference structure along with the rmsds of all the individual subunits were lower than 10 Å. A *partial positive* result was achieved when the global rmsd was lower than 10 Å but some of the subunits were misplaced, resulting in a rmsd greater than 10 Å for at least one subunit. A *negative* result was obtained when the global rmsd was greater than 10 Å. Out of the 21 complexes, we obtained 16 full positives (2UZX, 3R5D, 1CS4, 2WVY, 2DQJ, 1VCB, 2GC7, 2BO9, 2BBK, 1GPQ, 3V6D, 3SFD, 3PDU, 3NVQ, 2Y7H, and 1SUV), 3 partial positives (1Z5S, 3LU0, and 1MDA), and 2 negatives (3PUV and 1TYQ) (**Table 2** and **Fig. 4**).

The majority of the complexes belonged to the full positive category and were accurately modeled, with an average global rmsd from the reference structure equal to 2.2 Å. In the following, we discuss the few partial positive and negative results, highlighting the reasons behind their lower accuracy.

The best-scoring model of the 4-subunits 1Z5S had a rmsd of 9.0 Å, a p(10) of 0.86, APS of (1.8 Å, 32.1°), and a CC of 0.87. All best scoring models are grouped into a single cluster. The origin of the inaccuracy was subunit B, which was mis-rotated by almost 180°. The reason was that this subunit has a cylindrical shape and therefore the expected density is invariant under rotations around the main axis.

The best-scoring model of the 5-subunits 3LU0 had a rmsd of 9.3 Å, a p(10) of 0.85, APS of (4.3 Å, 5.7°), and a CC of 0.74. The best scoring models were grouped into two clusters. In the first cluster containing the best scoring model, the lower accuracy of the models was due to subunits A, B, and E. Subunits A and B, while positioned in the correct region of the density map, were displaced by 10.9 Å and 7.3 Å and mis-rotated by 33.7° and 24.1°, respectively. Subunit E was also displaced by 13.5 Å and mis-rotated by 34.6°. In the second cluster, the situation was similar, with subunit E displaced even farther apart in the incorrect region of the density, with a Placement Score of (72.6 Å, 151.7°).

The best-scoring model of the 6-subunits 1MDA had a rmsd of 8.1 Å, a p(10) of 0.85, APS of (3.2 Å, 34.0°), and a CC of 0.75. All best scoring models were grouped into a single cluster. The origin of the inaccuracy was the orientations of subunit A and M, which were both mis-rotated by almost 180°. The reason was that these subunits have cylindrical shape and therefore their expected densities are invariant under rotations around the main axis.

The best-scoring model of the 5-subunits 3PUV had a rmsd of 23.4 Å, a p(10) of 0.64, APS of (8.0 Å, 31.1°), and a CC of 0.74. All best scoring models were grouped into a single cluster. In the best-scoring model, subunits E, F, and G were correctly positioned. Subunits A and B were instead both misplaced and mis-rotated, with a APS of (17.6 Å, 17.7°) and (18.1 Å, 136.7°), respectively. The reason was that these subunits formed a closed dimer of roughly cubical shape, whose density could be fit also by an incorrect model in which the positions of the domain of subunits A and B were swapped.

The best-scoring model of the 7-subunits 1TYQ had a rmsd of 19.8 Å, a p(10) of 0.65, APS of (4.1 Å, 58.2°), and a CC of 0.65. The best scoring models were grouped into 3 clusters. In all clusters, subunits D and F were both misplaced and mis-rotated. The most likely reason for this inaccuracy was that these subunits formed an elongated helical bundle of about 40 residues in length, making sampling more challenging due to steric effects. In the first cluster containing the best scoring model, subunit E was also mis-rotated by almost 180°, due to its globular shape.

Integrative modeling of the RNA polymerase II

The yeast RNA polymerase II is a 12-subunit complex that catalyzes DNA transcription to synthesize mRNA strands [52]. To model this complex, we used the structures of all its subunits as determined in the RNA polymerase II X-ray structure (PDB code 1WCM) [52]. We incorporated a low-resolution cryo-EM map of the RNA polymerase II-Iwr1 complex (EMDB code 1883) [50] and two XL-MS datasets.

959 of the 1000 best scoring models grouped in the first cluster. The rmsd of the best scoring model with respect to the reference structure was 25.8 Å, with a p(20) of 0.77 and a data-model CC of 0.52. The major contribution for the inaccuracy was the misplacement of subunit Rpb8. The reason for the misplacement was that Rpb8 was not cross-linked with the rest of the complex. Excluding Rpb8 from the rmsd calculation yielded an rmsd of 21.4 Å.

We analysed the position of each subunit of the complex (**Fig. 5D**). Subunits 2, 3, 4, 5, 7, and 11 have an rmsd with respect to the reference structure inferior to 20 Å. The following subunits had an rmsd greater than 20 Å with respect to the reference structure: Rpb1 (25.6 Å rmsd), Rpb 6 (40 Å rmsd), Rpb 8 (149.5 Å rmsd), Rpb9 (67.3 Å rmsd), Rpb 10 (25.8 Å rmsd), and Rpb 12 (26.7 Å rmsd). Note that Rpb1 was correctly localized in the cryo-EM map but the domain corresponding to residues 1275-1733 was misplaced. Another reason for the inaccuracy is that subunits Rpb8, Rpb9, and Rpb12 were weakly cross-linked with the rest of the complex, forming 0, 1 and 3 cross-links respectively.

There was a total of 8 violated cross-links (3.1% of the total dataset), which involved the following subunits: Rpb1-Rpb2 (3 cross-links), Rpb1-Rpb4 (1 cross-link), Rpb1-Rpb5 (1 cross-link), Rpb1-Rpb6 (1 cross-link), and Rpb2-Rpb2 (2 cross-links).

Integrative modeling of the exosome complex

The 10-subunit yeast exosome complex is a macromolecular machine that is responsible for processing and degrading RNA in eukaryotic cells [59]. To model this complex, we used the structures of all subunits of the complex in one state, the crystal structure of the RNA-bound exosome (PDB code 4IFD, **Fig. S7**). We also incorporated new, independent data collected in another state, the low-resolution cryo-EM map of the RNA-free exosome (EMDB code 3367) [54] and a dataset of 98 cross-linked residue pairs obtained by XL-MS [53].

To the best of our knowledge, no high-resolution structure of the RNA-free 10-subunit exosome complex alone is available. We were forced to use the structure of the RNA-free exosome in complex with Ski7 (PDB code 5G06) as a reference in the analysis, to test the accuracy of our models. We expected our model to differ from the reference, especially in the region at the top, where the complex interacts with Ski7. Here, for instance, when one rigidly fits the entire reference structure to the RNA-free density map, the Csl4 subunit extends outside the density map (**Fig. 6D**), most likely as a consequence of the interaction with Ski7. On the other hand, the lower region in which Dis3 is located is expected to be structurally similar to the reference. In addition, cross-links were extracted from whole-cell lysate and therefore might reflect a mixture of different compositional and conformational states [53].

The 1000 best scoring models of the last step of the iterative sampling procedure grouped into a single cluster. The rmsd of the best scoring model with respect to the reference structure was 30.0 Å, with a p(10) of 0.42 and a data-model CC of 0.68. We analysed the position of each subunit of the complex (**Fig. 6D**). Strikingly, each domain of Dis3 was properly placed in its respective density region (**Fig. 6D**). Subunit Csl4 was localized entirely inside the density map, at variance with the reference structure. Subunits Rrp45, Ski6, Rrp46, Rrp40, and Rrp4 occupied the correct regions of the density map. Subunits Rrp42, Rrp43, and Mtr3 were misplaced, but still occupied the upper region of the density map. The majority of the cross-links were satisfied, with measured distance between cross-linked residue pairs always below 35 Å, with a few exceptions. The cross-link between residue 71 of Rrp43 and residue 104 of Rrp42 was violated (**Fig. 6E**). Two other cross-links involving Dis3 were found to be inconsistent with the cryo-EM map and therefore were violated. These distance restraints might be satisfied in the RNA-bound exosome form, which we expect to be present under the conditions in which the XL-MS data were collected.

Discussion and conclusions

A major difficulty in integrative structure modeling, in which data of different types are combined to model the structure of a biological complex, is to determine the relative weight of each piece of information. Inaccurate weighing results in models that are biased towards a particular type of data, thus reducing the accuracy of the model and under- or over-estimating its precision. To optimally weigh each piece of information, two factors need to be considered: 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. Noisy experimental data obviously provide less structural information than highly precise data at the same level of accuracy.

In the Bayesian framework, the problem of ranking alternative structural models based on multiple sources of information is achieved by a scoring function, which objectively quantifies the information content of each input data point. Here, we introduced a Bayesian approach that can be generally applied to integrative modeling based on cryo-EM maps at multiple resolutions. This approach addresses the problem of quantifying both the level of noise and data correlation. Experimental density maps are decomposed into a minimal Gaussian Mixture Model (GMM) that reproduces most structural features of the data. 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).

Similarly, 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. 14 then quantifies the discrepancy between data and model GMMs 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 fits the k-th component of the data-GMM when their overlap (Eq. 11) is close to the overlap between the data-GMM and the same component of the data-GMM (**Fig. S2B**). 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-th component of the data-GMM. Finally, because the components of the GMM are now 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 the number of score terms that are driving the model close to the target for each data-Gaussian, therefore corresponding to its weight. As desired, decreasing/increasing the resolution of the data (**Fig. 3B**), fewer/more Gaussians are needed to reproduce the density map, and as a consequence the restraint has a lower/higher weight.

The extensive benchmark of our modeling protocol based on the Bayesian cryo-EM scoring function demonstrated that in most cases our approach is able to position and orient the components of a macromolecular complex in a low-resolution cryo-EM density map with high accuracy. To compare our approach with state-of-the-art methods for modeling macromolecular complexes using low- and intermediate-resolution cryo-EM maps, we examined the results of the benchmark carried out with γ-TEMPy [26]. This method scores the models using mutual information between model and experimental densities and uses a genetic algorithm to accelerate sampling. In the following, we compare our approach with γ-TEMPy in terms of accuracy of the scoring function, sampling efficiency, and computational performances.

In order to compare the accuracy of the two approaches, we examined our best scoring model (**Table 2**) and their high scoring model (HC in **Table 1** of Ref. [26]) on a subset of 8 test cases of our

benchmark that were also included in the γ-TEMPy benchmark (PDB codes 2DQJ, 2BO9, 2BBK, 1VCB, 1TYQ, 1MDA, 1GPQ, and 1CS4). In all cases, our approach produced models that were significantly more accurate, with an averaged difference between the rmsd from the native structure calculated with the two approaches equal to 8.5 Å. Particularly striking is the case of the 3-subunits complex 1VCB, which our method and γ-TEMPy modeled with rmsd of 2.2 Å and 25.3 Å, respectively.

To assess sampling efficiency, we compared our best rmsd model with their best prediction (BP in **Table 1** of Ref. [26]). In 7 out of 8 cases, our approach was capable to sample more native-like models, with an averaged difference between the best rmsd from the native structure calculated with the two approaches equal to 3.9 Å. Only in the case of 1TYQ, our best rmsd model was less accurate (19.1 Å) than the BP model generated with γ-TEMPy (16.9 Å).

To assess the performances of the two approaches, we monitored the computational cost to run the two benchmarks, defined by the total number of core hours required to complete one test case. Our benchmark was executed in parallel on 48 cores on a compute cluster equipped with 2.50GHz Intel(R) Xeon(R) E5-2670 v2 processors. The minimum, maximum, and average computational cost across all test cases was 320, 6840, and 2166 core hours, respectively. Furthermore, this computational cost depends on the number of components of the data-GMM, which is particularly advantageous with over-sampled low-resolution maps. Each computation of the γ-TEMPy benchmark was instead run on 160 cores distributed on 40 AMD 4-core 2.6GHz processors (exact model not specified). Unfortunately, the time measurements for the benchmark with 10 Å-resolution maps were not reported. However, since the performances of γ-TEMPy depend on the number of voxels in the map and not necessarily on its resolution [26], we used the timings of the 20 Å-resolution benchmark to estimate the computational cost. The reported minimum, maximum, and average cost across all test cases was 640, 7840, and 2720 core hours, respectively.

In summary, the comparison of our approach with γ-TEMPy demonstrated that: a) our scoring function is capable of identifying significantly more accurate models; b) our sampling method is more efficient; c) the computational cost of our approach is not only inferior to γ-TEMPy, but also depends on the resolution of the map rather than on the number of voxels used to define it.

In the few cases in which our approach produced results of accuracy lower than the average, we identified two sources of error: a) positional ambiguity, where multiple placements result in the same score, and b) inefficient sampling of rigid body configurations in crowded environments. 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 can be rotated around their center of mass or swapped with only minimal penalty. In addition, the placement of DNA helices is degenerate, because their expected density is symmetric by rotation. Moreover, macromolecular complexes present a crowded environment in which sampling of rigid body configurations might be inefficient due to steric hindrance. A particularly pathological case is that of elongated helical bundles in crowded environment, where slight rotations of the helices results in large amplitude motions of the residues at their extremities.

In applications with real experimental density maps, we foresee an additional source of error that could affect the accuracy of the modeled complex. This consists in the fact that not all components of the modeled complex might have a corresponding experimental density or that the experimental density might represent more components than those explicitly modeled. Our approach currently assumes that all data-GMM components can be explained by a corresponding density of the model. This assumption is encoded in the scoring function by assigning the same total electron density to the data-GMM and the model-GMM, i.e. the two GMMs are normalized to the same value. It should be notated that this challenge, along with the two previously described, is not specific to the modeling protocol presented here, but it is faced by all the techniques to model architectures from low-resolution cryo-EM data.

The issue of mismatch between the overall model and experimental densities will be addressed in the future in two ways. First, we will assign to each component k of the data-GMM a different uncertainty parameter σ_k. In this way, regions of the experimental density map that correspond to components of the complex not explicitly modeled can be recognized as outliers, as they cannot be fit unless the representation of the model is improved. Second, we will introduce an unknown scaling parameter of the experimental (or model) density map. In this way, we will enforce the model overlaps $ov_{MD,k}$ to match the experimental overlap $ov_{DD,k}$ in Eq. 15 only after scaling by an unknown factor. In our Bayesian approach, this parameter will be sampled, along with the coordinates of the model particles, by a MC algorithm.

Our real-case examples were the yeast RNA polymerase II and the yeast exosome, for which we used 20.9 Å and 11.5 Å resolution cryo-EM maps and XL-MS data. In both cases, subunits were localized in the EM map with variable accuracy, depending mostly on the density of XL-MS data. For RNA polymerase II, we localized 7 of the subunits in the correct region of the EM map. In the exosome case, we could localize correctly the 6 subunits that were densely cross-linked. In particular, the 110 kDa subunit Dis3, which undergoes a large conformational change, was characterized in the RNA-unbound state starting from the structure in its RNA-bound state with an accuracy of 8.30 Å. Subunits with little cross-linking information were placed close (35.8 Å) to their native localization. Importantly, in these two examples we did not have to modulate the relative weights of XL-MS and cryo-EM data, as this was done automatically in the Bayesian framework, which is one of the greatest advantages of our method.

The modeling approach presented here introduces several technical features. First, building on the *gmconvert* utility [37], we developed a cryo-EM density processing tools that exploits a divide-and-conquer strategy to efficiently compute Gaussian Mixture Models with a large number of components. Second, we created a multi-scale modeling approach, as the scoring function adapts to any coarse-grained representation of the model and any resolution of the experimental density map. Third, we used a combination of flexible and rigid degrees of freedom: each domain with a known structure was constrained into a rigid-body, while all missing parts (loops, termini, or unknown regions) were represented by flexible strings of beads. Fourth, we developed an enhanced sampling technique based on an iterative Replica-Exchange strategy and a novel MC mover that randomly swaps rigid-bodies with similar shape. We implemented our modeling protocol in a series of scripts based on IMP.pmi

[60], a module of the Integrative Modeling Platform (IMP, http://integrativemodeling.org) [32] that can be used to build the system representation, setup the scoring function, define the degrees of freedom to sample, and finally analyze the solutions. Our approach was also implemented in the PLUMED-ISDB module [61] of the open-source PLUMED library (www.plumed.org) [62]. Thanks to the differentiability of the scoring function, this implementation can be used for real-space, flexible refinement of individual models using molecular dynamics at atomistic resolution or, in combination with metainference [14], to model ensemble of structures representing the conformational heterogeneity hidden in low-resolution areas of atomistic density maps [63].

In conclusion, the results presented here highlight the importance of integrative approaches when dealing with medium- to low-resolution data. Our Bayesian framework, implemented in the *Integrative Modeling Platform* and PLUMED library, enables the optimal integration of cryo-EM data with other sources of information, such as XL-MS, NMR spectroscopy, and small angle X-ray scattering data, in order to further improve the accuracy and precision of the output models.

Acknowledgements

We thank Dr. T. Huynh for assistance with the computer cluster at Institut Pasteur, and Dr. D. Saltzberg for useful discussions.

Authors contributions

S. H., M. B, R. P. performed research and analyzed the data. M. B., R. P. developed methodology. M. N. provided computing resources. All authors contributed to writing the manuscript.

EMDB code	reference	molecular Weight [kDa]	resolution [Å]	resolution method	optimal number of components	GMM resolution [Å]	CC
1439	[57]	300	23.0	FSC 0.5	64	21.6	0.97
1438	[57]	400	19.0	FSC 0.5	64	20.1	0.94
1708	[56]	400	14.0	FSC 0.5	64	13.7	0.92
3368	[54]	350	13.0	FSC 0.143	64	16.4	0.84
3367	[54]	350	11.5	FSC 0.143	256	13.1	0.93
3371	[54]	350	11.0	FSC 0.143	256	11.0	0.92
3370	[54]	350	6.7	FSC 0.143	256	7.3	0.81
3372	[54]	350	6.3	FSC 0.143	3709	6.3	0.95
3369	[54]	420	5.8	FSC 0.143	1024	5.7	0.85
3366	[54]	420	4.2	FSC 0.143	4012	4.4	0.91
6231	[64]	457	25.0	FSC 0.143	16	27.3	0.93
1753	[65]	700	21.0	FSC 0.5	64	20.8	0.97
1883	[50]	550	20.9	FSC 0.5	16	29.0	0.73
1711	[66]	380	13.0	FSC 0.5	64	15.2	0.94
3198	[67]	256	8.0	FSC 0.143	256	8.0	0.78
2923	[68]	540	7.2	FSC 0.143	256	7.8	0.78
2784	[69]	570	6.6	FSC 0.143	1024	6.4	0.84
3056	[70]	450	6.0	FSC 0.143	1024	6.0	0.83
3202	[67]	256	7.3	FSC 0.143	256	7.4	0.80
3219	[71]	590	3.6	FSC 0.143	4060	3.9	0.88

Table 1: Benchmark of the divide-and-conquer approach for GMM fitting. For each of the system studied, we report: the EMDB accession code, the reference paper, the molecular weight of the complex, the resolution of the map, the method used to quantify the experimental resolution, the optimal number of components, the resolution of the optimal GMM, the CC between the experimental cryo-EM map and the optimal GMM.

PDB code	reference	# subunits	# clusters	time / frame [s]	best rmsd model rank	best rmsd model rmsd [Å]	best scoring model rmsd [Å]	p(10)	CC	APS [Å,°]	
2UZX	[72]	2	1	1.6	551	1.1	1.5	1.00	0.90	0.7	2.9
3R5D	[73]	3	1	1.1	63	1.4	1.9	1.00	0.93	1.1	2.6
1CS4	[74]	3	1	0.7	159	2.6	3.5	1.00	0.79	0.8	1.1
2WVY	[75]	3	1	3.0	480	0.9	1.3	1.00	0.98	0.5	0.6
2DQJ	[76]	3	3	0.4	89	2.0	2.5	1.00	0.93	1.6	3.0
1VCB	[77]	3	1	0.2	453	1.8	2.2	1.00	0.82	1.0	2.7
2GC7	unpublished	4	1	0.6	817	1.3	2.0	1.00	0.94	0.9	1.8
2BO9	[78]	4	1	0.8	502	1.3	1.6	1.00	0.95	0.9	1.4
2BBK	[79]	4	1	0.9	857	2.1	2.4	1.00	0.90	1.7	1.0
1GPQ	[80]	4	1	0.5	805	1.7	2.2	1.00	0.90	1.4	1.5
3V6D	[81]	4	2	1.0	686	1.6	2.1	1.00	0.92	1.0	5.2
3SFD	[82]	4	1	1.3	385	1.4	1.5	1.00	0.94	0.7	1.0
3PDU	[83]	4	1	1.3	802	1.3	1.6	1.00	0.93	1.3	0.9
3NVQ	[84]	4	1	2.1	903	0.9	1.0	1.00	0.97	0.6	0.7
2Y7H	[85]	5	1	1.9	465	1.7	2.2	1.00	0.88	1.9	1.5
1SUV	[47]	6	1	1.4	247	5.2	5.3	1.00	0.77	5.2	0.4
1Z5S	[86]	4	1	0.3	892	8.7	9.0	0.86	0.87	1.8	32.1
3LU0	[87]	5	2	5.1	800	9.0	9.3	0.85	0.74	4.3	5.7
1MDA	[88]	6	2	1.8	4286	7.8	8.1	0.85	0.75	3.2	34.0
3PUV	[89]	5	1	3.1	698	22.7	23.4	0.64	0.74	8.0	31.1
1TYQ	[90]	7	3	5.1	599	19.1	19.8	0.65	0.65	4.1	58.2

Table 2: Results of the benchmark of the modeling protocol. For each of the system studied, we report: the PDB accession code, the reference paper, the number of subunits, the number of clusters, the average time needed to produce one model, the rank and rmsd of the model with minimum rmsd with respect to the reference structure (best rmsd model). We also report the following information about the best scoring model: rmsd, p(10), the data-model correlation coefficient (CC), and the average placement scores (APS) with respect to the reference structure.

References

1. Kuhlbrandt, W., *The resolution revolution*. Science, 2014. **343** 1443-4.
2. Nogales, E., *The development of cryo-EM into a mainstream structural biology technique*. Nat Methods, 2016. **13** 24-7.
3. Callaway, E., *The Revolution Will Not Be Crystallized*. Nature, 2015. **525** 172-174.
4. Bai, X.C., G. McMullan, and S.H. Scheres, *How cryo-EM is revolutionizing structural biology*. Trends in Biochemical Sciences, 2015. **40** 49-57.
5. Glaeser, R.M., *How good can cryo-EM become?* Nature Methods, 2016. **13** 28-32.
6. Li, X.M., P. Mooney, S. Zheng, C.R. Booth, et al., *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM*. Nature Methods, 2013. **10** 584-+.
7. Yu, X., L. Jin, and Z.H. Zhou, *3.88 A structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy*. Nature, 2008. **453** 415-9.
8. Bai, X.C., I.S. Fernandez, G. McMullan, and S.H. Scheres, *Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles*. Elife, 2013. **2** e00461.
9. Campbell, M.G., D. Veesler, A. Cheng, C.S. Potter, et al., *2.8 A resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy*. eLife, 2015. **4**.
10. Bai, X.C., C. Yan, G. Yang, P. Lu, et al., *An atomic structure of human gamma-secretase*. Nature, 2015. **525** 212-7.
11. Bartesaghi, A., A. Merk, S. Banerjee, D. Matthies, et al., *2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor*. Science, 2015. **348** 1147-51.
12. Schneidman-Duhovny, D., R. Pellarin, and A. Sali, *Uncertainty in integrative structural modeling*. Curr Opin Struct Biol, 2014. **28** 96-104.
13. Bonomi, M., G.T. Heller, C. Camilloni, and M. Vendruscolo, *Principles of protein structural ensemble determination*. Curr Opin Struct Biol, 2017. **42** 106-116.
14. Bonomi, M., C. Camilloni, A. Cavalli, and M. Vendruscolo, *Metainference: A Bayesian inference method for heterogeneous systems*. Sci Adv, 2016. **2** e1501177.
15. Ward, A.B., A. Sali, and I.A. Wilson, *Biochemistry. Integrative structural biology*. Science, 2013. **339** 913-5.
16. Schroder, G.F., *Hybrid methods for macromolecular structure determination: experiment with expectations*. Curr Opin Struct Biol, 2015. **31** 20-7.
17. Lopez-Blanco, J.R. and P. Chacon, *Structural modeling from electron microscopy data*. Wiley Interdisciplinary Reviews-Computational Molecular Science, 2015. **5** 62-81.
18. Pettersen, E.F., T.D. Goddard, C.C. Huang, G.S. Couch, et al., *UCSF Chimera--a visualization system for exploratory research and analysis*. Journal of Computational Chemistry, 2004. **25** 1605-12.
19. Rossmann, M.G., R. Bernal, and S.V. Pletnev, *Combining electron microscopic with X-ray crystallographic structures*. Journal of Structural Biology, 2001. **136** 190-200.
20. Sali, A. and T.L. Blundell, *Comparative Protein Modeling by Satisfaction of Spatial Restraints*. Journal of Molecular Biology, 1993. **234** 779-815.
21. Wriggers, W., *Conventions and workflows for using Situs*. Acta Crystallographica Section D-Biological Crystallography, 2012. **68** 344-351.
22. Lasker, K., M. Topf, A. Sali, and H.J. Wolfson, *Inferential Optimization for Simultaneous Fitting of Multiple Components into a CryoEM Map of Their Assembly*. Journal of Molecular Biology, 2009. **388** 180-194.

23. Zheng, W.S., *Accurate Flexible Fitting of High-Resolution Protein Structures into Cryo-Electron Microscopy Maps Using Coarse-Grained Pseudo-Energy Minimization*. Biophysical Journal, 2011. **100** 478-488.
24. Trabuco, L.G., E. Villa, K. Mitra, J. Frank, et al., *Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics*. Structure, 2008. **16** 673-683.
25. Topf, M., K. Lasker, B. Webb, H. Wolfson, et al., *Protein structure fitting and refinement guided by cryo-EM density*. Structure, 2008. **16** 295-307.
26. Pandurangan, A.P., D. Vasishtan, F. Alber, and M. Topf, *gamma-TEMPy: Simultaneous Fitting of Components in 3D-EM Maps of Their Assembly Using a Genetic Algorithm*. Structure, 2015. **23** 2365-2376.
27. Volkmann, N. and D. Hanein, *Quantitative fitting of atomic models into observed densities derived by electron microscopy*. Journal of Structural Biology, 1999. **125** 176-184.
28. Ratje, A.H., J. Loerke, A. Mikolajka, M. Brunner, et al., *Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites*. Nature, 2010. **468** 713-6.
29. Saha, M. and M.C. Morais, *FOLD-EM: automated fold recognition in medium- and low-resolution (4-15 A) electron density maps*. Bioinformatics, 2012. **28** 3265-3273.
30. DiMaio, F., M.D. Tyka, M.L. Baker, W. Chiu, et al., *Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta*. Journal of Molecular Biology, 2009. **392** 181-190.
31. Lindert, S., N. Alexander, N. Wotzel, M. Karakas, et al., *EM-Fold: De Novo Atomic-Detail Protein Structure Determination from Medium-Resolution Density Maps*. Structure, 2012. **20** 464-478.
32. Russel, D., K. Lasker, B. Webb, J. Velazquez-Muriel, et al., *Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies*. PLoS Biol, 2012. **10** e1001244.
33. Scheres, S.H., *RELION: implementation of a Bayesian approach to cryo-EM structure determination*. Journal of Structural Biology, 2012. **180** 519-30.
34. Habeck, M., *Bayesian Modeling of Biomolecular Assemblies with Cryo-EM Maps*. Front Mol Biosci, 2017. **4** 1-13.
35. Adams, P.D., P.V. Afonine, G. Bunkoczi, V.B. Chen, et al., *The Phenix software for automated determination of macromolecular structures*. Methods, 2011. **55** 94-106.
36. Rieping, W., M. Habeck, and M. Nilges, *Inferential structure determination*. Science, 2005. **309** 303-6.
37. Kawabata, T., *Multiple Subunit Fitting into a Low-Resolution Density Map of a Macromolecular Complex Using a Gaussian Mixture Model*. Biophysical Journal, 2008. **95** 4643-4658.
38. Robinson, P.J., M.J. Trnka, R. Pellarin, C.H. Greenberg, et al., *Molecular architecture of the yeast Mediator complex*. Elife, 2015. **4**.
39. Jonic, S., J. Vargas, R. Melero, J. Gomez-Blanco, et al., *Denoising of high-resolution single-particle electron-microscopy density maps by their approximation using three-dimensional Gaussian functions*. J Struct Biol, 2016. **194** 423-33.
40. Erzberger, J.P., F. Stengel, R. Pellarin, S. Zhang, et al., *Molecular architecture of the 40SeIF1eIF3 translation initiation complex*. Cell, 2014. **158** 1123-35.
41. Fernandez-Martinez, J., S.J. Kim, Y. Shi, P. Upla, et al., *Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform*. Cell, 2016. **167** 1215-+.
42. Swendsen, R.H. and J.S. Wang, *Replica Monte-Carlo Simulation of Spin-Glasses*. Physical Review Letters, 1986. **57** 2607-2609.

43. Frenkel, D. and B. Smit, *Understanding molecular simulation : from algorithms to applications*. 2nd ed. Computational science2002, San Diego, Calif. ; London: Academic. xxii, 638 p.
44. Rosenthal, P.B. and R. Henderson, *Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy*. Journal of Molecular Biology, 2003. **333** 721-745.
45. Prince, E., *International Tables for Crystallography Vol. C*. 3rd ed2004, Hoboken: Wiley. 1032 p.
46. Velazquez-Muriel, J., K. Lasker, D. Russel, J. Phillips, et al., *Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images*. Proceedings of the National Academy of Sciences of the United States of America, 2012. **109** 18821-18826.
47. Cheng, Y., O. Zak, P. Alsen, S.C. Harrison, et al., *Structure of the human transferrin receptor-transferrin complex*. Cell, 2004. **116** 565-576.
48. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, et al., *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 2011. **12** 2825-2830.
49. Johnson, S.C., *Hierarchical Clustering Schemes*. Psychometrika, 1967. **32** 241-254.
50. Czeko, E., M. Seizl, C. Augsberger, T. Mielke, et al., *Iwr1 Directs RNA Polymerase II Nuclear Import*. Molecular Cell, 2011. **42** 261-266.
51. Chen, Z.A., A. Jawhari, L. Fischer, C. Buchen, et al., *Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry*. Embo J, 2010. **29** 717-726.
52. Armache, K.-J., S. Mitterweger, A. Meinhart, and P. Cramer, *Structures of Complete RNA Polymerase II and its Subcomplex, Rpb4/7*. Journal of Biological Chemistry, 2005. **280** 7131.
53. Shi, Y., R. Pellarin, P.C. Fridy, J. Fernandez-Martinez, et al., *A strategy for dissecting the architectures of native macromolecular assemblies*. Nature Methods, 2015. **12** 1135-+.
54. Liu, J.J., C.Y. Niu, Y. Wu, D. Tan, et al., *CryoEM structure of yeast cytoplasmic exosome complex*. Cell Research, 2016. **26** 822-837.
55. Makino, D.L., M. Baumgartner, and E. Conti, *Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex*. Nature, 2013. **495** 70-75.
56. Malet, H., M. Topf, D.K. Clare, J. Ebert, et al., *RNA channelling by the eukaryotic exosome*. Embo Reports, 2010. **11** 936-942.
57. Wang, H.W., J. Wang, F. Ding, K. Callahan, et al., *Architecture of the yeast Rrp44-exosome complex suggests routes of RNA recruitment for 3 ' end processing*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104** 16844-16849.
58. Baker, M.L., T. Ju, and W. Chiu, *Identification of secondary structure elements in intermediate-resolution density maps*. Structure, 2007. **15** 7-19.
59. Houseley, J., J. LaCava, and D. Tollervey, *RNA-quality control by the exosome*. Nature Reviews Molecular Cell Biology, 2006. **7** 529-539.
60. Webb, B., S. Viswanath, M. Bonomi, R. Pellarin, et al., *Integrative structure modeling with the Integrative Modeling Platform*. Protein Science, 2018. **27** 245-258.
61. Bonomi, M. and C. Camilloni, *Integrative structural and dynamical biology with PLUMED-ISDB*. Bioinformatics, 2017. **33** 3999-4000.
62. Tribello, G.A., M. Bonomi, D. Branduardi, C. Camilloni, et al., *PLUMED 2: New feathers for an old bird*. Computer Physics Communications, 2014. **185** 604-613.
63. Bonomi, M., R. Pellarin, and M. Vendruscolo, *Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy*. bioRxiv, 2017. **doi: 10.1101/219972**.

64. Booth, D.S., Y.F. Cheng, and A.D. Frankel, *The export receptor Crm1 forms a dimer to promote nuclear export of HIV RNA*. Elife, 2014. **3**.
65. Vannini, A., R. Ringel, A.G. Kusser, O. Berninghausen, et al., *Molecular Basis of RNA Polymerase III Transcription Repression by Maf1*. Cell, 2010. **143** 59-70.
66. Julian, P., P. Milon, X. Agirrezabala, G. Lasso, et al., *The Cryo-EM Structure of a Complete 30S Translation Initiation Complex from Escherichia coli*. Plos Biology, 2011. **9**.
67. Fernandez-Leiro, R., J. Conrad, S.H.W. Scheres, and M.H. Lamers, *cryo-EM structures of the E. coli replicative DNA polymerase reveal its dynamic interactions with the DNA sliding clamp, exonuclease and tau*. Elife, 2015. **4**.
68. Martinez-Rucobo, F.W., R. Kohler, M. van de Waterbeemd, A.J.R. Heck, et al., *Molecular Basis of Transcription-Coupled Pre-mRNA Capping*. Molecular Cell, 2015. **58** 1079-1089.
69. Plaschka, C., L. Lariviere, L. Wenzeck, M. Seizl, et al., *Architecture of the RNA polymerase II-Mediator core initiation complex*. Nature, 2015. **518** 376-380.
70. des Georges, A., V. Dhote, L. Kuhn, C.U.T. Hellen, et al., *Structure of mammalian eIF3 in the context of the 43S preinitiation complex*. Nature, 2015. **525** 491-+.
71. Bernecky, C., F. Herzog, W. Baumeister, J.M. Plitzko, et al., *Structure of transcribing mammalian RNA polymerase II*. Nature, 2016. **529** 551-+.
72. Niemann, H.H., V. Jager, P.J.G. Butler, J. van den Heuvel, et al., *Structure of the human receptor tyrosine kinase met in complex with the Listeria invasion protein InIB*. Cell, 2007. **130** 235-246.
73. Schnell, R., W. Oehlmann, T. Sandalova, Y. Braun, et al., *Tetrahydrodipicolinate N-Succinyltransferase and Dihydrodipicolinate Synthase from Pseudomonas aeruginosa: Structure Analysis and Gene Deletion*. Plos One, 2012. **7**.
74. Tesmer, J.J.G., C.W. Dessauer, R.K. Sunahara, L.D. Murray, et al., *Molecular basis for P-site inhibition of adenylyl cyclase*. Biochemistry, 2000. **39** 14464-14471.
75. Zhu, Y.P., M.D.L. Suits, A.J. Thompson, S. Chavan, et al., *Mechanistic insights into a Ca2+-dependent family of alpha-mannosidases in a human gut symbiont*. Nature Chemical Biology, 2010. **6** 125-132.
76. Shiroishi, M., K. Tsumoto, Y. Tanaka, A. Yokota, et al., *Structural consequences of mutations in interfacial Tyr residues of a protein antigen-anti body complex - The case of HyHEL-10-HEL*. Journal of Biological Chemistry, 2007. **282** 6783-6791.
77. Stebbins, C.E., W.G. Kaelin, and N.P. Pavletich, *Structure of the VHL-ElonginC-ElonginB complex: Implications for VHL tumor suppressor function*. Science, 1999. **284** 455-461.
78. Pallares, I., R. Bonet, R. Garcia-Castellanos, S. Ventura, et al., *Structure of human carboxypeptidase A4 with its endogenous protein inhibitor, latexin*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102** 3978-3983.
79. Chen, L.Y., N. Doi, R.C.E. Durley, A.Y. Chistoserdov, et al., *Refined crystal structure of methylamine dehydrogenase from paracoccus denitrificans at 1.75 angstrom resolution*. Journal of Molecular Biology, 1998. **276** 131-149.
80. Abergel, C., V. Monchois, D. Byrne, S. Chenivesse, et al., *Structure and evolution of the Ivy protein family, unexpected lysozyme inhibitors in Gram-negative bacteria*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104** 6394-6399.
81. Das, K., S.E. Martinez, J.D. Bauman, and E. Arnold, *HIV-1 reverse transcriptase complex with DNA and nevirapine reveals non-nucleoside inhibition mechanism*. Nature Structural & Molecular Biology, 2012. **19** 253-259.

82. Zhou, Q.J., Y.J. Zhai, J.Z. Lou, M. Liu, et al., *Thiabendazole inhibits ubiquinone reduction activity of mitochondrial respiratory complex II via a water molecule mediated binding feature*. Protein & Cell, 2011. **2** 531-542.
83. Zhang, Y.F., Y. Zheng, L. Qin, S.H. Wang, et al., *Structural characterization of a beta-hydroxyacid dehydrogenase from Geobacter sulfurreducens and Geobacter metallireducens with succinic semialdehyde reductase activity*. Biochimie, 2014. **104** 61-69.
84. Liu, H.L., Z.S. Juo, A.H.R. Shim, P.J. Focia, et al., *Structural Basis of Semaphorin-Plexin Recognition and Viral Mimicry from Sema7A and A39R Complexes with PlexinC1*. Cell, 2010. **142** 749-761.
85. Kennaway, C.K., A. Obarska-Kosinska, J.H. White, I. Tuszynska, et al., *The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein*. Nucleic Acids Research, 2009. **37** 762-770.
86. Reverter, D. and C.D. Lima, *Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex*. Nature, 2005. **435** 687-692.
87. Opalka, N., J. Brown, W.J. Lane, K.A.F. Twist, et al., *Complete Structural Model of Escherichia coli RNA Polymerase from a Hybrid Approach*. Plos Biology, 2010. **8**.
88. Chen, L.Y., R. Durley, B.J. Poliks, K. Hamada, et al., *Crystal-Structure of an Electron-Transfer Complex between Methylamine Dehydrogenase and Amicyanin*. Biochemistry, 1992. **31** 4959-4964.
89. Oldham, M.L. and J. Chen, *Snapshots of the maltose transporter during ATP hydrolysis*. Proceedings of the National Academy of Sciences of the United States of America, 2011. **108** 15152-15156.
90. Nolen, B.J., R.S. Littlefield, and T.D. Pollard, *Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101** 15627-15632.

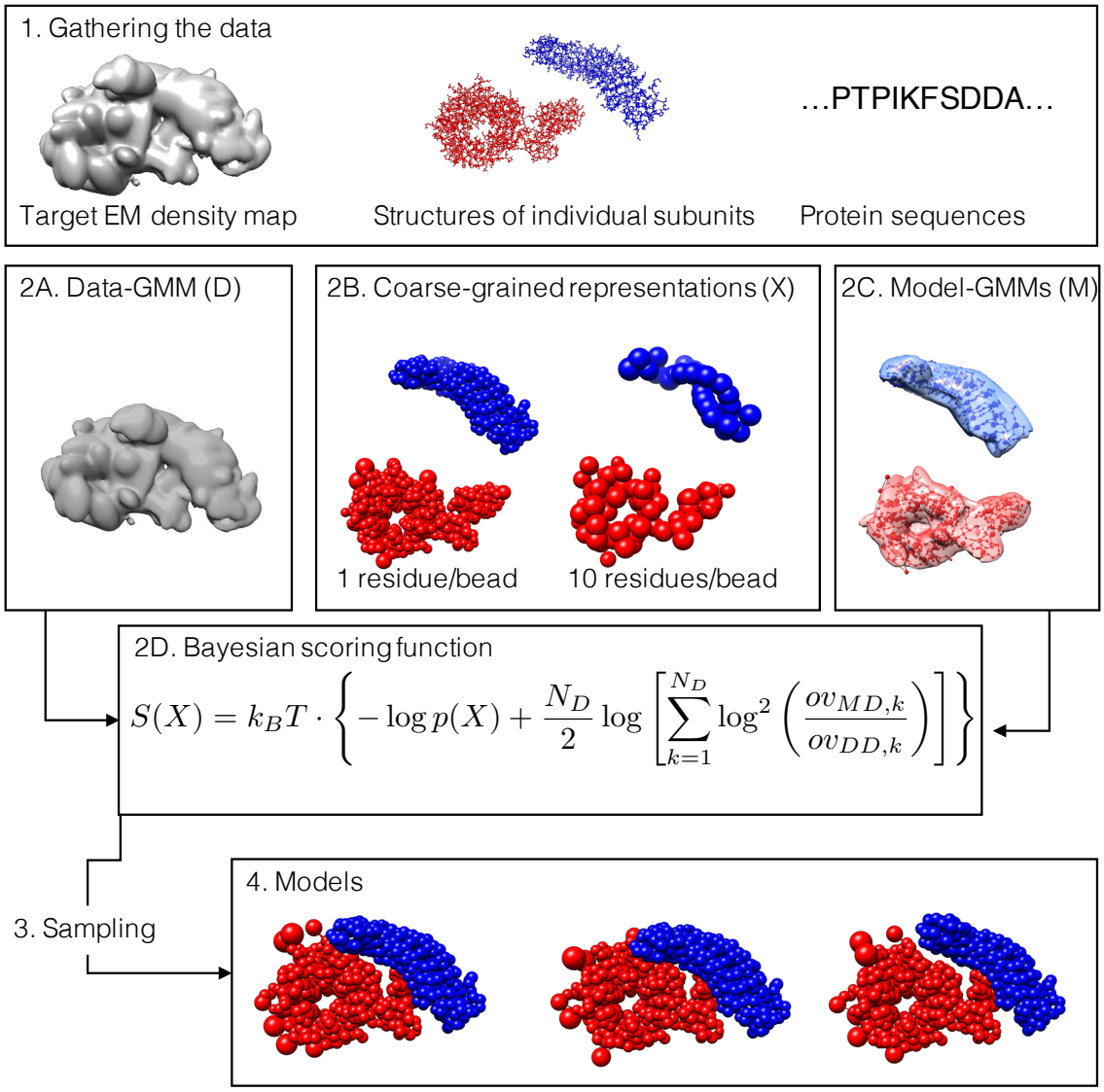

Figure 1: Workflow for multi-scale modeling of cryo-EM data. (1) The input information for the modeling protocol consists of: an experimental cryo-EM density map (left), the structures of the subunits (center), and the sequences of the subunits (right). (2) The density map is fitted with a GMM (*ie*, the data-GMM) using our divide-and-conquer approach. (3A) The atomistic coordinates of the subunits are suitably coarse-grained into large beads. Regions without a known atomistic structure are represented by a string of large beads, each representing a set of residues. (3B) GMM for the subunits (*ie*, the model-GMMs) are also computed from the atomistic coordinates. (4) The Bayesian scoring function encodes prior information about the system and measures the agreement between the data-GMM and the model-GMM. (5) Structural models are sampled by MC coupled with Replica Exchange, with or without the iterative sampling protocol. (6) The generated models are analysed.

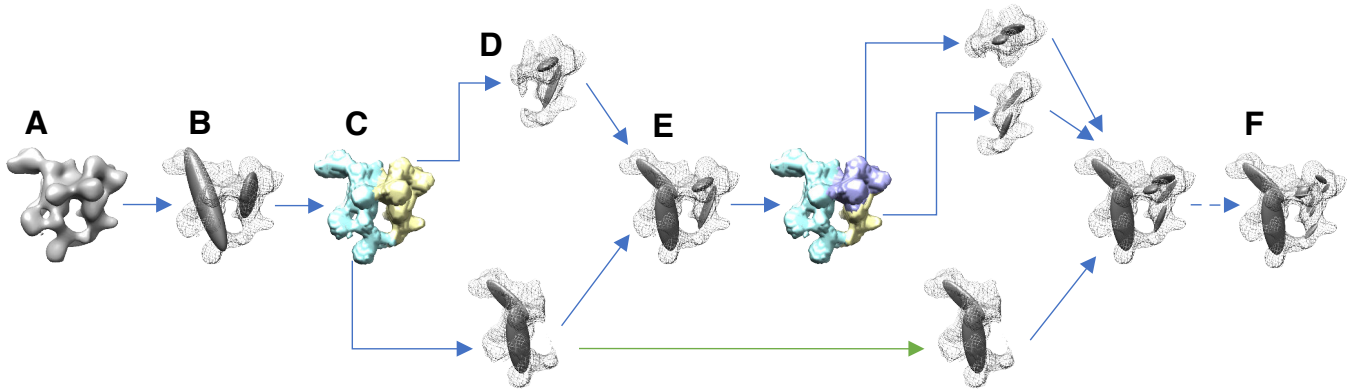

Figure 2: Divide-and-conquer approach for fitting cryo-EM density maps with a Gaussian Mixture Model (GMM). (A) The input map is threshold according to the recommended threshold. (B) The resulting map is initially fitted using a GMM with 2 components. (C) Each component of the GMM is used to partition the map into overlapping sub-maps. (D) Each sub-map is fitted using a GMM with 2 components, similarly to step B. (E) The sum of all the GMMs of the sub-maps results in a data-GMM that approximates the original map. The accuracy of approximation increases at every iteration. (F) The fitting procedure is iterated until the data-GMM reaches an optimal accuracy. The green arrow indicates a branch that was stopped because the local CC was higher than 0.95.

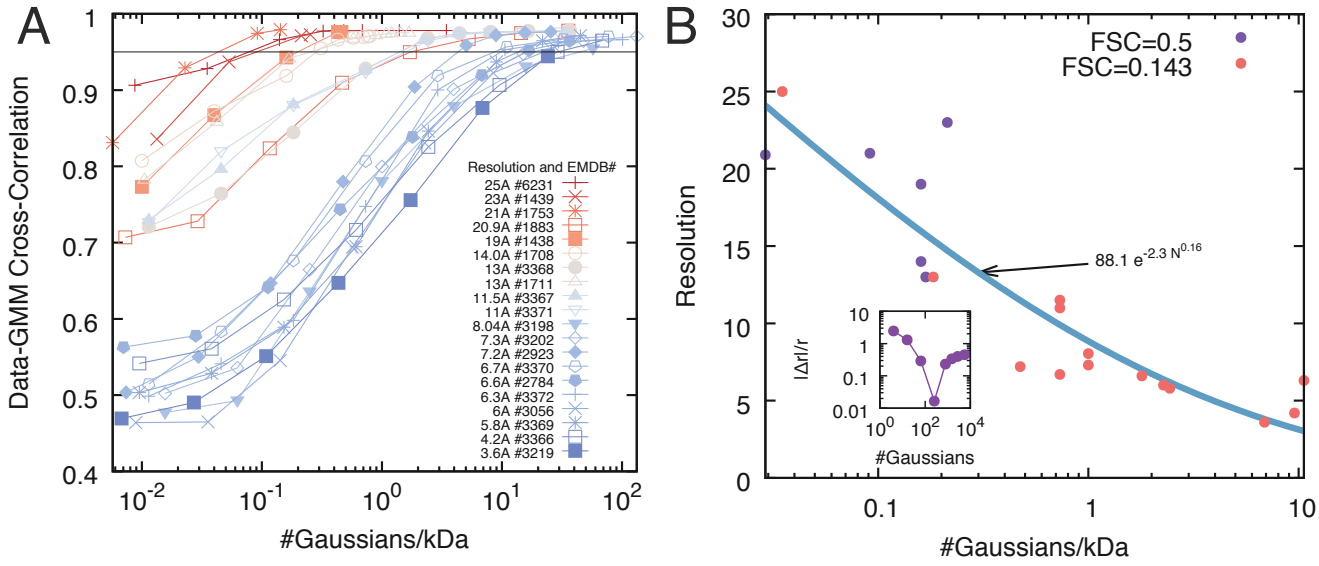

Figure 3: Benchmark of the divide-and-conquer fit of the data-GMM. (A) The accuracy of our divide-and-conquer approach is measured using the correlation coefficient between the input map and the corresponding data-GMMs obtained at different iterations. The accuracy increases with the number of components of the mixture, and the saturation point (*ie*, the number of components beyond which the accuracy does not increase significantly) depends on the resolution of the experimental map (red and blue curves are low and high resolutions, respectively). (B) Relationship between map resolution and number of components of the data-GMM. For all the density maps of panel A, the experimental resolution is plotted as a function of the optimal number of components of the data-GMM normalized by the molecular weight of the complex (solid circles). The points are fitted using a stretched exponential regression (blue line). The orange and purple circles correspond to maps whose resolution was determined by the Fourier Shell Correlation 0.143 and 0.5, respectively. (B, inset) For each density map, the optimal number of components is computed as the minimal absolute relative deviation $|\Delta r|/r$ between the data-GMM resolution and the density map resolution.

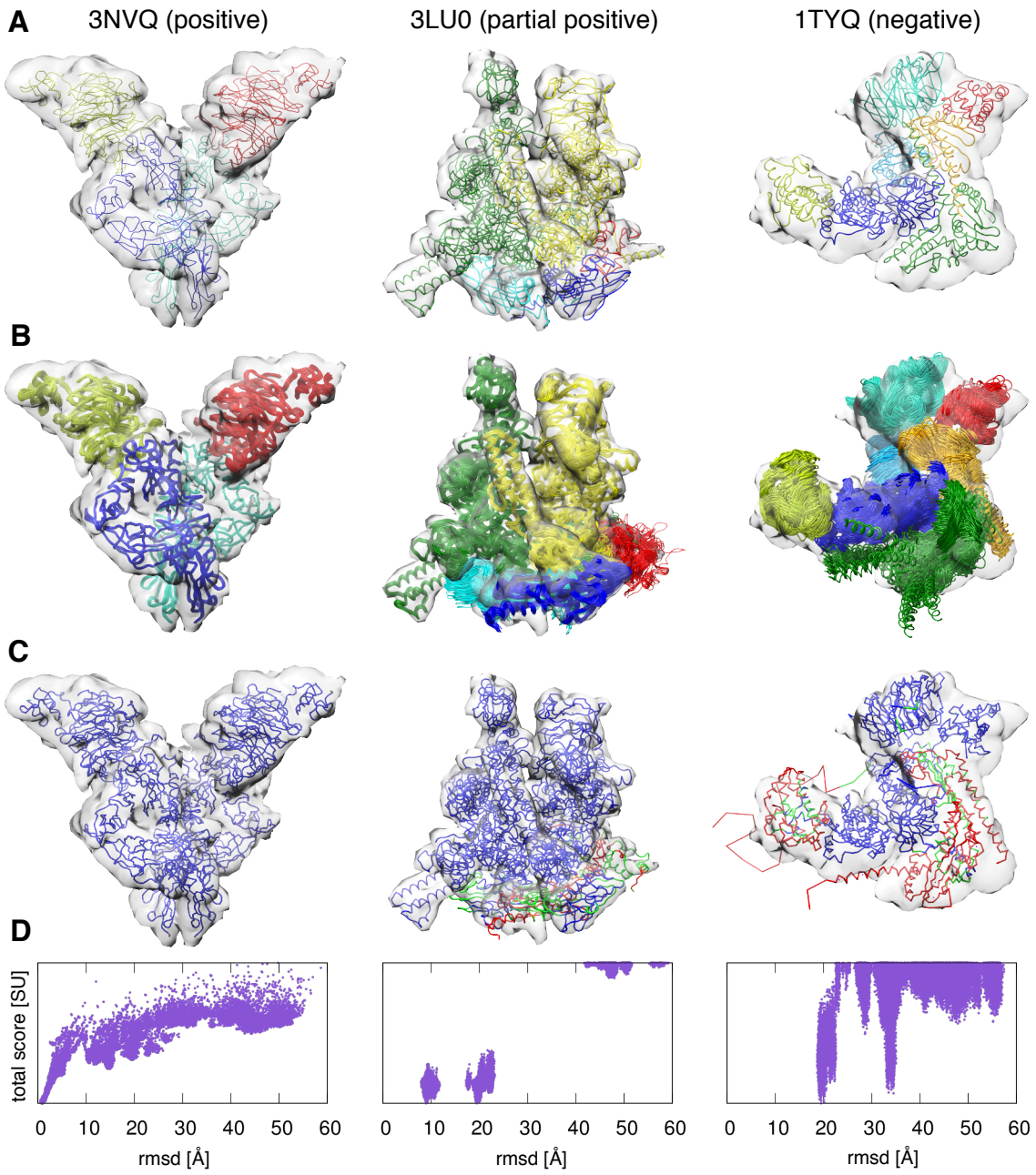

Figure 4: Benchmark of the modeling protocol. Examples of each of the three possible outcomes of the benchmark: positive (first column, PDB code 3NVQ), partial positive (second column, PDB code 3LU0), and negative (third column, PDB code 1TYQ). (A) Native structures and simulated 10 Å resolution cryo-EM density maps. (B) 50 best scoring models displayed with the simulated cryo-EM density maps. (C) Residue-wise accuracy of the best scoring models: residues whose positions deviate from the native structure less than 10 Å, between 10 and 20 Å, and above 20 Å are coloured in blue, green, and red, respectively. (D) Total score of all the sampled models as a function of the total rmsd from the native structure.

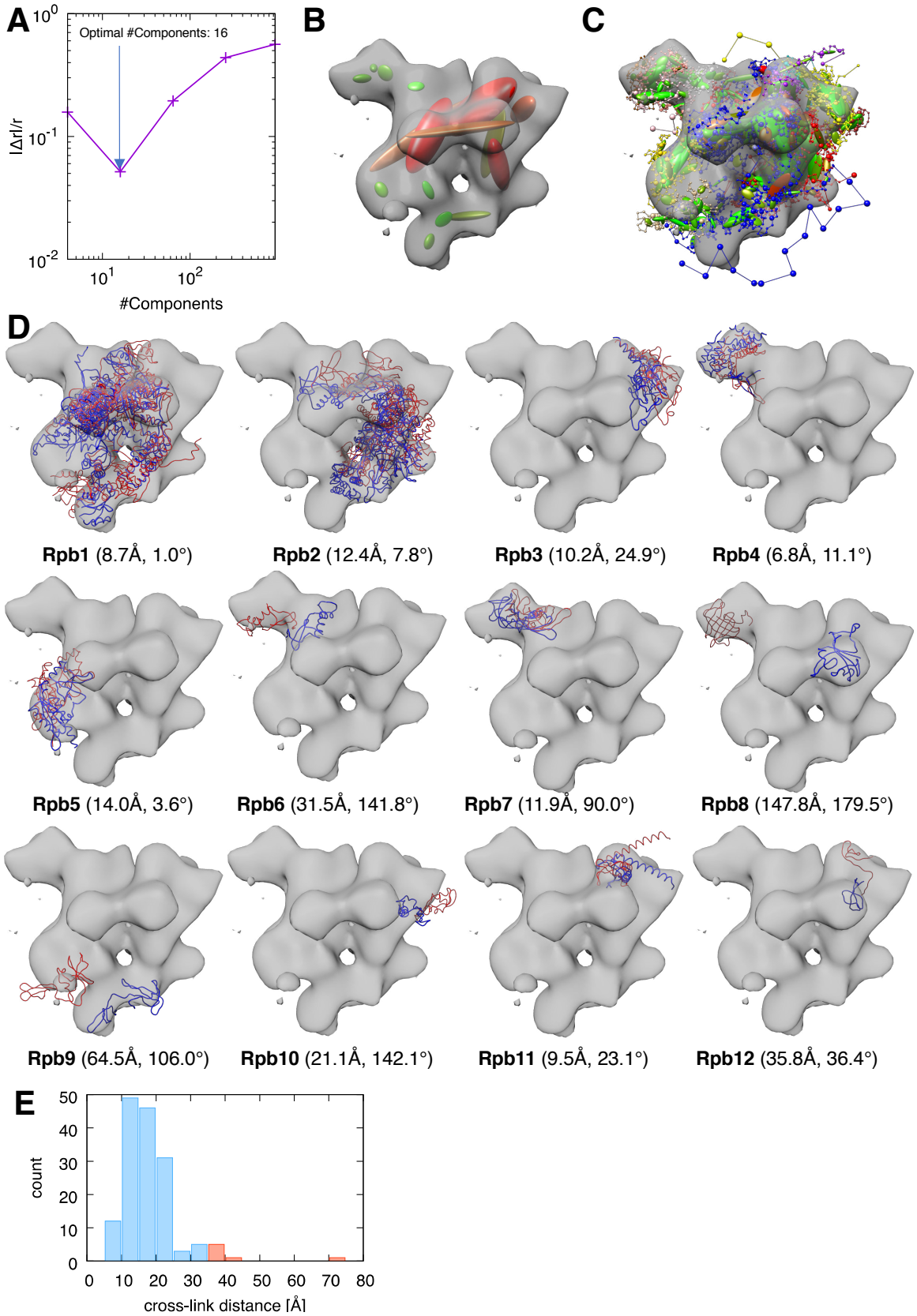
A
Optimal #Components: 16
|Δr|/r
#Components
B
C
D
Rpb1 (8.7Å, 1.0°)
Rpb2 (12.4Å, 7.8°)
Rpb3 (10.2Å, 24.9°)
Rpb4 (6.8Å, 11.1°)
Rpb5 (14.0Å, 3.6°)
Rpb6 (31.5Å, 141.8°)
Rpb7 (11.9Å, 90.0°)
Rpb8 (147.8Å, 179.5°)
Rpb9 (64.5Å, 106.0°)
Rpb10 (21.1Å, 142.1°)
Rpb11 (9.5Å, 23.1°)
Rpb12 (35.8Å, 36.4°)
E
count
cross-link distance [Å]

Figure 5: Integrative modeling of the RNA polymerase II. (A) Absolute relative deviation between data-GMM and experimental map resolutions $\frac{|\Delta r|}{r}$, plotted as a function of the number of components of the data-GMM. The minimum (blue arrow) corresponds to the optimal number of components used in the modeling (16 Gaussians). (B) The experimental cryo-EM density map (transparent grey surface) is represented with the optimal data-GMM (colored ellipsoids). The color gradient (from green to red) is proportional to the weight $\omega_{D,i}$ of the corresponding Gaussian (Eq. 2). The length of the three axes and their orientation represent the 3-dimensional covariance matrix $\Sigma_{D,i}$ (Eq. 3). (C) Representation of the best-scoring model. Coarse-grained subunits are represented by the strings of beads: the small beads and large beads represent 1- or 20-residue fragments, respectively. As for the data-GMM in panel B, the model-GMM is represented by ellipsoids. (D) All subunits of the model (red) and reference structure (PDB code 1WCM, blue) are represented along with the experimental cryo-EM map. For each panel, the name of the subunit is indicated in bold, together with the placement score of that subunit. (E) Histogram of the distance between cross-linked residues. The histogram bins corresponding to satisfied and violated cross-links are represented in blue and red, respectively.

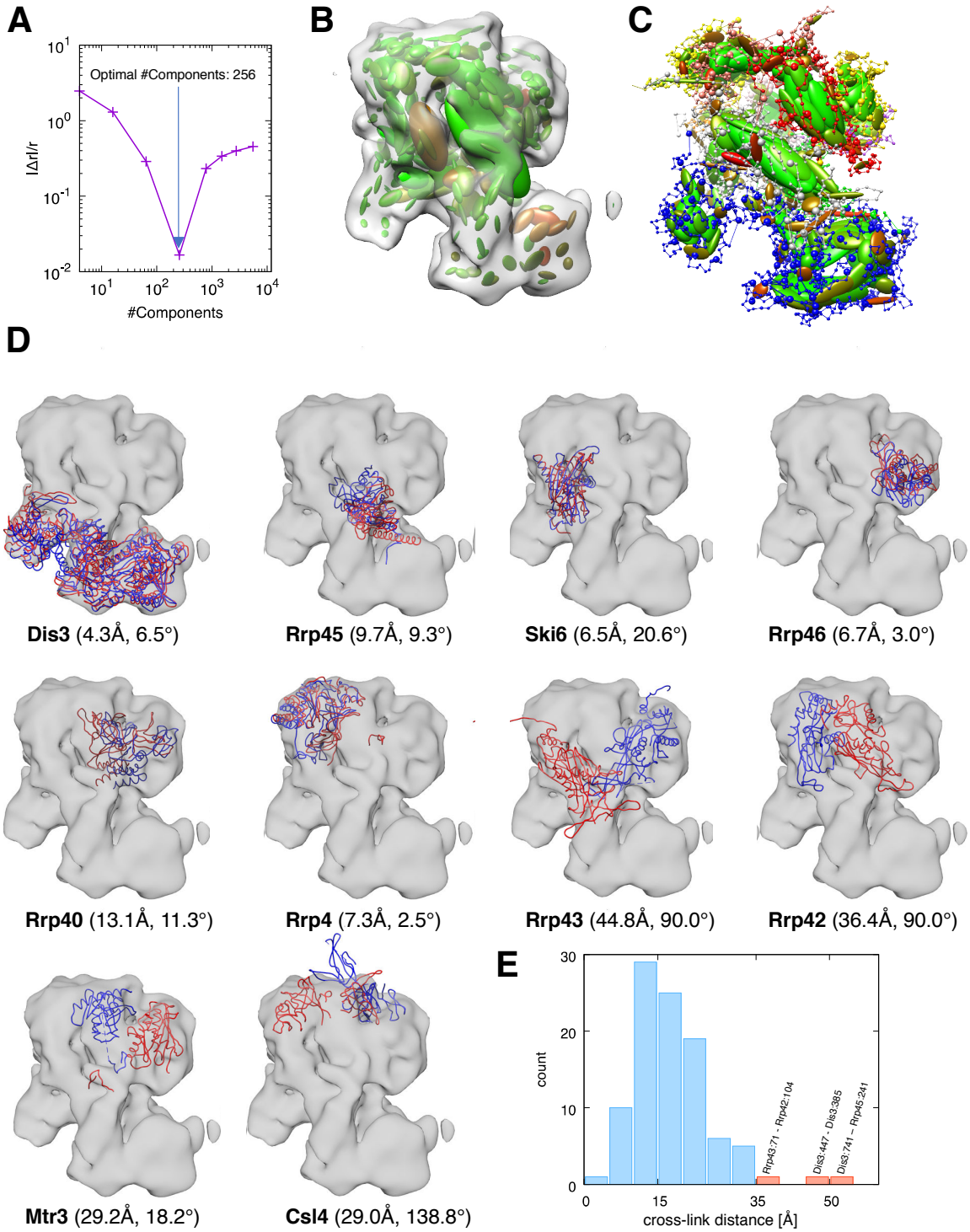

Figure 6: Integrative modeling of the yeast exosome complex. (A) Absolute relative deviation between data-GMM and experimental map resolutions $\frac{|\Delta r|}{r}$, plotted as a function of the number of components of the data-GMM. The minimum (blue arrow) corresponds to the optimal number of

components used in the modeling (256 Gaussians). (B) The original cryo-EM density map (transparent grey surface) is represented with the optimal data-GMM (colored ellipsoids). The color gradient (from green to red) is proportional to the weight $\omega_{D,i}$ of the corresponding Gaussian (Eq. 2). The length of the three axis and their orientation represent the 3-dimensional covariance matrix $\Sigma_{D,i}$ (Eq. 3). (C) Representation of the best-scoring model. Coarse-grained subunits are represented by the strings of beads: the small beads and large beads represent 1- or 10-residue fragments, respectively. As for the data-GMM in panel B, the model-GMM is represented by ellipsoids. (D) All subunits of the model (red) and reference structure (PDB code 5G06, blue) are represented along with the experimental cryo-EM map. For each panel, the name of the subunit is indicated in bold, together with the placement score of that subunit. (E) Histogram of the distance between cross-linked residues. The histogram bins corresponding to satisfied and violated cross-links are represented in blue and red, respectively.