



Bayesian Modeling of Biomolecular Assemblies with Cryo-EM Maps

Michael Habeck^{1,2*}

¹ Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, ² Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Göttingen, Germany

A growing array of experimental techniques allows us to characterize the three-dimensional structure of large biological assemblies at increasingly higher resolution. In addition to X-ray crystallography and nuclear magnetic resonance in solution, new structure determination methods such as cryo-electron microscopy (cryo-EM), crosslinking/mass spectrometry and solid-state NMR have emerged. Often it is not sufficient to use a single experimental method, but complementary data need to be collected by using multiple techniques. The integration of all datasets can only be achieved by computational means. This article describes Inferential structure determination, a Bayesian approach to integrative modeling of biomolecular complexes with hybrid structural data. I will introduce probabilistic models for cryo-EM maps and outline Markov chain Monte Carlo algorithms for sampling model structures from the posterior distribution. I will focus on rigid and flexible modeling with cryo-EM data and discuss some of the computational challenges of Bayesian inference in the context of biomolecular modeling.

Keywords: cryo-EM, modeling, Bayesian inference, Markov chain Monte Carlo, inferential structure determination

OPEN ACCESS

Edited by:

Slavica Jonic,
IMPMC, Sorbonne Universités -
CNRS UMR 7590, UPMC Univ Paris
6, MNHN, IRD UMR 206, France

Reviewed by:

Gunnar F. Schroeder,
Forschungszentrum Jülich, Germany
Riccardo Pellarin,
Centre National de la Recherche
Scientifique (CNRS), France

*Correspondence:

Michael Habeck
mhabeck@gwdg.de

Specialty section:

This article was submitted to
Biophysics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 01 December 2017

Accepted: 07 March 2017

Published: 22 March 2017

Citation:

Habeck M (2017) Bayesian Modeling
of Biomolecular Assemblies with
Cryo-EM Maps.
Front. Mol. Biosci. 4:15.
doi: 10.3389/fmolb.2017.00015

1. INTRODUCTION

Thanks to groundbreaking advances in experimental techniques it has become possible to study the structure of large biological assemblies at increasingly higher resolution. Traditionally, high-resolution biomolecular structure determination was only possible by X-ray crystallography or nuclear magnetic resonance (NMR) in solution (Berman et al., 2000). The application of NMR and X-ray crystallography to larger systems remained challenging due to the sheer size of the system and/or because it was difficult to find suitable crystallization conditions. More recently, emerging methods such as cryo-electron microscopy (cryo-EM) (Frank, 2002; Orlova and Saibil, 2004; Chiu et al., 2005), crosslinking/mass spectrometry (Gingras et al., 2007; Rappsilber, 2011) and solid-state NMR (Yan et al., 2013) have started to provide exciting insights into the structure of large macromolecular assemblies that was previously very difficult, if not impossible to obtain. In particular, cryo-EM has reached near-atomic and in some cases even atomic resolution over the last 5 years (Bai et al., 2015; Fischer et al., 2015; Khatter et al., 2015). The EM databank (EMDB) (Lawson et al., 2011) stores an increasing number of high-resolution EM reconstructions. Several biologically essential assemblies that resisted high-resolution studies have recently been characterized by cryo-EM including spliceosomal complexes (Yan et al., 2015; Agafonov et al., 2016; Galej et al., 2016; Rauhut et al., 2016; Wan et al., 2016), eukaryotic ribosomes (Anger et al., 2013; Khatter et al., 2015), and transcription initiation complexes (Plaschka et al., 2015).

Although several powerful experimental techniques are available that allow us to study the structure of large biomolecular systems, we need computational methods that assist us in

integrative modeling with diverse structural data (Sali et al., 2003; Robinson et al., 2007; Ward et al., 2013). The reasons for developing new computational methods are both of a principled and practical nature.

Structural models built from hybrid data should be as objective as possible and ideally not be biased by a human modeler, therefore automated computational modeling tools are indispensable (Karaca and Bonvin, 2013; Villa and Lasker, 2014; Schröder, 2015). The models should be compatible with all of the available data, which might come from different experimental sources. The modeling software should also be able to integrate data-independent prior information about the system.

Most existing refinement and modeling software focuses on structural data of a particular type. For example, a number of software packages for X-ray structure refinement or modeling with NMR restraints exist. To use these packages for modeling with hybrid data is often difficult and involves some sort of tweaking. We therefore need a versatile software that can integrate diverse types of structural information (Russel et al., 2012).

Every software for integrative modeling with hybrid data has to address the following questions: How much weight should the various pieces of information be given? How to deal with datasets that (partially) contradict some of the other datasets? Obviously, the weights can have a strong impact on the final structure (Brünger, 1992; Habeck et al., 2006), and it would be desirable to choose the weights in a data-driven, self-adaptive fashion. Because the individual datasets themselves typically provide only ambiguous structural information, we have to fit the model against all data simultaneously to obtain the least ambiguous result. What is a good representation of the remaining uncertainty about the structure? We need to represent the ambiguity of the structural model adequately.

The software should also be able to integrate data of varying resolution. A common scenario is that high-resolution information about the subunits in isolation is available (Esquivel-Rodríguez and Kihara, 2013), such that modeling the complex appears to be simple: we just need to put the pieces together. However, even in this seemingly simple situation several issues need to be considered.

The formation of the complex is often accompanied by a conformational change in the subunits (Gerstein et al., 1994). How much should we deviate from the known structures of the free subunits in order to fit the data of the complex? If the data is sparse (e.g., crosslinking or NMR data) or of a medium resolution, there is the risk of overfitting the data.

Another practical problem is the enormous size of the systems that can comprise tens of thousands up to millions of atoms. Is there enough information to determine the position of all atoms? Or should we rather lower our goal and aim for a coarse-grained, intermediate resolution model?

At the source of many of these issues is the question of how to deal with uncertainty in the data and about our model. We need a mathematical framework to quantitatively represent any uncertainty in the process that takes us from the input data to the final model. The framework should allow us to follow the propagation of the uncertainty about a biomolecular structure as

we combine data from diverse sources and to compute structural error bars that reflect the degree of uncertainty.

Bayesian probability theory is a unique and objective mathematical framework for quantitative inference from limited, diverse and uncertain information (Cox, 1946; Jaynes, 2003; MacKay, 2003). The essence of the Bayesian approach is that any probability should be interpreted as incomplete information about a quantity rather than a frequency of occurrence. Highly ambiguous and uncertain information results in multi-modal distributions that are spread out over many parameter values. Markov chain Monte Carlo (MCMC) methods (Liu, 2001) allow us to apply the Bayesian formalism in practice even to highly complex data and models.

More than a decade ago, Bayesian methods have been introduced for protein structure determination from solution NMR data (Rieping et al., 2005; Habeck, 2012). In this article, I will describe recent developments in Bayesian integrative modeling with hybrid data.

2. METHODS

2.1. Inferential Structure Determination

Inferential structure determination (ISD) is the first strictly statistical approach to biomolecular modeling (Habeck et al., 2005a; Rieping et al., 2005). Originally ISD was developed for solution NMR data on small protein domains (Rieping et al., 2008; Habeck, 2012). But the basic principle can be applied to large systems and diverse structural data (Bayrhuber et al., 2008; Shahid et al., 2012; Habenstein et al., 2015).

At the core of the ISD approach is a probabilistic formulation of the structure determination problem. We have to distinguish two principal types of information that guide us in the modeling of a biomolecular structure: the experimental data D and data-independent prior information I about biomolecular structures. All the information is encoded statistically through conditional probabilities. The probability:

$$\Pr(D|\theta, I)$$

quantifies how probable it is to observe data D if the actual configuration of the system is θ . $\Pr(D|\theta, I)$ is called the *likelihood* function. The prior probability:

$$\Pr(\theta|I)$$

expresses what we know about reasonable system configurations θ without observing any data.

Probability calculus allows us to combine both types of information and to derive a *posterior* distribution over all conformational degrees of freedom by invoking Bayes' theorem (Jaynes, 2003):

$$\Pr(\theta|D, I) = \frac{1}{\Pr(D|I)} \Pr(D|\theta, I) \Pr(\theta|I).$$

The posterior $\Pr(\theta|D, I)$ expresses what we know about the unknown structure given the experimental data D and our prior knowledge I . The probability $\Pr(D|I)$ (the so-called model

evidence) can be ignored if we are only interested in estimating θ , because $\Pr(D|I)$ does not depend on θ . However, if we aim to compare different prior or modeling assumptions, it will be important to calculate $\Pr(D|I)$ (Habeck, 2011; Mechelke and Habeck, 2012, 2014; Knuth et al., 2015).

Often, we need to introduce additional unknown parameters to express our prior information or to model the experimental data. Let's denote these parameters by ξ ; in statistical parlance, ξ are *nuisance parameters*. It is straightforward to infer both θ and ξ from the experimental data. All we need to do is to introduce a prior probability for the model parameters ξ and to invoke Bayes' theorem on the joint parameter space:

$$\Pr(\theta, \xi|D, I) = \frac{1}{\Pr(D|I)} \Pr(D|\theta, \xi, I) \Pr(\theta|I) \Pr(\xi|I).$$

where we assumed that θ and ξ are independent *a priori*: $\Pr(\theta, \xi|I) = \Pr(\theta|I) \Pr(\xi|I)$. It is straightforward to relax this assumption if necessary.

The posterior probability $\Pr(\theta, \xi|D, I)$ encodes all available information about the unknown parameters. In biomolecular structure determination, the posterior is typically too complex to do any further analytical calculations. By drawing Monte Carlo samples from $\Pr(\theta, \xi|D, I)$ we generate a finite approximation of the posterior (Liu, 2001). These samples can be used to compute expectations and variances over the unknown parameters and thereby estimate the parameters and compute error bars.

2.2. Probabilistic Models for Hybrid Data

Before we can launch an ISD calculation, we need to choose a likelihood $\Pr(D|\theta, \xi, I)$ and the priors $\Pr(\theta|I)$ and $\Pr(\xi|I)$. The application of ISD to multiple datasets D_i is straightforward: $\Pr(D|\theta, \xi, I) = \prod_i \Pr(D_i|\theta, \xi)$. Each dataset is described independently with an appropriate probabilistic model; all datasets are integrated by simply multiplying all factors representing the various datasets. Because probabilities for different datasets are calibrated (they all normalize to one), there is no issue of weighing the different datasets relative to each other.

We use a Boltzmann distribution as a prior over the conformational degrees of freedom:

$$\Pr(\theta|I) = \frac{1}{Z} \exp\{-E(\theta)\} \quad (1)$$

where $E(\theta)$ is a force field. ISD currently supports two force fields: a quartic repulsion term that lacks any attractive interaction, and a linearly ramped Lennard-Jones potential (see Habeck, 2011; Mechelke and Habeck, 2012 for more details). The prior distribution $\Pr(\theta|I)$ allows us to restrict the conformational degrees of freedom such that reasonable model structures are preferred (for example, structures that are free of atom-atom clashes and have well-packed interfaces). The prior distribution over the model parameters $\Pr(\xi|I)$ is typically of a standard form and chosen such that sampling with MCMC is straightforward.

2.2.1. Probabilistic Model for EM Maps

The result of a cryo-EM study is a 3D reconstruction of the structure, which typically comes in the form of a regular cubic

grid with equal grid spacing in all three spatial directions. To construct a probabilistic model for 3D reconstructions, we first need a mathematical relation that allows us to compute a theoretical density map from a given structure θ . ISD's current model for density maps is quite simple. The theoretical map is obtained from an atomic model by placing spherical Gaussians of the same size and weight at each atom. The theoretical density at 3D position x is:

$$\rho(x; \theta, \sigma) = \sum_k \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\left\{-\frac{1}{2\sigma^2}\|x - x_k(\theta)\|^2\right\} \quad (2)$$

where the index k runs over all atoms that contribute to the density and $x_k(\theta)$ is the 3D position of the k -th atom in the structure parameterized by the conformational degrees of freedom θ . The theoretical density map can be interpreted as a blurred version of an atomic map with infinite resolution:

$$\rho(x; \theta, \sigma) = g_\sigma * \rho(x; \theta, 0) \quad \text{with} \quad \rho(x; \theta, 0) = \sum_k \delta[x - x_k(\theta)]$$

where δ is the Dirac delta function, g_σ is a Gaussian blur kernel with bandwidth σ and $*$ denotes a 3D convolution. Model (2) is admittedly simplistic and valid only for modeling protein complexes at intermediate to low resolutions. For high-resolution maps and/or the modeling of protein/nucleic acid complexes the model should also incorporate atom-wise weights (proportional to atom mass) as well as scattering and temperature factors.

Let us assume that experimental values ρ_n are available at positions x_n ($n = 1, \dots, N$) which are typically the centers of voxels that make up a cubic grid. The discrepancy between the experimental map ρ_n and the theoretical map $\rho(x_n; \theta, \sigma)$ can be assessed with a Gaussian distribution. Alternative error models for density maps have been proposed (Vasishatan and Topf, 2011), but the Gaussian model is still the most widely used model.

The likelihood function resulting from a Gaussian model is:

$$\begin{aligned} \Pr(\rho|\theta, \xi, I) &= \prod_{n=1}^N \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2} [\rho_n - \alpha\rho(x_n; \theta, \sigma)]^2\right\} \\ &= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2} \sum_n [\rho_n - \alpha\rho(x_n; \theta, \sigma)]^2\right\} \end{aligned} \quad (3)$$

where the calibration factor α was introduced. There are three nuisance parameters $\xi = (\sigma, \alpha, \lambda)$. Typically, the bandwidth of the blur kernel σ is set to a constant value which depends on the resolution of the map. For example, the default value in Chimera (Pettersen et al., 2004) is $\sigma = 0.225 \times \text{resolution}$. For this fixed choice of the bandwidth, σ can be absorbed into the background information I . However, it is also possible to estimate σ along with the other nuisance parameters and the conformational degrees of freedom.

To estimate the scaling parameter, we have to look at the conditional posterior distribution:

$$\Pr(\alpha | \lambda, \theta, D, I) \propto \Pr(\alpha | I) \times \exp \left\{ -\frac{\lambda \|\rho(\theta, \sigma)\|^2}{2} \left(\alpha - \frac{\sum_n \rho_n \rho(x_n; \theta, \sigma)}{\|\rho(\theta, \sigma)\|^2} \right)^2 \right\}$$

where $\|\rho\| = \sqrt{\sum_n \rho_n^2}$. The second factor is a Gaussian centered about the estimator:

$$\hat{\alpha}(\theta, \sigma) = \frac{\sum_n \rho_n \rho(x_n; \theta, \sigma)}{\|\rho(\theta, \sigma)\|^2} \quad (4)$$

which is the slope of a straight line relating the calculated volume $\rho(x_n; \theta, \sigma)$ to the observed density ρ_n .

The Gaussian model is directly related to the cross-correlation coefficient, which is often used to compare EM maps. To see this, let's integrate out the unknown scaling factor α . If we ignore the fact that α should be positive and choose a uniform (improper) prior over α (i.e., $\Pr(\alpha | I) = \text{const}$), we can analytically integrate out α to obtain a new likelihood that no longer depends on α (this procedure is also called marginalization in Bayesian statistics, Habeck et al., 2005a):

$$\Pr(\rho | \theta, \lambda, I) = \int d\alpha \Pr(\rho | \theta, \alpha, \lambda, I) \Pr(\alpha | I) \propto \lambda^{(N-1)/2} \exp \left\{ -\frac{\lambda \|\rho\|^2}{2} [1 - C^2(\theta)] \right\} \quad (5)$$

where

$$C(\theta) = \frac{\sum_n \rho_n \rho(x_n; \theta, \sigma)}{\|\rho\| \|\rho(\theta, \sigma)\|}$$

is the cross-correlation between the experimental and the theoretical map. The effective likelihood function (Equation 5) attains its maximum when the cross-correlation coefficient is one. Whenever we assess the goodness of fit between the model and the experimental map by means of the cross-correlation coefficient, we implicitly assume that the error of the EM map follows a Gaussian distribution.

The parameter λ is the inverse variance of the Gaussian likelihood (Equation 3) and called the *precision* of the model (Bernardo and Smith, 2009). It is also possible to estimate the precision λ of the fit between the experimental and the theoretical density map. The parameter λ assesses how well the experimental and theoretical map agree on average. For large λ , the experimental map is very reliable and imposes a strong force on the model to adapt itself such that the calculated map reproduces the observed map as closely as possible. Assuming Jeffreys's prior for the precision, i.e., $\Pr(\lambda | I) = 1/\lambda$, the conditional posterior of the precision is a Gamma distribution (Habeck et al., 2006):

$$\Pr(\lambda | \theta, \alpha, \rho, I) \propto \lambda^{N/2-1} \exp\{-\lambda E_{\text{map}}(\theta, \alpha)\} \quad (6)$$

where the least-squares residual

$$E_{\text{map}}(\theta, \alpha) = \frac{1}{2} \sum_n [\rho_n - \alpha \rho(x_n; \theta, \sigma)]^2$$

is the restraint energy resulting from the Gaussian model of the experimental EM map. The expected value of the precision given the experimental map ρ and all unknown parameters is the inverse mean-squared error:

$$\hat{\lambda}(\theta, \alpha) \approx \frac{N}{2 E_{\text{map}}(\theta, \alpha)}. \quad (7)$$

Estimator (Equation 7) tells us that the precision of the map increases when the fit between the observed map and the calculated map improves. This seems reasonable, but there is a problem.

Typically, EM maps are surrounded by bordering layers of low density voxels ($\rho_n \approx 0$). If we classify all voxels into N_1 voxels that contain density of the biomolecular assembly and N_0 voxels that carry only noise or zero density, we have $N = N_0 + N_1$. By increasing N_0 (e.g., by zero padding) the goodness of fit E_{map} does not change or changes only very little, such that we can artificially increase the apparent precision of the density map simply by increasing N_0 :

$$\hat{\lambda}(\theta, \alpha) \approx \frac{N_0 + N_1}{2 E_{\text{map}}(\theta, \alpha)} \geq \frac{N_1}{2 E_{\text{map}}(\theta, \alpha)}.$$

To obtain a realistic estimate of λ , we should only fit those voxels that carry real density.

In principle, the task of classifying voxels into noise and non-noise voxels is an inference problem in itself: we would have to introduce a mask that tells us whether a voxel carries true signal or not. For the sake of simplicity we do not introduce an adaptive mask that we estimate along with the model parameters, but restrict the fitting to voxels that are likely to carry the true signal. These voxels are identified in a couple of preparatory steps, which I will outline in the next section.

If we look at the conditional posterior of the conformational degrees of freedom θ , we find that:

$$\Pr(\theta | \xi, \rho, I) \propto \exp\{-E(\theta) - \lambda E_{\text{map}}(\theta, \alpha)\}. \quad (8)$$

By taking the negative logarithm of the posterior probability, we obtain a hybrid energy function (Jack and Levitt, 1978; Brünger and Nilges, 1993; Habeck et al., 2005a):

$$E_{\text{hybrid}}(\theta) = E(\theta) + \lambda E_{\text{map}}(\theta, \alpha). \quad (9)$$

The precision acts as a weighting factor for the EM map (Habeck et al., 2006). If λ is too large, the forces from the EM term can bias the final structure (overfitting). Therefore, it is important to obtain a realistic estimate of λ .

2.2.2. Preparation of EM Maps

ISD carries out several preparatory steps before modeling with EM maps starts: thresholding, cropping, decimation, and masking. These steps improve the speed of fitting and are necessary to obtain a meaningful estimate of the precision of the density map.

Typically the user provides a threshold ρ_{\min} above which the density shows the particle. ISD clips the density at ρ_{\min} , i.e., all values greater than the threshold are set to the threshold. After clipping, the density is shifted by subtracting the threshold such that the smallest experimental density is zero:

$$\rho_n \leftarrow \begin{cases} \rho_n - \rho_{\min} & ; \rho_n \geq \rho_{\min} \\ 0 & ; \rho_n < \rho_{\min} \end{cases} \quad (10)$$

After thresholding all $\rho_n \geq 0$. To reduce the map to those voxels that carry the real signal, a cropping operation is applied to reduce the 3D grid to a minimum size. Cropping removes bordering layers which only contain zero-density voxels analogous to an auto crop in image processing programs.

To represent the assumption that the structure is entirely covered by the thresholded density map, ISD introduces a box prior, which confines the system to lie inside the interior of a cubic box that coincides with the boundary of the 3D map. The box is parameterized by its lower left and upper right corner where the lower left corner is located at the origin of the 3D grid on which the thresholded EM map is evaluated. The box has a soft boundary which is implemented as a logistic function with finite steepness γ :

$$s_\gamma(x) = \frac{1}{1 + e^{-\gamma x}} \quad (11)$$

where typically $\gamma = 1 \text{ \AA}^{-1}$. The complete prior over the conformational degrees of freedom is:

$$\Pr(\theta|I) \propto \exp\{-E(\theta)\} \prod_k \prod_{d=1}^3 s_\gamma(x_{kd}(\theta) - l_d) s_\gamma(u_d - x_{kd}(\theta)) \quad (12)$$

where l_d, u_d are the spatial coordinates of the lower left / upper right corner of the bounding box of the EM map and $x_{kd}(\theta)$ are the spatial coordinates of the k -th atom.

The Gaussian likelihood (Equation 3) is only valid for voxels that carry signal. Let us introduce a binary mask $m_n \in \{0, 1\}$ which indicates for each voxel, if it carries signal ($m_n = 1$) or noise ($m_n = 0$). The modified Gaussian likelihood is:

$$\Pr(\rho|\theta, \xi, I) = \left(\frac{\lambda}{2\pi}\right)^{\sum_n m_n/2} \exp\left\{-\frac{\lambda}{2} \sum_n m_n [\rho_n - \alpha \rho(x_n; \theta, \sigma)]^2\right\}. \quad (13)$$

As mentioned above, the mask m_n should in principle be also considered an unknown parameter and therefore be estimated along with the other unknown quantities. However, this is currently not implemented in ISD and therefore m is part of the background information I .

Another parameter that we have to consider is the spacing of the EM map. The Gaussian likelihood assumes that the discrepancy between the experimental and calculated map is independent from voxel to voxel and shows no spatial correlations. However, this assumption is violated when the size of the voxels becomes too small. By resampling the experimental map on a finer grid, we could artificially increase the number of data points, which would result in an increase of the estimated weight λ . Therefore, EM maps are typically downsampled in ISD such that the spacing is roughly $2 \times \sigma$. A more rigorous treatment that accounts for spatial correlations between neighboring voxels is currently under development.

2.2.3. Conformational Degrees of Freedom

ISD supports multiple parameterizations for biomolecular systems. ISD typically decouples internal degrees of freedom from rigid external degrees of freedom, although modeling based on Cartesian coordinates is also supported. In case we want to model the internal flexibility of the subunits of a biomolecular assembly, ISD uses dihedral angles to parameterize the atom positions. The external degrees of freedom are three translational and three rotational degrees of freedom. To parameterize the rotation matrices, ISD uses a Lie group representation (Gallego and Yezzi, 2015). It is also possible to model symmetric assemblies by using virtual copies of the symmetry mates. ISD supports cyclic, dihedral and helical symmetry. The parameters of a helical symmetry can be estimated along with the conformational degrees of freedom.

To sample the conformational degrees of freedom θ , ISD uses the gradient of the log posterior probability (i.e., the gradient of the hybrid energy). Typically it is straightforward to compute the gradient with respect to the Cartesian coordinates. The Cartesian gradient is mapped onto the conformational degrees of freedom by virtue of the chain rule. This requires us to evaluate the Jacobian of the parameterization. In case of dihedral angles, there is an efficient recursive algorithm that avoids building up the full Jacobian matrix by traversing the tree of covalent bonds.

2.3. Markov Chain Monte Carlo for Biomolecular Modeling

The posterior probability $\Pr(\theta, \xi|D, I)$ encodes everything that can be said about the conformational degrees of freedom θ and the nuisance parameters ξ in the light of the experimental data D and our modeling assumptions I . Because $\Pr(\theta, \xi|D, I)$ is a high-dimensional probability distribution that is not suited for analytical computations, we explore $\Pr(\theta, \xi|D, I)$ by drawing random samples from it. Sampling from $\Pr(\theta, \xi|D, I)$ is based on Markov chain Monte Carlo (MCMC) (Liu, 2001). An MCMC algorithm simulates a Markov chain over (θ, ξ) space whose stationary distribution is the posterior $\Pr(\theta, \xi|D, I)$. After convergence of the Markov chain, the generated θ, ξ are valid samples from $\Pr(\theta, \xi|D, I)$. The samples can be used to compute expected values, variances and other statistics that characterize the posterior distribution. If we were to construct a multi-dimensional histogram from the θ, ξ samples, it would approximate the posterior distribution. The longer we run the Markov chain, the closer we get to the posterior distribution.

2.3.1. Gibbs Sampling

Gibbs sampling (Geman and Geman, 1984) is an iterative MCMC algorithm that decomposes sampling from $\Pr(\theta, \xi | D, I)$ into two successive steps, which are repeated:

$$\begin{aligned}\theta^{(t+1)} &\sim \Pr(\theta | \xi^{(t)}, D, I) \\ \xi^{(t+1)} &\sim \Pr(\xi | \theta^{(t+1)}, D, I)\end{aligned}\quad (14)$$

where t is an iteration index (pseudo time) and the superindex (t) marks samples generated in the t -th iteration; the notation \sim means “sampled from.” It can be shown that the Gibbs sampler (Equation 14) generates valid samples from the joint distribution $\Pr(\theta, \xi | D, I)$.

To implement a Gibbs sampler, we need to compute the conditional posterior distributions $\Pr(\theta | \xi, D, I)$ and $\Pr(\xi | \theta, D, I)$. The conditional posterior over the conformational degrees of freedom involves the hybrid energy (Equation 9):

$$\Pr(\theta | \xi, D, I) \propto \exp\{-\lambda E_{\text{map}}(\theta, \alpha) - E(\theta)\}. \quad (15)$$

Sampling of the nuisance parameters is most easily done by applying a Gibbs sampling strategy to $\Pr(\xi | \theta, D, I)$ itself. We break down the second step in scheme (14) into the generation of α and λ samples according to:

$$\begin{aligned}\alpha^{(t+1)} &\sim \Pr(\alpha | \lambda^{(t)}, \theta^{(t+1)}, D, I) \\ \lambda^{(t+1)} &\sim \Pr(\lambda | \alpha^{(t+1)}, \theta^{(t+1)}, D, I)\end{aligned}\quad (16)$$

The conditional posteriors for the individual nuisance parameters, e.g., $\Pr(\lambda | \alpha, \theta, D, I)$, have been discussed in the previous section. Often these distributions are of a standard form and can be sampled directly using random number generators. For example, the conditional posterior of the precision λ is a Gamma distribution (Equation 6). Efficient algorithms for generating variates from a Gamma distribution exist (Devroye, 1986).

2.3.2. Hamiltonian Monte Carlo

Sampling the conformational degrees of freedom θ from the conditional posterior (Equation 9) is the most challenging step in an ISD calculation. Typically, the conformational degrees of freedom are highly coupled, and $\Pr(\theta | \xi, D, I)$ exhibits multiple peaks. A powerful variant of Metropolis Monte Carlo (Metropolis et al., 1957) is the Hybrid Monte Carlo method, also known as Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2010). The improvement over the simple Metropolis sampler is achieved by using a more efficient proposal step. In the standard version of Metropolis Monte Carlo, new candidate structures are proposed by randomly perturbing a conformational degree of freedom. The perturbation is either accepted or rejected depending on whether it produced an acceptable change in the hybrid energy or not. This kind of proposal results in a random walk in conformational space, which explores the space very inefficiently, because typically we can only apply small perturbations to the structure without increasing the hybrid energy by an unacceptable amount.

HMC proposes the candidate structure by running a short molecular dynamics trajectory where the hybrid energy plays the role of a force field. This has the advantage that the moves in structure space are adapted to the shape of the posterior distribution and that the conformational degrees of freedom change conjointly rather than one by one. HMC is several orders of magnitude more efficient than random walk Metropolis Monte Carlo, but comes at an additional computational cost. To run the proposal trajectory, one needs to calculate the gradient of the hybrid energy with respect to the conformational degrees of freedom. Since ISD uses non-Cartesian parameterizations, the gradient can be quite involved. Thanks to the chain rule we can break the computation of the gradient into two steps: First, the Cartesian gradient is calculated. In a second step, the Cartesian gradient is projected into the space of the conformational degrees of freedom. ISD implements this projection for dihedral angles and the rotational degrees of freedom of a rigid-body transformation.

2.3.3. Replica-Exchange Simulation

The posterior distribution arising in an application of ISD, is quite complex and typically shows multiple modes. As we will see in Section 3.3, the posterior distribution encountered in integrative modeling with cryo-EM data is often sharply peaked and exhibits isolated peaks. It is highly challenging to draw conformational samples from such a posterior distribution. ISD uses replica-exchange simulations (also known as parallel tempering) (Swendsen and Wang, 1986; Geyer, 1991) to address the sampling problem.

There are two factors that contribute to the posterior, the prior and the likelihood, and both are difficult to simulate in their own right. Therefore, ISD controls the complexity of each factor independently by introducing two “temperatures” (Habeck et al., 2005b). The first parameter, the inverse temperature $\beta \in [0, 1]$, scales the likelihood:

$$[\Pr(D|\theta, \xi, I)]^\beta;$$

for $\beta = 1$ we obviously recover the original likelihood, for $\beta = 0$ we completely switch off the data.

The second parameter controls the shape of the conformational prior. Because the non-bonded interactions $E(\theta)$ span many orders of magnitude, it is highly inefficient to work with the standard Boltzmann ensemble which scales down the non-bonded energy when the temperature is increased. Instead of the Boltzmann ensemble, ISD uses the Tsallis ensemble to smooth out non-bonded interaction (Habeck et al., 2005b) and simulates:

$$\left[1 + (q-1)(E(\theta) - E_{\min})\right]^{-q/(q-1)}$$

where $q \geq 1$ is the so-called Tsallis q and E_{\min} has to be chosen such that $E(\theta) > E_{\min}$ for all structures. For $q = 1$, we recover the standard Boltzmann prior (Equation 1).

The choice of the tempering schedule (i.e., the sequence of β and q) is difficult and crucial. We have to trade-off efficiency vs.

ergodicity of sampling. With increasing number of temperatures, the overlap between the replicas increases which results in an elevated swapping rate. But with increasing number of replicas the time for round trips increases quadratically, because states diffuse across different temperatures (i.e., there is no directed exchange of states that would aim for rapid mixing of states across different temperatures) (Earl and Deem, 2005). Therefore, we would rather choose a minimal number of replicas such that the smallest swapping rate is maintained.

3. RESULTS

In this section, I will illustrate Bayesian integrative modeling with hybrid data focusing on EM maps.

3.1. Flexible Fitting with Hamiltonian Monte Carlo

ISD can fit known structures and structural models into EM maps. In flexible fitting, we are trying to change the internal structure of a biomolecule so as to better fit an experimental EM map. A couple of software packages for flexible fitting has been published. Normal mode and elastic network methods (Delarue and Dumas, 2004; Tama et al., 2004; Hinsen et al., 2005; Schröder et al., 2007; Jolley et al., 2008; Tan et al., 2008) boost transitions along the principal directions of structural change. Molecular dynamics (MD) based methods (Orzechowski and Tama, 2008; Trabuco et al., 2008) combine a density fitting score with a full-fledged force field. Real-space refinement in Cartesian and internal coordinates, originally developed for X-ray crystallographic data, has been adapted to cryo-EM maps (Fabiola and Chapman, 2005). Rigid-body modeling with Flex-EM (Topf et al., 2008) freezes secondary structure elements and keeps just the linker regions flexible. Fragment-based structure prediction methods such as Rosetta has been combined with density map refinement (DiMaio et al., 2009).

ISD uses dihedral angles to parameterize the structures of the subunits of a macromolecular complex. In addition to the dihedral angles, each subunit has six external degrees of freedom that describe a rigid transformation of the subunit (three translational and three rotational degrees of freedom). The complete list of dihedral angles as well as the translational and rotational degrees of freedom from all subunits makes up the conformational degrees of freedom θ .

To study flexible fitting with ISD, let us first look at a specific example. Adenylate kinase (AK) is a widely used test system to predict and simulate conformational changes in proteins (see e.g., Orzechowski and Tama, 2008; Beckstein et al., 2009; Whitford et al., 2009). AK adopts two conformational states: an open state in which no ligands are bound and a closed state. The overall difference between both states is an RMSD of ~ 7 Å. The conformational change can be understood as a rigid-body movement of three domains relative to each other: CORE, LID, and NMP-bind. During the conformational change, these three domains maintain their internal structure (Müller et al., 1996; Whitford et al., 2009).

I ran local posterior sampling with HMC starting from the open state (PDB code 4ake) and fitted it into a simulated EM

map of the closed state (PDB code 1ake) at 10 Å resolution. **Figure 1A** shows the evolution of the RMSD to the initial and target structures during flexible fitting. The simulation starts at an RMSD of about 7 Å and rapidly improves it by optimizing the agreement with the experimental and theoretical maps. This is reflected by the evolution of the cross-correlation coefficient (see **Figure 1B**), which increases as the RMSD to the target structure decreases. After less than 200 steps of HMC sampling the fitted structure has an RMSD < 1 Å to the target structure and a cross-correlation of almost 100%. During flexible fitting, the structure of the three domains remains intact. This is reflected by the fact that the RMSD restricted to those $C\alpha$ atoms that belong to the same domain changes only little compared to the change in the overall RMSD (see **Figure 1C**). Thus, the HMC sampler preserves the integrity of the input structure and introduces larger scale changes only in a few hinge regions.

3.2. Flexible Fitting Benchmark

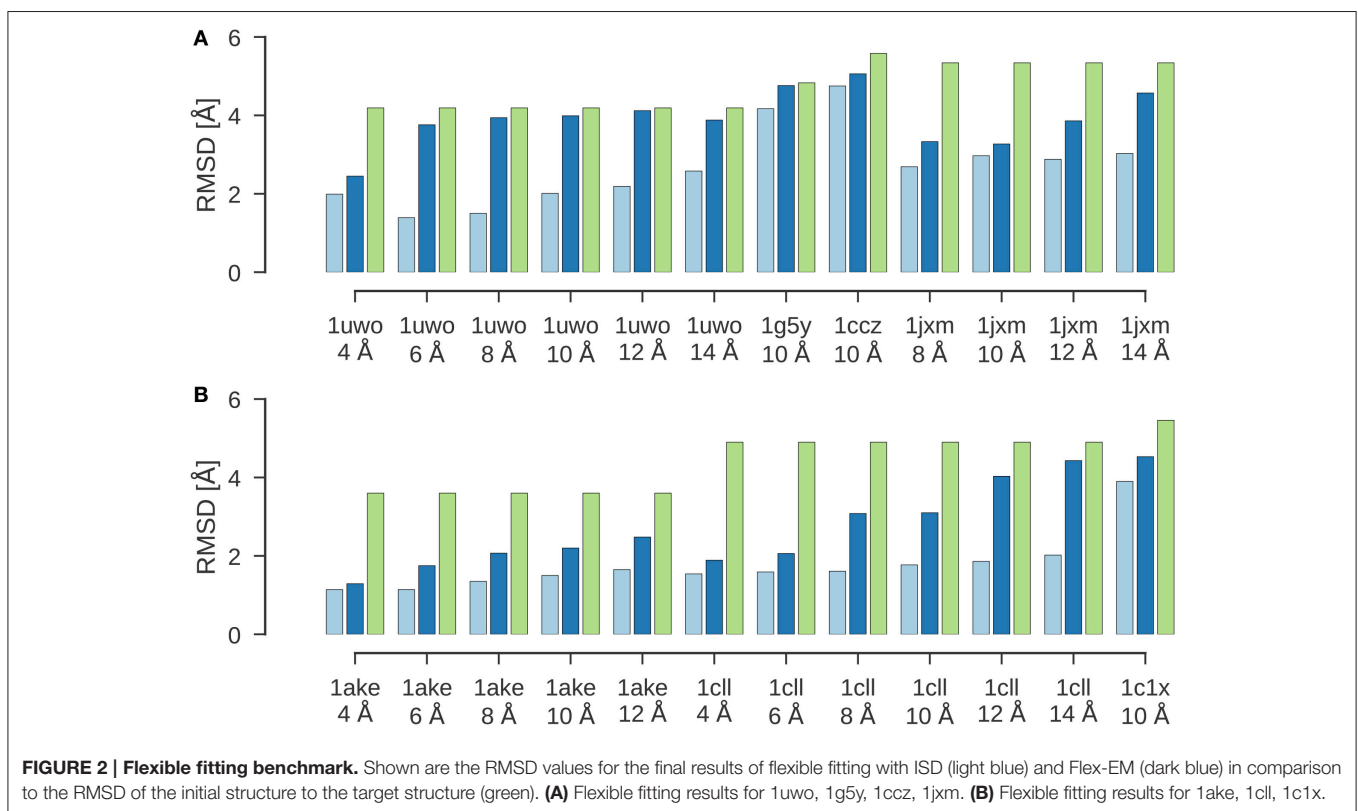
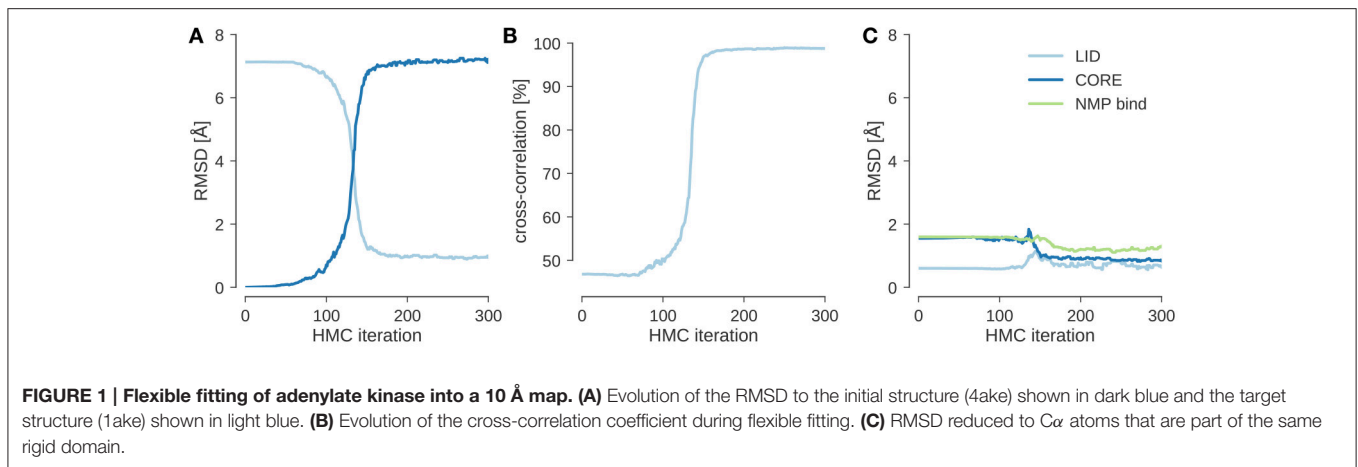
To systematically validate local flexible fitting of EM maps with ISD, I applied HMC sampling of the posterior distribution to a benchmark proposed by Topf et al. (2008) to test their Flex-EM method. The Flex-EM benchmark comprises various medium sized proteins and simulated EM maps at different resolutions ranging from 4 to 14 Å. For each flexible fitting task of the single-domain subset, I launched an HMC sampler starting from the initial structure as provided by the benchmark. The initial structure was obtained by homology modeling based on a template structure that shows an alternative conformational state. The task is to deform the homology model such that it better agrees with a simulated EM map showing a different conformational state.

Figure 2 shows the results of a flexible fitting benchmark from Topf et al. (2008). In all cases, ISD improves the fit of the initial structure quite significantly and achieves cross-correlation coefficients above 95%. Moreover, the RMSDs of the final structures fitted with ISD are systematically better than the fits obtained with Flex-EM.

Although flexible fitting with HMC performs well in practice, there are still conceptual problems with this approach. Sampling with HMC does not explore the full posterior distribution, but stays in the vicinity of the initial structure. A truly Bayesian approach, however, aims to explore the entire posterior distribution by using, for example, a full-blown replica simulation. However, global sampling of the posterior will result in many alternative fits of the EM map that will show a large RMSD to the target structure, because the force fields implemented in ISD cannot distinguish between the target structure and other globular structures that fit the density map. A remedy is to not only use the known structure that is fitted against the EM map as the initial structure, but also to develop a probabilistic model that allows for deformations of the known structure. Such a model is currently under development.

3.3. Global Fitting of Symmetric Assemblies

Global sampling of the posterior distribution is currently only possible in ISD, if the internal structure of the subunits is kept



fixed. The only degrees of freedom are the six external degrees of freedom parameterizing a global rotation and translation of each subunit. The sampling problem arising in global fitting of EM maps is quite severe. To see this, let us first study sampling from the prior (Equation 12), which is the Boltzmann ensemble confined by a soft box containing the experimental density map. Sampling from this prior is a sort of toy version of the density fitting problem. Instead of fitting the assembly against the density map, our aim is to generate non-clashing configurations that lie inside a box which contains the thresholded map. This is an instance of a 3D packing problem, which is NP-hard.

Let us look at a specific example: The symmetric chaperonin GroEL has been studied extensively by cryo-EM, X-ray crystallography and NMR. A 3D reconstruction of GroEL at a resolution of 4.1 Å is available (EMD-6422). The original map spans 240^3 voxels. The EMDB entry suggests a user-defined threshold of $\rho_{\min} = 3.5$ for visualizing the map. After thresholding (Equation 10) and cropping, the grid has $135 \times 133 \times 133$ voxels, i.e., only $\sim 17\%$ of the original volume carries information that is useful for structural modeling. The 3D cropping operation results in a box that spans a volume of $144.5 \times 142.3 \times 142.3 \text{ \AA}^3$. This example illustrates that thresholding and cropping can achieve a drastic reduction in the

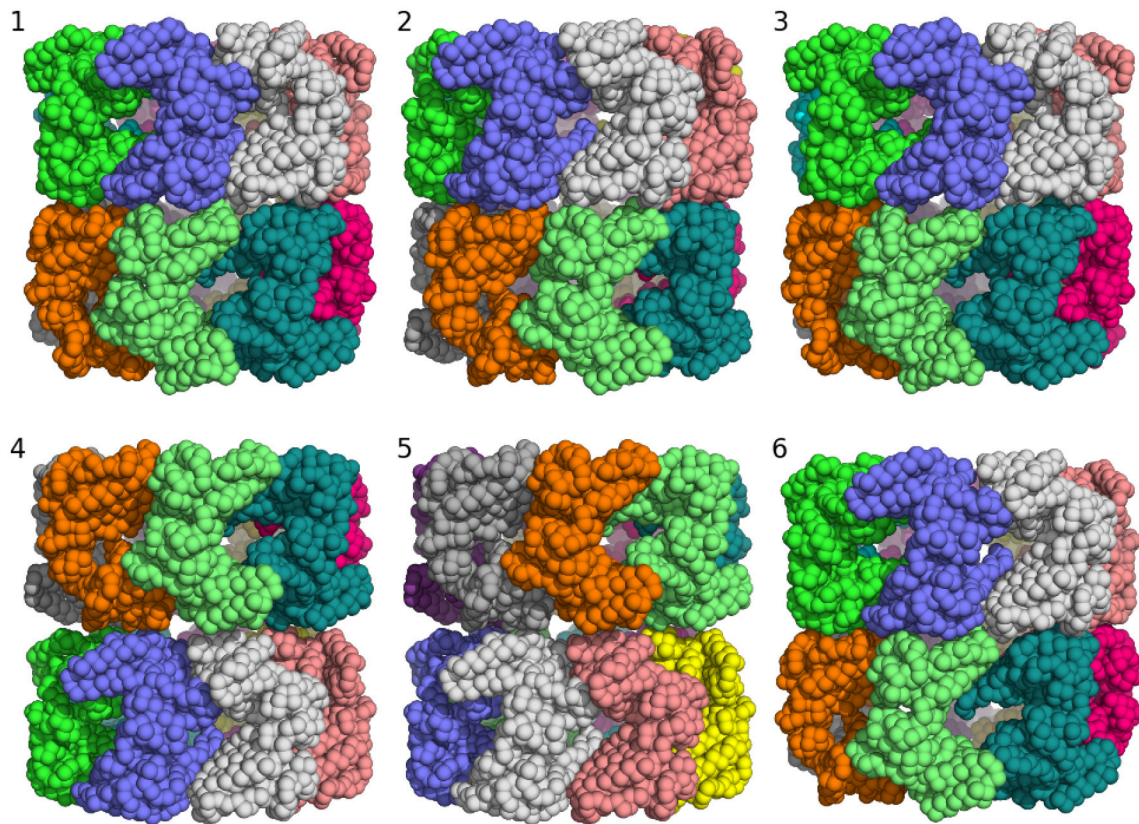


FIGURE 3 | Major structural clusters of the GroEL 14-mer generated from the prior distribution confined to a box. Subunits are color coded. The lowest energy clusters are shown on top (structures 1–3). The second lowest energy structures are clusters 4 and 5. Structure 6 is a rare high energy configuration that is also generated by replica-exchange Monte Carlo.

number of grid points that have to be evaluated during density fitting.

GroEL exhibits a seven-fold tetrahedral symmetry (D7). Therefore, our task is to sample configurations of the 14-mer that fit inside the box and minimize the overlap between atoms from different subunits. I used a Tsallis replica simulation to sample structures of the GroEL 14-mer. There are only six conformational degrees of freedom: three rotational and three translational degrees of freedom, which determine the position and orientation of a single GroEL subunit. The positions and orientations of the other 13 subunits are generated by the action of the D7 symmetry operator.

Although this is a low-dimensional sampling problem, it turns out to be surprisingly hard. I needed 59 replicas in the Tsallis ensemble to achieve an average swap rate of 38%. If the non-bonded interactions are fully switched on, there are only few arrangements that fit into the box without producing significant clashes between atoms from different subunits. As a consequence, the box prior exhibits a few isolated peaks. The shape of the prior distribution is reminiscent of a golf-course energy landscape and quite different from the funnel-shaped energy landscape imposed by distance restraints.

Clustering of the sampled rigid-body degrees of freedom yields six groups of symmetric assemblies that fit into the box

(see **Figure 3** and **Table 1**). Each group is defined very precisely with an ensemble RMSD ranging between 0.13 and 0.23 Å over the entire 14-mer. The tightness of the clusters shows that there is only a discrete set of arrangements that fits into the box. The first three clusters achieve the lowest non-bonded energies $E(\theta)$. The energy of the next two clusters is elevated by 70 units. Replica-exchange Monte Carlo occasionally also samples a high-energy structure (cluster 6). The first five clusters show the same arrangement of the seven-membered ring formed by chains A–G. The RMSD of these chains to the arrangement in the crystal structure is below 0.8 Å; only the last cluster shows a higher RMSD of 4.7 Å. The major difference between the clusters is in how the rings are arranged relative to each other. In clusters 1, 2, 3, and 6, the two rings are oriented in the same fashion as in the crystal structure (with the termini facing each other), whereas clusters 4 and 5 show an inverted orientation.

Posteriors based on distance data such as those arising in NMR applications exhibit a continuum of high-probability structures. The Markov chain is guided to the most likely structures by a funnel-shaped probability landscape. The distributions arising in EM fitting problems show a very different landscape with multiple isolated peaks that carry similar probability mass and therefore all contribute significantly to the posterior. Rigid-body modeling with EM maps can be viewed as

TABLE 1 | Summary of a clustering analysis of the prior ensemble of GroEL.

| Cluster | av. energy | Population [%] | Ensemble RMSD | RMSD (7-mer) [Å] | RMSD (14-mer) [Å] |
|---------|------------|----------------|---------------|------------------|-------------------|
| 1 | 228.8 | 22.8 | 0.2 | 0.8 | 7.8 |
| 2 | 234.0 | 23.1 | 0.2 | 0.7 | 9.0 |
| 3 | 234.1 | 23.1 | 0.1 | 0.7 | 13.4 |
| 4 | 301.7 | 19.3 | 0.2 | 0.8 | 71.5 |
| 5 | 301.7 | 11.5 | 0.2 | 0.8 | 80.2 |
| 6 | 995.5 | 0.2 | 0.1 | 4.7 | 8.6 |

Six major clusters have been identified. Listed are their average non-bonded energy, the RMSD to the average structure within each cluster (precision) and the RMSD (accuracy) to the crystal structure (PDB code 1oel) for a single ring (chains A–G) and the entire 14-mer (chains A–N).

a 3D packing problem. In case of GroEL, the packing constraint from the prior box and the D7 symmetry already determine the overall structure of the assembly to a large degree without any use of the density map. But the tests also show that even sampling from the prior alone can be quite challenging.

The minimum energy assembly sampled from the prior fits the density map only poorly with a cross-correlation of $\sim 10\%$. Refining the assembly in the presence of the map improves the cross-correlation to 55% and decreases the RMSD of the entire 14-mer to 1.1 Å.

3.4. Multi-Body Modeling of GroEL/ES

In general rigid-body modeling applications, we have to fit multiple rigid bodies into an EM map. I will use the GroEL/ES complex to illustrate multi-body fitting with ISD. GroEL/ES is formed by GroEL and the cochaperonin GroES. GroES interacts with one of the seven-membered rings formed by GroEL after a conformational change has been induced in the subunits. Therefore, the structures of the two GroEL 7-mers are no longer identical, and we have to fit three rigid bodies: one subunit of free GroEL (PDB code 1aon, chain A), one subunit of GroEL in complex with GroES (1aon, chain H), and one subunit of GroES (1aon, chain O). Each of the three subunits is duplicated by the action of a 7-fold cyclic symmetry. The symmetry mates are not represented explicitly, but generated from each of the three rigid bodies. Forces that act on the symmetry mates are backprojected onto the subunit. Therefore, we have a total of 18 conformational degrees of freedom.

I used ISD to fit GroEL/ES into a 23.5 Å map (Ranson et al., 2001) (EMD-1046). To shortcut the convergence of posterior sampling, I first ran a replica simulation with a C α representation of the subunits and switched off the non-bonded interactions. With this strategy, the sampler rapidly generates models that achieve a cross-correlation of 96% (see Figure 4D). Inspection of the structures shows that there are two clusters which differ only in the structure of the GroES subunit. The structure of the two GroEL rings is already very close to the crystal structure (1aon) with an RMSD of 3.5 ± 0.5 Å over the 14-mer formed by the GroEL subunits (Figure 4A). The GroES 7-mer arranges in two versions of the ring: One is the correct structure with an RMSD of 2.1 ± 0.6 Å to the crystal structure. The second structure is incorrect with an RMSD of 20.0 ± 0.3 Å. Both structures

are almost equally populated. The correct structure is adopted by 51.3% of the structures; the population of the incorrect assembly is 47.7% (see Figure 4B). There is a tiny fraction with a population of $\sim 1\%$ that shows a third arrangement of the GroES subunit (RMSD 9.17 ± 0.51 Å). Figure 4C shows the distribution of the RMSD over the entire assembly.

In a refinement step, I used a full-atom representation of the subunits and switched on the non-bonded energy terms. The RMSD to the crystal structure drops to 1.4 Å without compromising the fit to the EM map: the cross-correlation coefficient of the full-atom structure is still 96%.

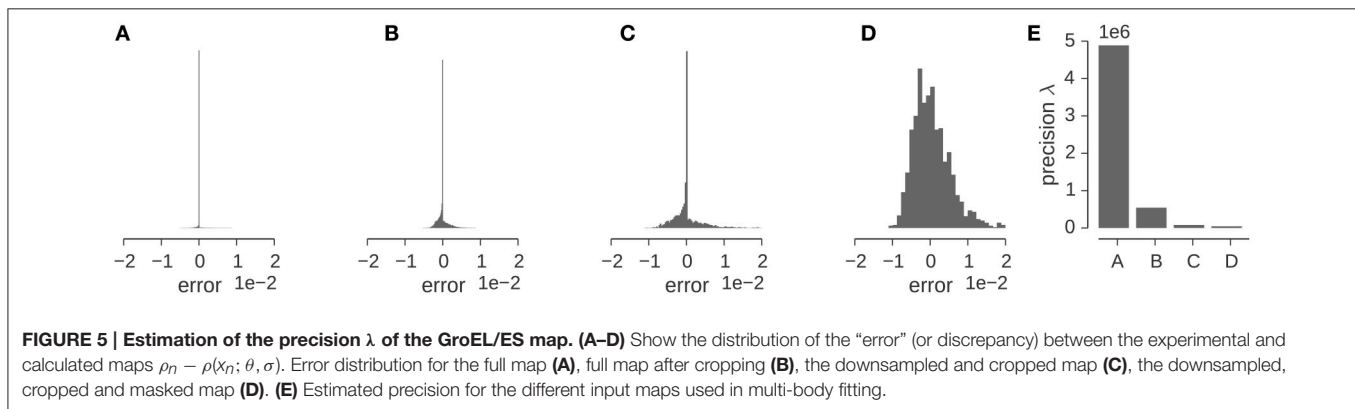
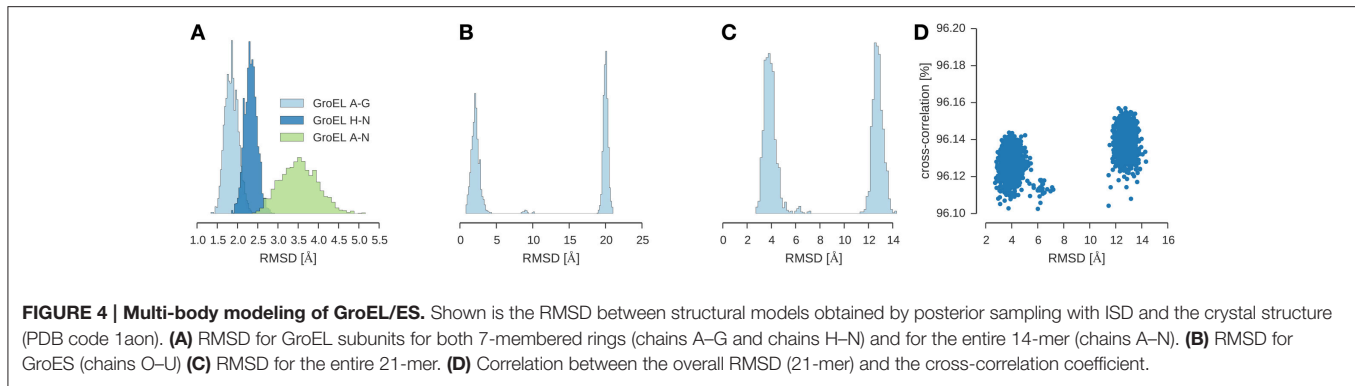
3.5. Estimation of the Precision of an EM Map

As outlined in Section 2.2.1, it is challenging to obtain a good estimate of the precision of an EM map, because an EM map typically contains many zero-density voxels in addition to the non-noise voxels, but only voxels carrying a real signal should contribute to the precision. To identify which voxels carry true signal, we would have to first solve the fitting problem. Therefore, both problems, the estimation of a well-fitting structure and the construction of a good mask, are highly related. Moreover, the errors (i.e., the discrepancy between the experimental and calculated maps) are spatially correlated, but the Gaussian model (3) treats them as completely independent observations, which also results in an artificial increase in the precision. The reason for the latter effect is the following: If errors are correlated, the effective number of data points is smaller than the number of voxels (Sivia, 2004). According to Equation (7) the precision of the map is proportional to the number of voxels for the simple Gaussian model, the precision will therefore be overestimated, if the errors between neighboring voxels are correlated.

Let us illustrate the various factors that influence the precision for a concrete example. Figure 5 shows the distribution of the discrepancy between the experimental and the calculated density map for the GroEL/ES map analyzed in the previous section. The Gaussian likelihood assumes that this distribution has a bell-shaped curve whose width is determined by the precision λ . The distribution of the discrepancy $\epsilon_n = \rho_n - \rho(x_n; \theta, \sigma)$ is shown in (Figures 5A–D) for various stages of preprocessing. The original map contains many low-density voxels that lead to a very sharp, dominating peak at zero in the distribution of ϵ_n (Figure 5A). Cropping (Figure 5B) and subsequent decimation (Figure 5C) chops away many of the zero-density voxels and decreases the detrimental effect of the low-density voxels. However, the distribution of ϵ_n is only captured well by a Gaussian, if we mask out low-density voxels (see Figure 5D). The effect of the preprocessing steps on the estimated precision is shown in Figure 5E. Each of the preparation steps lowers the estimated precision by orders of magnitude.

4. CONCLUSION

This article discusses how ISD incorporates EM maps into a structure calculation and demonstrates some aspects of Bayesian integrative modeling with EM data. The Bayesian framework is



highly suited to address issues in structural modeling with hybrid data such as how to weigh multiple datasets relative to each other. The major bottleneck of an inferential structure determination is conformational sampling. The posterior distribution arising in EM fitting poses a challenging sampling problem, which can be overcome with replica-exchange Monte Carlo.

The article does not cover crosslinking/mass spectrometry and solid-state NMR, which are complementary methods for characterizing the structure of large assemblies. ISD has also been used to model biomolecular assemblies from solid-state NMR data. For example, we have used ISD to compute the structure of the membrane domain of the trimeric autotransporter adhesin YadA (Shahid et al., 2012). We modeled a fully flexible subunit in the presence of a cyclic trimer symmetry. Although the data are highly ambiguous due to the imprecision of solid-state NMR restraints and the trimer symmetry, ISD was able to determine the correct structure of the YadA membrane anchor domain. Another example is our recent structure of a type 1 pilus FimA from *E. coli* (Habenstein et al., 2015). Here solid-state NMR and scanning electron microscopy data were combined with solution NMR data to estimate the internal structure of the subunit as well as the parameters of the helical symmetry of the FimA pilus. Also modeling with crosslinking data is possible with ISD, e.g., Carstens et al. (2016) discuss chromosome structure modeling. However, the use of crosslinking data for modeling macromolecular complexes still needs to be benchmarked thoroughly. A common scenario is to combine cryo-EM with

crosslinking data, which also needs to be tested systematically with ISD. A Bayesian approach to modeling macromolecular assemblies with crosslinking data has been proposed recently by Ferber et al. (2016).

Future work will focus on various aspects of modeling with hybrid data. One goal is to develop a better model for EM maps that incorporates the various preprocessing steps discussed in Section 2.2.2. The model will incorporate a mask that will be estimated along with the other unknown parameters. Moreover, we will develop a likelihood function that accounts for spatial correlations between errors in the density map. Another goal is to support modeling with coarse-grained representations of biomolecular systems (Tozzini, 2005; Saunders and Voth, 2013). Especially, for very large systems it will be critical to work with a multiscale representation to enable exhaustive conformational sampling. We are already using highly coarse-grained models for modeling the 3D structure of chromosomes and genomes from chromosome conformation capture data (Carstens et al., 2016).

AUTHOR CONTRIBUTIONS

MH designed and performed research and wrote the manuscript.

FUNDING

The author acknowledges funding from the German Research Foundation (DFG) (SFB 860, Project B09).

REFERENCES

- Agafonov, D. E., Kastner, B., Dybkov, O., Hofe, R. V., Liu, W. T., Urlaub, H., et al. (2016). Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science* 351, 1416–1420. doi: 10.1126/science.aad2085
- Anger, A. M., Armache, J. P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D. N., et al. (2013). Structures of the human and Drosophila 80S ribosome. *Nature* 497, 80–85. doi: 10.1038/nature12104
- Bai, X. C., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., et al. (2015). An atomic structure of human γ -secretase. *Nature* 525, 212–217. doi: 10.1038/nature14892
- Bayrhuber, M., Meins, T., Habeck, M., Becker, S., Giller, K., Villinger, S., et al. (2008). Structure of the human voltage-dependent anion channel. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15370–15375. doi: 10.1073/pnas.0808115105
- Beckstein, O., Denning, E. J., Perilla, J. R., and Woolf, T. B. (2009). Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open closed transitions. *J. Mol. Biol.* 394, 160–176. doi: 10.1016/j.jmb.2009.09.009
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235
- Bernardo, J. M., and Smith, A. F. M. (2009). *Bayesian Theory*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons.
- Brünger, A. T. (1992). The free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355, 472–474.
- Brünger, A. T., and Nilges, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Q. Rev. Biophys.* 26, 49–125. doi: 10.1017/S0033583500003966
- Carstens, S., Nilges, M., and Habeck, M. (2016). Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS Comput. Biol.* 12:e1005292. doi: 10.1371/journal.pcbi.1005292
- Chiu, W., Baker, M. L., Jiang, W., Dougherty, M., and Schmid, M. F. (2005). Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure* 13, 363–372. doi: 10.1016/j.str.2004.12.016
- Cox, R. T. (1946). Probability, frequency and reasonable expectation. *Am. J. Phys.* 14, 1–13. doi: 10.1119/1.1990764
- Delarue, M., and Dumas, P. (2004). On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6957–6962. doi: 10.1073/pnas.0400301101
- Devroye, L. (1986). *Non-uniform Random Variate Generation*. New York, NY: Springer Verlag.
- DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., and Baker, D. (2009). Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.* 392, 181–190. doi: 10.1016/j.jmb.2009.07.008
- Duane, S., Kennedy, A. D., Pendleton, B., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* 195, 216–222. doi: 10.1016/0370-2693(87)91197-X
- Earl, D. J., and Deem, M. W. (2005). Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7, 3910–3916. doi: 10.1039/b509983h
- Esquivel-Rodríguez, J., and Kihara, D. (2013). Computational methods for constructing protein structure models from 3D electron microscopy maps. *J. Struct. Biol.* 184, 93–102. doi: 10.1016/j.jsb.2013.06.008
- Fabiola, F., and Chapman, M. S. (2005). Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure* 13, 389–400. doi: 10.1016/j.str.2005.01.007
- Ferber, M., Kosinski, J., Ori, A., Rashid, U. J., Moreno-Morcillo, M., Simon, B., et al. (2016). Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nat. Methods* 13, 515–520. doi: 10.1038/nmeth.3838
- Fischer, N., Neumann, P., Konevega, A. L., Bock, L. V., Ficner, R., Rodnina, M. V., et al. (2015). Structure of the *E. coli* ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* 520, 567–570. doi: 10.1038/nature14275
- Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* 31, 303–319. doi: 10.1146/annurev.biophys.31.082901.134202
- Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J., and Nagai, K. (2016). Cryo-EM structure of the spliceosome immediately after branching. *Nature* 537, 197–201. doi: 10.1038/nature19316
- Gallego, G., and Yezzi, A. (2015). “A compact formula for the derivative of a 3-d rotation in exponential coordinates”. *J. Math. Imaging Vis.* 51, 378–384. doi: 10.1007/s10851-014-0528-x
- Geman, S., and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596
- Gerstein, M., Lesk, A. M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry* 33, 6739–6749. doi: 10.1021/bi00188a001
- Geyer, C. J. (1991). “Markov chain Monte Carlo maximum likelihood,” in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (Fairfax Station, VA: Interface Foundation of North America), 156–163.
- Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 8, 645–654. doi: 10.1038/nrm2208
- Habeck, M. (2011). Statistical mechanics analysis of sparse data. *J. Struct. Biol.* 173, 541–548. doi: 10.1016/j.jsb.2010.09.016
- Habeck, M. (2012). “Inferential structure determination from nmr data,” in *Bayesian Methods in Structural Bioinformatics* eds T. Hamelryck, K. Mardia, and J. Ferkinghoff-Borg (Berlin; Heidelberg: Springer), 287–311. doi: 10.1007/978-3-642-27225-7_12
- Habeck, M., Nilges, M., and Rieping, W. (2005a). Bayesian inference applied to macromolecular structure determination. *Phys. Rev. E* 72:031912. doi: 10.1103/PhysRevE.72.031912
- Habeck, M., Nilges, M., and Rieping, W. (2005b). Replica-exchange Monte Carlo scheme for Bayesian data analysis. *Phys. Rev. Lett.* 94, 0181051–0181054. doi: 10.1103/PhysRevLett.94.018105
- Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1756–1761. doi: 10.1073/pnas.0506412103
- Habenstein, B., Loquet, A., Hwang, S., Giller, K., Vasa, S. K., Becker, S., et al. (2015). Hybrid structure of the type 1 pilus of uropathogenic *Escherichia coli*. *Angew. Chem. Int. Ed. Engl.* 54, 11691–11695. doi: 10.1002/anie.201505065
- Hinsen, K., Reuter, N., Navaza, J., Stokes, D. L., and Lacapère, J. J. (2005). Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.* 88, 818–827. doi: 10.1529/biophysj.104.050716
- Jack, A., and Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Cryst. Sect. A* 34, 931–935. doi: 10.1107/S0567739478001904
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
- Jolley, C. C., Wells, S. A., Fromme, P., and Thorpe, M. F. (2008). Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophys. J.* 94, 1613–1621. doi: 10.1529/biophysj.107.115949
- Karaca, E., and Bonvin, A. M. (2013). Advances in integrative modeling of biomolecular complexes. *Methods* 59, 372–381. doi: 10.1016/j.ymeth.2012.12.004
- Khatyer, H., Myasnikov, A. G., Natchiar, S. K., and Klaholz, B. P. (2015). Structure of the human 80S ribosome. *Nature* 520, 640–645. doi: 10.1038/nature14427
- Knuth, K. H., Habeck, M., Malakar, N. K., Mubeen, A. M., and Placek, B. (2015). Bayesian evidence and model selection. *Digit. Signal Process.* 47, 50–67. doi: 10.1016/j.dsp.2015.06.012
- Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., et al. (2011). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39, D456–D464. doi: 10.1093/nar/gkq880
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York, NY: Springer.
- MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.
- Mechelke, M., and Habeck, M. (2012). Calibration of Boltzmann distribution priors in Bayesian data analysis. *Phys. Rev. E* 86:066705. doi: 10.1103/PhysRevE.86.066705
- Mechelke, M., and Habeck, M. (2014). Bayesian weighting of statistical potentials in NMR structure calculation. *PLoS ONE* 9:e100197. doi: 10.1371/journal.pone.0100197

- Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A., and Teller, E. (1957). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114
- Müller, C. W., Schlauderer, G. J., Reinstein, J., and Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 4, 147–156. doi: 10.1016/S0969-2126(96)00018-4
- Neal, R. M. (2010). “MCMC using hamiltonian dynamics,” in *The Handbook of Markov Chain Monte Carlo* eds S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Chapman & Hall/CRC Press), 113–162.
- Orlova, E. V., and Saibil, H. R. (2004). Structure determination of macromolecular assemblies by single-particle analysis of cryo-electron micrographs. *Curr. Opin. Struct. Biol.* 14, 584–590. doi: 10.1016/j.sbi.2004.08.004
- Orzechowski, M., and Tama, F. (2008). Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* 95, 5692–5705. doi: 10.1529/biophysj.108.139451
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084
- Plaschka, C., Larivière, L., Wenzel, L., Seitz, M., Hemann, M., Tegunov, D., et al. (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature* 518, 376–380. doi: 10.1038/nature14229
- Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B., Fenton, W. A., Horwich, A. L., et al. (2001). ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell* 107, 869–879. doi: 10.1016/S0092-8674(01)00617-1
- Rappilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* 173, 530–540. doi: 10.1016/j.jsb.2010.10.014
- Rauhut, R., Fabrizio, P., Dybkov, O., Hartmuth, K., Pena, V., Chari, A., et al. (2016). Molecular architecture of the *Saccharomyces cerevisiae* activated spliceosome. *Science* 353, 1399–1405. doi: 10.1126/science.aag1906
- Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science* 309, 303–306. doi: 10.1126/science.1110428
- Rieping, W., Nilges, M., and Habeck, M. (2008). ISD: a software package for Bayesian NMR structure calculation. *Bioinformatics* 24, 1104–1105. doi: 10.1093/bioinformatics/btn062
- Robinson, C. V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. *Nature* 450, 973–982. doi: 10.1038/nature06523
- Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10:e1001244. doi: 10.1371/journal.pbio.1001244
- Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. *Nature* 422, 216–225. doi: 10.1038/nature01513
- Saunders, M. G., and Voth, G. A. (2013). Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* 42, 73–93. doi: 10.1146/annurev-biophys-083012-130348
- Schröder, G. F. (2015). Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr. Opin. Struct. Biol.* 31, 20–27. doi: 10.1016/j.sbi.2015.02.016
- Schröder, G. F., Brunger, A. T., and Levitt, M. (2007). Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* 15, 1630–1641. doi: 10.1016/j.str.2007.09.021
- Shahid, S. A., Bardiaux, B., Franks, W. T., Krabben, L., Habeck, M., van Rossum, B. J., et al. (2012). Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. *Nat. Methods* 9, 1212–1217. doi: 10.1038/nmeth.2248
- Sivia, D. S. (2004). “Some thoughts on correlated noise,” in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23RD International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 707 (Melville, NY: AIP Publishing), 303–313. doi: 10.1063/1.1751374
- Swendsen, R. H., and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* 57, 2607–2609. doi: 10.1103/PhysRevLett.57.2607
- Tama, F., Miyashita, O., and Brooks, C. L. (2004). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.* 147, 315–326. doi: 10.1016/j.jsb.2004.03.002
- Tan, R. K., Devkota, B., and Harvey, S. C. (2008). YUP.SCX: coaxing atomic models into medium resolution electron density maps. *J. Struct. Biol.* 163, 163–174. doi: 10.1016/j.jsb.2008.05.001
- Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure* 16, 295–307. doi: 10.1016/j.str.2007.11.016
- Tozzini, V. (2005). Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15, 144–150. doi: 10.1016/j.sbi.2005.02.005
- Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683. doi: 10.1016/j.str.2008.03.005
- Vasishtan, D., and Topf, M. (2011). Scoring functions for cryoEM density fitting. *J. Struct. Biol.* 174, 333–343. doi: 10.1016/j.jsb.2011.01.012
- Villa, E., and Lasker, K. (2014). Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Curr. Opin. Struct. Biol.* 25, 118–125. doi: 10.1016/j.sbi.2014.04.001
- Wan, R., Yan, C., Bai, R., Huang, G., and Shi, Y. (2016). Structure of a yeast catalytic step I spliceosome at 3.4 resolution. *Science* 353, 895–904. doi: 10.1126/science.aag2235
- Ward, A. B., Sali, A., and Wilson, I. A. (2013). Biochemistry. Integrative structural biology. *Science* 339, 913–915. doi: 10.1126/science.1228565
- Whitford, P. C., Noel, J. K., Gosavi, S., Schug, A., Sanbonmatsu, K. Y., and Onuchic, J. N. (2009). An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* 75, 430–441. doi: 10.1002/prot.22253
- Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C., and Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* 349, 1182–1191. doi: 10.1126/science.aac7629
- Yan, S., Suiter, C. L., Hou, G., Zhang, H., and Polenova, T. (2013). Probing structure and dynamics of protein assemblies by magic angle spinning NMR spectroscopy. *Acc. Chem. Res.* 46, 2047–2058. doi: 10.1126/science.aac7629

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Habeck. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.