GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Computational methods for random tomography with applications to cryo-EM data

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm PEI
of Georg-August University School of Science (GAUSS)

vorgelegt von
Nima Vakili
aus Tehran, Iran

Göttingen 2021

Referent:
Prof. Dr. Michael Habeck
Universitätsklinikum Jena

Korreferent:
Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Stephan Waack
Institut für Informatik, Universität Göttingen

Prof. Dr. Carsten Damm
Institut für Informatik, Universität Göttingen

Prof. Dr. Florin Manea
Institut für Informatik, Universität Göttingen

Prof. Dr. Daniel Rudolf
Institut für Mathematische Stochastik, Universität Göttingen

Tag der Mündlichen Prüfung: 29.10.2021

I hereby declare that I have written this thesis independently without any help from others and without the use of documents or aids other than those stated. I have mentioned all used sources and cited them correctly according to established academic citation rules.

Göttingen 2021

Abstract

Over the last decade, Cryo Electron Microscopy (cryo-EM) has emerged as a powerful and reliable technique to determine the three-dimensional (3D) structure of large macromolecular assemblies at near atomic resolution. Tens of thousands of two-dimensional (2D) noisy images from copies of a macromolecule are captured at different unknown orientations to characterize the 3D volume of biological complexes. The reconstruction problem is typically formulated as a nonlinear, non-convex optimization task and multi-modality causes many of the conventional reconstruction techniques that work based on local optimization to become trapped in local modes with poor initialization. These difficulties necessitate the development of adequate statistical models to describe the whole reconstruction process.

At the core of all algorithms proposed in this thesis, we use grid-free coarse grained representation of molecular densities rather than their atomic models in real space. To this end, we employ radial basis functions with spherical Gaussian kernels. Each Gaussian kernel is centered on a particle characterized by a 3D position and a non-negative weight.

The contribution of this dissertation can be divided into three major parts. The first contribution is introducing a new ab initio reconstruction method for estimating an initial model in cryo-EM based on a mixture of spherical Gaussian model. Here, we consider the full reconstruction problem when structure as well as orientations are unknown. We adopt Markov Chain Monte Carlo algorithms within a Bayesian framework to estimate the parameters of the model. Sampling from the posterior distribution is challenging due to the high-dimensionality and multi-modality. We address these difficulties by using Hamiltonian Monte Carlo and a global rotational sampling approach.

As we mentioned earlier, there are two major challenges in the reconstruction problems, unknown structure and unknown projection directions. If either of two quantities is known, then the problem boils down into two simpler tasks, which we address as two sub-problems in this thesis.

The first sub-problem is the tomographic reconstruction problem that occurs when in cryo-EM reconstruction, the projection directions are known. The second contribution of the thesis is the development of three kernel-based algorithms to characterize the 3D volume from 2D tomographic images in real space. In our first method, the particle positions cab be located at a fixed grid, such as a hexagonal grid and only updates their weights by using an iterative non-negative least-squares algorithm. The other two algorithms are

mesh-free approaches in which for the first method, we use Expectation Maximization to find the weighted particle positions and for the second method, we develop a fully Bayesian framework by assigning equally weighted particles.

The second sub-problem is a rigid registration (pose estimation) problem, and it happens when in cryo-EM reconstruction, the structure is known. The last contribution of the thesis is dedicated to find a unifying framework for assessing the match between biomolecular structures. We propose a kernel-correlation to compute the rigid-transformation between two complexes. We use an upper bound method (sometimes augmented by a deterministic annealing or a global search strategy), to solve the kernel-correlation.

Finally, we quantify our proposed approaches on various simulated and real datasets.

This thesis is based on the following publications and manuscripts, respectively:

- Vakili N, Habeck M. Bayesian Random Tomography of Particle Systems. Frontiers in molecular biosciences. 2021 May 21;8:399.
- Vakili N, Habeck M. Kernel-based tomographic reconstruction on and off the grid, in preparation.
- Vakili N, Habeck M. Matching biomolecular structures by registration of point clouds, in preparation.

Acknowledgment

First and foremost, I would like to express my sincere gratitude to my main supervisor, Michael Habeck. I'm truly proud of and grateful for my time working with Michael. I appreciate all his contributions of time, helpful feedback, his great ideas, his patience, his guidance and, of course, funding to make my PhD experience stimulating and productive. I learned a lot from you, Michael.

Also, I greatly appreciate my second supervisor, Prof. Munk, for his great feedback and his excellent advice and his support.

I would like to thank my thesis committee, Prof. Waach, Prof. Damm, Prof. Manea and Prof. Rudolf, for their great feedback and agreeing to be on my committee.

I gratefully acknowledge the funding sources DFG that made my PhD work possible.

Deepest thanks to my great team mates, Thach, Simeon, Paul, Katha and Christian. I'm grateful for every moment spent with you at work and in Göttingen.

I was very fortunate to meet so many wonderful people and make friends with them during my doctoral studies. Babak (you both Nasour and Nashour), Mohsen, Sayedeh, Pari, Amir, Leila, and my buddy Milan, because of you guys, I feel at home, here in Göttingen.

My lovely parents, Ladan, Davood and my dearest sister Newsha deserve special thanks for their continued support, endless love and countless sacrifices to help me get to this point.

Lastly, my beautiful girlfriend, Newsha, deserves endless gratitude for her cheerful support and her invaluable care. I am truly grateful for every moment we spend together.

Contents

Chapter 1

Introduction

1.1 Overview of single particle cryo-electron microscopy (cryo-EM)

1.1.1 General remarks

Understanding how a protein behaves as an individual protein or as a complex which interacts with other proteins or with membranes plays a key role in medical treatments especially in processing of drug development. Determining the 3D atomic structure of biomolecular assemblies from 2D images (acquired from different techniques such as X-ray crystallography, NMR spectroscopy) can ease the research development and lead to discover more efficient drugs. Over the last few years, cryo-electron microscopy (cryo-EM) has established itself as an alternative powerful technique to characterize the structure of large macromolecular assemblies (Frank, 2006; Cheng and Walz, 2009). Although development of the technique initiated in the 1970s, due to recent advances in imaging hardware, image processing software and improvements in sample preparation, cryo-EM nowadays is able to determine the 3D structure of macromolecular assemblies at near-atomic resolution. Moreover, the ability to work with non-crystallized particles and large molecules staying at their native hydrated state, distinguishes cryo-EM from standard structure determination techniques like X-ray crystallography or nuclear magnetic resonance (NMR) and renders cryo-EM a viable tool for studying macromolecular structures. In cryo-EM, many copies of the desired specimen are absorbed onto a thin continuous carbon grid (Thompson *et al.*, 2016). To prevent the formation of ice crystals, the grid is rapidly frozen

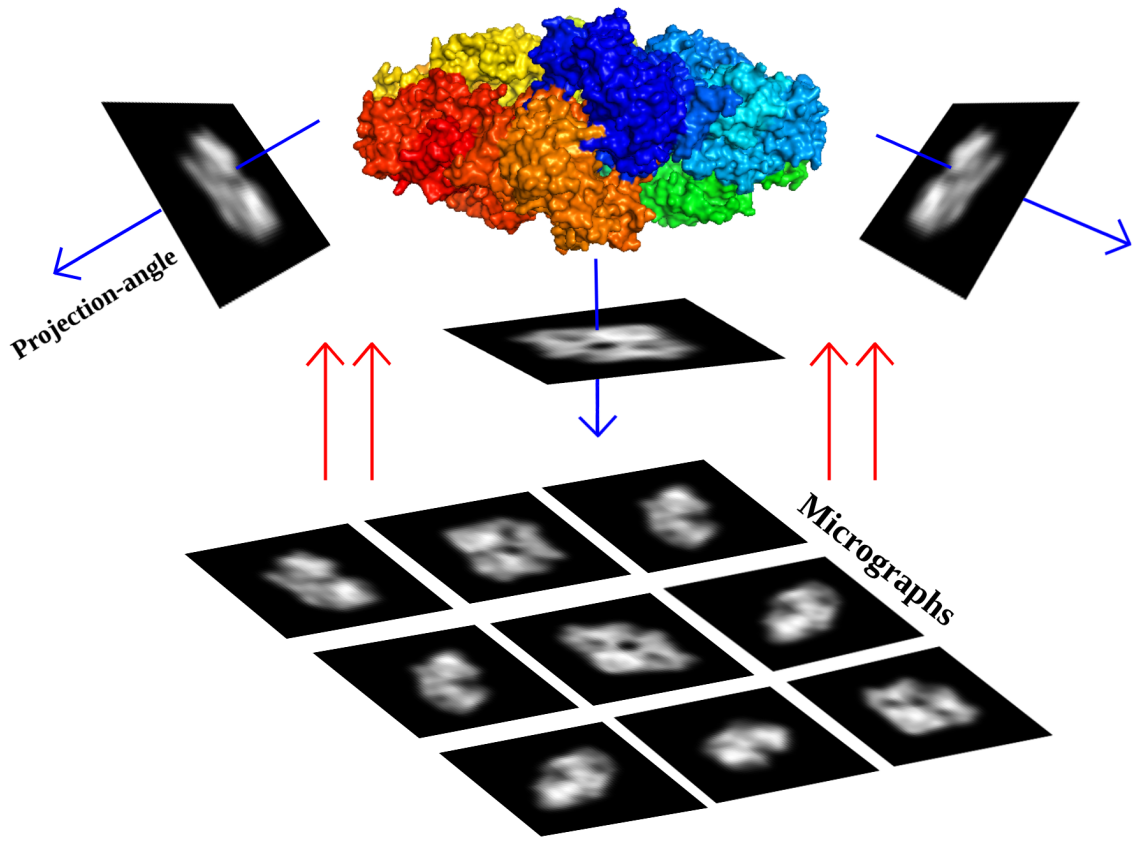


Figure 1.1: Schematic of single-particle cryo-EM reconstruction, from data collection to 3D volume generation of beta-galactosidase sample. While the blue arrows depict the forward model, the red ones show the backward model or the reconstruction problem. Characterizing the 3D volume from 2D projection images which involves the main contribution of my thesis.

at a rate of $106°C/s$ by plunging it into a cryogen (mostly liquid ethane or propane). The frozen randomly orientated particles are imaged by electrons to generate thousands of projection images of the specimen at random, unknown directions (called micrographs). The crucial part in cryo-EM is to reconstruct the shape of biomolecular assemblies from its two-dimensional projection images, and what makes it hard is that the relative orientation of these particle images are unknown. Fig1.1 depicts a quick schematic of generating 2D images from a known density map (called forward model and indicated by blue arrows) versus reconstructing a 3D volumetric shape of a density map from 2D projection images (called backward model indicated by red arrows).

1.1.2 2D image processing pipeline

In cryo-EM, the 2D projection images are convolved with the point spread function (PSF) of the microscope and corrupted by noise with very low signal-to-noise ratio (SNR). The PSF describes how aberrations of the lenses in a transmission electron microscope modulate the recorded image. Before a 3D structure can be estimated, these large micrograph images must be pre-processed to select the subsections of the images that correspond to the projections of the individual macromolecules. This process is called particle picking(Nicholson and Glaeser, 2001; Zhu *et al.*, 2004; Roseman, 2004). Particle images then should be clustered based on their common projection directions (Scheres *et al.*, 2005) and finally, all the particle images in each cluster are aligned to remove in-plane rotations and translations among cluster members. The preprocessing algorithm, in the end, results in smaller set of around 10 to 100 class averages with higher SNR. Fig 1.2 illustrates the diverse steps in cryo-EM data processing pipeline, starting from acquired micrographs to inferring an initial 3D structure.

As we stated, 2D projection images are modified by the contrast transfer function (CTF) of the microscope (The CTF is the Fourier transform of the PSF). The CTF is an oscillatory function and the frequency of the oscillations depends on the spherical aberration constant and the defocus. These oscillations affect contrast changes and spectrum amplitudes. An additional envelope attenuates high-resolution frequencies. To obtain the best result from the acquired images, estimation of the CTF[1] is indispensable (Rohou and Grigorieff, 2015).

The image formation model describes how 2D projection images are generated from macromolecular complexes. Mathematically, image formation can be formulated as,

$$g_n(\boldsymbol{u}) = h_n \star \int f(\boldsymbol{R_n^T r})\mathrm{dz} + "noise", \quad n = 1, \ldots, N \tag{1.1}$$

where $f(\boldsymbol{r})$ is a 3D volume for $\boldsymbol{r} \in \mathbb{R}^3$ and $f : \mathbb{R}^3 \mapsto \mathbb{R}_+$, $\boldsymbol{u} \in \mathbb{R}^2$ denotes a position in n-th projection image, $\boldsymbol{r}$ is a position in the volume, $\boldsymbol{R_n} \in SO(3)$ is a 3D rotation matrix and finally, h_n denotes a PSF that convoluted with the z-direction projected rotated molecule. The term "noise" is supposed to be i.i.d. Gaussian (sometimes, a Poisson distribution) noise which leads to a least-squares solution.
The goal of cryo-EM reconstruction methods is to invert this process and the crucial step is to reconstruct a volumetric representation of the assembly from the 2D images

[1] So far, we only use the class averages with a high SNR, but our model is able to be adopted easily with the CTF term.

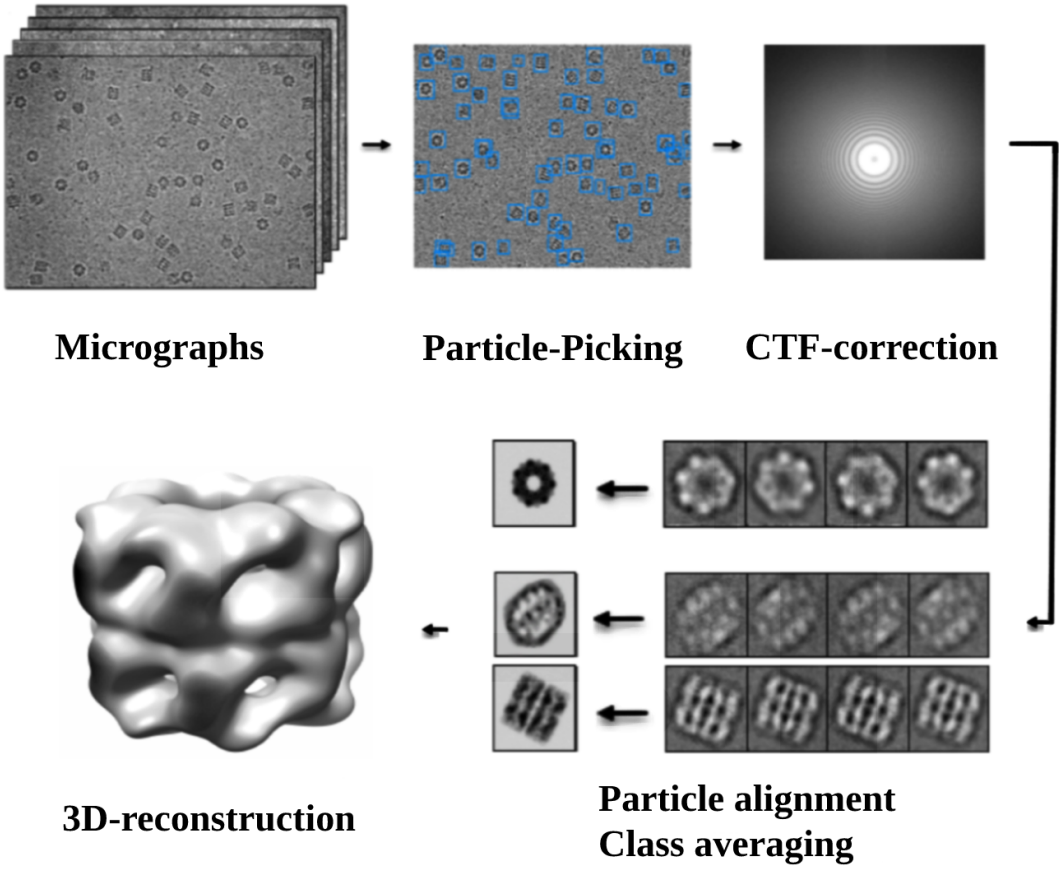


Figure 1.2: Schematic of single-particle cryo-EM reconstruction, from data collection to 3D volume generation. The sample of interest is imaged tens of thousands of times by collecting movie frames that are aligned for motion-correction. The acquired frames are then averaged to gain higher resolution. Next step, individual particles are picked from defocused CTF corrected images. To obtain class averages, particles are extracted, cleaned and normalized, and then subjected to 2D classification and averaging. Class averages with improved SNR are used to build an ab initio 3D structure.

i.e., estimating $f(\boldsymbol{r})$ given $g_1, \ldots, g_N$ and $h_1, \ldots, h_N$. Lack of information regarding the relative orientations of all particles ($\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n$) and existence of noise in projection images (to prevent radiation damage), make the reconstitution task prone to over-fitting which mathematically can be classified as an ill-posed problem. In chapter 2 we will show how to tackle the over-fitting problem by introducing a regularizer, which is called a prior probability, in a Bayesian framework.

1.1.3 3D reconstruction

The problem of determining the relative orientations of 2D projections of a 3D structure can be approached using the so-called common-lines technique, also known as angular reconstitution technique, which is based on the projection-slice theorem. The theorem states that, in Fourier space, the 2D Fourier transform of a 2D projection of a 3D object forms a central section of the object's 3D Fourier transform. This technique has been applied several times by (Penczek *et al.*, 1996; Van Heel, 1987; Elmlund *et al.*, 2008; Singer and Shkolnisky, 2011). To speed up the search process, the Single-particle IMage Processing Linux Engine (SIMPLE) software (Elmlund and Elmlund, 2012) uses bijection orientation search and other optimization techniques, such as simulated annealing and differential evolution optimizers. Reconstruction methods based on common-lines mostly suffer from either time-consuming calculations, noise sensitivity, initial model bias or even suffering from how to model and report the errors. These kinds of difficulties necessitate the development of statistical models to describe the data generation process and the 3D structures.

Ab initio reconstruction methods do not rely on an input reference volume for approximating orientations. Most of the statistical methods for ab inition reconstruction that exist to date have used a likelihood formulation to characterize the unknown 3D reconstruction. Statistical approaches include maximum-likelihood methods (Scheres *et al.*, 2007; Scheres, 2010), maximum a posteriori (MAP) methods (Jaitly *et al.*, 2010; Scheres, 2012), regularized maximum likelihood (RELION) in a Bayesian framework with assumption that the Fourier components of the signal are also independent, zero-mean and Gaussian distributed with unknown and resolution-dependent variance (Scheres, 2012), Probabilistic Initial 3D Model (PRIME) which scores the orientations by introducing weighted factors proportional to the correlation (Elmlund *et al.*, 2013), Bayesian method augmented by deterministic annealing by (Jaitly *et al.*, 2010), and more recently a Bayesian approach using stochastic gradient descent (SGD) with branch and bound maximum-likelihood optimization steps (Punjani *et al.*, 2017).

Almost all of the above discussed methods adopt iterative optimization algorithms to reconstruct the unknown volume. However, using local optimization techniques critically relies on the quality of initialization, which is in conflict with the idea of having an unbiased ab initio reconstruction. Moreover, these methods are not able to report any uncertainty quantification which can help better understanding of the result. Furthermore, there exist many ad hoc parameters in most of these methods which need to be set by the

user.

To address these difficulties, we propose a fully Bayesian method in real space by imposing a physically realistic prior over model parameters. Our method is able to quantify the uncertainties of the particle models. Moreover, it does not depend on the initial parameter settings. At the center of our approach is a grid-free coarse grained representation of the molecular complexes rather than their atomic models (called kernel representation). We use radial basis functions with spherical Gaussian kernels. Each Gaussian kernel can be interpreted as a weighted (non-negative) particle, such that the unknown volume is modeled by a particle cloud (more details can be found in section 1.2). By introducing the kernel representation of the unknown volume, the 3D reconstruction problem boils down to finding point cloud configurations and rotations that match the cryo-EM images. To this end, if our forward model is considered as,

$$g_{nm} = \alpha_n + \gamma_n \sum_k \phi_2(\boldsymbol{u}_{nm}; \boldsymbol{P}\boldsymbol{R}_{\boldsymbol{n}}\boldsymbol{x}_{\boldsymbol{k}} + \boldsymbol{t}_{\boldsymbol{n}}, \sigma^2) + "noise" \tag{1.2}$$

by assuming i.i.d Gaussian noise with variance $\frac{1}{\tau_n}$, we can build the likelihood function of our model as,

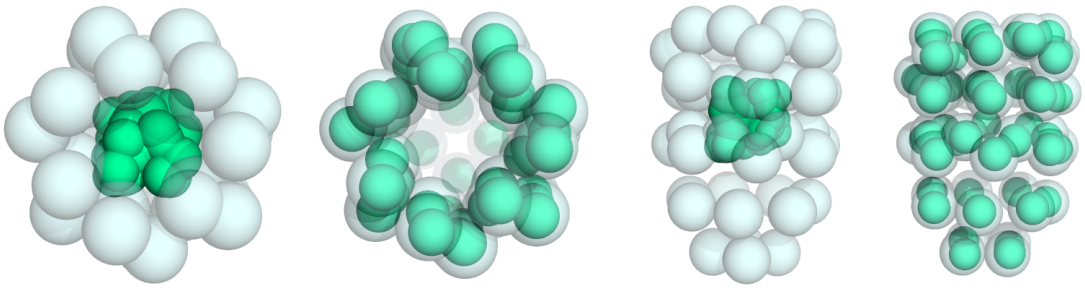
$$\Pr(g_n|\boldsymbol{x},\boldsymbol{R}_n,\boldsymbol{t}_n,\gamma_n,\alpha_n,\tau_n)=\left(\tfrac{\tau_n}{2\pi}\right)^{M_n/2}\exp\left\{-\tfrac{\tau_n}{2}\sum_{m=1}^{M_n}\left[g_{nm}-\alpha_n-\gamma_n\sum_k\phi_2(\boldsymbol{u}_{nm};\boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k+\boldsymbol{t}_n,\sigma^2)\right]^2\right\} \tag{1.3}$$

where $\tau_n > 0$ denotes the precision of the image, $\boldsymbol{x} = \{\boldsymbol{x}_k; k = 1, \ldots, K\}$ are particle positions, $\boldsymbol{P}$ is the 2×3 projection matrix, α_n and γ_n are an offset and a scaling factor, $\boldsymbol{t}_n$ and $\boldsymbol{R}_n$ are shifts and rotations respectively, the intensities are $g_{nm} = g_n(u_{nm})$ for $n = 1, \ldots, N$ images and $m = 1, \ldots, M_n$ is the number of pixels in the n-th image and finally, ϕ_2 is the Gaussian kernel that we adopt for the volume representation (more information can be found in section 1.2).
By denoting all 2D projection images by the symbol D and incorporating the prior beliefs $\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi)$ about our model where, $\xi = \{(\alpha_n, \gamma_n, \tau_n, \boldsymbol{t}_n; n = 1, \ldots, N)\}$, we derive the posterior distribution based on Bayes' law,

$$\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi|D) = \frac{\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi)}{\Pr(D)} \simeq \Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi) \tag{1.4}$$

In previous work, Joubert and Habeck (2015) used a Bayesian formulation to obtain a posterior distribution over the 3D volume (expressed as a mixture of Gaussians (more

details can be found in 1.2)) and the unknown rotations. The component means of the Gaussian mixture are interpreted as positions of coarse-grained particles that represent the unknown atomic structure at a lower resolution. With the help of conjugate priors and introducing latent assignment variables, they could derive analytical updates for Gibbs sampler that infers the particle positions and rotations. The Gibbs sampler bears a high resemblance with standard Gibbs sampling used in Gaussian mixture model estimation. A serious drawback, however, is that Gibbs sampling suffers from slow convergence and depends strongly on the initial conditions. Therefore, to locate the posterior mode, many restarts of the Gibbs sampler from varying initial conditions were necessary. Another limitation is that the likelihood is restricted to a Poisson model, otherwise the Gibbs updates are no longer valid. However, the Poisson model is limited in that, ignores the effect of the point spread function and correlations in the noise. Another problem is that the prior over the component means (particle positions) must be conjugate, i.e. a Gaussian distribution, which again is unrealistic for particle clouds representing biomolecular structures, because excluded-volume effects are ignored. Later in chapter 2, we overcome



(a) initial (b) estimated (c) initial (d) estimated

Figure 1.3: Results for estimating the particle positions from GroEL/GroES dataset with correct rotations. The transparent spheres show the true coarse-grained representation of GroEL/GroES that was used to generate the projection data. Green solid spheres indicate the particle positions in the initial and final stage of HMC sampling. Panels **(a,c)** show side and top views of the inital particle configuration; panels **(b,d)** show the particle configuration after 100 steps of HMC.

these limitations by developing a more general probabilistic model for particle systems and their projection images. We no longer aim to develop analytical updates for the Gibbs sampler, but explore the use of Markov chain Monte Carlo (MCMC) algorithms to infer both the particle positions as well as the unknown rotations. Drawing conformations of the particle system for fixed rotations can be reached with Hamiltonian Monte Carlo (HMC)(Neal, 2011) (Fig1.3). Inferring the rotations is more challenging because of the

multi-modality of the conditional posterior over the rotational parameters. To sample the rotations, we use a Metropolis-Hastings algorithm that explores the unit quaternions parameterizing the unknown projection directions. Since Metropolis-Hastings samples a probability distribution only locally, we occasionally run a global sampling step that is computationally more expensive. We quantify our proposal on diverse simulated and real experimental data sets and results show reliability of the method in 3D reconstructions. In this section, we considered the full 3D reconstruction problem in which both the 3D volume as well as projection directions are unknown. If either of the two quantities, structure or orientations, are known, then the reconstruction problem can boil down into two sub-problem which are easier to tackle. The first sub-problems is the tomographic reconstruction problem. Consider the case in which, for given orientations (for example a tilt series in tomography), the goal is to reconstruct the 3D structure. More details can be found in section 1.3 and chapter 3. The second sub-problem is a rigid registration problem (also known as pose estimation). This problem can be categorized either by 3D to 2D pose estimation task, when we search for a rigid transformation that matches the cryo-EM image and the projected 3D structure, or by 3D to 3D superimposing task, when we aim to assess the match between two given structures. More details can be found through the section 1.4 and chapter 4.

1.2 Kernel representation of images and volumes

Usually, both the input images as well as the unknown 3D volume are represented on grids: a pixel grid in case of 2D images, a voxel grid in case of the reconstructed 3D structure. Here, we investigate the use of grid-free radial basis function (RBF) kernel for representing the 2D projection data as well as the 3D structure.

$$f(\boldsymbol{r}) = \sum_{k=1}^{K} w_k \, \phi(\|\boldsymbol{r} - \boldsymbol{x}_k\|) \tag{1.5}$$

where K is the number of basis functions, w_k a weight (if $w_k > 0$), $\boldsymbol{x}_k \in \mathbb{R}^3$ a position vector that specifies the center of the k-th kernel and finally, $\|\cdot\|$ is the Euclidean norm. The kernel representation has been widely used in Machine Learning, computer vision, image processing and medical imaging (Schölkopf *et al.*, 2002; Takeda *et al.*, 2007; Zhang *et al.*, 2012; Rabbani *et al.*, 2006; Rusu *et al.*, 2008; Jin *et al.*, 2014; Jonić and Sorzano, 2015; Nogales-Cadenas *et al.*, 2013).

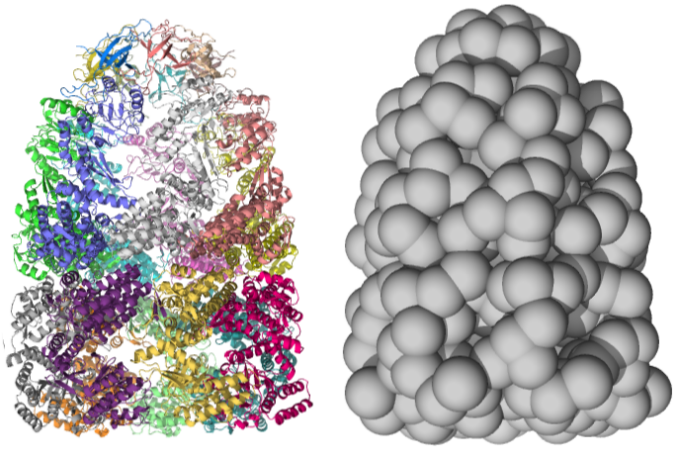


Figure 1.4: Representation of GroEL/GroES by using pseudo-atomic particles (point cloud). (left) atomic structure (PDB code 1aon).(right) The coarse-grained models 300 particles.

We propose to use a mixture of spherical Gaussian kernels to represent the 3D volume rather than a voxel grid. This is a rational choice to simulate the unknown density volumes with a smooth and blobby appearance. Moreover, using the RBF kernel can significantly reduce the amount of computations by employing far fewer parameters than voxel based algorithms require. Each Gaussian kernel is centered on a particle characterized by a 3D position and a non-negative weight. This is equivalent to using a grid-free particle system or coarse-grained representation of the unknown structures rather that an atomic-resolution model. Fig 1.4 depicts the pseudo-atomic representation of the GroEL/GroES density with 300 particles.

There are many options for constructing the kernels, but in this thesis we use d-dimensional spherical Gaussian RBF kernels.

$$\phi_d(\boldsymbol{r}, \boldsymbol{x}, \sigma^2) := \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \exp\{-\frac{1}{2\sigma^2} \|\boldsymbol{r} - \boldsymbol{x}\|^2\} \tag{1.6}$$

where $\sigma > 0$ is the bandwidth of the kernel. For $\sigma \to 0$, the kernel collapses to an atomic measure and for $\sigma > 0$, the kernel representation is particularly suited for smooth densities such as molecular densities that do not exhibit sharp edges.
As a short notice, we propose two different likelihoods for our model in chapter 2. The first likelihood model works directly with 2D images as inputs, while for the second likelihood model, images should be first converted to the 2D point-cloud and then be used

as inputs. To convert 2D grayscale images to weighted point clouds, we can either use the Expectation Maximization algorithm to fit a finite mixture of 2D Gaussians with a pre-chosen number of components to an image or, one can use the DP-means algorithm (Kulis and Jordan, 2012) as an alternative, which requires choosing a resolution like parameter specifying how many pixels should be represented by a single point.

1.3 Computational methods for tomographic reconstruction

In section 1.1.2, we outlined the typical pipeline used in cryo-EM for reconstructing a high-resolution 3D structure from projection images where the particle orientations are unknown. In this section, as a sub-problem, we consider the simpler problem where the particle orientations are given. This imaging modality is relevant to standard tomography where a sample is imaged from different known projection directions.

Electron Tomography (ET) is a widely used technique for determining the 3D structure of cellular samples from nanometer scale to atomic-resolution (Frank, 2006; Herman, 2009; Midgley and Dunin-Borkowski, 2009; Saghi and Midgley, 2012; Leary *et al.*, 2012; Miao *et al.*, 2016; Tian *et al.*, 2020). Generally, in tomography a 3D model of an object is characterized by a collection of 2D projection images (called tilt-series). These noisy images are acquired by rotating the specimen at a limited range of angles, usually from -70° to $+70^\circ$. The missing wedge artifacts (refers to limits on the ET's range of specimen tilts) and the noisy images are two common problems in this field. The goal in tomography is to reconstruct a 3D structure $f(\boldsymbol{r})$ from the X-ray transform g,

$$g(\boldsymbol{u}) = \int f(\boldsymbol{R}^T \boldsymbol{r}) dz \tag{1.7}$$

in which $\boldsymbol{r} \in \mathbb{R}^3$ is a position, $\boldsymbol{u} \in \mathbb{R}^2$ is a position in the image plane, and $\boldsymbol{R} \in SO(3)$ is known rotation matrix.
The most widely used image reconstruction methods in tomography (Frank, 2008) are analytical methods such as Weighted Back Projection (WBP) (Radermacher, 2007) and Filtered Back Projection(FBP) (Kak and Slaney, 2001). Many of the conventional algorithms in tomographic reconstruction use the Fourier slice theorem in reciprocal space (Penczek *et al.*, 1996; Van Heel, 1987; Elmlund *et al.*, 2008). These methods mostly suffer from existing the missing wedges or lack of enough projection images. To overcome these difficulties, iterative reconstruction techniques were proposed. Generally, two types of

iterative reconstruction methods have been proposed: 1-Real space approaches such as Algebraic Reconstruction Technique (ART) (Gordon *et al.*, 1970) or its major refinement of it, the simultaneous ART (SART) (Anishinraj *et al.*, 2012) and the Simultaneous Iterative Reconstruction Techniques (SIRT) which suffers from filling the missing wedge (Trampert and Leveque, 1990). 2-Hybrid-space approaches, which iterate between real space and reciprocal space such as Equal Slope Tomography (EST) (Miao *et al.*, 2005) or GENeralized Fourier Iterative REconstruction (GENFIRE) (Pryor *et al.*, 2017). Very recently, (Pham *et al.*, 2020; Yang *et al.*, 2021) proposed REal Space Iterative Reconstruction Engine (RESIRE) algorithm which solves the least square problem by a gradient descent method combined with the Fourier slice theorem and variational calculus.

In chapter 3, we propose three kernel-based reconstruction methods in real space. All the three algorithms are based on underlying probabilistic models. Besides, we use the coarse-grained representation of the 3D structure as a collection of weighted spherical particles (more details can be found in 1.2).

In our first method, we consider the particle positions to be located on a hexagonal grid and only update their weights by employing an iterative non-negative least squares algorithm.

The second reconstruction method is based on Expectation Maximization algorithm similar to the one used in Gaussian mixture fitting. It involves a subsampling step in which a chosen number of pixels is randomly selected from the projection images. Our statistical model is,

$$\boldsymbol{y}_{nm} \sim \sum_{k=1}^{K} w_k \phi(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{r}_k, \sigma), \quad m = 1, \ldots, M \tag{1.8}$$

where $\boldsymbol{y}_{nm}$ is m-th pixel position of the n-th image, $\boldsymbol{r}_k$ is the k-th particle position, $\boldsymbol{P}$ is the projection matrix and $\boldsymbol{R} \in SO(3)$ is known rotation matrices.

And finally, our thirt proposal is a fully Bayesian inference technique which is solved by HMC method. The algorithm is similar to our Bayesian reconstruction approach, which discussed in chapter 2 but with known orientations. Incorporating an excluded volume term as a prior and the ability to determine the uncertainties in our Bayesian model, distinguish this method from the other two approaches.

We demonstrate our method by using simulated and real data taken from cryo-ET and scanning transmission electron microscopy (STEM) (Levin *et al.*, 2016) and the results depict reliability of our approaches in comparison with the other standard tomographic

reconstruction methods.

1.4 3D Registration problem

In this subsection, we discuss the second sub-problem of cryo-EM reconstruction, namely the estimation of projection directions when the 3D structure is known. This sub-problem is a rigid registration or pose estimation problem, which is a very common task in computer vision. It is a process of inferring the optimal transformation of a person or an object in an given image or video from a reference position. Because of its fundamental importance, it arises as a subtask in diverse applications, e.g., animation, robotics, gaming, medical image alignment, structural bioinformatics, etc. In this thesis, we consider two versions of the 3D rigid registration problem that are relevant to cryo-EM reconstruction and, more generally, the analysis of biomolecular structures. First, finding a rigid transformation from 3D to 2D that matches the cryo-EM image and the projected 3D structure. Second, rigid docking of biomolecular structures from 3D to 3D which is one of the basic application of pose estimation in bioinformatics. More details can be found in chapter 4.

1.4.1 Rigid registration of biomolecular structures

Increasing in the amount of data sets provided daily by Electron Microscopy/Electron Tomography necessitates to develop methods to compare and classify them. In this section, we propose a kernel correlation method to find the match between two biomolecular structures. Superimposing and comparing structures can play a key role in diverse applications, from increasing the resolution of cryo-EM by fitting atomic models into volume, to the study of conformational changes by comparing densities at different stages. (Volkmann and Hanein, 1999; Roseman, 2000; Villa and Lasker, 2014; Chacón and Wriggers, 2002; Topf and Sali, 2005).

Except for the manual docking strategy, which is limited by the experience of the user (Wriggers and Chacon, 2001; Volkmann and Hanein, 2003), automated approaches in superposition problems are typically based on search method. Search methods usually optimize six degrees of freedom, three parameters for the rotation and three translation parameters. The simplest search method is an exhaustive search, in which all the parameters are systematically searched using a given step size. The grid search is computationally expensive, but many efforts have been made to accelerate it, from employing the Fast

Fourier Transform (Roseman, 2000; Chacón and Wriggers, 2002) to stochastic sampling, such as Monte Carlo and simulated annealing methods (Topf *et al.*, 2005; Alber *et al.*, 2007). Garzón *et al.* (2007) used spherical harmonic combined with a translational scanning and Wriggers *et al.* (1998) proposed the vector quantization method which used Root-Mean-Square-Deviation (RMSD) as a scoring function. Moreover, some approaches use a correlation coefficient between a given density map and a low-resolution density map of atomic structures. A local cross-correlation measure was introduced to avoid the influence of non-overlapping density. Roseman (2000) uses computation over a local area, along with a local normalization or using key vectors that capture local surface information in (Ceulemans and Russell, 2004), comparisons of the cross-correlation function (CCF) to other metrics like those borrowing from machine learning techniques, enable systematic and objective evaluation of scoring functions (Vasishtan and Topf, 2011) and adding molecular contour information to the fitting criterion by (Chacón and Wriggers, 2002). A major disadvantage of cross-correlation maximization of either density or structure factor amplitude is the computational cost. GMfit (Kawabata, 2008) employs a representation of the density map in terms of a Gaussian Mixture Model (GMM), a linear combination of 3D anisotropic Gaussian distribution functions. A score based on the overlap of two Gaussian mixtures is optimized to find the match.

In chapter 4, we develop a unifying probabilistic framework to assess the match between biomolecular structures by applying ideas from computer vision. The structures, as we discussed earlier, are represented by the coarse-grained model (1.2) and compared by measuring the overlap between the particle clouds. We employ the kernel correlation to measure the overlap. We quantify our method in multiple examples, from protein structural alignment to comparison of circularly permuted structures and rigid modeling with EM maps.

Our strategy is to minimize the Gaussian kernel correlation (KC). The KC of two point clouds represented by position matrices $\vec{X}$ and $\vec{Y}$ and weight vectors $\vec{q}$ and $\vec{p}$, viewed as a function of the rigid transformation is defined as (Tsin and Kanade, 2004),

$$\mathrm{KC}(\boldsymbol{R},\boldsymbol{t}) = \sum_{i=1}^{M}\sum_{j=1}^{N} q_i\, p_j\, \phi(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|) = \boldsymbol{q}^T \boldsymbol{\Phi}(\boldsymbol{R},\boldsymbol{t})\boldsymbol{p} \tag{1.9}$$

where, $\|\cdot\|$ is the Euclidean norm and $\boldsymbol{\Phi}$ is the $M \times N$ kernel matrix with elements

$\phi_{ij} = \phi(\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t})$. Throughout this study, we use the Gaussian kernel,

$$\phi_\sigma(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\{-\tfrac{1}{2\sigma^2}\boldsymbol{x}^2\} \tag{1.10}$$

where, the positive parameter σ is the *bandwidth* of the kernel.

We develop an upper bound minimization technique to optimize the kernel correlation. This is a local minimization problem which is prone to be trapped in a local mode. To address this difficulty, we employ the deterministic annealing method or a global rotational/translational search technique.

In the next sub-section 1.4.2, we will briefly talk about the ICP approach, which is widely used in computer vision, especially in pose estimation. We then continue in sub-section 1.4.3 by explaining the quaternion representation, one of the common approaches to parameterize the orientations in rigid docking.

1.4.2 Iterative Closest Point (ICP) method

One of the common algorithm used in rigid registration is the Iterative Closest Point (ICP) method (Besl and McKay, 1992). In this section, we briefly describe the algorithm. Consider $\vec{X}$ and $\vec{Y}$ as two point clouds. We are looking for the rigid transformation with the best correspondence between these two clouds by solving the following L_2 minimization problem,

$$E(\boldsymbol{R}, \boldsymbol{t}) = \sum_{i=1}^{M} \left\| \boldsymbol{R}\boldsymbol{x_i} + \boldsymbol{t} - \boldsymbol{y_{j^*}} \right\|^2 \tag{1.11}$$

where $\boldsymbol{R} \in SO(3)$ is rotation, $\boldsymbol{t} \in R^3$ is translation, $\vec{X} = \{\boldsymbol{x_i}, i = 1, \ldots, M\}$ and $\vec{Y} = \{\boldsymbol{y_j}, j = 1, \ldots, N\}$. Finally,

$$j^* = \underset{j \in \{1,\ldots,N\}}{arg\,min} \left\| \boldsymbol{R}\boldsymbol{x_i} + \boldsymbol{t} - \boldsymbol{y_j} \right\| \tag{1.12}$$

By giving initial transformation R and t, the algorithm iteratively solve the above minimization problem. ICP suffers from local minima and its performance highly relies on the quality of initialization. $E(\boldsymbol{R}, \boldsymbol{t})$ can be minimized in closed form using a singular value decomposition.

1.4.3 Tessellation of $SO(3)$ using unit quaternions

Quaternion provides a convenient representation for rotations of object in 3D space. They offer a simple and compact representation of the rotational parameters in space. A quaternion can be represented as either a scalar plus a 3-component vector or a vector with four components or a complex number with 3 imaginary variables. A quaternion q is defined as,

$$q = q_0 + q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} \tag{1.13}$$

where q_0 is a scalar and $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are three imaginary numbers with the basic rules of,

$$\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1 \tag{1.14}$$

In the following, we briefly review some properties of the quaternion presentation.
If q be a quaternion, then the complex conjugate of q, is defined as,

$$q^* = q_0 - q_1\boldsymbol{i} - q_2\boldsymbol{j} - q_3\boldsymbol{k} \tag{1.15}$$

besides,

$$\begin{aligned} (q^*)^* &= q \\ q^*q &= qq^* \\ q + q^* &= 2q_0 \\ (pq)^* &= q^*p^* \end{aligned} \tag{1.16}$$

The norm of a quaternion q is denoted by $|q| = \sqrt{q^*q}$ and a quaternion is called a unit quaternion if the norm is equal to 1. The inverse of a quaternion is defined as $q^{-1} = \frac{q^*}{|q|^2}$ and its clear that, $q^{-1}q = qq^{-1} = 1$. And finally for unit quaternion, $q^{-1} = q^*$. We represent rotations using the unit quaternions in $\mathbb{S}^3$, the four-dimensional sphere. A great advantage using unit quaternions to represent rotations is that it leads to a clear idea of simulating the space of rotations $SO(3)$. We apply the idea of tessellating the rotational space from computer vision (Straub *et al.*, 2017) which uniformly approximates $SO(3)$. This can be achieved with the 600-cell, a tessellation of the unit sphere using 4D tetrahedra. To achieve a finer discretization, each tetrahedron can be split into eight tetrahedra whose corners again lie on the unit sphere. By projecting the center of a tetrahedron onto the unit sphere, we obtain a valid 3D rotation matrix. Using this method, we discretize the space or rotations in a way that can be refined to increasingly higher precision. At the coarsest level,

this construction results in 330 rotation matrices (only the upper half-sphere has to be considered due to symmetry arguments). A finer representation comprises $8 \times 330 = 2640$ rotations. As we stated earlier, a challenging task in cryo-EM reconstruction problems is to estimate the rotations. A brute force strategy for sampling the rotations is to choose a fixed resolution of discretized rotation space, evaluate the conditional posterior for all rotations, and sample a rotation from the discretized conditional posterior. This property provides us possibility of sampling the space of rotations in a more efficient and systematic way.

1.5 Computational approaches used in this thesis

Due to high dimensionality and multi-modality, solving the Bayesian model 1.4 that we introduced in section 1.1.3, is numerically intractable. We use MCMC algorithms to address these difficulties. In the following, we will concisely talk about MCMC algorithms that are used in this thesis.

1.5.1 Markov Chain Monte Carlo algorithms

In this section, we give a brief overview of Markov Chain Monte Carlo (MCMC) algorithms. More information can be found in (Bishop, 2006; Neal *et al.*, 2011).

The posterior distribution derived from Bayesian inference is quite complex. High dimensionality and existence of several local modes make solving the reconstruction problem intractable and we need to resort to some form of approximation. In this thesis, we use approximate inference methods based on numerical sampling techniques (Monte Carlo techniques). Markov Chain Monte Carlo is a powerful framework to sample from a target distribution. A Markov Chain is formed by a sequence of random variables $x^1, x^2, \ldots, x^T$ which meets the following conditional independence property,

$$p(x^{t+1}|x^1, \ldots, x^t) = p(x^{t+1}|x^t), \qquad for \quad t \in \{1, \ldots, T-1\} \tag{1.17}$$

$p(x^{t+1}|x^t)$ are called transition probabilities and a Markov Chain is homogeneous if the transition probabilities are the same for all t. A homogeneous Markov chain with transition

probabilities $\boldsymbol{T}$, the distribution $p^*(x)$ is invariant if,

$$p^*(x) = \sum_{x'} \boldsymbol{T}(x', x) p^*(x') \tag{1.18}$$

Our goal is to construct a Markov Chain for which the distribution we wish to sample from is invariant. We also require that the Markov Chain be ergodic, which means for $t \to \infty$, the distribution $p(x^t)$ converges to the required invariant distribution $p^*(x)$ regardless the choice of initial probabilities $p(x^0)$. An ergodic Markov Chain can only have on invariant distribution which is called equilibrium distribution.

1.5.2 The Metropolis-Hastings algorithm

The Metropolis-Hastings sampler is one of the most popular MCMC technique which used to draw samples from complex probability distributions.The algorithm works by proposing a new state x^* based on a previous state x^{t-1} according to the proposal distribution $q(x^*|x^{t-1})$. An overview of the Metropolis-Hasting sampler is presented in algorithm 1,

Algorithm 1 Metropolis Hastings sampler

Require: Define target density p and proposal density q
Require: Choose initial sample x^0
for $t = 1, \ldots, T$ **do**
  $x = x^{t-1}$
  $x' \sim q(.|x)$
  $c = min(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)})$
  $u \sim U[0, 1]$
  if $u <= c$ **then**
    $x^t = x'$
  else
    $x^t = x$
  end if
end for

1.5.3 The Gibbs sampler

The Gibbs sampler is another popular MCMC technique. If we take $p(\boldsymbol{x}) = p(x_1, \ldots, x_T)$ as the target distribution from which we want to sample, by choosing an initial state

for the Markov Chain, each step of the Gibbs algorithm is to replace the value of one variable by a value sampled from the distribution of that variable given on the values of the remaining variables. The algorithm 2 depicts an overview of the method.

Algorithm 2 Gibbs sampler

Require: Define target distribution p
Require: Choose initial sample $x^0 = (x_1^0, x_2^0, \ldots, x_d^0)$
  for $t = 1, \ldots, T$ **do**
    for $k = 1, \ldots, d$ **do**
      $x_k^t \sim p(x_k | x_1^t, \ldots, x_{k-1}^{t-1}, x_{k+1}^{t-1}, \ldots, x_d^{t-1})$
    end for
  end for

1.5.4 The Hamiltonian Monte Carlo algorithm

The Hamiltonian Monte Carlo (HMC) algorithm is another MCMC technique which is based on analogy with physical systems (Hamiltonian dynamics) to exploit samples more effectively. Using the gradient of the log probability distribution as well as the distribution itself, make the Hamiltonian dynamics method more efficient and faster compared to the other MCMC techniques. If we denote position $\boldsymbol{z} \in \mathbb{R}^d$ and momentum $\boldsymbol{r} \in \mathbb{R}^d$, the total energy of the system (Hamiltonian) is,

$$H(\boldsymbol{z}, \boldsymbol{r}) = U(\boldsymbol{z}) + K(\boldsymbol{r}) \tag{1.19}$$

where $U(\boldsymbol{z})$ is a potential term and $K(\boldsymbol{r})$ is a kinetic energy term and a common choice for that, is to use a zero-mean Gaussian distribution with unit variance, $K(\boldsymbol{r}) = \frac{\boldsymbol{r}^T \boldsymbol{r}}{2}$. To describe how the system changes in time, Hamiltonian dynamics are characterized by,

$$\frac{\partial z_i}{\partial t} = \frac{\partial H}{\partial r_i} \qquad \frac{\partial r_i}{\partial t} = -\frac{\partial H}{\partial z_i} \tag{1.20}$$

In order to simulate Hamiltonian dynamics numerically, it is necessary to approximate the Hamiltonian equations by discretizing time. This is done by splitting the interval T up into a series of smaller intervals of length ϵ. We use the Leapfrog algorithm to this end (see 3).

Algorithm 3 Leapfrog integrator

for $l = 1, \ldots, L$ **do**
$r_i(l + \frac{\epsilon}{2}) = r_i(l) - (\frac{\epsilon}{2})\frac{\partial U}{\partial z_i}(\boldsymbol{z}(l))$
$z_i(l + \epsilon) = z_i(l) + \epsilon r_i(l + \frac{\epsilon}{2})$
$r_i(l + \epsilon) = r_i(l + \frac{\epsilon}{2}) - (\frac{\epsilon}{2})\frac{\partial U}{\partial z_i}(\boldsymbol{z}(l + \epsilon))$
$l = l + \epsilon$
end for

Generally, the Hamiltonian dynamical method is an iterating between a series of leapfrog updates and a resampling of the momentum variables from their probability distributions. An overview of the HMC approach is presented in Algorithm 4.

Algorithm 4 HMC sampler

Generate an initial position $z^{(0)}$
for $t = 1, \ldots, T$ **do**
$t = t + 1$
$z_0 = z^{(t-1)}$
$r_0 \sim N(0, M)$
update $[z_0, r_0]$ through the leapfrog to obtain q^*,r^*
$\alpha = min(1, \exp(-U(z^*) + U(z_0) - K(r^*) + K(r_0)))$
$c \sim Unif(0, 1)$
if $c \leqslant \alpha$ **then**
$z^{(t)} = z^*$
else
$z^{(t)} = z^{(t-1)}$
end if
end for

1.6 Synopsis

This thesis comprises five chapters. In chapter 2, we introduce a fully Bayesian approach to reconstruct the 3D structure of macromolecular assemblies by using 2D projection images derived from Cryo-EM. There are two challenges in the reconstruction problem that we have to deal with, unknown structure and unknown projection directions. Knowing either of these two quantities results in simpler sub-problems that we tackle in chapters 3 and 4.

In chapter 3, we assume that the orientation of projected images are given. To characterize

the unknown structure, we propose 3 different kernel-based reconstruction methods which the first one considers the particle positions on a regular grid and only updates the weights but the other two methods are mesh-free approaches where the particle positions can move freely during the reconstruction.

In chapter 4, we assume that the structure is given, and the goal is to the find rigid transformation (pose estimation) between two structures from 3D to 2D or from 3D to 3D. To this end, we propose a mathematical expression based on kernel-correlation to find the match between biomolecular structures.

Finally, we conclude our efforts in chapter 5 by summarizing our achievement and by offering ideas for future research.

Chapter 2

Bayesian Random Tomography of Particle Systems

This chapter is the first research result from my Ph.D. study. Here we present a probabilistic model for 3D volume reconstruction of cryo-EM datasets. This chapter was published in Journal: Frontiers in molecular biosciences, 2021. Cited as: Vakili, N., & Habeck, M. (2021). Own contribution:

- Concept and implementation of the algorithm and the code.
- All figures, tables.
- Manuscript in parts.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 21 May 2021
doi: 10.3389/fmolb.2021.658269

Bayesian Random Tomography of Particle Systems

Nima Vakili[1] *and Michael Habeck*[1,2]*

[1]Microscopic Image Analysis Group, Jena University Hospital, Jena, Germany, [2]Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Random tomography is a common problem in imaging science and refers to the task of reconstructing a three-dimensional volume from two-dimensional projection images acquired in unknown random directions. We present a Bayesian approach to random tomography. At the center of our approach is a meshless representation of the unknown volume as a mixture of spherical Gaussians. Each Gaussian can be interpreted as a particle such that the unknown volume is represented by a particle cloud. The particle representation allows us to speed up the computation of projection images and to represent a large variety of structures accurately and efficiently. We develop Markov chain Monte Carlo algorithms to infer the particle positions as well as the unknown orientations. Posterior sampling is challenging due to the high dimensionality and multimodality of the posterior distribution. We tackle these challenges by using Hamiltonian Monte Carlo and a global rotational sampling strategy. We test the approach on various simulated and real datasets.

Keywords: 3D Reconstruction, random tomography, cryo-EM, bayesian inference, coarse-grained modeling, markov chain Monte Carlo, inferential structure determination

OPEN ACCESS

Edited by:
Edina Rosta,
King's College London,
United Kingdom

Reviewed by:
Takanori Nakane,
MRC Laboratory of Molecular Biology (LMB), United Kingdom
Slavica Jonic,
UMR7590 Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), France

***Correspondence:**
Michael Habeck
michael.habeck@uni-jena.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *25 January 2021*
Accepted: *26 April 2021*
Published: *21 May 2021*

Citation:
Vakili N and Habeck M (2021) Bayesian Random Tomography of Particle Systems. Front. Mol. Biosci. 8:658269. doi: 10.3389/fmolb.2021.658269

1 INTRODUCTION

Many different imaging techniques acquire two-dimensional (2D) projection data of an unknown three-dimensional (3D) object. If the projection directions are known, tomographic reconstruction methods can be used to recover the 3D structure of the object (Natterer, 2001). An additional complication arises, if the projection directions are unknown. This imaging modality is of particular relevance to single-particle cryo-electron microscopy (cryo-EM). In recent years, cryo-EM has emerged as a powerful technique to determine the structure of large biomolecular assemblies at near atomic resolution (Frank, 2006). In cryo-EM, many copies of the particle of interest are first applied to a carbon grid and then plunge-frozen to prevent the formation of ice crystals. The frozen randomly orientated particles are imaged with electrons resulting in thousands to millions of noisy projection images. Similar reconstruction problems arise in cryo-electron tomography as well as single-particle diffraction experiments at free-electron lasers (von Ardenne et al., 2018). A completely different field of application is *in situ* microscopy of various specimens such as mesoscopic organisms (Levis et al., 2018).

The reconstruction problem common to all of these imaging methods is to recover a 3D volume from 2D images acquired in random projection directions and has been termed random tomography (Panaretos, 2009). Since the projection directions are unknown, we have to estimate them in the course of the reconstruction. Moreover, to avoid model bias, the desired reconstruction method should not rely on an initial guess of the volume (ab initio reconstruction).

Various ab initio reconstruction methods have been proposed (Bendory et al., 2020) including maximum likelihood via expectation maximization (Scheres et al., 2007) and maximum a posteriori (MAP) estimation (Jaitly et al., 2010; Scheres, 2010, 2012a), regularized maximum likelihood (Scheres, 2012b), stochastic gradient descent (Punjani et al., 2017), common lines (Vainshtein and Goncharov, 1986; Van Heel, 1987; Penczek et al., 1996; Elmlund et al., 2008; Singer and Shkolnisky, 2011; Elmlund and Elmlund, 2012; Lyumkis et al., 2013), the method of moments (Kam, 1980; Levin et al., 2018), random-model methods (Yan et al., 2007; Sanz-Garcia et al., 2010), methods using stochastic hill climbing (Elmlund et al., 2013) or nonlinear dimensionality reduction (Vargas et al., 2014) and frequency marching (Barnett et al., 2017).

These approaches typically reconstruct the unknown volume by solving an optimization problem. However, optimization approaches do not offer any uncertainty quantification. Another drawback is that many reconstruction algorithms are iterative procedures that critically depend on the initialization, which counteracts the idea of achieving an unbiased ab initio reconstruction. Moreover, most algorithms employ a number of ad hoc parameters that need to be tuned by the user and impact the final result in a way that is not always obvious.

Our goal is to develop a fully Bayesian approach to 3D reconstruction using a meaningful model of the unknown structure (including a physically realistic prior) and utilizing sampling algorithms for parameter estimation and uncertainty quantification. In our previous work (Joubert and Habeck, 2015), we already took the first step towards this goal. We considered the reconstruction problem in random tomography as a density estimation problem utilizing a mixture of Gaussians. With the help of conjugate priors and the introduction of latent assignment variables, we could derive analytical updates for a Gibbs sampler that infers the unknown rotations and component means.

However, there are various problems with our previous Gibbs sampling approach. First, Gibbs sampling suffers from slow convergence and depends strongly on the initial conditions. Therefore, to locate the posterior mode many restarts of the Gibbs sampler from varying initial conditions are necessary. Second, our Gibbs sampling algorithm is restricted to a Poissonian likelihood. The Poisson model is limited in that it ignores the effect of the point spread function and correlations in the noise. Third, the prior over the component means (particle positions) is chosen to be a conjugate, zero-centered Gaussian distribution, which is not realistic for biomolecular structures, because it ignores excluded-volume effects.

Here, we overcome these limitations by developing a more general probabilistic model for particle systems and their projection images. We no longer aim to develop analytical updates for the Gibbs sampler, but use of Markov chain Monte Carlo (MCMC) algorithms to infer both the particle positions as well as the unknown rotations. Sampling conformations of the particle system for fixed rotations can be achieved with Hamiltonian Monte Carlo (HMC). To sample the rotations, we use a Metropolis-Hastings algorithm that explores the unit quaternions parameterizing the unknown projection directions. Since Metropolis-Hastings samples a probability distribution only locally, we occasionally run a global sampling step that is computationally more expensive. Using simulated and real experimental data, we demonstrate that our Bayesian approach to random tomography is capable of estimating physically plausible coarse-grained models.

2 PROBABILISTIC MODEL AND POSTERIOR SAMPLING

We aim to reconstruct a 3D volume $f(\boldsymbol{r})$ for $\boldsymbol{r} \in \mathbb{R}^3$ and $f : \mathbb{R}^3 \mapsto \mathbb{R}_+$. We do not observe $f(\boldsymbol{r})$ directly but only projection images

$$g(\boldsymbol{u}) = \int f(\boldsymbol{R}^T \boldsymbol{r})\, \mathrm{d}z = \int f(\boldsymbol{\theta}^\perp \boldsymbol{u} + \boldsymbol{\theta} z)\, \mathrm{d}z =: \mathcal{X}_{\boldsymbol{\theta}}[f](\boldsymbol{u}) \tag{1}$$

where $\boldsymbol{R} \in SO(3)$ is a 3D rotation matrix whose last row $\boldsymbol{\theta} \in \mathbb{R}^3$ is a unit vector pointing into the projection direction, and $\boldsymbol{\theta}^\perp \in \mathbb{R}^{3\times 2}$ is the matrix whose columns span the plane orthogonal to $\boldsymbol{\theta}$ such that $\boldsymbol{R}^T = [\boldsymbol{\theta}^\perp, \boldsymbol{\theta}]$. Throughout this article, $\boldsymbol{u} \in \mathbb{R}^2$ denotes a position in the projection image, and $\boldsymbol{r} \in \mathbb{R}^3$ a position in the volume. The integral transform $\mathcal{X}_{\boldsymbol{R}}[f]$ (**Eq. 1**) is known as the X-ray transform or John transform (Natterer, 2001). In 2D, the X-ray transform is identical to the Radon transform. The reconstruction problem in random tomography is to estimate $f(\boldsymbol{r})$ from N random projection directions $\boldsymbol{\theta}_n$, or equivalently $\boldsymbol{R}_n$, such that

$$g_n(\boldsymbol{u}) = \mathcal{X}_{\boldsymbol{\theta}_n}[f](\boldsymbol{u}) + n(\boldsymbol{u}), \quad n = 1, \ldots, N \tag{2}$$

where $n(\boldsymbol{u})$ is the noise.

2.1 Kernel Expansion of Images and Volumes

The standard discretization of images and volumes is based on pixels and voxels placed on regular 2D and 3D grids. Instead, we expand images and volumes into sums of basis functions that can be centered at irregular positions (as in meshless methods). We use a radial basis function (RBF) kernel ϕ such that the kernel expansion of the volume becomes

$$f(\boldsymbol{r}) = \sum_{k=1}^{K} w_k \boldsymbol{\phi}(\boldsymbol{r} - \boldsymbol{x}_k) \tag{3}$$

where K is the number of basis functions, $\|\cdot\|$ is the Euclidean norm, w_k a coefficient or weight (if $w_k > 0$) and $\boldsymbol{x}_k \in \mathbb{R}^3$ a position vector that determines the center of the kth kernel. We can represent members of a reproducing kernel Hilbert space using this expansion. RBF representations are widely used in machine learning (Schölkopf and Smola, 2002), image processing (Takeda et al., 2007) and numerical applications (Schaback and Wendland, 2006).

A physical interpretation of the kernel representation is that we model the object as a collection of K particles at positions $\boldsymbol{x}_k$ with mass $w_k > 0$. The model **(3)** can then be interpreted as the blurred version of a particle system:

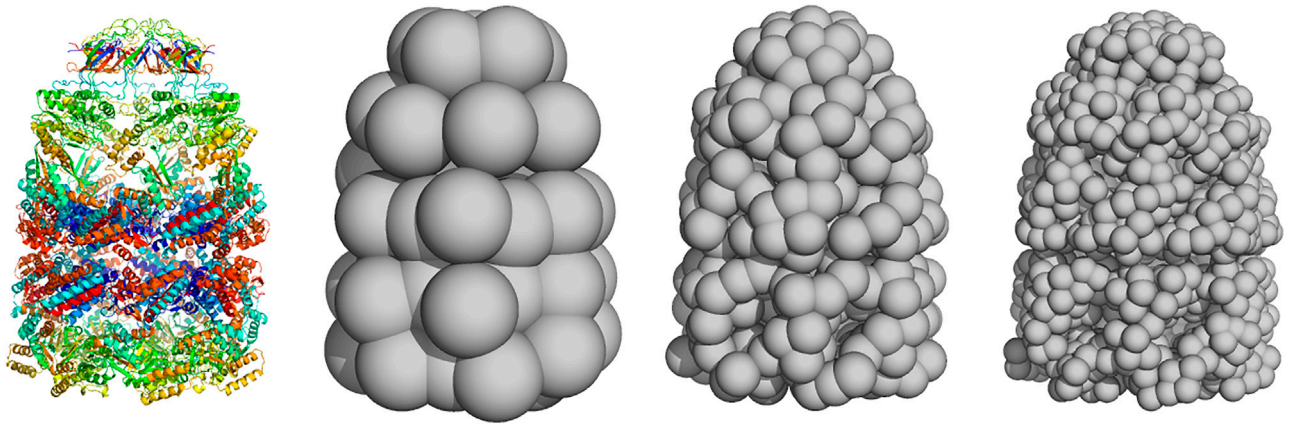


FIGURE 1 | Coarse-grained representation of GroEL/GroES using a varying number of particles **(left)** atomic structure (PDB code 1aon). Panels **(middle left)** to **(right)** show coarse-grained models using K = 50, 300, and 1000 particles.

$$f(\boldsymbol{r}) = \left(\phi * \sum_{k=1}^{K} w_k \delta_{\boldsymbol{x}_k} \right)(\boldsymbol{r}) \tag{4}$$

where $\delta_{\boldsymbol{x}_k}$ is the delta function centered a $\boldsymbol{x}_k$ and the particle density, $\sum_k w_k \delta_{\boldsymbol{x}_k}$, is blurred by a convolution (denoted by $*$) with the RBF kernel. The particle locations and weights $\{(\boldsymbol{x}_k, w_k); k = 1, \ldots, K\}$ can also be viewed as a weighted point cloud. The component means $\boldsymbol{x}_k$ could be fixed to a regular 3D grid. But we will consider particle systems that are not tied to a grid and can be distributed in an irregular fashion (similar to meshless or meshfree methods used in numerical analysis). Typically, the particle system is a coarse-grained representation of the unknown structure rather than an atomic-resolution representation. Therefore, 3D reconstruction from 2D projection data provides a pseudo-atomic representation whose resolution depends on the number of particles K (**Figure 1** for an illustration).

One motivation for our choice of the volume representation (**Eq. 3**) are its efficient transformation properties. Rigid transformations of $f(\boldsymbol{r})$ involve a shift by the translation vector $\boldsymbol{t}$ and a reorientation brought about by the rotation matrix $\boldsymbol{R}$. Under the RBF expansion these transformations reduce to rigid transformations of the particle positions:

$$f(\boldsymbol{r}) \overset{R,t}{\mapsto} f\left(\boldsymbol{R}^T(\boldsymbol{r} - \boldsymbol{t})\right) = \sum_k w_k \, \phi(\boldsymbol{r} - \boldsymbol{R}\boldsymbol{x}_k - \boldsymbol{t}) = \sum_k w_k \, \phi(\boldsymbol{r} - \boldsymbol{x}'_k) \tag{5}$$

where $\boldsymbol{x}'_k = \boldsymbol{R}\boldsymbol{x}_k + \boldsymbol{t}$.

There are many options for $\phi(r)$. We will restrict ourselves to Gaussian RBF kernels. The d-dimensional spherical Gaussian is defined by

$$\phi_d(\boldsymbol{r}; \boldsymbol{x}, \sigma^2) := \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left\| \boldsymbol{r} - \boldsymbol{x} \right\|^2 \right\} \tag{6}$$

where $\sigma > 0$ is the bandwidth of the kernel. The volume representation that we will use throughout this paper is a mixture of K spherical Gaussians:

$$f(\boldsymbol{r}) = \sum_{k=1}^{K} w_k \, \phi_3(\boldsymbol{r}; \boldsymbol{x}_k, \sigma^2) \tag{7}$$

This representation is very common in statistics, in particular in density estimation where $\boldsymbol{x}_k$ are observed samples resulting in a kernel density estimate of an unknown probability density function. Indeed, our original motivation (Joubert and Habeck, 2015) to choose this representation of $f(\boldsymbol{r})$ was mainly driven by viewing 3D reconstruction from random projections as an instance of a density estimation problem. Other examples for uses of (Gaussian) particle representations in cryo-EM data analysis such as denoising or the analysis of continuous conformational changes have been proposed by Jin et al. (2014); Jonić et al. (2016); Jonić and Sorzano (2016).

A convenient property of the spherical Gaussian kernel is its behavior under the X-ray transform (**Eq. 1**):

$$\mathcal{X}_{\boldsymbol{\theta}}[\phi_d](\boldsymbol{u}) = \int \phi_d(\boldsymbol{\theta}^\perp \boldsymbol{u} + \boldsymbol{\theta} z; \boldsymbol{x}, \sigma^2) \, \mathrm{d}z = \phi_{d-1}(\boldsymbol{u}; \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}, \sigma^2) \tag{8}$$

where again $\boldsymbol{R} = [\boldsymbol{\theta}^\perp, \boldsymbol{\theta}]^T \in SO(3)$ and the 2×3 projection matrix $\boldsymbol{P}$ is

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{9}$$

Spherical Gaussians are closed under the X-ray transform, and the projected volume **(7)** is again a K component mixture of spherical Gaussians

$$\mathcal{X}_{\boldsymbol{\theta}}[f](\boldsymbol{u}) = \sum_{k=1}^{K} w_k \, \phi_2(\boldsymbol{u}; \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}_k, \sigma^2) \tag{10}$$

with centers $\boldsymbol{x}'_k = \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}_k \in \mathbb{R}^2$. This fact motivates us to also represent the input images as mixtures of spherical Gaussians in 2D (see ***Representation of Projection Images by Point Clouds*** for a concrete application).

2.2 Probabilistic Model

The unknown parameters of our model are the particle positions $\boldsymbol{x}_k$ and weights w_k as well as the unknown rotation matrices $\boldsymbol{R}_n$. Since we interpret the Gaussian components as particles of equal mass, we fix the weights: $w_k = K^{-1}$, such that the main inference parameters are $\boldsymbol{x}_k$ and $\boldsymbol{R}_n$.

2.2.1 Likelihoods

We tested two probabilistic models for the input data. The first model uses the input images $\{g_n; n = 1, \ldots, N\}$ directly. For each image, the intensities are $g_{nm} = g_n(\boldsymbol{u}_{nm})$ at pixel positions $\boldsymbol{u}_{nm}$ where $m = 1, \ldots, M_n$ with M_n being the number of pixels in the nth image. Typically, the number of pixels M_n is identical for all projection images.

A simple image model is to assume pixelwise identically and independently distributed Gaussian noise in the image formation **(2)**, such that the likelihood of the nth image is

$$\Pr\left(g_n \middle| \boldsymbol{x}, \boldsymbol{R}_n, \boldsymbol{t}_n, \gamma_n, \alpha_n, \tau_n\right) = \left(\frac{\tau_n}{2\pi}\right)^{M_n/2} \exp\left\{ -\frac{\tau_n}{2} \sum_{m=1}^{M_n} \left[g_{nm} - \alpha_n - \gamma_n \sum_k \phi_2\left(\boldsymbol{u}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma^2\right)\right]^2 \right\} \tag{11}$$

where $\tau_n > 0$ is the precision of the image, and α_n, γ_n are an offset and a scaling factor (the constant weight $w_k = 1/K$ has been absorbed by the scaling factor γ_n). The two-dimensional translation $\boldsymbol{t}_n$ accounts for a shift of the image. These three to five nuisance parameters per image (depending on whether shifts $\boldsymbol{t}_n$ are fitted or not) need to be estimated in addition to the particle positions $\boldsymbol{x} = \{\boldsymbol{x}_k; k = 1, \ldots, K\}$ and the rotations $\boldsymbol{R} = \{\boldsymbol{R}_n; n = 1, \ldots, N\}$. Model **(11)** is an idealized image formation model. It ignores important effects such as the CTF or correlated noise that are highly relevant for cryo-EM applications.

The second model also uses a kernel expansion of the input image motivated by the fact that ideally, according to our image model, the projection image should also be a mixture of spherical Gaussians (**Eq. 10**). In a preprocessing step, we fit a point cloud $Y_n = \{\boldsymbol{y}_{nm} \in \mathbb{R}^2; m = 1, \ldots, M_n\}$ to the nth input image g_n such that

$$g_n(\boldsymbol{u}) \approx \alpha_n + \gamma_n \sum_{m=1}^{M_n} \phi_2\left(\boldsymbol{u}; y_{nm}, \sigma_n^2\right) \tag{12}$$

Typically, we choose $M_n = M$ but this is not a requirement. Again, model **(12)** does not account for the CTF or other important effects in cryo-EM image formation. In each projection direction, the 2D point cloud can be blurred to a different degree captured by the width σ_n. The Supplementary Material details how projection images can be converted to point clouds; ***Representation of projection images by point clouds*** in Results shows a practical example for further illustration.

As in Joubert and Habeck (2015), we model the 2D point clouds as samples from the projected 3D volume:

$$\Pr\left(Y_n | \boldsymbol{x}, \boldsymbol{R}_n, \boldsymbol{t}_n, \sigma_n\right) = \prod_{m=1}^{M_n} \frac{1}{K} \sum_{k=1}^{K} \phi_2\left(\boldsymbol{y}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma_n^2\right) \tag{13}$$

In the following, we will denote all nuisance parameters, i.e. all parameters except particle positions and rotations, collectively by ξ. In case of the image likelihood **(11)**, we have $\xi = \{(\alpha_n, \gamma_n, \tau_n, \boldsymbol{t}_n); n = 1, \ldots, N\}$. In case of the point cloud likelihood (**Eq. 13**), we have $\xi = \{(\sigma_n, \boldsymbol{t}_n); n = 1, \ldots, N\}$. Moreover, we will denote both likelihoods as $\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)$ where D are the data (projection images or 2D point clouds).

2.2.2 Priors

After incorporating our prior beliefs about the model parameters, we are able to derive the posterior distribution by invoking Bayes' theorem:

$$\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi | D) = \frac{\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi)}{\Pr(D)} \tag{14}$$

where $\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi)$ is the prior which we assume to factor into

$$\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi) = \Pr(\boldsymbol{x})\Pr(\boldsymbol{R})\Pr(\xi) \tag{15}$$

The normalization factor $\Pr(D)$ is the model evidence, which can be ignored if we are only interested in parameter estimation.

We use standard priors for the nuisance parameters: Jeffreys priors for precisions τ_n and $1/\sigma_n^2$. The prior for the scaling factors and offsets are flat. Note that these priors are improper (i.e., not normalizable). Since we are only interested in parameter estimation, this does not pose a problem. The priors for the scaling factor and offset could be improved. For example, cryo-EM images are often normalized such that the mean intensity is zero and the standard deviation is one. It is possible to express this information as a prior on the offset and scaling factor. The Supplementary Material provides more details about these priors. For the image shifts $\boldsymbol{t}_n$, a zero-centered two-dimensional Gaussian distribution is a reasonable choice.

Typically, biomolecules orient themselves randomly in the ice layer that is imaged by cryo-EM. Therefore, we choose a uniform distribution over $SO(3)$:

$$\Pr(\boldsymbol{R}) = \prod_{n=1}^{N} \Pr(\boldsymbol{R}_n) \propto 1 \tag{16}$$

These priors are proper, because the rotation group is compact.

In our previous work (Joubert and Habeck, 2015), we used a zero-centered Gaussian prior for all particle positions $\boldsymbol{x}_k$ to ensure that prior and likelihood are conjugate, which enabled the derivation of closed-form updates for the component means. However, this prior is very unrealistic, if we think of the Gaussian basis functions as massive particles that should not occupy the same region in space (excluded volume), but rather repel each other. Since the packing of biomolecular structures is reminiscent of fluids (Liang and Dill, 2001), the prior should favor particle configurations that show similar packing characteristics. To model repulsive interactions between particles, we use a

Boltzmann distribution over the positions $\boldsymbol{x}_k$ involving a soft repulsive interaction potential $E(\boldsymbol{x})$:

$$\Pr(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \propto \exp\{-\beta E(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)\} \tag{17}$$

Furthermore, the particles are confined to a box with soft boundaries (Habeck, 2017). Pairs of particles repel each other if the distance is smaller than the particle diameter $2R$ where R is the effective particle radius. We choose a quartic repulsion which is commonly used in NMR structure calculation:

$$E(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) = \sum_{k<k'} \left[\left|\left|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\right|\right| \le 2R\right] \left(1 - \frac{\left\|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\right\|}{2R}\right)^4 \tag{18}$$

where $[\cdot]$ is the Iverson bracket. Given the total number of atoms L of the system, the particle radius can be predicted for a desired number of particles K by using the relation

$$R \approx 0.92\ (L/K)^{0.42}\ \text{Å}. \tag{19}$$

Using a configurational temperature estimator (Mechelke and Habeck, 2013), the inverse temperature is estimated to $\beta \approx 175$. The estimates for R and β are based on an analysis of several biomolecular structures at different levels of coarse graining. See Supplementary Material for details.

Since the excluded-volume term (**Eq. 18**) is purely repulsive, we add a radius of gyration term such that the overall prior for particle positions is

$$\Pr(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \propto \exp\{-\beta E(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)\} \exp\{-\alpha R_g(\boldsymbol{x})\} \tag{20}$$

where $R_g(\boldsymbol{x})$ is the radius of gyration of the coarse-grained structure $\boldsymbol{x}$ and α a positive constant. The radius of gyration term imposes a weak preference for compact structures and prevents configurations with isolated particles that do not contact another particle. In our experiments, we set $\alpha = 10$ Å; in principle, we could estimate α by using techniques similar to those used in the estimation of β. But since α does not have a strong impact on the final structure, we restricted ourselves to a single fixed value for α.

2.3 Inference

Bayesian random tomography employs MCMC sampling from the posterior distribution **(14)**. We use a Gibbs sampling strategy (Geman and Geman, 1984) where each group of parameters, the particle positions $\boldsymbol{x}$, the rotations $\boldsymbol{R}$ and the nuisance parameters ξ, is updated separately while clamping the other parameters to their current values. To update the nuisance parameters, we use standard samplers for generating Gamma variates and normally distributed random variables (more details can be found in the Supplementary Material). However, the conditional posteriors of the particle positions $\boldsymbol{x}$ and the rotations $\boldsymbol{R}$ are not of a standard form and need to be updated with more sophisticated algorithms.

2.3.1 Sampling Particle Positions With Hamiltonian Monte Carlo

To sample the particle positions, we use Hamiltonian Monte Carlo (HMC) (Neal, 2011). The conditional posterior distribution over particle positions is

$$\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D) \propto \Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi) \Pr(\boldsymbol{x})$$

In HMC, $-\log\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D)$ defines a potential energy over configuration space that is composed of an attractive term $-\log\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)$ matching particle positions to the projection data, and a repulsive contribution $-\log\Pr(\boldsymbol{x})$ stemming from the excluded-volume term **(18)**. For fixed rotations and nuisance parameters, the particle positions undergo Hamiltonian dynamics following the gradient of $-\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D)$ during a short leapfrog integration. The resulting configuration is accepted or rejected according to the Metropolis criterion.

2.3.2 Sampling Rotational Parameters With Metropolis-Hastings

A challenging problem is to estimate the rotations. Because the projection images are statistically independent of each other, the problem decomposes into N subproblems:

$$\Pr(\boldsymbol{R}_n|\boldsymbol{x}, \xi, D) \propto \exp\left\{-\frac{\tau_n}{2} \sum_{m=1}^{M_n} \left[g_{nm} - \alpha_n - \gamma_n \sum_{k=1}^{K} \phi_2\left(\boldsymbol{u}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma^2\right)\right]^2\right\} \tag{21}$$

if projection images g_n are fitted directly, or

$$\Pr(\boldsymbol{R}_n|\boldsymbol{x}, \xi, D) \propto \prod_{m=1}^{M_n} \sum_{k=1}^{K} \phi_2\left(\boldsymbol{y}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma_n^2\right) \tag{22}$$

if we fit 2D point clouds. In Joubert and Habeck (2015), we introduced assignment variables such that the conditional posterior **(22)** is replaced by the matrix von Mises-Fisher distribution, which can be simulated in a straightforward fashion (Habeck, 2009). However, because the assignment variables are highly coupled to the other parameters, this strategy converges only slowly to the next local minimum. Moreover, there is no flexibility regarding the likelihood function.

We use the Metropolis-Hastings (MH) algorithm (Liu, 2001) to estimate the rotation matrices. We parameterize rotation matrices using unit quaternions (Horn, 1987) and propose new quaternions by adding a random perturbation that is sampled from a uniform distribution. We run 10 MH steps to update the quaternions representing each projection direction in every Gibbs sampling iteration and adapt the step-size automatically: Upon acceptance, the step-size increases by multiplying it with a factor of 1.02; in case of rejection, the step-sizes decreases by a factor of 0.98. This rule results in an acceptance rate of approximately 50%. We use this sampling algorithm to simulate both types of conditional posteriors **(21)** and **(22)**.

2.3.3 Global Sampling of Rotational Parameters

Since the MH algorithm achieves only local sampling of probability distributions, we occasionally scan all rotations systematically. The unit quaternions are elements of the 3-sphere, the unit sphere embedded in the four-dimensional

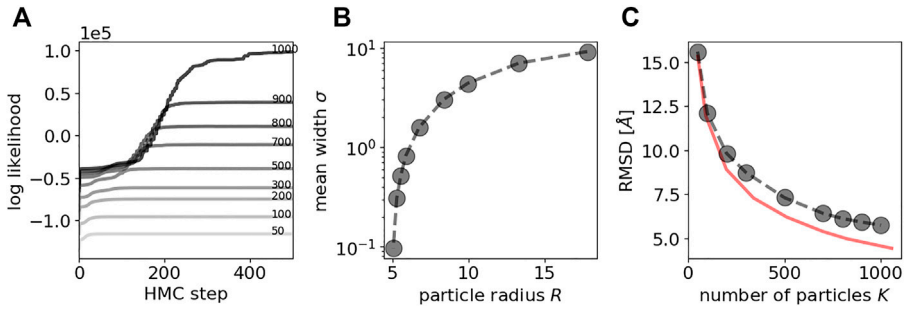


FIGURE 2 | HMC sampling of particle positions with fixed rotations for a simulated data set of GroEL/ES. A Evolution of the log likelihood during HMC sampling. The larger the number of particles K, the higher is the final log likelihood. Increasing darkness indicates larger number of particles. Line annotations also indicate the number of particles. B Average standard deviation (computed over all 35 input point clouds) vs. the size of the particle R. C RMSD between Carbon-alpha positions of the crystal structure and the coarse-grained models inferred with HMC. As a reference, the RMSD between the Carbon-alpha positions and the coarse-grained versions of the crystal structures is shown as red curve.

space. To evenly cover rotation space, we discretize the 3-sphere using the 600-cell (Coxeter, 1973). The 600-cell is composed of even sized tetrahedra whose corners lie on the unit sphere. By projecting the center of a tetrahedron onto the unit sphere we obtain a unit quaternion parameterizing a valid rotation matrix. Due to the degeneracy of the quaternions we only have to consider the upper half of the 4D sphere that is covered by 330 tetrahedra at the coarsest level of discretization. To obtain a finer tessellation of $SO(3)$, we can split each tetrahedron into eight tetrahedra whose corners again lie on the 4D unit sphere. By default, we use a frequency of 0.1 to run a global rotation scan. The conditional posterior is evaluated for all rotations and then sampled from the discrete distribution.

The source code and scripts for reproducing the tests are available at github.com/michaelhabeck/bayesian-random-tomography.

3 RESULTS

3.1 Sampling Tests

To test MCMC strategies for inferring particle positions and rotations, we use the structure of the GroEL/GroES complex. This system has been studied extensively with cryo-EM. Since our focus is mainly on algorithmic aspects, we first use simulated data that exactly follow our probabilistic model. To generate input point clouds in 2D, we use the crystal structure of GroEL/GroES (PDB code 1aon; 58,674 atom coordinates in total). The 2D point clouds are generated by projecting the 3D positions of every 10th Carbon-alpha atom (802 points in total) along 35 random directions into 2D. We also generated corresponding projection images by blurring the point clouds with a Gaussian filter of width 5 Å.

3.1.1 Sampling Particle Positions and Precisions With Fixed Rotations

We first studied the performance of sampling particle positions by fixing the rotations to the correct values and sampling only the particle positions and the precisions of the projection data. HMC sampling of particle positions started from a random initial configuration for K ranging between 50 and 1,000 particles. In all of our HMC experiments, the number of leapfrog steps was set to 10, whereas the step-size was adjusted automatically. The precisions $1/\sigma_n^2$ follow Gamma distributions and can be sampled directly.

Figure 2A shows the evolution of the log likelihood achieved by the particle system during HMC. After roughly 200 to 500 HMC steps (depending on K), the particle cloud reproduces the input data well, which is reflected in high values of the log likelihood. The sampled particle configurations are very similar to the true structure at the same level of coarse graining. Successful sampling of $\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D)$ with HMC is observed reliably for many different initial particle configurations.

It is clear that an increasing number of particles K results in a higher goodness of fit, which is obvious from **Figures 2A,B** showing the average standard deviation σ_n of the point cloud likelihood (**Eq. 13**) as a function of particle radius: A higher number of particles K results in more flexible models that result in a better goodness of fit and higher precision. These findings indicate that HMC is highly suited to sample particle configurations.

Figure 2C shows the accuracy of the coarse-grained models inferred from the projection data with HMC. The accuracy is quantified by the root mean square deviation (RMSD) between corresponding positions in a reference structure and a coarse-grained model. Here, our reference structure is the atomic structure of GroEL/ES reduced to the positions of 8,015 Carbon-alpha atoms listed in the PDB entry 1aon. To compare this structure with a coarse-grained model, positions in the atomic structure are assigned to positions in the coarse-grained model that are closest in 3D space. There are two factors that contribute to this measure of accuracy: the level of coarse graining as well as the performance of posterior sampling based on the 2D projection data. To disentangle both contributions, we also show the accuracy between the crystal structure and its coarse-grained versions (obtained with the DP-means algorithm by Kulis and Jordan (2012); also see the Supplementary Material).

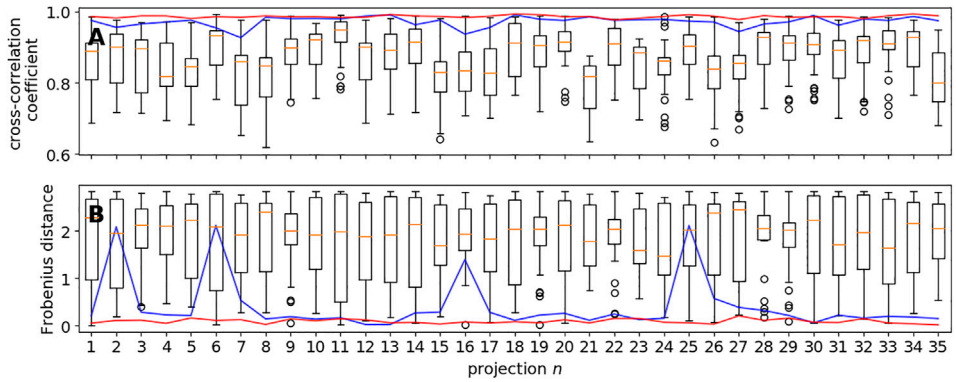


FIGURE 3 | Global vs. local sampling of orientational parameters. Shown are the cross-correlation coefficients **(panel A)** and Frobenius distances **(panel B)** for each of the 35 input directions achieved with local sampling based on the MH algorithm and global sampling using a regular discretization of the 3-hemisphere. The blue curve shows the results obtained with the coarsest covering based on 330 unit quaternions; the red curve shows the results obtained with a finer covering (2,460 quaternions). The box plots show the variability within 30 trials of MH starting from random rotations.

This curve shows that coarse-grained models of GroEL/ES using 1,000 particles achieve an accuracy of about 4.6 Å, whereas an ultra coarse-grained model based on only 50 particles is on average 15.5 Å away from any Carbon-alpha atom in the crystal structure. For very high levels of coarse graining (small K), the models inferred with HMC reach the maximum accuracy that is possible at this level of coarse graining. With increasing number of particles K, the gap in accuracy widens but is still similar to the maximum attainable value. For example, with K = 1000 the model obtained with HMC achieves an RMSD of 5.7 Å, whereas the coarse grained model obtained directly from the crystal structure achieves an accuracy of 4.6 Å.

If we estimate particle configurations from projection images instead of point clouds, we obtain similar results. **Supplementary Figure S4** shows the log likelihood and cross-correlation coefficients obtained with different numbers of particles, again ranging between 50 and 1,000. The evolution of the log likelihood indicates that the HMC sampler seems to converge even faster compared to a simulation based on point cloud data: within 20–150 HMC steps the log likelihood plateaus. The accuracy of the structure after 500 HMC steps is similar to or better than the accuracy of the particle models fitted against 2D point clouds and almost reaches the accuracy of the coarse-grained models derived from the crystal structure. **Supplementary Figure S5** shows FSC curves for all 3D models. For the same number of particles, the FSC curves are similar with a slight preference for the image-based models when using larger numbers of particles. The resolution ranges from 12.2 Å (50 particles) to 4.5 Å (1,000 particles). **Supplementary Table S1** shows resolution estimates for all models.

3.1.2 Sampling Rotational Parameters and Precisions With Fixed Particle Positions

To test our rotational sampling approach, we fixed the particle positions to an ultra coarse-grained structure (K = 200) of GroEL/ES. Although each rotation can be updated independently of the other rotations, and each conditional posterior (given either by **Eqs. 21** or **22**) is only a four-dimensional probability distribution over the quaternions, the sampling problem is still challenging due to its multimodality. Since Metropolis-Hastings (MH) is a local sampling algorithm, it tends to become trapped in subordinate modes of the conditional posterior, which are typical for rigid registration problems. As a result, running MH on the conditional posteriors is not sufficient to reliably recover the rotation matrices.

Figure 3A shows the cross-correlation coefficients for the 35 projection images obtained with global rotational sampling in comparison with MH runs starting from 30 random rotations. Global rotational sampling was based on the first two discretizations of the 3-hemisphere using 330 and 2,640 quaternions, respectively. The number of local sampling attempts was set to 30 so as to match the speed of global sampling at the finer level. That is, the coarse sampling based on 330 quaternions is approximately 8 times faster than the 30 local sampling trials. As evidenced by **Figure 3A**, global sampling is capable of finding rotation matrices that yield high cross-correlation coefficients, whereas MH alone fails to do so in a systematic fashion. **Figure 3B** shows the Frobenius distances (ranging from 0 to a maximum of $2\sqrt{2}$) between the true rotation matrix and the estimated rotation matrices. Again, global rotational sampling achieves more accurate rotations, whereas the distances scatter largely for the local MH trials. These findings suggest that global rotational sampling is indispensable for Bayesian random tomography in agreement with our previous findings (Joubert and Habeck, 2015) where we had to resort to repeated Gibbs sampling runs.

Before we study sampling of the full posterior distribution (all parameters $\boldsymbol{R}$, $\boldsymbol{x}$ and ξ are unknown), we will first outline how experimental projection images can be converted to 2D point clouds that are suitable for our approach to random tomography.

3.2 Representation of Projection Images by Point Clouds

Experimental projection data are typically presented as projection images rather than point clouds. In this subsection, we discuss

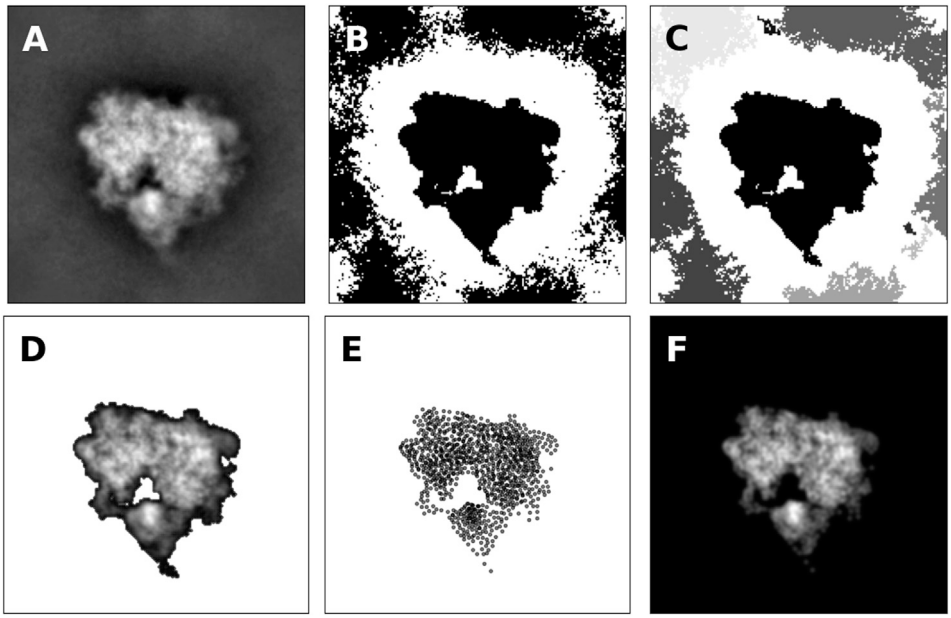


FIGURE 4 | Representation of projection images by 2D point clouds **(A)** Class average of the 80S ribosome **(B)** Mask obtained by thresholding image intensities greater than the median intensity. Black pixels are part of the mask **(C)** Clustering of pixels that are part of the mask. Pixels that form a connected component are grouped together and shown in different grayscale colors **(D)** Pixels that form the most central connected components with shifted image intensities **(E)** 2D point cloud composed of 1,000 particles obtained by running the Expectation Maximization algorithm **(F)** Model image according to **Eq. 10**. The cross-correlation coefficient between the model and the original image is 95.8%. If only pixels are considered that are part of the mask indicating the central connected component, the cross-correlation coefficient increases to 99.6%.

how to convert 2D projection images to 2D point clouds that are suitable for our Bayesian random tomography approach. We discuss this for a cryo-EM data set, but similar techniques are also applicable to other data, as we will demonstrate later.

The projection properties of mixtures of spherical Gaussians (**Eq. 10**) suggest to also represent the projection image as a mixture of Gaussians. Our model can only capture nonnegative intensities. Therefore, we first have to choose a suitable threshold θ above which image intensities are considered real signal. The threshold will be used to construct a binary mask: the intensities of pixels that are part of the mask will be shifted by θ such that their shifted intensities are nonnegative; the intensities of pixels that are not part of the mask will be set to zero (i.e., they will be ignored in the construction of the point cloud). A simple choice of θ for class averages from cryo-EM is the median intensity, but a different choice might be more suitable for other types of images.

An example of the thresholding procedure is shown in **Figure 4B** for a class average showing the projection of the 80S ribosome (shown in **Figure 4A**). Black pixels indicate pixels with intensity above the median. By looking at the mask, it is clear that only the central pixels forming a connected component carry signal.

Next, we identify pixels that form connected components. Again this applies to cryo-EM images; other types of images might require a different treatment to construct a suitable mask. To identify signal pixels that form a connected component, we convert the thresholded image to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the pixels with intensities above the threshold are the vertices $\mathcal{V} = \{\boldsymbol{u}_m; g(\boldsymbol{u}_m) > \theta, m = 1, \ldots, M\}$. Edges are introduced between all pairs of pixels that are nearest neighbors on the 2D square lattice, i.e. their Euclidean distance is smaller than or equal to one pixel:

$$\mathcal{E} = \left\{ (i,j) \in \{1, \ldots, |\mathcal{V}|\}^2; \parallel \boldsymbol{u}_i - \boldsymbol{u}_j \parallel \leq 1 \right\}.$$

As shown in **Figure 4C**, multiple connected components are typically found in the masked pixels. Since cryo-EM class averages are often centered, we pick the connected component whose center of mass is closest to the image center. The selected pixels including their intensity (shifted by θ) are shown in **Figure 4D**.

To obtain a particle-based representation of the central connected component, we run the Expectation Maximization algorithm (details in Supplementary Material). **Figure 4E** shows the estimated point cloud using 1,000 particles. The estimated standard deviation of the Gaussian is 1.34 pixels. The density generated by the 2D particles is shown in **Figure 4** and correlates highly with the original image and the masked image. **Supplementary Figure S1** shows more examples of class averages represented as 2D point clouds.

3.3 3D Reconstruction by Sampling the Full Posterior Distribution

We applied Bayesian random tomography to three real datasets, two cryo-EM datasets and one dataset from stochastic microscopy experiments visualizing marine microorganisms.

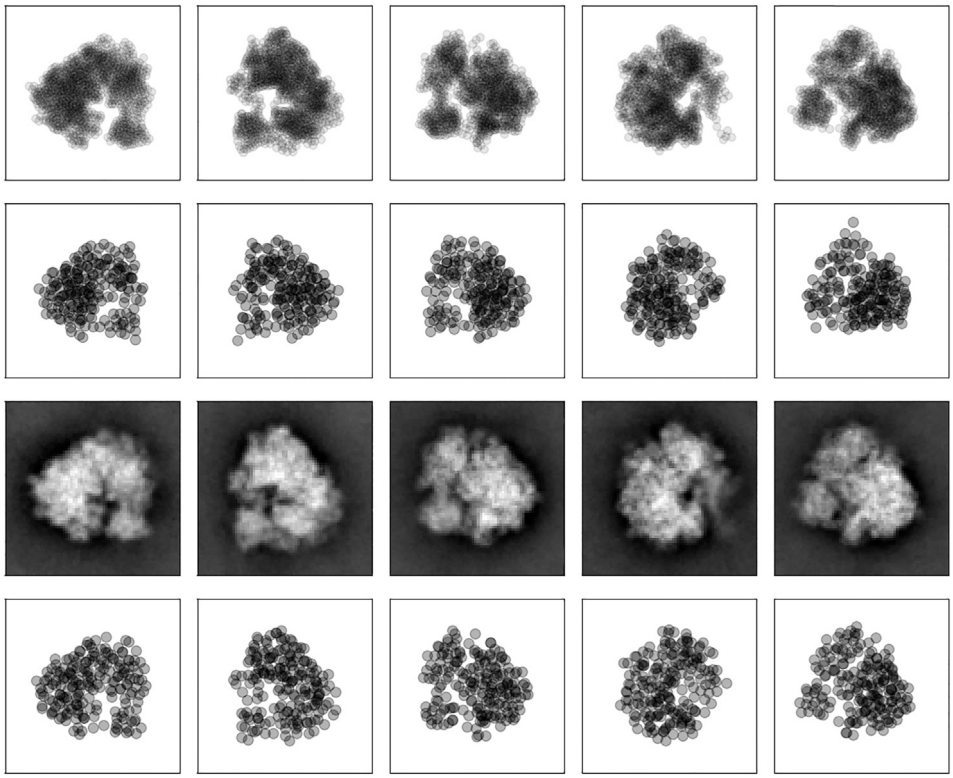


FIGURE 5 | 2D projections of the 80S ribosome. First row: point clouds derived from class averages. Each projection image is represented by 1,000 points. Second row: 2D projections of the coarse-grained model calculated with Bayesian random tomography based on 2D point clouds. Third row: Class averages. Bottom row: 2D projections of the coarse-grained model calculated with Bayesian random tomography based on class averages.

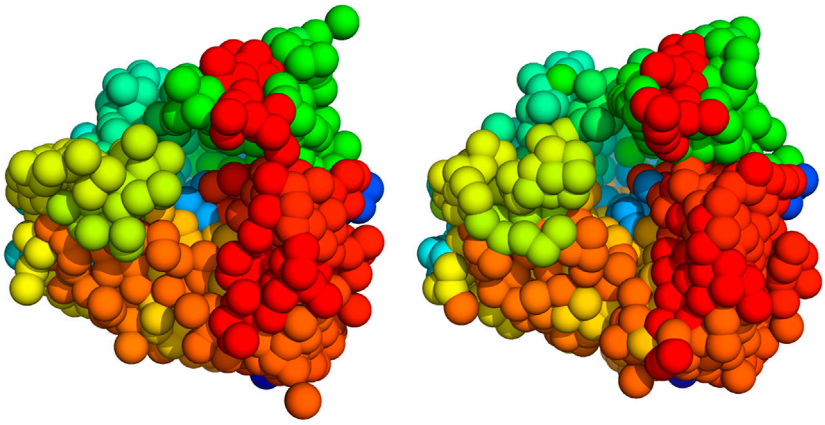


FIGURE 6 | 3D models of the 80S ribosome **(Left)** 1,000 particle model inferred with Bayesian random tomography **(Right)** Initial model computed with PRIME. The particles are sorted such that spatially close particles have similar indices. By using Pymol's chainbow command, we can then visualize the particle models such that substructures are better visible.

In these applications, we sampled the joint posterior distribution of all unknown parameters, particle positions $\boldsymbol{x}_k$, rotations $\boldsymbol{R}_n$ and nuisance parameters ξ, with the MCMC techniques discussed above. We started our reconstruction simulations from spherical random structures and random rotations and did not observe any dependence on the initial values.

The first dataset is comprised of 400 2D class averages of the 80S ribosome computed with SIMPLE2 (Elmlund and Elmlund, 2012) from cryo-EM micrographs (EMPIAR-10028); the size of the images is 80×80 pixels, the pixel size is 2.68 Å. The class averages are part of a SIMPLE2 tutorial and publicly available at https://simplecryoem.com/SIMPLE3.0/old_pages/2.5/data/

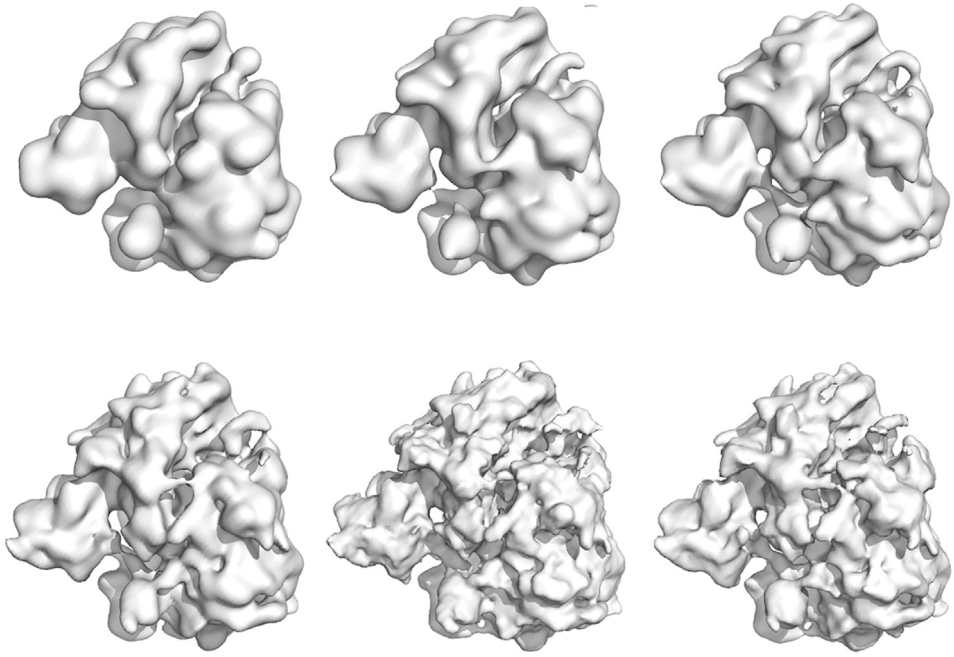


FIGURE 7 | Density maps of the 80S ribosome obtained with Bayesian random tomography using 50 class averages as input. Top row: 200, 1,000, 2000 particles **(left to right)**. Bottom row: 4,000, 8,000, 12,000 particles **(left to right)**.

simple2.5tutorials.tgz. **Figure 4** and **Supplementary Figure S1** show some example images and the 2D point clouds that were generated with the procedure outlined in **subsection 3.2**. Class averages were converted to 2D point clouds each composed of 1,000 points. Because the dataset is highly redundant, we only used the first 50 class averages and point clouds in the posterior simulations.

We used $K = 200$ and $K = 1000$ particles with a radius of $R = 16.4$ and $R = 8.4$ Å, respectively to fit the ribosome point clouds. We ran 500 iterations of Gibbs sampling with the global strategy for the rotational parameters and HMC for the particle positions. **Figure 5** shows five input point clouds and the projected model after convergence. We observe a good agreement between the experimental point clouds and the model point clouds with an RMSD ranging between 6.4 Å and 9.8 Å and an average of 7.7 ± 0.7 Å.

We also compared our 3D coarse-grained model of the 80S ribosome with a structure obtained with PRIME (Elmlund et al., 2008). To simplify the comparison, we converted the density map obtained with PRIME to a structure made up of 1,000 particles. The indices of the particle models were ordered such that spatially close particles have similar particle indices (which can be achieved, for example, by solving a traveling salesman problem using the matrix of inter-particle distances as input). Both structures show similar features (**Figure 6**); an FSC analysis reveals a resolution of 15.5 Å using the 0.143 criterion (**Supplementary Figure S6**).

We also ran simulations based on the first 50 class averages rather than 2D point clouds using 200 up to 12,000 particles. Again, we ran 500 steps of Gibbs sampling where the rotational parameters were updated globally with a frequency of 0.1. Projections of the 200 particle model are shown in the bottom rows of **Figure 5**. The cross-correlation coefficient between the class averages and the model images ranges between a minimum and maximum value of 90%–96% with an average of 94 ± 1%. For comparison, we also report the RMSDs to the particle clouds which range between 6.1 Å and 13.1 Å and an average of 8.3 ± 3.0 Å.

Using the last 100 particle configurations, we also generated density maps for each simulation and compared them to the high-resolution reconstruction EMD-2660 (Wong et al., 2014). The density maps are shown in **Figure 7**. To assess the quality of the particle models, we computed the FSC between the high-resolution map and the model maps (**Supplementary Figure S6**). Based on the 0.143 criterion, the resolution of the particle models ranges from 23.6 Å (200 particles) to 10.6 Å (12,000 particles). For comparison, the reconstruction obtained with SIMPLE reaches a resolution of 6.2 Å based on 200 class averages. More details about the quality of the reconstruction and computation times can be found in the Supplementary Material (**Supplementary Tables S2, S3**).

The posterior samples can be also used to assess the uncertainty of the particle models in the form of structural error bars. To carry out uncertainty quantification, the particle models first need to be superimposed and a correspondence between particles across different samples has to be established. We solve these two tasks by using the Iterative Closed Point (ICP) method followed by a linear assignment step where particle distances between superimpose clouds are used as a cost. **Supplementary Figure S7** shows an example for

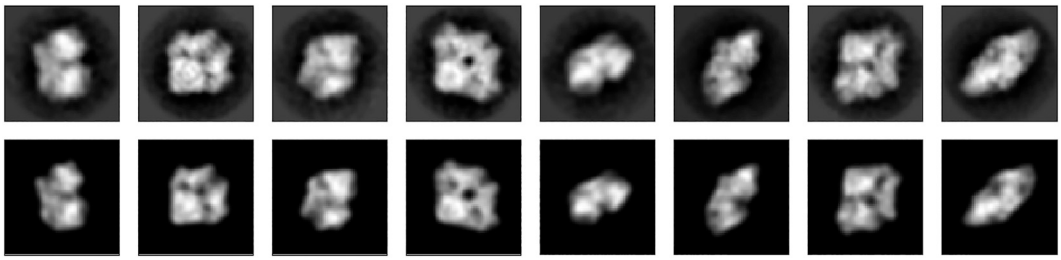


FIGURE 8 | 2D projections of beta-galactosidase. **Top row:** eight (out of 16) projection images (RELION class averages). **Bottom row:** Projection images calculated with Bayesian random tomography using 500 particles.

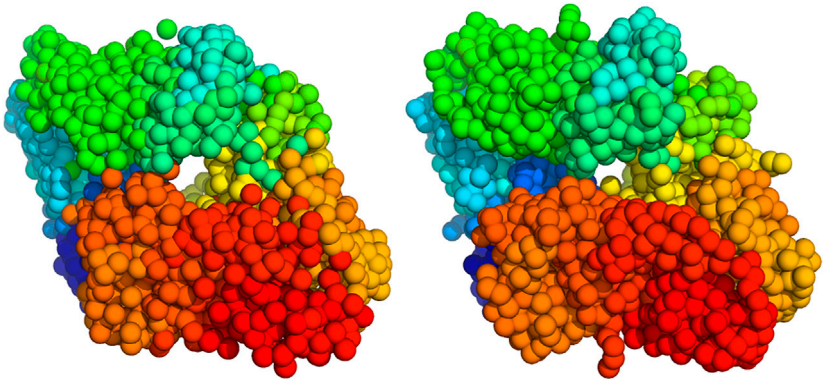


FIGURE 9 | 3D models of beta-galactosidase **(Left)** 2000 particle model inferred with Bayesian random tomography **(Right)** Coarse-grained model of the atomic structure (PDB code 1jz8).

structures based on 200 and 2000 particles. The distribution of uncertainties is inhomogeneous. Highly uncertain particles tend localize on the surface of the 200-particle model. The 2000-particle model shows smaller variations in the uncertainty of particle positions. So the large variations in the uncertainties of the 200-particle model might also be caused by the small number of particles.

The second cryo-EM dataset comprises 16 class averages of beta-galactosidase. These images are part of a RELION tutorial and available at ftp://ftp.mrc-lmb.cam.ac.uk/pub/scheres/relion31_tutorial_precalculated_results.tar.gz. The class average based on the data from EMPIAR-10204. The size of the images is 60×60 pixels, the pixel size is 3.54 Å. In this test, we inferred the structure from the images directly using likelihood **(11)** without converting the class averages to 2D point clouds.

Similar to the ribosome simulations we used 500 steps of Gibbs sampling with occasional global sampling of the rotational parameters to infer the coarse-grained structure of beta-galactosidase. We inferred structural models for systems with 100 up to 2000 particles.

The top row of **Figure 8** shows the first eight class averages that were used as an input for particle-based random tomography. The bottom row shows the projection images of a model composed of 500 particles that was obtained with sampling the full posterior distribution. Starting from a random initial structure and rotations, our sampling algorithm estimates a model structure and orientations that reproduce the experimental images closely with cross-correlation coefficients ranging between 94.7% and 97.5% and an average of $95.9 \pm 0.01\%$.

We compared the structure inferred with Bayesian random tomography against a high-resolution crystal structure (PDB code 1jz8) and a near-atomic cryo-EM reconstruction (EMD-5995). To enable this comparison, we converted the PDB structure to a 3D point cloud composed of 2000 particles. Correspondences between particles in our model and the model based on the crystal structure were established as in the calculation of the RMSD. **Figure 9** shows both models. The RMSD between our particle model and the Carbon-alpha atoms of the high-resolution structure 1jz8 is 3.4 Å. For comparison, we also report the RMSD between 1jz8 and its coarse-grained version (shown on the right of **Figure 9**) which is 2.4 Å. Bayesian random tomography achieves a similar accuracy by inferring a 3D model from the class averages as direct coarse graining of the high-resolution structure. **Supplementary Figure S8** shows density maps for all of the five simulations. By comparison with the high-resolution reconstruction (EMD-5995) we assess the resolution of the models to range between 25 Å (100 particles) and 11.5 Å (2000 particles). For comparison, the initial model from

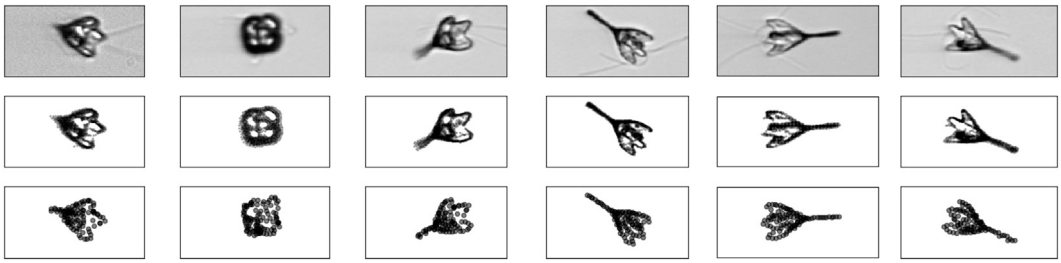


FIGURE 10 | Stochastic microscopy images of a plankton species. **Top row:** six (out of 16) projection images. **Middle row:** 2D point clouds representing the image data. **Bottom row:** 2D projections of the particle model calculated with Bayesian random tomography.

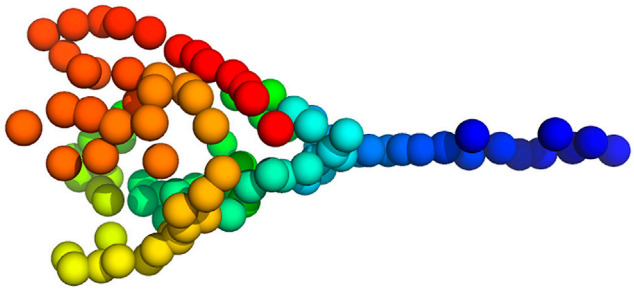


FIGURE 11 | 3D model of Pyramimonas Longicauda using 100 particles inferred from the point clouds shown in **Figure 10**.

RELION achieves a resolution of 9.8 Å (**Supplementary Figure S9** shows the corresponding FSC curves).

To assess the impact of the Boltzmann prior (**Eq. 17**), we ran two posterior simulations using 200 and 1,000 particles with the inverse temperature set to zero (i.e. the repulsive inter-particle energy is switched off). The quality of the reconstructed density map is largely unaffected by this change. For the 200 particles model, the average cross-correlation with Boltzmann prior is 94.7 ± 1.1%; without the Boltzmann prior we have 95.7 ± 0.9%. For the 1,000 particles model, these averages are 95.5 ± 1.5% (with Boltzmann prior) and 95.9 ± 1.5% (without Boltzmann prior). A comparison of the FSC curves obtained with and without Boltzmann prior confirms this finding (**Supplementary Figure S11**). The estimated resolution of the 200-particle model is 20.5 (19.4) Å with (without) Boltzmann prior; the 1000-particle model achieves a resolution of 12.0 (11.6) Å with (without) Boltzmann prior.

However, the Boltzmann prior has a strong effect on the packing of particles as assessed by the radial distribution functions (**Supplementary Figure S11**). With Boltzmann prior, the radial distribution shows a prominent peak close to the particle diameter, which is indicative of local order similar to a fluid. Without the Boltzmann prior, this peak disappears and we observe an enrichment of very short distances indicating a physically unrealistic particle packing. If our goal is to reconstruct a single 3D density from a homogeneous dataset, introducing the Boltzmann prior is not harmful, but dispensable. Turning the argument around, we find that the Boltzmann prior is compatible with the data and does not result in a severe loss of fitting quality. We expect that the prior will become essential in more advanced 3D reconstruction tasks, in particular when facing conformational heterogeneity.

Finally, we applied our random tomography approach to a dataset that shows structures on length scales that are much larger than the length scales imaged in cryo-EM. Following the work by Levis et al. (2018), we downloaded *in situ* microscopy images of the marine plankton species Pyramimonas Longicauda; the data are available at https://darchive.mblwhoilibrary.org/handle/1912/7341. These mesoscopic organisms are transparent and therefore allow for 3D reconstruction from 2D microscopic images. Since the organism seems to be quasi symmetric, we selected out of the 121 projection images recorded in 2013, 16 representative images. The selected images cover most of the views that are present in the dataset.

The intensity of microscopic images g_n is proportional to the transmissivity, which is related to the optical density of the object via an exponential transform. Therefore, to convert the images to 2D point clouds, we use the expectation maximization approach (see Supplementary Material) with weights proportional to $-\log g_n > 0$, since $g_n \in (0, 1)$. The six out of the 16 selected images and their point cloud representations are shown in **Figure 10**. Each microscopic image was converted to 2D cloud composed of 1,000 points.

The fact that the magnification can vary from image to image requires that we extend the likelihood for 2D point clouds **(13)** (also **Supplementary Equations S1, S2** in the Supplementary Material). These variations are accounted for by an additional factor that scales the coordinates of the projected model so as to match the 2D point cloud derived from the microscopic image. Moreover, we need to account for shifts in the image plane. These extensions increase the number of unknown parameters per image from four to eight: four quaternions parameterizing the unknown orientation, two translation parameters accounting for a shift, a scaling factor compensating variations in the magnification and a precision.

Inference of a 3D particle model proceeded as before. We estimated a model composed of 100 particles from the 16 2D point clouds starting from a random structure and random rotations (the initial values for the scaling factors and

translations were one and zero, respectively). **Figure 11** shows a 3D model of the plankton species inferred with Bayesian random tomography.

4 DISCUSSION

We outlined a Bayesian approach to random tomography, the problem of reconstructing a 3D structure from 2D views along unknown random directions. At the core of our approach is a representation of 3D volumes using a radial basis function kernel whose centers are our main inference parameters. We interpret the kernel centers as particle positions and use an excluded-volume prior to ensure that estimated particle configurations show a physically plausible packing. We demonstrated that coarse-grained models can be inferred from projection data (images or point clouds) with MCMC algorithms such as HMC and global sampling of the rotations.

In cryo-EM applications, our approach can be used to generate an initial model that can be refined further. So far, we tested the method only an class averages that displayed a high SNR. In future applications, we plan to explore the use of Bayesian random tomography from raw cryo-EM images and include the effect of the CTF into our model. Another route for extending the approach is account for conformational heterogeneity, which is one of the major bottlenecks in cryo-EM data processing. An interesting approach to characterize conformational variability in the presence of continuous flexibility has been proposed recently by Chen and Ludtke (2021) who use an autoencoder network with a Gaussian mixture model to represent conformational changes in a low dimensional latent space.

In all applications discussed in this paper, the number of particles K was fixed. An interesting question for future research is to estimate the number of particles based on the projection data. This might also provide a new way of measuring the resolution of the input data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://darchive.mblwhoilibrary.org/handle/1912/7341 ftp://ftp.mrc-lmb.cam.ac.uk/pub/scheres/relion31_tutorial_precalculated_results.tar.gz https://simplecryoem.com/SIMPLE3.0/old_pages/2.5/data/simple2.5tutorials.tgz.

AUTHOR CONTRIBUTIONS

MH designed research. NV and MH performed research. NV and MH contributed new analytic tools. NV and MH analyzed data. NV and MH wrote the paper.

ACKNOWLEDGMENTS

MH acknowledges funding from the German Research Foundation (DFG) under project SFB 860, TP B09 as well as funding from the Carl-Zeiss foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.658269/full#supplementary-material

REFERENCES

Barnett, A., Greengard, L., Pataki, A., and Spivak, M. (2017). Rapid Solution of the Cryo-Em Reconstruction Problem by Frequency Marching. *SIAM J. Imaging Sci.* 10, 1170–1195. doi:10.1137/16m1097171

Bendory, T., Bartesaghi, A., and Singer, A. (2020). Single-particle Cryo-Electron Microscopy: Mathematical Theory, Computational Challenges, and Opportunities. *IEEE Signal. Process. Mag.* 37, 58–76. doi:10.1109/msp.2019.2957822

Chen, M., and Ludtke, S. (2021). Deep Learning Based Mixed-Dimensional Gmm for Characterizing Variability in Cryoem.arXiv preprint arXiv:2101.10356

Coxeter, H. S. M. (1973). *Regular Polytopes*. New York, NY: Courier Corporation.

Elmlund, D., and Elmlund, H. (2012). SIMPLE: Software for ab initio Reconstruction of Heterogeneous Single-Particles. *J. Struct. Biol.* 180, 420–427. doi:10.1016/j.jsb.2012.07.010

Elmlund, H., Elmlund, D., and Bengio, S. (2013). PRIME: Probabilistic Initial 3D Model Generation for Single-Particle Cryo-Electron Microscopy. *Structure* 21, 1299–1306. doi:10.1016/j.str.2013.07.002

Elmlund, H., Lundqvist, J., Al-Karadaghi, S., Hansson, M., Hebert, H., and Lindahl, M. (2008). A New Cryo-EM Single-Particle ab initio Reconstruction Method Visualizes Secondary Structure Elements in an ATP-Fueled AAA+ Motor. *J. Mol. Biol.* 375, 934–947. doi:10.1016/j.jmb.2007.11.028

Frank, J. (2006). *Three-dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press. doi:10.1093/acprof:oso/9780195182187.001.0001

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-6, 721–741. doi:10.1109/tpami.1984.4767596

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

Habeck, M. (2009). Generation of Three-Dimensional Random Rotations in Fitting and Matching Problems. *Comput. Stat.* 24, 719–731. doi:10.1007/s00180-009-0156-x

Horn, B. K. P. (1987). Closed-form Solution of Absolute Orientation Using Unit Quaternions. *J. Opt. Soc. Am. A.* 4, 629–642. doi:10.1364/josaa.4.000629

Jaitly, N., Brubaker, M. A., Rubinstein, J. L., and Lilien, R. H. (2010). A Bayesian Method for 3D Macromolecular Structure Inference Using Class Average Images from Single Particle Electron Microscopy. *Bioinformatics* 26, 2406–2415. doi:10.1093/bioinformatics/btq456

Jin, Q., Sorzano, C. O. S., de la Rosa-Trevín, J. M., Bilbao-Castro, J. R., Núñez-Ramírez, R., Llorca, O., et al. (2014). Iterative Elastic 3d-To-2d Alignment Method Using Normal Modes for Studying Structural Dynamics of Large Macromolecular Complexes. *Structure* 22, 496–506. doi:10.1016/j.str.2014.01.004

Jonić, S., Vargas, J., Melero, R., Gómez-Blanco, J., Carazo, J. M., and Sorzano, C. O. (2016). Denoising of High-Resolution Single-Particle Electron-Microscopy Density Maps by Their Approximation Using Three-Dimensional Gaussian Functions. *J. Struct. Biol.* 194, 423–433. doi:10.1016/j.jsb.2016.04.007

Jonic, S., and Sanchez Sorzano, C. O. (2016). Coarse-graining of Volumes for Modeling of Structure and Dynamics in Electron Microscopy: Algorithm to

Automatically Control Accuracy of Approximation. *IEEE J. Sel. Top. Signal. Process.* 10, 161–173. doi:10.1109/JSTSP.2015.2489186

Joubert, P., and Habeck, M. (2015). Bayesian Inference of Initial Models in Cryo-Electron Microscopy Using Pseudo-atoms. *Biophysical J.* 108, 1165–1175. doi:10.1016/j.bpj.2014.12.054

Kam, Z. (1980). The Reconstruction of Structure from Electron Micrographs of Randomly Oriented Particles. *J. Theor. Biol.* 82, 15–39. doi:10.1016/0022-5193(80)90088-0

Kulis, B., and Jordan, M. I. (2012). "Revisiting K-Means: New Algorithms via Bayesian Nonparametrics," in Proceedings of the 29th International Conference on Machine Learning (ICML-12). Editors J. Langford and J. Pineau (New York, NY, USA), 513–520.

Levin, E., Bendory, T., Boumal, N., Kileel, J., and Singer, A. (2018). "3d ab initio Modeling in Cryo-Em by Autocorrelation Analysis,"in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 1569–1573.

Levis, A., Schechner, Y. Y., and Talmon, R. (2018). Statistical Tomography of Microscopic Life. *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 6411–6420.

Liang, J., and Dill, K. A. (2001). Are Proteins Well-Packed? *Biophys. J.* 81, 751–766. doi:10.1016/s0006-3495(01)75739-6

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

Lyumkis, D., Vinterbo, S., Potter, C. S., and Carragher, B. (2013). Optimod - an Automated Approach for Constructing and Optimizing Initial Models for Single-Particle Electron Microscopy. *J. Struct. Biol.* 184, 417–426. doi:10.1016/j.jsb.2013.10.009

Mechelke, M., and Habeck, M. (2013). Estimation of Interaction Potentials through the Configurational Temperature Formalism. *J. Chem. Theor. Comput.* 9, 5685–5692. doi:10.1021/ct400580p

Natterer, F. (2001). *The Mathematics of Computerized Tomography*. Philadelphia, Pa: SIAM. doi:10.1137/1.9780898719284

Neal, R. M. (2011). *Handbook of Markov Chain Monte Carlo*, 113–162.Mcmc Using Hamiltonian Dynamics

Panaretos, V. M. (2009). On Random Tomography with Unobservable Projection Angles. *Ann. Stat.* 37, 3272–3306. doi:10.1214/08-aos673

Penczek, P. A., Zhu, J., and Frank, J. (1996). A Common-Lines Based Method for Determining Orientations for N > 3 Particle Projections Simultaneously. *Ultramicroscopy* 63, 205–218. doi:10.1016/0304-3991(96)00037-x

Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. (2017). cryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination. *Nat. Methods* 14, 290–296. doi:10.1038/nmeth.4169

Sanz-García, E., Stewart, A. B., and Belnap, D. M. (2010). The Random-Model Method Enables ab initio 3D Reconstruction of Asymmetric Particles and Determination of Particle Symmetry. *J. Struct. Biol.* 171, 216–222. doi:10.1016/j.jsb.2010.03.017

Schaback, R., and Wendland, H. (2006). Kernel Techniques: from Machine Learning to Meshless Methods. *Acta numerica* 15, 543–639. doi:10.1017/s0962492906270016

Scheres, S. H. W. (2012a). A Bayesian View on Cryo-EM Structure Determination. *J. Mol. Biol.* 415, 406–418. doi:10.1016/j.jmb.2011.11.010

Scheres, S. H. W., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P. B., Frank, J., et al. (2007). Disentangling Conformational States of Macromolecules in 3D-EM through Likelihood Optimization. *Nat. Methods* 4, 27–29. doi:10.1038/nmeth992

Scheres, S. H. W. (2010). Maximum-likelihood Methods in Cryo-EM. Part II: Application to Experimental Data. *Methods Enzymol.* 482, 295–320. doi:10.1016/s0076-6879(10)82012-9

Scheres, S. H. W. (2012b). RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination. *J. Struct. Biol.* 180, 519–530. doi:10.1016/j.jsb.2012.09.006

Schölkopf, B., and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT.

Singer, A., and Shkolnisky, Y. (2011). Three-Dimensional Structure Determination from Common Lines in Cryo-EM by Eigenvectors and Semidefinite Programming. *SIAM J. Imaging Sci.* 4, 543–572. doi:10.1137/090767777

Takeda, H., Farsiu, S., and Milanfar, P. (2007). Kernel Regression for Image Processing and Reconstruction. *IEEE Trans. Image Process.* 16, 349–366. doi:10.1109/tip.2006.888330

Vainshtein, B. K., and Goncharov, A. B. (1986). Determination of the Spatial Orientation of Arbitrarily Arranged Identical Particles of Unknown Structure from Their Projections. *Soviet Phys. Doklady* 31, 278.

Van Heel, M. (1987). Angular Reconstitution: A Posteriori Assignment of Projection Directions for 3D Reconstruction. *Ultramicroscopy* 21, 111–123. doi:10.1016/0304-3991(87)90078-7

Vargas, J., Álvarez-Cabrera, A.-L., Marabini, R., Carazo, J. M., and Sorzano, C. O. S. (2014). Efficient Initial Volume Determination from Electron Microscopy Images of Single Particles. *Bioinformatics* 30, 2891–2898. doi:10.1093/bioinformatics/btu404

von Ardenne, B., Mechelke, M., and Grubmüller, H. (2018). Structure Determination from Single Molecule X-Ray Scattering with Three Photons Per Image. *Nat. Commun.* 9, 1–9. doi:10.1038/s41467-018-04830-4

Wong, W., Bai, Xc., Brown, A., Fernandez, I. S., Hanssen, E., Condron, M., et al. (2014). Cryo-em Structure of the Plasmodium Falciparum 80s Ribosome Bound to the Anti-protozoan Drug Emetine. *Elife* 3, e03080. doi:10.7554/eLife.03080

Yan, X., Dryden, K. A., Tang, J., and Baker, T. S. (2007). Ab Initio random Model Method Facilitates 3D Reconstruction of Icosahedral Particles. *J. Struct. Biol.* 157, 211–225. doi:10.1016/j.jsb.2006.07.013

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

Supplementary Material

1 PROBABILISTIC MODEL AND INFERENCE ALGORITHMS

The likelihoods for both types of input data are detailed in the main text of the manuscript (Eqs. 11 and 13). In case we have to account for an additional image-wise magnification factor s_n such as in the *in situ* microscopy application, these likelihoods generalize to

$$\Pr(g_n|\boldsymbol{x}, \boldsymbol{R}_n, \xi) = \left(\frac{\tau_n}{2\pi}\right)^{M_n/2} \exp\left\{-\frac{\tau_n}{2}\sum_{m=1}^{M_n}\left[g_{nm} - \alpha_n - \gamma_n \sum_k \phi_2\big(\boldsymbol{u}_{nm}; s_n(\boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n), \sigma^2\big)\right]^2\right\} \quad \text{(S1)}$$

for images where $\xi_n = (\boldsymbol{t}_n, s_n, \gamma_n, \alpha_n, \tau_n)$ and

$$\Pr(Y_n|\boldsymbol{x}, \boldsymbol{R}_n, \xi_n) = \prod_{m=1}^{M_n} \frac{1}{K}\sum_{k=1}^{K} \phi_2\big(\boldsymbol{y}_{nm}; s_n(\boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n), {\sigma_n}^2\big)\,. \quad \text{(S2)}$$

for 2D point clouds where $\xi_n = (\boldsymbol{t}_n, s_n, \sigma_n)$.

The Jeffreys' priors for the precision parameters are

$$\tau_n \sim 1/\tau_n$$

where in case of the point cloud likelihood (Eq. S2) $\tau_n = 1/\sigma_n^2$. The conditional posteriors of these parameters are Gamma distributions:

$$\tau_n \sim \tau_n^{M_n/2-1} \exp\left\{-\tau_n\chi_n^2/2\right\} \quad \text{(S3)}$$

where

$$\chi_n^2 = \sum_{m=1}^{M_n}\left[g_{nm} - \alpha_n - \gamma_n \sum_k \phi_2\big(\boldsymbol{u}_{nm}; s_n(\boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n), \sigma^2\big)\right]^2$$

Efficient random number generators for Gamma-variates are offered by libraries such as `numpy.random` from the Python array processing library NumPy. The conditional posterior of the offset and scale is a two-dimensional Gaussian; to update these parameters, we generate a sample from the bivariate Normal distribution by calling `numpy.random.multivariate_normal`. The shifts $\boldsymbol{t}_n$ and magnification factors s_n can be sampled with the Metropolis-Hastings algorithm.

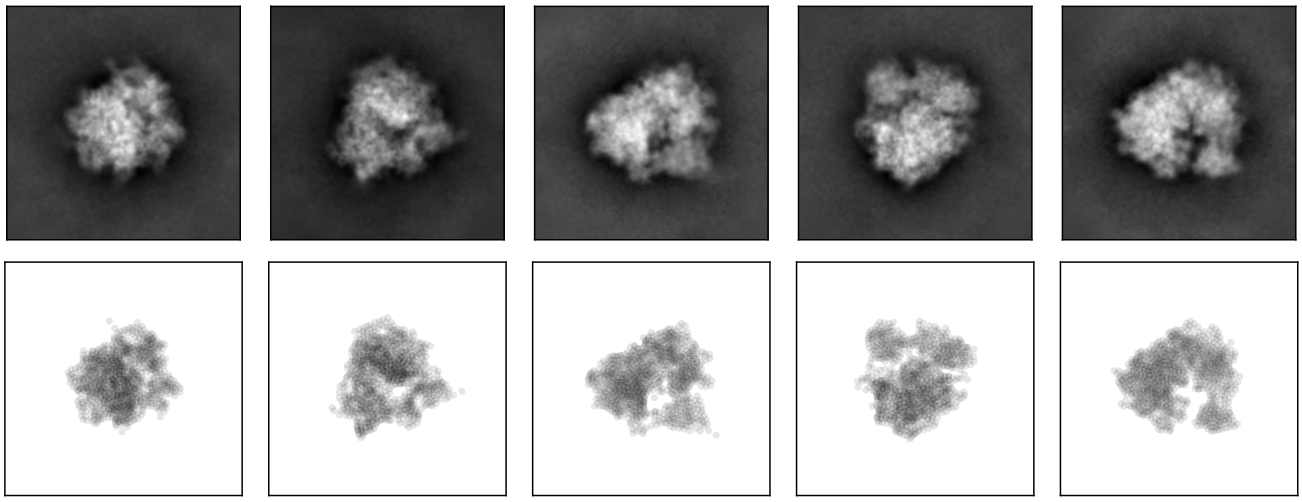


Figure S1. Representation of class averages by 2D point clouds. Top row: Class averages of the 80S ribosome. Bottom row: Particle representations obtained by Expectation Maximization of the central connected component.

2 FITTING POINT CLOUDS TO IMAGES WITH EXPECTATION MAXIMIZATION

Given a collection of pixels $\boldsymbol{u}_i$ with associated intensities g_i ($i = 1, \ldots, I$), we aim to represent this information by a 2D point cloud $\{\boldsymbol{y}_m; m = 1, \ldots, M\}$. The model to relate the image intensities and pixels to a point cloud is a mixture of Gaussians (Eq. 12):

$$g_i \approx \alpha + \gamma \sum_{m=1}^{M} \phi_2\big(\boldsymbol{u}_i; \boldsymbol{y}_m, \sigma^2\big)$$

where α is a suitable background parameter, γ is a scaling factor and σ the width of the Gaussian components. Using the approach described in subsection 3.2, we correct for the background by masking out the intensities that are greater than a suitable threshold and subtracting the threshold from the intensity such that $\alpha = 0$. Without loss of generality we can set $\gamma = 1$. So we are trying to fit $\boldsymbol{y}_m$ and σ such that the shifted intensities g_i at pixels $\boldsymbol{u}_i$ are reproduced as closely as possible with the Gaussian mixture model:

$$g_i \approx \sum_{m=1}^{M} \phi_2\big(\boldsymbol{u}_i; \boldsymbol{y}_m, \sigma^2\big)\,.$$

This can be achieved with an Expectation Maximization (EM) algorithm that cycles over the following updates:

- Soft assignment: Each pixel $\boldsymbol{u}_i$ with associated intensity g_i is assigned to a 2D point $\boldsymbol{y}_m$ with probability p_{im}:

$$p_{im} = \frac{\phi_2\big(\boldsymbol{u}_i; \boldsymbol{y}_m, \sigma^2\big)}{\sum_{m'=1}^{M} \phi_2\big(\boldsymbol{u}_i; \boldsymbol{y}_{m'}, \sigma^2\big)}$$

- Particle positions $\boldsymbol{y}_m$ are the centers of mass of the assigned pixels weighted by the (positive) image intensity g_i and the assignment probabilities p_{im} computed in the previous step:

$$\boldsymbol{y}_m = \frac{\sum_{i=1}^{I} p_{im}\, g_i \boldsymbol{u}_i}{\sum_{i=1}^{I} p_{im}\, g_i}$$

  where the denominator is the total image intensity represented by the m-th particle.
- The width of the Gaussians is estimated by

$$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{I} g_i} \sum_{i=1}^{I} \sum_{m=1}^{M} p_{im}\, g_i \, \|\boldsymbol{u}_i - \boldsymbol{y}_m\|^2}$$

For the entire series of 400 class averages provided by SIMPLE, we fitted 1000 particles to each image. The reasoning for choosing the same number of particles is that our model predicts the same total intensity which is identical to the number of points that represent the projection images. Figure S1 shows five representative class averages of the 80S ribosome and the their point cloud representation.

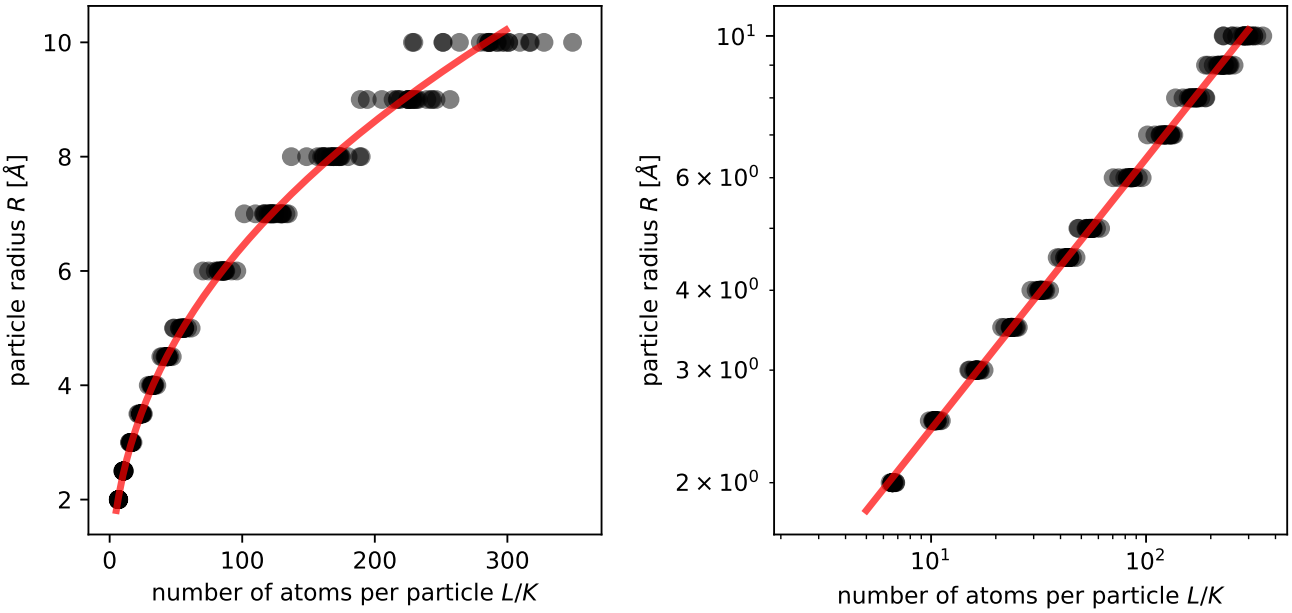


Figure S2. Relation between the average number of atoms per particle (L/K) and the particle radius.

3 CHOICE OF PARAMETERS IN BOLTZMANN PRIOR

To enforce a reasonable packing of particle structures, we use a Boltzmann distribution (Eq. 17) with a repulsive pairwise distance potential (Eq. 18). The only parameters that need to be set are the particle radius R and the inverse temperature β of the Boltzmann ensemble.

3.1 Particle radius

Biomolecules pack in a way that is reminiscent of fluids (Liang and Dill, 2001). To design a prior distribution that favors particle configurations with similar packing characteristics, we analyzed a number of biomolecular structures at different degrees of coarse graining. The coarse-grained models were derived from atomic structures by running the DP-means algorithm (Kulis and Jordan, 2012) for different distance cutoffs $\lambda = 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20$ Å, rather than using a computationally more demanding coarse graining approach proposed by us (Chen and Habeck, 2017).

The DP-means algorithm is a variant of the K-means clustering algorithm. DP-means assigns points to cluster centers, if the Euclidean distance of a point from a center does not exceed a distance cutoff. If no cluster exists that is close enough to a given point, a new cluster is created. This process is repeated multiple times until the iterations converge. In contrast to K-means, the number of clusters varies and depends on the cutoff value λ, which is the only input parameter: smaller cutoffs result in a larger number of clusters. The clusters centers form the particle positions; the distance cutoff λ corresponds to the particle diameter. We use the implementation of DP-means available at `https://github.com/michaelhabeck/DP-means`.

The coarse-grained structures obtained with DP-means allow us to establish a relation between the number of particles K and the distance cutoff λ. In our implementation of Bayesian particle-based tomography, we work with a fixed number of particles and want to choose the particle radius R such that volume-exclusion observed in known biomolecular structures is enforced. Since different biomolecular structures pack similarly, we expect that the average number of atoms per particle can be related to the particle radius

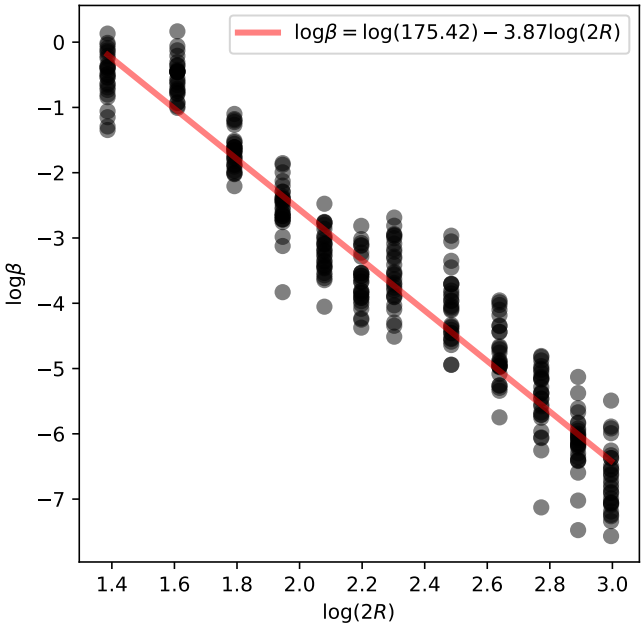


Figure S3. Configurational temperature as a function of particle radius.

independent of the specific structure. If L denotes the total number of atoms in a biomolecular structure, then L/K is the average number of atoms per particle. We use half the DP-means cutoff as a proxy for the particle radius $R = \lambda/2$. Figure S2 shows that there is indeed a high correlation between L/K and R that can be captured with a linear relation between the logarithms of both quantities:

$$\log R \approx -0.087 + 0.423 \log(L/K)$$

Mapping this back, we can predict a particle radius

$$R \approx 0.92 \times (L/K)^{0.42}$$

from the number of atoms L and the chosen number of particles K. For a given biomolecular system, we compute L based on the amino acid sequence.

3.2 Inverse temperature

The inverse temperature β is estimated from coarse grained structures using the configurational temperature formalism (Mechelke and Habeck, 2013). For each coarse-grained model that was produced to estimate the particle radii, we computed the configurational temperature. Figure S3 shows the relation between particle diameter and the configurational temperature, which can be fitted with a straight line in log space. The inverse temperature corresponds to the slope of the line and is set to $\beta = 175$.

4 SAMPLING PARTICLE POSITIONS AND PRECISIONS WITH FIXED ROTATIONS

Particle positions are sampled with HMC. To test the performance of HMC, we fitted coarse-grained models of GroEL/ES against 35 simulated projection images and 2D point clouds. For both types of data, HMC generates high-quality reconstructions.

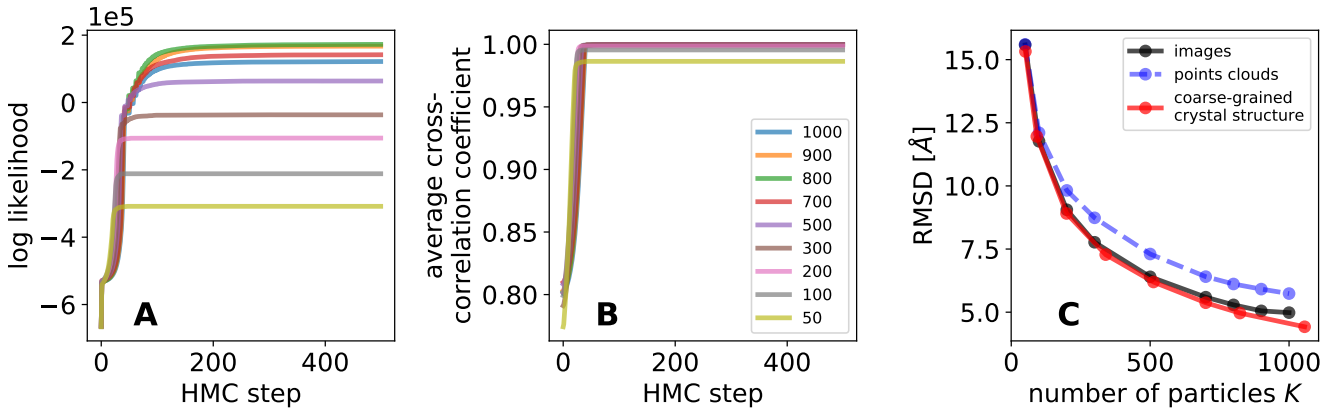


Figure S4. HMC sampling of particle positions with fixed rotations for a simulated class averages of GroEL/ES. **A** Evolution of the log likelihood during HMC sampling. **B** Evolution of the average cross-correlation coefficient. **C** RMSD between Carbon-alpha positions of the crystal structure and the coarse-grained models inferred with HMC. As a reference, the RMSD between the Carbon-alpha positions and the coarse-grained versions of the crystal structures is shown as red curve.

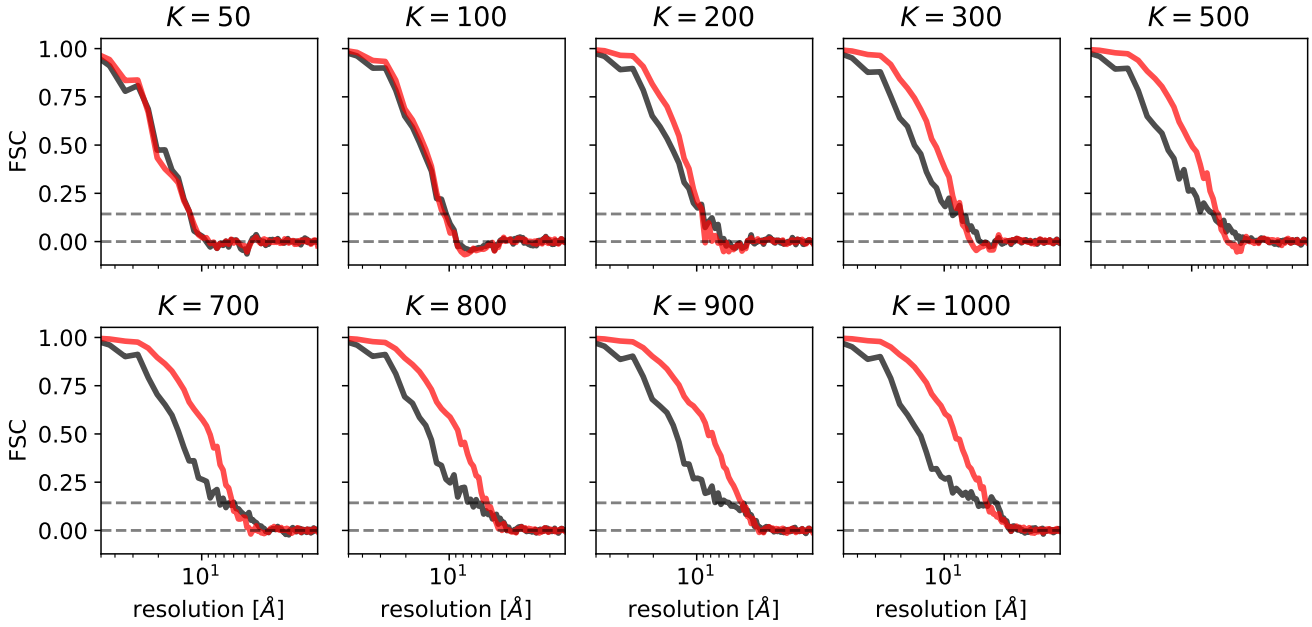


Figure S5. FSC curves for particle models of GroEL/ES obtained by fitting 2D point clouds (black curves) and projection images (red curves).

#particles	**resolution (FSC at 0.143) [Å]**		**RMSD [Å]**	
	point clouds	images	point clouds	images
50	12.2	12.2	15.6	15.5
100	10.5	10.5	12.1	11.7
200	7.8	9.3	9.8	9.1
300	8.3	8.3	8.7	7.8
500	7.1	6.5	7.3	6.4
700	6.5	6.2	6.4	5.6
800	6.0	5.8	6.1	5.3
900	6.0	4.8	5.9	5.0
1000	5.8	5.0	5.7	5.0

Table S1. Resolution estimates and RMSD values for particle-based models of GroEL/ES using point clouds and projection images as input data.

5 POSTERIOR SAMPLING BASED ON CLASS AVERAGES OF THE 80S RIBOSOME

5.1 Resolution assessment

In addition to the tests based on 2D point clouds derived from the projection images, we also used the first 50 class averages themselves as an input. We assessed the accuracy of the particle models inferred from both types of input data by computing the FSC correlating the high-resolution structure EMD-2660 with the initial model generated by our method. Figure S6 shows the FSCs and Table S2 lists the resolutions derived from these curves.

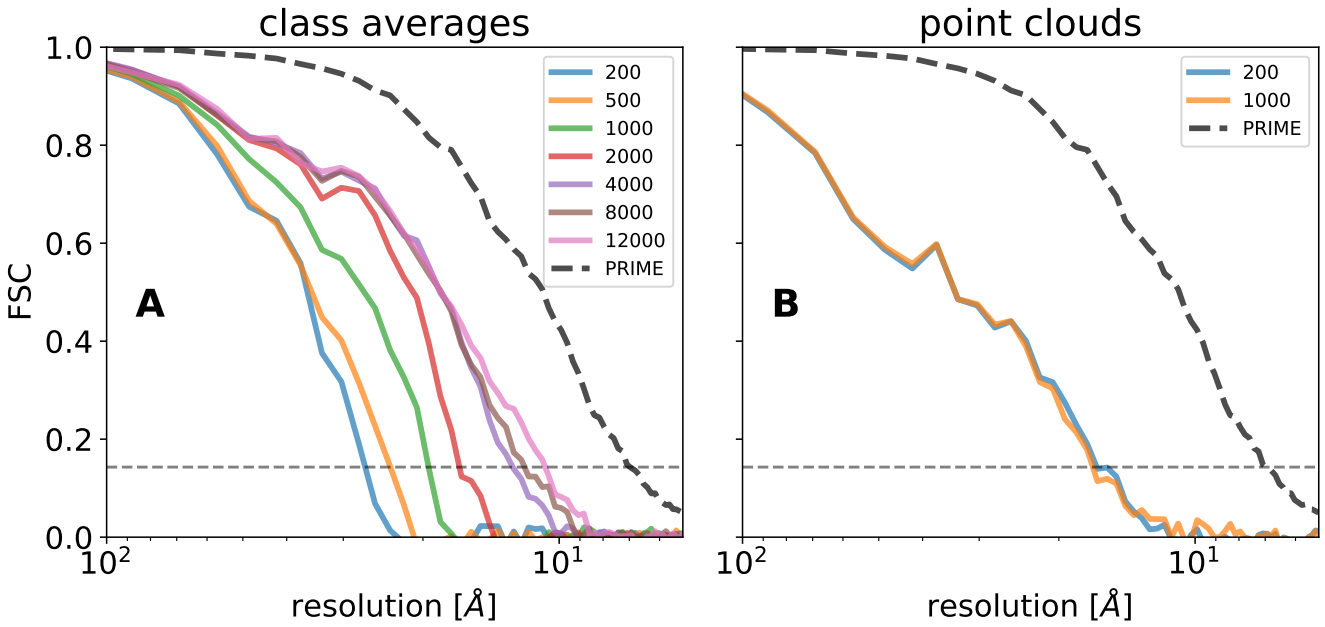


Figure S6. FSCs of particle models inferred from 50 class averages (**A**) and 2D points clouds (**B**).

#particles	diameter [Å]	resolution (FSC at 0.143) [Å]
200	32.8	27.6
500	22.3	23.5
1000	16.7	19.7
2000	12.5	16.9
4000	9.3	12.5
8000	7.0	11.9
12000	5.9	10.6

Table S2. Resolution estimates for particle-based models of the 80S ribosome obtained with random tomography. The FSCs are computed by comparing the high-resolution reconstruction (EMD-2660) with the density map generated from the last 100 sampled particle configurations.

5.2 Computation times

Computation times for 100 steps of Gibbs sampling using 50 class averages as input. Tests were run on an intel i5 processor (1.6 GHz oct-core using a single thread only).

#particles	computation time [h]
200	0.1
1000	0.3
4000	0.9
6000	1.2
8000	1.6
10000	2.0
12000	2.4

Table S3. Computation times for a 3D reconstruction from 50 class averages of the 80S ribosome using 100 Gibbs sampling steps.

5.3 Uncertainty Quantification

Bayesian methods and posterior sampling allow for uncertainty quantification. Since particle models can differ by a rigid transformation and a permutation of particle indices, we first need to superimpose the configurations generated by posterior sampling and establish a correspondence between particle positions across different samples. We do this by using the ICP (iterative closest point) method followed by linear assignment. After these steps, we can compute the standard deviation of each particle position which provides a structural error bar. Figure S7 shows a particle model where the error bar is encoded in the bead radius.

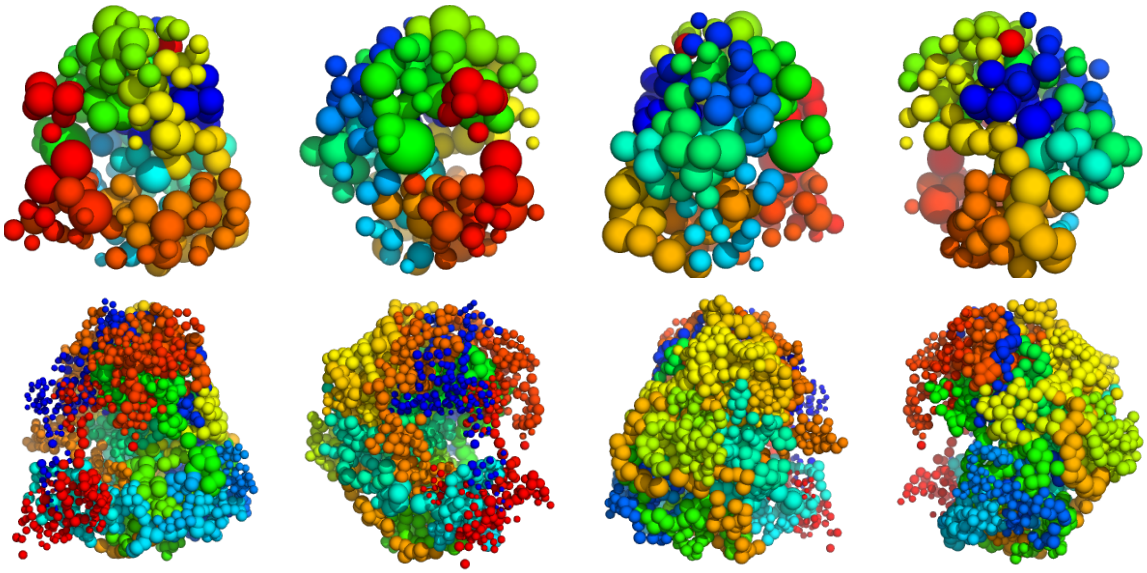


Figure S7. Uncertainty quantification of the ribosome model. The size of the spheres is proportional to the standard deviation of particle positions after superposition and assignment. The top row shows a model based on 200 particles. The bottom row shows the structure based on 2000 particles. Shown are side views of the 80S ribosome that differ by a 90-degree rotation about the z axis.

6 POSTERIOR SAMPLING BASED ON CLASS AVERAGES OF BETA-GALACTOSIDASE

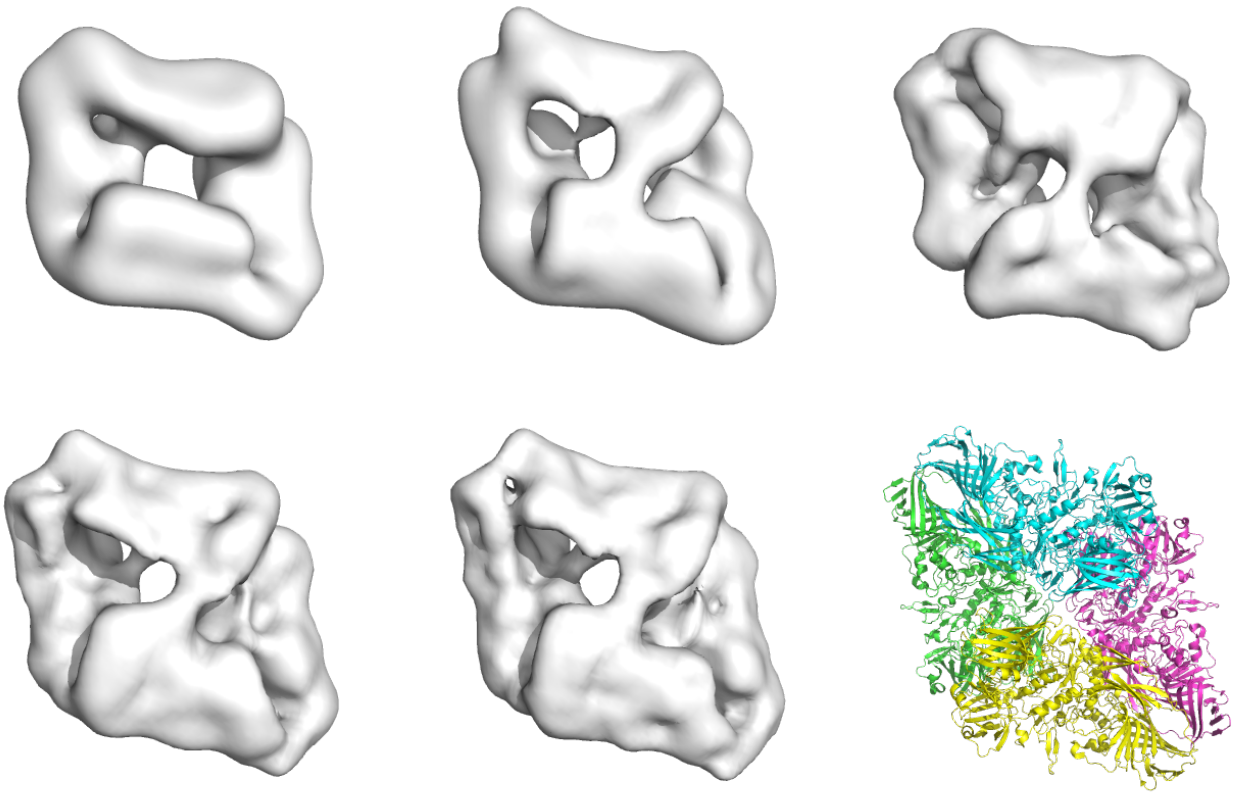


Figure S8. Density maps of beta-galactosidase obtained with Bayesian random tomography using 16 class averages as input. Top row: 100, 200, 500 (left to right), Bottom row: 1000, 2000, PDB code 1jz8 (left to right).

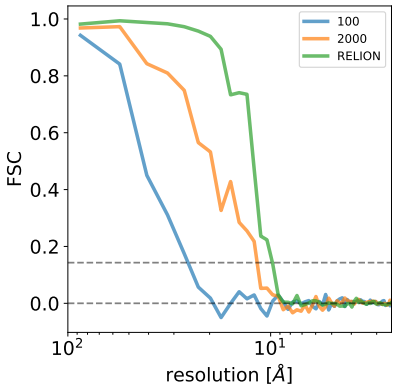


Figure S9. FSC curves comparing the high-resolution reconstruction EMD-5995 with Bayesian models using 100 and 2000 particles, respectively, as well as an initial structure obtained with RELION.

7 IMPACT OF BOLTZMANN PRIOR

To assess the impact of the Boltzmann prior (Eq. 17), we ran simulations with the beta-galactosidase images where the Boltzmann prior was switched on and off ($\beta = 0$).

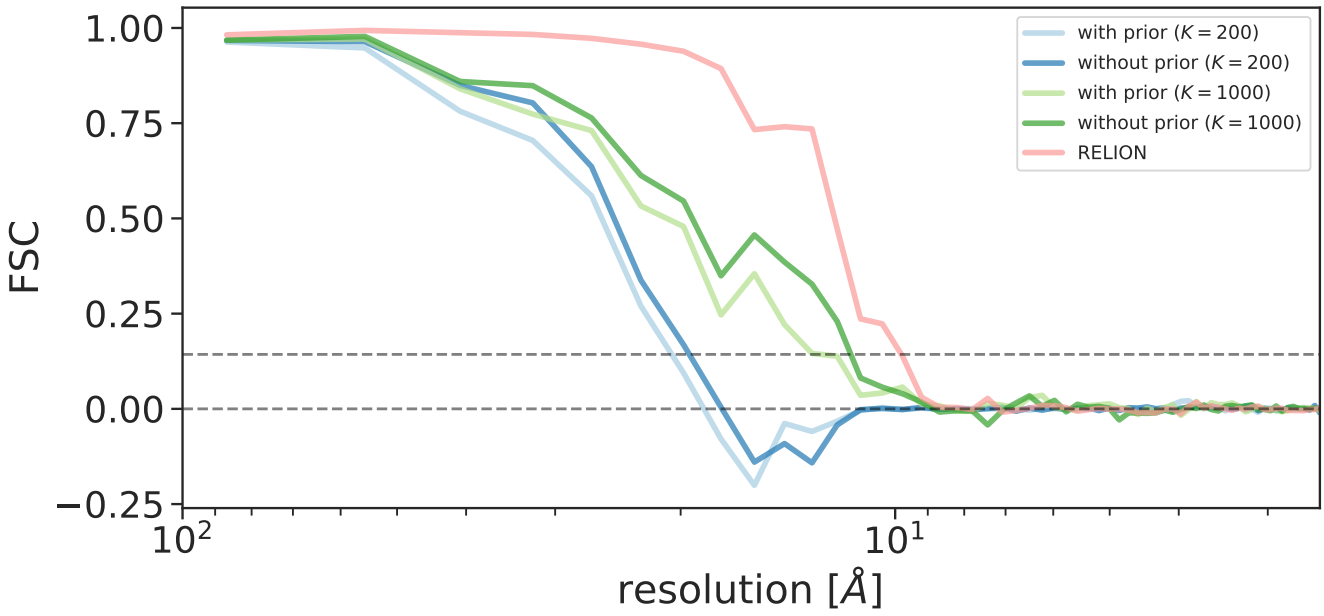


Figure S10. Impact of Boltzmann prior on the accuracy of the 3D reconstruction. FSC between high-resolution reconstruction EMD-5995 and models calculated with and without Boltzmann prior for various numbers of particles. The FSC curve for the RELION's initial model is shown in yellow.

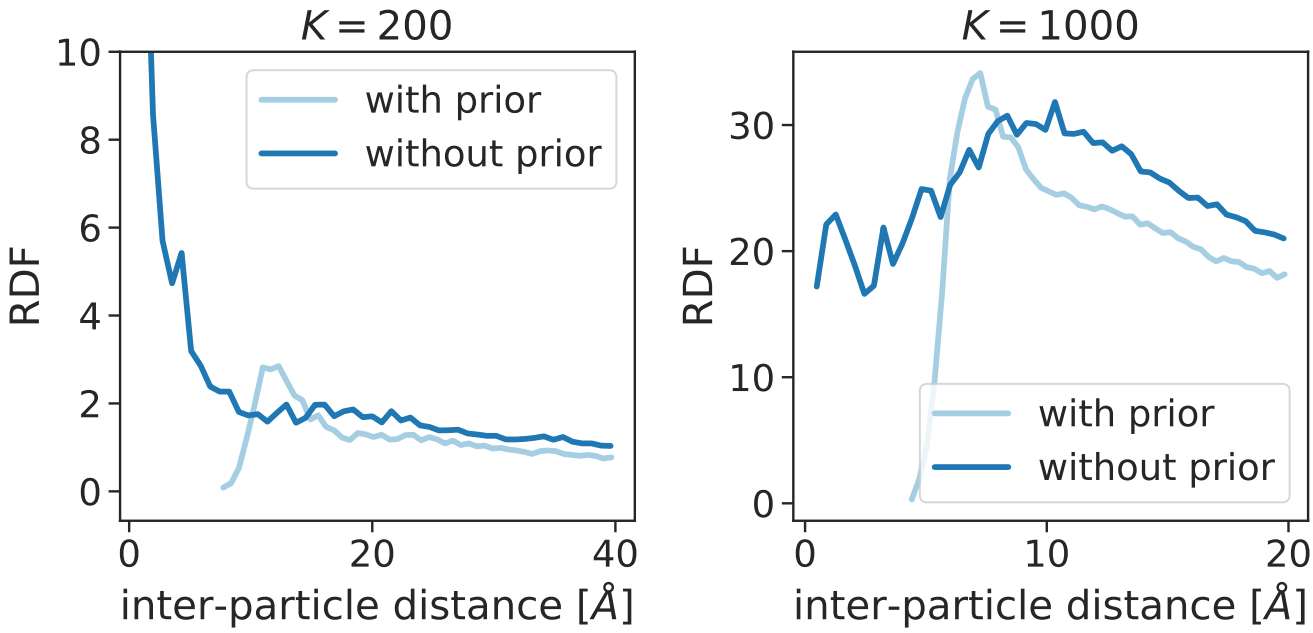


Figure S11. Impact of Boltzmann prior on particle packing. Shown are the radial distribution functions computed from sampled particle models of beta-galactosidase with (light blue line) and without Boltzmann prior (dark blue line).

REFERENCES

Liang J, Dill KA. Are proteins well-packed? *Biophys. J.* **81** (2001) 751–766.

Kulis B, Jordan MI. Revisiting k-means: New algorithms via bayesian nonparametrics. Langford J, Pineau J, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (New York, NY, USA) (2012), 513–520.

Chen YL, Habeck M. Data-driven coarse graining of large biomolecular structures. *PLoS ONE* **12** (2017) e0183057.

Mechelke M, Habeck M. Estimation of Interaction Potentials through the Configurational Temperature Formalism. *J Chem Theory Comput* **9** (2013) 5685–5692. doi:10.1021/ct400580p.

Chapter 3

Kernel-based tomographic reconstruction on and off the grid

This chapter is the second research result from my Ph.D. study. Here we present three grid-free probabilistic models for 3D volume reconstruction of tomography datasets which the relative orientation of the particles are known. This manuscript is in preparation. Own contribution:

- Concept and implementation of the algorithm and the code.
- All figures, tables.
- Manuscript in parts.

Kernel-based tomographic reconstruction on and off the grid

Nima Vakili [1,2], Michael Habeck[1,2,*]

October 2, 2021

[1]Microscopic Image Analysis Group, Jena University Hospital, 07743 Jena, Germany.
[2]Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany.

*E-Mail: michael.habeck@uni-jena.de

Abstract

Tomography is a major technique in imaging science and has made wide-ranging contributions to physics, material science, biology and medicine. Independent of the specific application domain, the reconstruction problem in tomography is to recover a 3D object from 2D projection images. Although a large array of reconstruction methods has been proposed, there is a continued interest in developing new algorithms. Here, we propose three reconstruction algorithms to characterize the 3D object from a limited set of projection images. At the core of all three algorithms is a kernel-based representation of the 3D structure as a collection of weighted spherical particles. This representation differs from the voxel grid representation used by most of the existing methods. The particle representation is a natural choice for smooth molecular densities, allows for fast and accurate projection, and has built-in positivity and sparsity. We benchmark the particle-based reconstruction algorithms on simulated and real data sets.

1 Introduction

Tomography is a prominent imaging technique in the biological and physical sciences. Thanks to advances in imaging hardware, it is now possible to record images with resolutions ranging from the nano to the atomic scale (Miao et al., 2016; Frank, 2006; Herman, 2009; Midgley & Dunin-Borkowski, 2009; Saghi & Midgley, 2012; Leary et al., 2012; Tian et al., 2020). A series of 2D projection images, also called a tilt series, is acquired by rotating the specimen over a limited angular range. A 3D structure is computed from the projection images with reconstruction algorithms. Two factors mainly determine the quality of tomographic reconstructions: Radiation damage and the missing wedge caused by a limited angular sampling (notice that specimens cannot be tilted beyond $\pm 70^{\circ}$).

Direct tomographic reconstruction methods include Back Projection (BP), Weighted Back Projection (WBP) (Herman, 1980; Radermacher, 2007) and Filtered Back Projection (FBP) which involves several filtering steps on the projection data followed by BP. Another approach is to use a common lines technique based on the Fourier slice theorem in reciprocal space (Kak & Slaney, 2001). All of these algorithms suffer from missing data and can produce severe artifacts as well as systematic deviations from the ground truth.

Iterative reconstruction algorithms aim to overcome these shortcomings and can be categorized into three classes: real-space, reciprocal-space and hybrid approaches. Among the real-space methods, the Algebraic Reconstruction Technique (ART) (Gordon et al., 1970), its refined version Simultaneous ART (SART) (Andersen & Kak, 1984), and the Simultaneous Iterative Reconstruction Techniques (SIRT) (Gilbert, 1972) are most widely used. These methods use multiplicative and additive updates to iteratively solve a linear inverse reconstruction problem. A disadvantage of iterative methods are their elevated computational costs that are roughly doubled compared to WBP per iteration.

Fourier-based forward and back-projection methods reduce the computation time considerably. In particular, approaches based on the Non-Uniform Fast Fourier Transform (NUFFT) achieve a substantially lower approximation error (Fessler & Sutton, 2001; Matej et al., 2004; O'Connor & Fessler, 2006). These methods mostly face challenges when reconstructing large multi-dimensional datasets and have high memory demands.

Another family of reconstruction methods follows a hybrid approach by iterating between real and reciprocal space. Among the hybrid reconstruction methods is Equal Slope Tomography (EST) (Miao et al., 2005) which is limited by the requirement that the tilt angles need to be consistent with equal slope increments. GENeralized Fourier Iterative REconstruction (GENFIRE) (Pryor et al., 2017) is a powerful hybrid method that produces high-resolution reconstructions from a limited number of 2D projections. The major drawbacks are a high memory usage due to an oversampling strategy and numerical errors due to interpolation. Pham et al. (2020) as well as Yang et al. (2021) propose REal Space Iterative Reconstruction Engine (RESIRE) which combines gradient descent methods with the Fourier slice theorem to solve the reconstruction problem.

In this paper, we propose three kernel-based real-space algorithms to address the aforementioned drawbacks. At the core of our algorithms is the representation of the unknown structure as a weighted sum of Gaussian radial basis functions (RBFs) that share the same smoothing parameter (the kernel bandwidth σ). Each Gaussian kernel can be interpreted as a *particle* characterized by a 3D position and a non-negative weight (Vakili & Habeck, 2021). The particle positions can be located at fixed positions, such as a regular grid, in which case the particle weights are the only free parameters. Another interesting option is to also estimate the particle positions. In this case, the particle positions are not tied to a regular grid, but can move freely during the reconstruction process. The reconstruction method is then mesh-free and operates "off the grid." Mesh-free methods pay the prize of introducing non-linear model parameters, the particle positions, whereas the weights are linear parameters.

We present three algorithms for kernel-based tomographic reconstruction operating on and off the grid. Our first method places the particle positions on a hexagonal grid and only updates the weights using an iterative non-negative least-squares algorithm. The second reconstruction method is based on an Expectation Maximization (EM) algorithm similar to the one used in Gaussian mixture fitting and updates both the particle weights and positions, i.e. it is a mesh-free method. EM involves a subsampling step in which a chosen number of pixels is randomly selected from the projection images. The third method is a fully Bayesian approach that assigns equal weight to all particles and fits the particle positions with Hamiltonian Monte Carlo. This approach is the most flexible one in that it allows for the use of alternative fitting criteria and the incorporation of prior information such as excluded volume. All reconstruction algorithms operate

in real space and do not require gridding or interpolation on a grid. To speed up the reconstruction of particle positions, we can benefit from fast methods for nearest neighbor search and efficient data structures such as KD-trees. Finally, we benchmark our methods by using simulated and real data sets acquired from Cryo-ET and scanning transmission electron microscopy (STEM).

2 Methods

2.1 Volume representation

The standard representation of a volume $f(\boldsymbol{x})$ is based on gridding where the volume is discretized using a regular cubic grid composed of voxels. Gridding renders the reconstruction problem finite. The voxel grid offers a non-parametric representation of the volume that can be very flexible, depending on the voxel size. This flexibility comes at the cost that any natural feature such as blobbiness or differentiability is hard to impose. Moreover, the representation is quite dense and typically involves many voxel values (in the order of millions).

Our approach models the unknown volume $f(\boldsymbol{x})$ by using the kernel representation

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \, \phi((\boldsymbol{x} - \boldsymbol{x}_k)/\sigma) = \sum_{k=1}^{K} w_k \, \rho(\|\boldsymbol{x} - \boldsymbol{x}_k\|/\sigma) \,. \tag{1}$$

The *kernel* function $\phi : \mathbb{R}^3 \mapsto \mathbb{R}_+$ maps from a 3D position $\boldsymbol{x} \in \mathbb{R}^3$ to a non-negative density value. Typically, the kernel function has a bump-shaped appearance and is centered at positions $\boldsymbol{x}_k$. The scale parameter or *bandwidth* σ serves as a unit for measuring the extent to which a position deviates from the kernel center and determines the smoothness of the volume: The smaller σ, the rougher will be the representation of the volume (1). In general, ϕ could exhibit any suitable shape, but here we restrict ourselves to shapes of spherical symmetry by letting $\phi(\boldsymbol{x}) = \rho(\|\boldsymbol{x}\|)$ where $\|\cdot\|$ denotes the Euclidean norm and $\rho : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a one-dimensional function that maps a non-negative scalar argument, the distance from the kernel center, to a non-negative density value. Due to the spherical symmetry of the kernel, it is also called a radial basis function (RBF) kernel, and representation (1) is known as a RBF network (Bishop et al., 1995). If we restrict the weights w_k associated with each center $\boldsymbol{x}_k$ to a positive range, then the output of the RBF network will be non-negative and therefore suitable for representing 3D densities.

The Gaussian RBF kernel $\phi_{D,\sigma}(\boldsymbol{x})$ in $\mathbb{R}^D$ is defined as

$$\phi_{D,\sigma}(\boldsymbol{x}) = (2\pi\sigma^2)^{-D/2} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}\|^2\right\}, \quad \boldsymbol{x} \in \mathbb{R}^D. \tag{2}$$

That is, for the Gaussian RBF kernel we have $\rho(r) \propto \exp(-r^2/2)$. With our choice (2) representation (1) is a mixture of K (isotropic) Gaussian densities located at centers $\boldsymbol{x}_k \in \mathbb{R}^3$:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \, \phi_{3,\sigma}(\boldsymbol{x} - \boldsymbol{x}_k) \tag{3}$$

This representation has a rich tradition in statistics where it is commonly used in density estimation (Friedman et al., 2001).

Park & Sandberg (1991) have shown that the Gaussian RBF expansion has universal approximation capabilities. So the class of volumes that can be represented by (3) is rather broad. For $K \to \infty$, we can approximate smooth 3D densities arbitrarily closely. For $\sigma \to 0$, the kernel collapses to an atomic measure. For $\sigma > 0$, the kernel representation is particularly suited for smooth densities such as molecular densities that do not exhibit sharp edges.

A physical interpretation of model (3) is that we represent the volume as a blurry density generated by K particles. The particles form a collection of points with masses $w_k > 0$ located at 3D positions $\boldsymbol{x}_k$. The atomic density is a sum of weighted point sources positioned at $\boldsymbol{x}_k$:

$$\sum_{k=1}^{K} w_k \delta_{\boldsymbol{x}_k}$$

and blurring is achieved by a convolution with the RBF kernel ϕ.

The model parameters are the particle positions and weights $\{(\boldsymbol{x}_k, w_k)\}_{k=1}^{K}$, known in computer vision as a weighted 3D point cloud. A major difference to voxel-based representations is that the particle locations can be off the grid. Therefore, the kernel representation (3) can in principle achieve sub-pixel resolution whereas the resolution of a voxel representation is limited by the Nyquist frequency. The kernel representation (3) is linear in the weights w_k, and nonlinear in the particle positions $\boldsymbol{x}_k$. For finite K and $\sigma > 0$, model (3) is non-negative, smooth and sparse. Another advantage over grid-based representations is that computation of projection images is straightforward and efficient.

2.2 Modeling projection images

In our application of the grid-free kernel representation to tomographic data, we restrict ourselves to tilt series recorded by parallel-beam projection, which is equivalent to the X-ray transform (Natterer, 2001). The parallel-beam projection of the 3D Gaussian RBF kernel is a 2D Gaussian RBF kernel:

$$\int \phi_{3,\sigma}(\boldsymbol{R}\boldsymbol{x})\,\mathrm{d}z = \phi_{2,\sigma}(\boldsymbol{P}\boldsymbol{R}\boldsymbol{x}) \tag{4}$$

where $\boldsymbol{R} \in SO(3)$ is a rotation matrix and $\boldsymbol{P}$ the 2×3 projection matrix

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} . \tag{5}$$

Property (4) is very convenient, because it allows us to evaluate the parallel-beam projection of model (3) by performing a rigid transformation on the particle locations $\boldsymbol{x}_k$, omitting the last coordinate of the resulting 3D positions, and finally blurring the projected 2D point cloud with a Gaussian kernel:

$$(\boldsymbol{P}\boldsymbol{R}f)(\boldsymbol{y}) = \sum_{k=1}^{K} w_k\, \phi_{2,\sigma}(\boldsymbol{y} - \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}_k) \tag{6}$$

where $\boldsymbol{P}\boldsymbol{R}f$ denotes the rotated and projected model volume f and $\boldsymbol{y} \in \mathbb{R}^2$ is a position in the image plane.

A tilt series is composed of a set of images g_n with associated rotation matrices $\boldsymbol{R}_n \in SO(3)$. Each projection image is made up of pixels (enumerated by index m) located at 2D coordinates $\boldsymbol{y}_{nm} \in \mathbb{R}^2$ and pixel values g_{nm}. That is, a single projection image g_n is encoded by $g_n = \{(\boldsymbol{y}_{nm}, g_{nm})\}_{m=1}^{M_n}$ where M_n is the number of pixels in the n-th projection image. In total, we have N such projection images $\{(\boldsymbol{y}_{nm}, g_{nm})\}_{n=1,m=1}^{N,M_n}$ with associated rotation matrices $\{\boldsymbol{R}_n\}_{n=1}^{N}$.

2.3 Optimization of the particle weights by non-negative least squares

Let us first discuss the problem of estimating the particle weights $\{w_k\}_{k=1}^{K}$ for given particle positions $\{\boldsymbol{x}_k\}_{k=1}^{K}$. To estimate the weights, we match the projected model $(\boldsymbol{P}\boldsymbol{R}_n f)(\boldsymbol{y}_{nm})$ at pixel positions $\boldsymbol{y}_{nm}$ against the pixel intensities g_{nm} of the n-th projection

image. This can be achieved by minimizing the least-squares error:

$$\ell(\boldsymbol{w}) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} (g_{nm} - (\boldsymbol{P}\boldsymbol{R}_n f)(\boldsymbol{y}_{nm}))^2 = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left(g_{nm} - \sum_{k=1}^{K} w_k \, \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n \boldsymbol{x}_k) \right)^2 . \tag{7}$$

For fixed particle positions, the elements K_{nmk} of the $M_n \times K$ kernel matrix $\boldsymbol{K}_n$ encoding the forward operator of the n-th parallel-beam projection are given by

$$K_{nmk} = \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n \boldsymbol{x}_k) \, . \tag{8}$$

To keep the kernel matrix $\boldsymbol{K}_n$ sparse, we neglect entries involving distances $\|\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n \boldsymbol{x}_k\| > 5\sigma$. The projection is a matrix-vector multiplication $\boldsymbol{K}_n \boldsymbol{w}$ where $\boldsymbol{w}$ is the K-dimensional vector of particle weights. Using matrix notation, the least-squares cost function simplifies to

$$\ell(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w} - 2\boldsymbol{b}^T \boldsymbol{w} + \text{const.} \tag{9}$$

where the $K \times K$ normal matrix $\boldsymbol{A}$ and the K-size vector $\boldsymbol{b}$ are defined as:

$$\boldsymbol{A} = \sum_{n=1}^{N} \boldsymbol{K}_n^T \boldsymbol{K}_n, \quad \boldsymbol{b} = \sum_{n=1}^{N} \boldsymbol{K}_n^T \boldsymbol{g}_n$$

and $\boldsymbol{g}_n$ is the M_n-size vector of pixel values in the n-th projection image. The normal matrix $\boldsymbol{A}$ is element-wise non-negative because the Gaussian RBF kernel is non-negative.

Minimization of objective function (9) is a non-negative least-squares (NNLS) problem, because particle masses should be non-negative $w_k \geq 0$. We minimize $\ell(\boldsymbol{w})$ under the non-negativity constraint by applying the majorization-minimization algorithm proposed by Sha et al. (2007). Since $\boldsymbol{A}$ is element-wise non-negative, the multiplicative update is very simple for our NNLS problem:

$$w_k \leftarrow w_k \cdot \frac{|b_k|}{(\boldsymbol{A}\boldsymbol{w})_k} \, . \tag{10}$$

The NNLS algorithm is an iterative procedure reminiscent of the Richardson-Lucy algorithm. If the initial weights are non-negative, then they will remain non-negative throughout the iterations, because the correction factors $|b_k|/(\boldsymbol{A}\boldsymbol{w})_k$ are non-negative. Typically, we start with uniform weights. There is no algorithmic parameter except for the number of iterations. Instead of fixing the number of iterations, we stop when the change in the least-squares loss (9) is negligible.

2.4 Expectation maximization

The iterative NNLS solver assumes that the particle positions are known. We will next discuss two grid-free methods that also estimate the particle positions. The observation that our model (3) is a finite mixture of isotropic Gaussians suggests a simple algorithm for learning the particle weights and positions as well as the kernel bandwidth. To do so, we approximately represent the information contained in the n-th projection image by a 2D point cloud. The positions of the 2D points are the centers of M randomly selected pixels. The image intensities g_{nm} enter this subsampling process as probabilities according to which pixels are selected. To account for a homogeneous background, we also introduce a threshold θ such that the probability of choosing the m-th pixel of the n-th projection image is proportional to $\max\{0, g_{nm} - \theta\}$ (i.e. pixels with an intensity smaller than θ cannot be chosen).

Our statistical model for the 2D point cloud representing the n-th projection image is the mixture of 2D spherical Gaussian densities (Eq. 6) obtained by projecting the volume (3) using a parallel-beam geometry:

$$\boldsymbol{y}_{nm} \sim \sum_{k=1}^{K} w_k \, \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k), \quad m = 1, \ldots, M \tag{11}$$

where M is a suitable number of representative pixels per projection image. When choosing M, we have to trade-off accuracy and computational complexity. The representation of a projection image by a cloud of M 2D points can be considered a subsampling step similar to subsampling used in, for example, stochastic gradient descent.

The log likelihood of all sampled pixels $\{\boldsymbol{y}_{nm}\}_{n=1,m=1}^{N,M}$ is

$$L(\{\boldsymbol{x}_k\}, \{w_k\}, \sigma) = \sum_{n=1}^{N} \sum_{m=1}^{M} \log \sum_{k=1}^{K} w_k \, \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k)\,. \tag{12}$$

Using standard arguments (Jensen inequality), we can derive an Expectation Maximization (EM) algorithm to iteratively learn the model parameters:

1. *E-step:* Compute the probability p_{nmk} that pixel position $\boldsymbol{y}_{nm}$ has been generated from the k-th particle:

$$p_{nmk} = \frac{w_k \, \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k)}{\sum_{k'} w_{k'} \, \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_{k'})}\,. \tag{13}$$

2. *M-step:* Compute the summary statistics $\hat{\boldsymbol{y}}_k \in \mathbb{R}^3$ and $\boldsymbol{A}_k \in \mathbb{R}^{3\times 3}$ according to

$$\hat{\boldsymbol{y}}_k = \sum_{n=1}^{N} \boldsymbol{R}_n^T \boldsymbol{P}^T \sum_{m=1}^{M} p_{nmk} \boldsymbol{y}_{nm}, \quad \boldsymbol{A}_k = \sum_{n=1}^{N} \boldsymbol{R}_n^T \boldsymbol{P}^T \boldsymbol{P} \boldsymbol{R}_n \sum_{m=1}^{M} p_{nmk}$$

and update the model parameters as well as the bandwidth:

$$\boldsymbol{x}_k = \boldsymbol{A}_k^{-1} \hat{\boldsymbol{y}}_k, \quad w_k = \frac{\sum_{n,m} p_{nmk}}{N\,M}, \quad \sigma^2 = \frac{\sum_{n,m,k} p_{nmk} \|\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n \boldsymbol{x}_k\|^2}{2\,M\,N}. \tag{14}$$

In each iteration, we resample the pixel positions $\boldsymbol{y}_{nm}$ that are used in the E- and M-step. This approach is reminiscent of stochastic optimization methods based on subsampling such as stochastic gradient descent. The larger the number of representative pixels M, the more faithful is the representation of the information of the projection images. We typically start with a fairly small M and increase M in the course of the EM iterations until M is in the order of K (number of basis functions / particles).

2.5 Bayesian refinement with Hamiltonian Monte Carlo

Finally, we apply a grid-free Bayesian reconstruction algorithm that aims to compute reconstructions from random tomography images acquired under unknown projection directions (Vakili & Habeck, 2021). Here, we are dealing with a simpler reconstruction problem, because the projection directions are known in the parallel-beam setup. Moreover, we assume equal weights $w_k = 1$ such that the particle positions $\{\boldsymbol{x}_k\}_{k=1}^K$ are the only free parameters. The reconstruction is based on the posterior distribution

$$\Pr(\{\boldsymbol{x}_k\}, \xi \mid D) \propto \Pr(D \mid \{\boldsymbol{x}_k\}, \xi)\ \Pr(\{\boldsymbol{x}_k\}, \xi) \tag{15}$$

where ξ denotes all additional parameters and D symbolizes all projection images $\{g_n\}_{n=1}^N$ and their associated rotation matrices $\{\boldsymbol{R}_n\}_{n=1}^N$.

The posterior distribution (15) is composed of two factors, the likelihood function $\Pr(D \mid \{\boldsymbol{x}_k\}, \xi)$ and the prior $\Pr(\{\boldsymbol{x}_k\}, \xi)$. To assign the likelihood function, we first need to establish a forward model for the projection images g_n:

$$g_{nm} \approx \alpha_n + \gamma_n \sum_{k=1}^{K} \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n \boldsymbol{x}_k) \tag{16}$$

where α_n is an unknown offset and γ_n a scaling factor. Relation (16) holds only approximately due to noise and possible model mismatch. Here, we assume Gaussian

pixel-wise independent noise of variance τ_n^{-1}. The probability of observing the n-th projection image is therefore

$$\Pr(g_n \mid \{\boldsymbol{x}_k\}, \gamma_n, \alpha_n, \tau_n) = \left(\frac{\tau_n}{2\pi}\right)^{M_n/2} \exp\left\{-\frac{\tau_n}{2}\sum_{m=1}^{M_n}\left[g_{nm} - \alpha_n - \gamma_n \sum_{k=1}^{K} \phi_{2,\sigma}(\boldsymbol{y}_{nm} - \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k)\right]^2\right\}. \tag{17}$$

There are three nuisance parameters per image: α_n (offset), γ_n (scaling factor) and τ_n (precision); these parameters are unknown and need to be estimated in addition to the particle positions. We combine the nuisance parameters from all images in a single symbol ξ. The likelihood function $\Pr(D \mid \{\boldsymbol{x}_k\}, \xi)$ is the product of all probabilities (17).

To set up the prior, we assume the factorization $\Pr(\{\boldsymbol{x}_k\}, \xi) = \Pr(\{\boldsymbol{x}_k\})\ \Pr(\xi)$ and use uniform priors for $\log \tau_n$, γ_n and α_n. To motivate the prior over the kernel centers $\boldsymbol{x}_k$, we refer to our physical interpretation of the kernel representation according to which the 3D structure arises from a cloud of particles. The particles have a finite size and should not occupy the same regions in space (excluded volume). We implement these restrictions via the Boltzmann distribution

$$\Pr(\{\boldsymbol{x}_k\}) \propto \exp\{-\beta E(\{\boldsymbol{x}_k\})\} \tag{18}$$

with a Lennard-Jones potential energy:

$$E(\{\boldsymbol{x}_k\}) = 4\epsilon \sum_{k<k'} \left(\frac{\sigma}{\|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\|}\right)^{12} - \left(\frac{\sigma}{\|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\|}\right)^{6} \tag{19}$$

that imposes a long-range attraction between particles resulting in compact structures, but enforces volume exclusion via a repellent term (Vakili & Habeck, 2021).

To sample the particle positions, we use Hamiltonian Monte Carlo (HMC) (Neal, 2011). The conditional posterior distribution over particle positions is

$$\Pr(\{\boldsymbol{x}_k\} \mid \xi, D) \propto \Pr(D \mid \{\boldsymbol{x}_k\}, \xi)\ \Pr(\{\boldsymbol{x}_k\})\,.$$

In HMC, $-\log \Pr(\{\boldsymbol{x}_k\} \mid \xi, D)$ defines a potential energy over configuration space that is composed of an attractive term $-\log \Pr(D \mid \{\boldsymbol{x}_k\}, \xi)$ matching particle positions to the projection data, and a repulsive contribution $-\log \Pr(\{\boldsymbol{x}_k\})$ stemming from the excluded-volume term (19). For fixed rotations and nuisance parameters ξ, the particle positions undergo Hamiltonian dynamics following the gradient of $-\Pr(\{\boldsymbol{x}_k\} \mid \xi, D)$ during a short leapfrog integration. The resulting configuration is accepted or rejected according to the Metropolis criterion. More details can be found in Vakili & Habeck (2021).

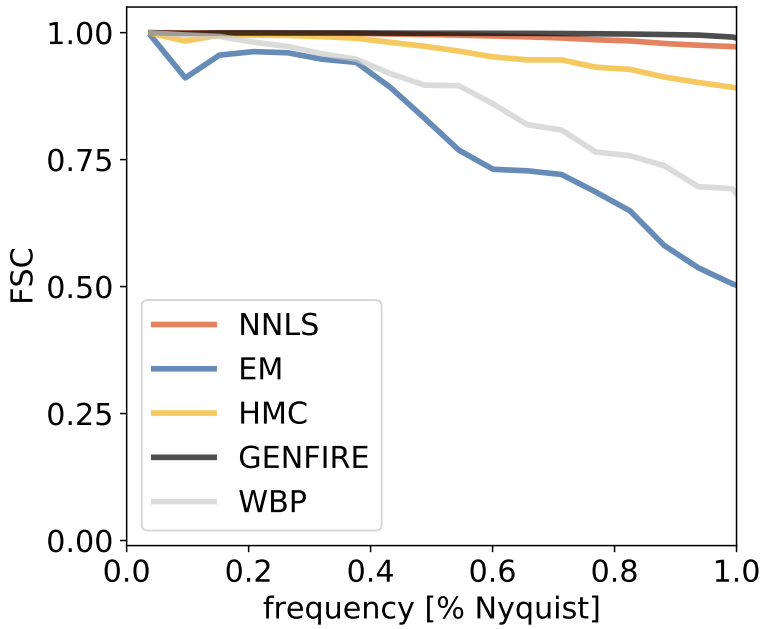


Figure 1: FSC curves of reconstructions obtained from a complete, noise-free tilt series of the vesicle structure.

3 Results

We tested our kernel-based reconstruction algorithms on simulated and real data and compared it to other reconstruction methods.

3.1 Reconstruction from a complete noise-free tilt series of a biological vesicle

As a first test, we ran the reconstruction algorithms outlined in Methods on the simulated tilt series of a biological vesicle. The phantom as well as the simulation of the tilt series was obtained with the Python version of the GENFIRE package. The first set of simulated data covers the entire range of angles between $\pm 90^{o}$ sampled with an angular increment of 2^{o} (91 projection images in total) and is free of noise.

3D reconstructions from the complete, noise-free data set provide a baseline for the reconstruction algorithms proposed in this article under perfect conditions. We first ran the iterative NNLS algorithm using fixed kernel centers that we placed on a hexagonal grid. The distance to the nearest neighbors was chosen to be identical to the pixel size. The kernel bandwidth was set to $\sigma = 0.4\times$ pixel size.

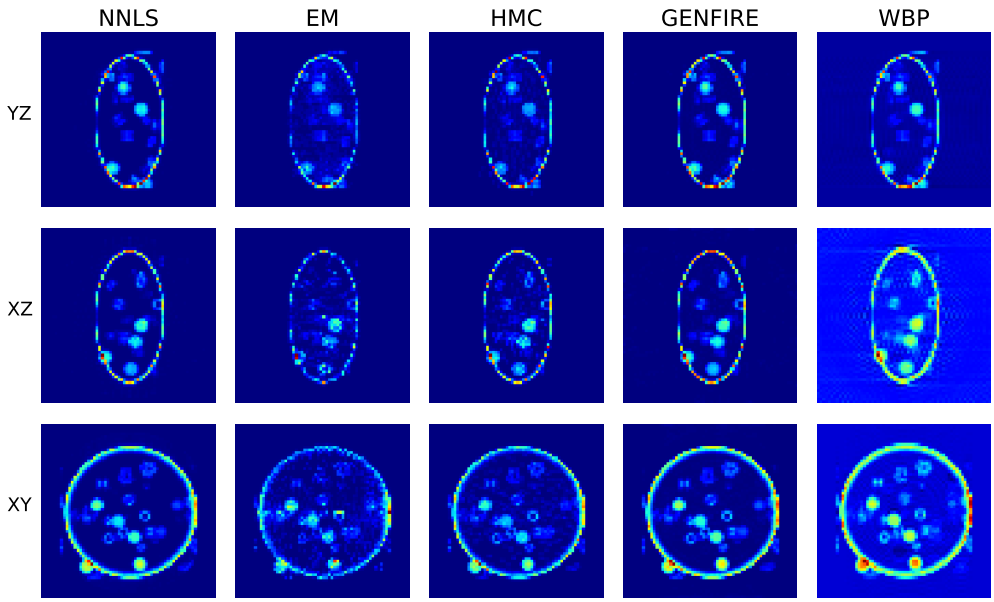


Figure 2: Central slices (thickness 10 voxels) of the reconstructions computed with a complete, noise-free tilt series of a vesicle.

We assess the quality of the reconstruction via FSC curves (Fig. 1). The NNLS approach achieves an accuracy that is comparable to the accuracy obtained with GENFIRE for almost all frequencies. A slight decrease in the correlation with the ground truth is only observed at high frequencies. Notice that GENFIRE was used to simulate the tilt series and therefore has an advantage in that it uses the correct forward model, whereas our kernel-based approach uses a different forward model. The near perfect quality of the NNLS reconstruction illustrates that the RBF model has the capacity to represent 3D structures accurately. Moreover, the NNLS approach uses a total of 112931 RBF kernels which is less than half the number of voxels ($64^3 = 262144$) used by GENFIRE. Therefore, even though the number of adjustable parameters is significantly reduced, NNLS achieves almost the same accuracy as GENFIRE.

We also ran the grid-free kernel approaches, EM and HMC, with $K = 5000$ particles. Compared to the NNLS approach with fixed RBF centers, the number of adjustable parameters ($5000 \times 4 = 20000$) is reduced by a factor of 5/10 compared to NNLS/GENFIRE. As shown by the FSC analysis, the reconstructions obtained with the two mesh-free methods are not as accurate as the NNLS reconstruction, but still of a very high quality. For comparison, we also computed a reconstruction with the

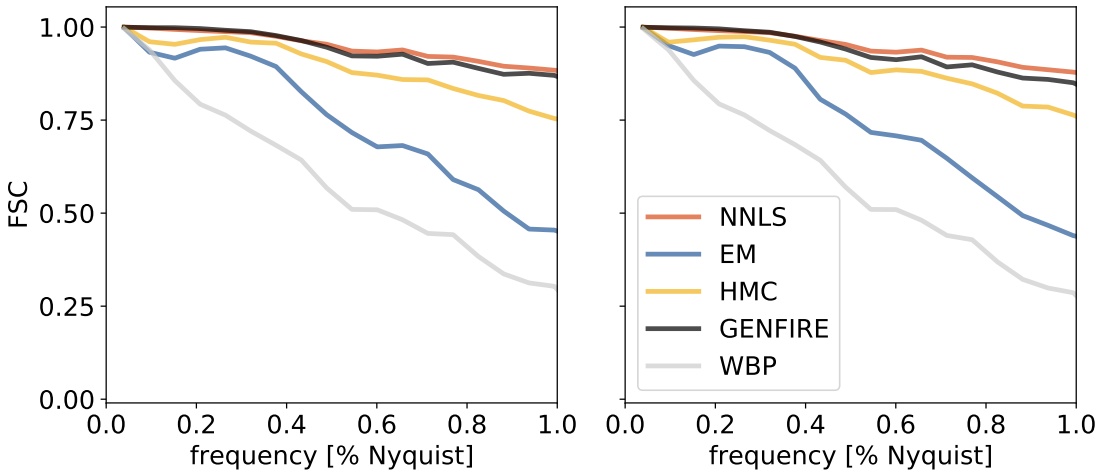


Figure 3: FSC curves of reconstructions obtained from incomplete, noise-free tilt series of the vesicle structure. Left panel: Results obtained with 71 projection images (angular increment is 2^{o}). Right panel: Results obtained with 41 projection images (angular increment is 3.5^{o}.

weighted back-projection (WBP) algorithm. The EM approach is less accurate than WBP, whereas HMC clearly outperforms WBP, even though the number of adjustable parameters is an order of magnitude smaller. Central slices of the reconstructions are shown in Fig. 2.

3.2 Reconstruction from an incomplete noise-free tilt series suffering from a missing wedge

Next, we looked at two tilt series that cover projection angles between $\pm 70^{o}$ and are therefore suffering from a missing wedge. The first series uses an angular increment of 2^{o} resulting in 71 projection images. The second tilt series uses an angular increment of 3.5^{o} resulting in 41 projection images. Again, the projection images where generated with the GENFIRE software.

We ran all reconstruction algorithms using the same settings as in the previous test. The FSC analysis (Fig. 3) suggests that in the presence of a missing wedge, the kernel-based reconstruction methods start to outperform existing methods. NNLS achieves a slightly better accuracy than GENFIRE, and the gap widens at high frequencies with increasing data sparseness. Both grid-free methods are more accurate than WBP in

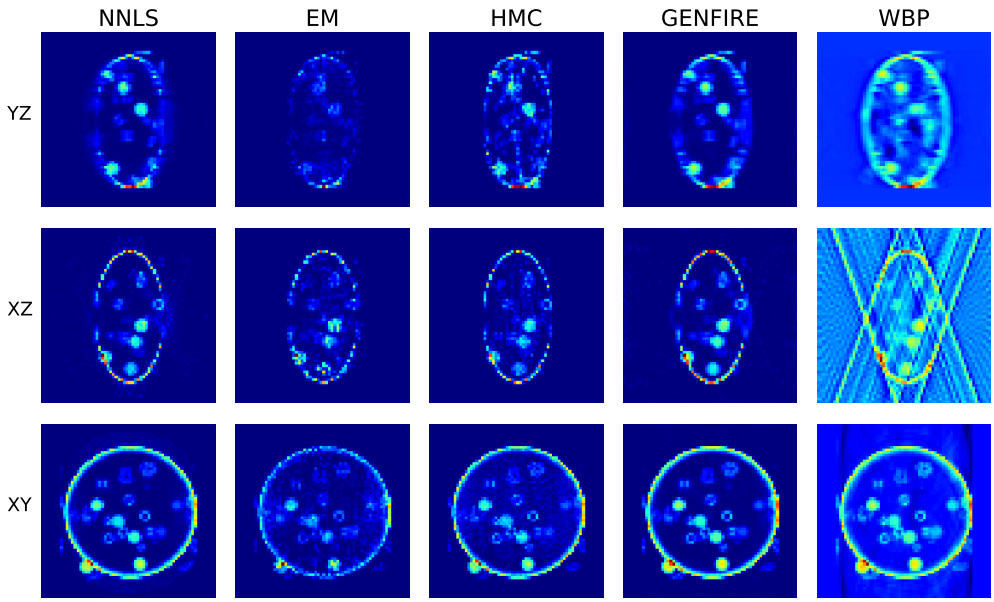


Figure 4: Central slices (thickness 10 voxels) of the reconstructions computed with an incomplete, noise-free tilt series between $\pm 70^{o}$ sampled with an angular increment of 3.5^{o}.

the presence of a missing wedge, where the reconstruction by HMC reaches almost the same accuracy as GENFIRE and NNLS.

The central slices of the reconstructions obtained with the sparser data set (41 projection images) are shown in Fig. 4. The central slice in the XY plane is only weakly affected in all reconstructions, because the missing wedge points in the z-direction. However, the central slices in the ZY and XZ plane show missing wedge artifacts for WBP and to a small extent also for GENFIRE where the YZ slice appears to be slightly blurred.

The analysis of a tilt series that suffers from an extended missing wedge is shown in Fig. 5. The tilt series comprises 21 projection images that cover a tilt range of $\pm 15^{o}$ sampled with an angular increment of 1.5^{o}. The FSC analysis shows all reconstruction methods fail to recover high-resolution details. The grid-based methods GENFIRE and NNLS show the best FSC curves. The grid-free reconstruction obtained with HMC is only slightly worse. WBP shows the worst FSC curve. However, the central slices reveal that all grid-based approaches suffer from severe streak artifacts where the kernel-based reconstruction obtained with NNLS seems be less affected than the struc-

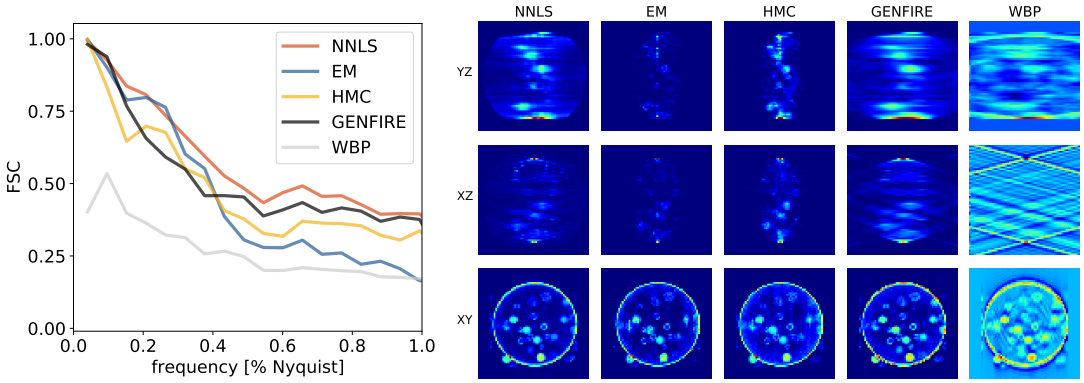


Figure 5: Reconstructions obtained from a noise-free tilt series suffering from a large missing wedge. Left: FSC curves of reconstructions obtained from an incomplete and noisy tilt series. Orientations between $\pm 15^{o}$ were sampled with an angular increment of 1.5^{o}. Right: Central slices through tomographic reconstructions.

ture obtained with GENFIRE and WBP. Although the FSC curves favor the grid-based reconstructions, the grid-free approaches do not exhibit an elongation in the missing direction. This is clear from the volumetric representations shown in Fig. 6.

Figure 7 shows the particle models computed with the grid-free Bayesian approach using HMC to sample particle positions. Notice that this representation is meaningful, because all particles have the same weight (or occupancy). The particle models accurately capture the shape of the vesicle phantom even with a restricted tilt range. We do not observe streak artifacts arising from the missing wedge even for the model obtained from a very narrow tilt range.

3.3 Reconstruction from a simulated noisy tilt series of a biological vesicle

We also studied the effect of noise by adding pixel-wise independent Gaussian noise to the simulated tilt series covering a range of $\pm 70^{o}$ sampled with an angular increment of 3.5^{o} (41 projection images). The signal-to-noise ratio (SNR) was set to one. Figure 8 shows the FSC curves and central slices. GENFIRE outperforms all other reconstruction methods, because it incorporates a denoising step, whereas the other reconstruction methods do not include such a step.

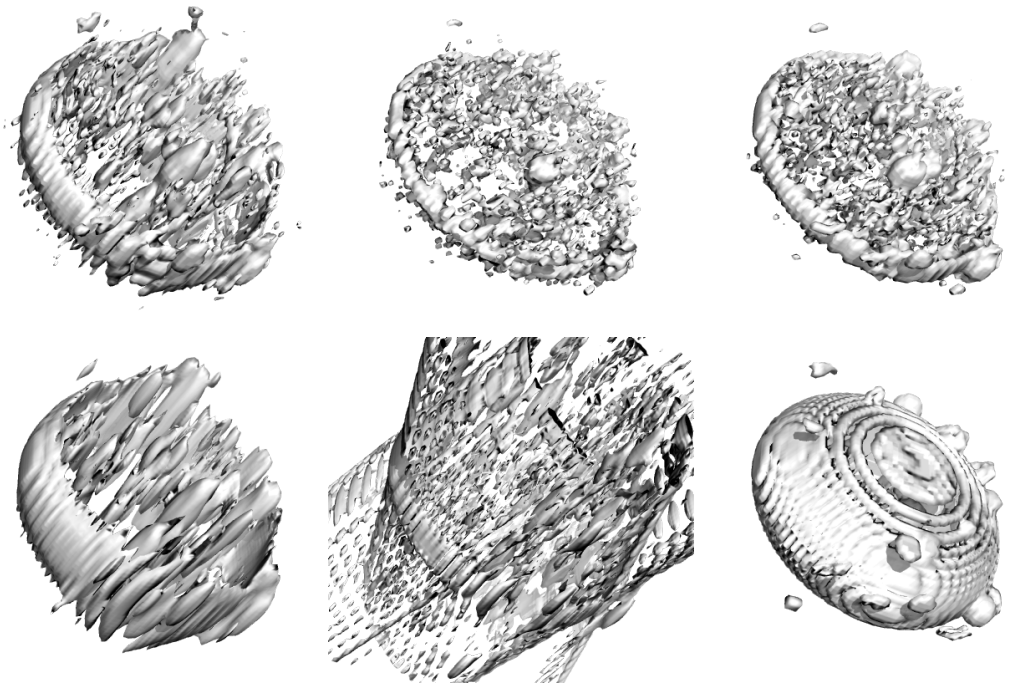


Figure 6: Density maps obtained with various reconstruction algorithms from a tilt series covering a range of $\pm 15^{o}$ sampled with an angular increment of 1.5^{o}. Top row (from left to right): reconstructions obtained with kernel-based methods NNLS, EM, and HMC. Bottom row: The left panel shows the reconstruction obtained with GENFIRE. Middle panel: WBP result. The right panel shows the ground truth (vesicle phantom).

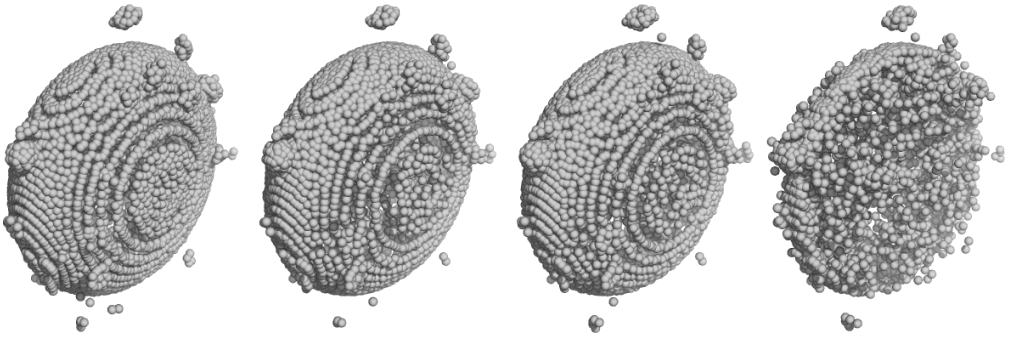


Figure 7: Particle models computed with HMC. The 3D positions of 5000 particles were adapted with HMC so as to match a simulated tilt series of a vesicle phantom. Left: Complete tilt series (tilt range $\pm 90^{o}$, increment 2^{o}). Center left: Tilt range $\pm 70^{o}$, increment 2^{o}. Center right: Tilt range $\pm 70^{o}$, increment 3.5^{o}. Right: Tilt range $\pm 15^{o}$, increment 1.5^{o}.

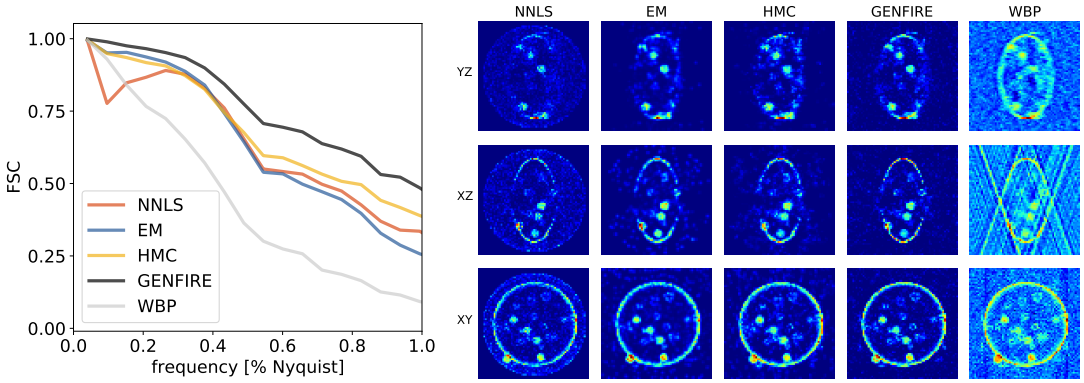


Figure 8: Left: FSC curves of reconstructions obtained from an incomplete and noisy tilt series. Orientations between $\pm 70^{\circ}$ were sampled with an angular increment of 3.5°; the SNR is one. Right: Central slices of of the reconstructions obtained with kernel-based reconstruction methods, WBP and GENFIRE.

3.4 Application to STEM data

We applied the kernel-based reconstruction algorithms to real STEM data measured on Co_2P nanocrystals.[1] The tilt range spans 150° with an angular increment of 2°. Each of the 74 projection images has size 1157×1157 pixels; the pixel size is 0.71 nm. The tilt axis points in y-direction and the specimen was rotated about the z-direction. We resized the images to a size of 136×136 pixels. Again, we set the number of particles to 5000.

Figure 9 shows particle models obtained with the kernel-based reconstruction methods. All reconstructions exhibit a star-shaped volume. In case of HMC, we observe particles that are ejected from the main body of the reconstruction due to the repulsive excluded-volume forces.

To assess our models, we compare them to a SIRT reconstruction (Levin et al., 2016), which is available together with the tilt series. We down-sampled the reconstruction to a size of $136 \times 136 \times 136$ and computed the cross-correlation coefficients (CCC) between the SIRT reconstruction and our models. The NNLS approach achieves the highest value with a cross-correlation of 95.7%. The grid-free methods show also a high agreement with the SIRT reconstruction: 93.7% (EM) and 92.8% (HMC).

[1]The data set was downloaded from `https://dx.doi.org/10.6084/m9.figshare.c.2185342`.

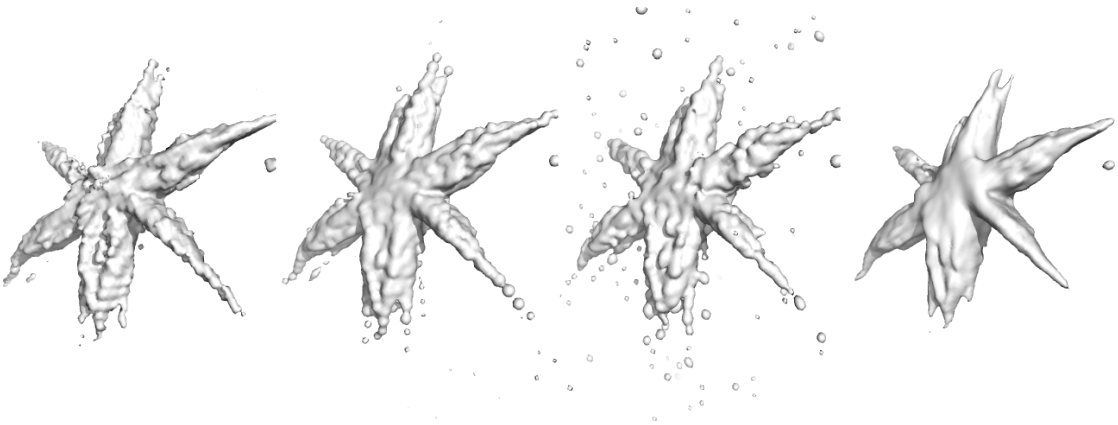


Figure 9: Density maps of a hyperbranched Co_2P nanoparticle obtained with various reconstruction algorithms from a STEM data set. Left to middle right panels: Reconstructions obtained with kernel-based methods NNLS, EM, and HMC. Right panel: SIRT reconstruction.

4 Conclusion

In this article, we propose three kernel-based algorithms for tomographic reconstruction from a tilt series acquired in the parallel-beam setup. Our first reconstruction algorithm uses fixed kernel centers that were placed on a hexagonal grid in our tests. The only adjustable parameters are the particle weights that are fitted with an iterative non-negative least-squares approach. The other two reconstruction methods adjust the particle positions and are therefore grid-free methods. The first of the two grid-free methods is an Expectation Maximization algorithm that estimates the particle positions and weights as well as the bandwidth of the Gaussian kernel function. The second grid-free algorithm uses a Bayesian formulation of the reconstruction problem and adjusts the particle positions with Hamiltonian Monte Carlo, whereas the weights are set to one. In addition to a data fitting term (the likelihood function), the Bayesian model also incorporates a prior that imposes excluded volume interactions between the particles by means of a Lennard-Jones term.

The kernel representation has several advantages. It allows for a fast evaluation of the forward model without any interpolation. Moreover, the kernel representation can be interpreted as a particle system, which is particularly suited to model molecular structures. When working with a small number of particles, the kernel representation

has a built-in sparsity. Other constraints that are automatically enforced by the kernel representation are smoothness and positivity.

We tested the kernel-based reconstruction algorithms on a simulated and real data set. Our tests show that the kernel representation can achieve the same accuracy as a voxel-grid representation. In the presence of a missing wedge, the grid-free reconstructions using a limited number of particles are less susceptible to streak artifacts along the direction of the missing wedge. Reconstructions from noisy data are not as accurate as structures obtained with state-of-the art methods such as GENFIRE. So there is the need to make the kernel-based algorithms more robust. Other future directions that could be pursued are the use of a multi-scale representation and a faster implementation. For the grid-free algorithms, it would be interesting to develop strategies for estimating the number of particles from the data.

Funding

MH acknowledges funding from the German Research Foundation (DFG) under project SFB 860, TP B09 as well as funding from the Carl Zeiss Foundation.

References

Andersen, A. H. and Kak, A. C. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrasonic imaging*, 6(1): 81–94, 1984.

Bishop, C. M. et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

Fessler, J. A. and Sutton, B. P. A min-max approach to the multidimensional nonuniform FFT: Application to tomographic image reconstruction. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 1, pp. 706–709. IEEE, 2001.

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

Friedman, J., Hastie, T., Tibshirani, R., et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Gilbert, P. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theoretical Biology*, 36(1):105–117, 1972.

Gordon, R., Bender, R., and Herman, G. T. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology*, 29(3):471–481, 1970.

Herman, G. T. Image reconstruction from projections. *The fundamental of computerized tomography*, pp. 260–276, 1980.

Herman, G. T. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer Science & Business Media, 2009.

Kak, A. C. and Slaney, M. *Principles of computerized tomographic imaging*. SIAM, 2001.

Leary, R., Midgley, P. A., and Thomas, J. M. Recent advances in the application of electron tomography to materials chemistry. *Accounts of chemical research*, 45(10): 1782–1791, 2012.

Levin, B. D., Padgett, E., Chen, C.-C., Scott, M., Xu, R., Theis, W., Jiang, Y., Yang, Y., Ophus, C., Zhang, H., et al. Nanomaterial datasets to advance tomography in scanning transmission electron microscopy. *Scientific Data*, 3(1):1–11, 2016.

Matej, S., Fessler, J. A., and Kazantsev, I. G. Iterative tomographic image reconstruction using Fourier-based forward and back-projectors. *IEEE transactions on medical imaging*, 23(4):401–412, 2004.

Miao, J., Förster, F., and Levi, O. Equally sloped tomography with oversampling reconstruction. *Physical Review B*, 72(5):052103, 2005.

Miao, J., Ercius, P., and Billinge, S. J. Atomic electron tomography: 3D structures without crystals. *Science*, 353(6306), 2016.

Midgley, P. A. and Dunin-Borkowski, R. E. Electron tomography and holography in materials science. *Nature Materials*, 8(4):271–280, 2009.

Natterer, F. *The mathematics of computerized tomography*. SIAM, 2001.

Neal, R. M. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, pp. 113–162, 2011.

O'Connor, Y. and Fessler, J. A. Fourier-based forward and back-projectors in iterative fan-beam tomographic image reconstruction. *IEEE transactions on medical imaging*, 25(5):582–589, 2006.

Park, J. and Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

Pham, M., Yuan, Y., Rana, A., Miao, J., and Osher, S. RESIRE: real space iterative reconstruction engine for Tomography. *arXiv preprint arXiv:2004.10445*, 2020.

Pryor, A., Yang, Y., Rana, A., Gallagher-Jones, M., Zhou, J., Lo, Y. H., Melinte, G., Chiu, W., Rodriguez, J. A., and Miao, J. GENFIRE: A generalized Fourier iterative reconstruction algorithm for high-resolution 3D imaging. *Scientific Reports*, 7(1):1–12, 2017.

Radermacher, M. Weighted back-projection methods. In *Electron tomography*, pp. 245–273. Springer, 2007.

Saghi, Z. and Midgley, P. A. Electron tomography in the (S)TEM: from nanoscale morphological analysis to 3D atomic imaging. *Annual Review of Materials Research*, 42: 59–79, 2012.

Sha, F., Lin, Y., Saul, L. K., and Lee, D. D. Multiplicative Updates for Nonnegative Quadratic Programming. *Neural Comput.*, 19(8):2004–2031, 2007. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.2007.19.8.2004.

Tian, X., Kim, D. S., Yang, S., Ciccarino, C. J., Gong, Y., Yang, Y., Yang, Y., Duschatko, B., Yuan, Y., Ajayan, P. M., et al. Correlating the three-dimensional atomic defects and electronic properties of two-dimensional transition metal dichalcogenides. *Nature Materials*, 19(8):867–873, 2020.

Vakili, N. and Habeck, M. Bayesian random tomography of particle systems. *Frontiers in Molecular Biosciences*, 8:399, 2021.

Yang, Y., Zhou, J., Zhu, F., Yuan, Y., Chang, D. J., Kim, D. S., Pham, M., Rana, A., Tian, X., Yao, Y., et al. Determining the three-dimensional atomic structure of an amorphous solid. *Nature*, 592(7852):60–64, 2021.

Chapter 4

Matching biomolecular structures by registration of point clouds

This chapter is the third research result from my Ph.D. study. Here we develop a kernel correlation to compute the rigid transformation between two molecular structures. This manuscript is in preparation. Own contribution:

- Concept and implementation of the algorithm and the code.
- All figures, tables.
- Manuscript in parts.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Research Article

OXFORD

Structural Bioinformatics

Matching biomolecular structures by registration of point clouds

Nima Vakili [1,2] and Michael Habeck [1,2]*

[1] Microscopic Image Analysis Group, Jena University Hospital, Jena, Germany and
[2] Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Assessing the match between two biomolecular structures is at the heart of many structural analyses, including superposition, alignment and docking. These tasks are typically solved with specialized structure-matching techniques implemented in software for protein structural alignment, rigid-body docking, or rigid fitting into cryo-EM maps.
Results: This article presents a unifying framework to compare biomolecular structures by applying ideas from computer vision. The structures are represented as three-dimensional point clouds and compared by measuring the overlap of the point clouds. We use the kernel correlation to measure point cloud overlap, and discuss local and global optimization methods for maximizing the kernel correlation over the space of rigid transformations. We derive a majorization-minimization procedure that can be used to register two point clouds without establishing a point-to-point correspondence. We demonstrate that the majorization-minimization algorithms outperform the commonly used Iterative Closest Point registration algorithm. We illustrate the approach on various 3D fitting problems such as the comparison of circularly permuted structures and rigid fitting of cryo-EM maps or bead models from small-angle scattering.
Availability:
Contact: michael.habeck@uni-jena.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Superposition and comparison of biomolecular structures are common tasks in structural biology. Structures are compared and aligned to predict the conformation and function of proteins, for example by homology modeling. To assess the conformational diversity of structures revealed by NMR or other structure determination methods, multiple conformations need to be superimposed in a meaningful fashion. Rigid docking approaches optimize the match between an experimental shape revealed by cryo-electron microscopy or solution scattering with a known structure.

Aside from manual superposition approaches requiring user intervention, the standard approach to superimpose and compare different conformations of the same biomolecule is to minimize their root-mean-square deviation (RMSD). To compute the RMSD, six rigid degrees of freedom are determined by a singular value decomposition (Kabsch, 1976). This assumes a least-squares criterion for assessing the match between two structures.

Robust variants of the RMSD take account of the fact that the RMSD will be less meaningful if the structures undergo large conformational changes (Hirsch and Habeck, 2008; Mechelke and Habeck, 2010) or show varying degrees of structural heterogeneity (Theobald and Wuttke, 2006). Variants of the standard least-squares superposition such as the weighted RMSD (Damm and Carlson, 2006) have been proposed to tackle more challenging superposition tasks.

Optimization of the (weighted) RMSD is analytically feasible, but restricted to cases where the correspondence between the positions in both structures is known. This requirement holds, for example, for NMR ensembles. In the more general case, the correspondence between positions in both structures is unknown. Both problems, 3D superimposition and establishing a correspondence between positions, are intertwined and cannot be solved independently of each other.

This situation occurs in structural alignment, where we need to solve two problems: First, establish the correspondence between evolutionary related amino acids. Second, find the best superposition of the three-dimensional coordinates such that related amino acids are close in space. However, the approach is restricted to situations where the correspondence is consistent with the sequential order of the amino acids (such that dynamic programming approaches can used to solve the alignment problem). Already when dealing with circularly permuted structures, a standard alignment approach is no longer applicable.

When working with reconstructions from cryo-electron microscopy (cryo-EM), we need to dock high-resolution structures into 3D density maps that might only achieve an intermediate or low resolution (Villa and Lasker, 2014). Rigid docking could either work by using a voxel-based representation of both structures (which is the standard representation of 3D reconstructions from cryo-EM), or a particle-based representation such as an atomic structure or bead model. In the first case, the high-resolution structure needs to be converted to a density map. In the second case, the density map has to be converted to a (pseudo)-atomic structure. Kawabata (2008) introduced a decomposition of cryo-EM density maps into a mixture of anisotropic Gaussians that are characterized by a center position, a weight and a covariance matrix (corresponding to a ellipsoidal shape). Omokage search (Suzuki *et al.*, 2016) uses this representation to rapidly compare the overall shape of two or more structures obtained with X-ray crystallography, cryo-EM or small-angle scattering (SAS).

Here, we approach all of the mentioned comparison and superposition tasks within a common framework. We use a particle-based representation of biomolecular structures including atomic structures, cryo-EM density maps, and bead models from SAS. To compare two biomolecular structures, we use the *kernel correlation* which has been introduced in computer vision to register point clouds. We discuss local and global strategies to optimize the kernel correlation over rigid transformations of one structure against the other structure. Finally, we illustrate our approach on various comparison and superposition tasks.

2 Methods

2.1 Representation of biomolecular structures by weighted point clouds

We first need to find a common representation for biomolecular structures from different experimental sources. The most widespread and detailed representation of a biomolecular structure is an array of three-dimensional atomic positions that can be obtained from the PDB (Berman *et al.*, 2000), but other representations are also relevant. Cryo-EM typically reconstructs three-dimensional volumes represented as voxelized density values over a cubic grid. Structural manipulations of 3D volumes such as rigid transformation necessitate the interpolation of density values, which is time consuming and prone to artifacts. Therefore, we will represent volumetric data as weighted point clouds (Vakili and Habeck, 2021).

A weighted point cloud comprises a collection of N points in 3D space at positions $\boldsymbol{x}_i$. The positions can be stored in an $N \times 3$ matrix $\boldsymbol{X}$ whose rows are the 3D coordinates of the points. Each point has an associated weight p_i; the weights form an N-dimensional vector $\boldsymbol{p}$.

This representation is very natural for atomic structures where the points are representative atoms such as alpha carbons and weights can be chosen proportional to the atomic mass or occupancy. In case of large structures such as multi-domain proteins or macromolecular complexes, we may want to reduce the size of the point cloud by coarse graining. In this case, a single point represents more than one amino acid.

A simple and powerful algorithm to obtain weighted point clouds from atomic structures as well as cryo-EM reconstructions is DP-means (Kulis and Jordan, 2012), a non-parametric variant of the K-means algorithm. In DP-means, the radius of the bead particle is chosen by the user and the optimal number of points is estimated automatically.

Point clouds derived from atomic structures are typically ordered in the same fashion as the underlying sequence of residues. However, in case of EM reconstructions or bead models from SAS there is no such order. To establish an order, we solve the traveling salesman problem using the distance matrix of the points as input. This will cause an ordering of the points where spatially close points have similar indices. Ordering of point clouds helps to compare them visually, but can also result in a speed-up of registration algorithms (see e.g. Rusu *et al.* (2009)).

2.2 Assessing the match between two point clouds by kernel correlation

Let us now compare two point clouds represented by position matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ and weight vectors $\boldsymbol{q}$ and $\boldsymbol{p}$. The number of points in both clouds will differ in general; let M denote the number of points in cloud $(\boldsymbol{X}, \boldsymbol{q})$ and N the number of points in $(\boldsymbol{Y}, \boldsymbol{p})$. Both point clouds are typically represented in different frames of reference. To allow for a meaningful comparison of their shapes, we first need to find a common frame of reference by rigidly transforming one of the two points in space (*rigid registration*). A rigid transformation involves a rotation encoded by a 3×3 matrix $\boldsymbol{R}$ and a translation represented by a vector $\boldsymbol{t}$. To find the best pose (encoded by the optimal $\boldsymbol{R}$ and $\boldsymbol{t}$), we need a quantitative measure of how well two point clouds match.

The *kernel correlation* (KC) between two point clouds viewed as a function of the rigid transformation is defined as (Tsin and Kanade, 2004):

$$\mathrm{KC}(\boldsymbol{R}, \boldsymbol{t}) = \sum_{i=1}^{M} \sum_{j=1}^{N} q_i \, p_j \, \phi(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|) = \boldsymbol{q}^T \boldsymbol{\Phi}(\boldsymbol{R}, \boldsymbol{t}) \, \boldsymbol{p} \quad (1)$$

where ϕ is a suitable kernel function, $\|\cdot\|$ the Euclidean norm and $\boldsymbol{\Phi}$ the $M \times N$ kernel matrix with elements $\phi_{ij} = \phi(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)$. Throughout this paper, we will use the Gaussian kernel

$$\phi_\sigma(r) = (2\pi\sigma^2)^{-3/2} \exp\left\{-\tfrac{1}{2\sigma^2} r^2\right\} \quad (2)$$

for point cloud comparison, but in principle we could also use a different kernel. The positive parameter σ is the *bandwidth* of the kernel and determines how tolerant the kernel correlation is against mismatches. For larger values of σ, $\mathrm{KC}(\boldsymbol{R}, \boldsymbol{t})$ is rather indifferent against variations in the transformation parameters, whereas small values of σ focus on common sub-structures that fit almost perfectly.

A convenient aspect of using the kernel correlation as a metric for point cloud comparison is that it does not require a correspondence between the points in both clouds. In fact, the kernel correlation is invariant under permutation of both point clouds. By optimizing KC as a function of the rigid body degrees of freedom, we can align two point clouds without establishing a point-to-point correspondence.

The kernel correlation can be motivated in several ways. The auto-correlation of the Gaussian kernel is

$$\int \phi_{\sigma_1}(\|\boldsymbol{x} - \boldsymbol{\mu}_1\|) \, \phi_{\sigma_2}(\|\boldsymbol{x} - \boldsymbol{\mu}_2\|) \, d\boldsymbol{x} = \phi_\sigma(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)$$

where $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. Due to this self-reproducing property of the Gaussian kernel, KC can be viewed as the inner product of two kernel density estimators (KDEs):

$$q(\boldsymbol{x}) = \sum_{i=1}^{M} q_i \, \phi_{\frac{\sigma}{\sqrt{2}}}(\|\boldsymbol{x} - \boldsymbol{x}_i\|), \quad p(\boldsymbol{x}) = \sum_{j=1}^{N} p_j \, \phi_{\frac{\sigma}{\sqrt{2}}}(\|\boldsymbol{x} - \boldsymbol{y}_j\|) \, .$$

The kernel correlation is the scalar product (or overlap) of both KDEs:

$$\mathrm{KC}(\boldsymbol{R},\boldsymbol{t}) = \int q(\boldsymbol{x})\, p\big(\boldsymbol{R}^T(\boldsymbol{x}-\boldsymbol{t})\big)\, d\boldsymbol{x}$$

where one of the two KDEs has been rigidly transformed. By maximizing the kernel correlation, we minimize the L2 distance between both kernel density estimates. The kernel correlation is also known as correntropy (Liu *et al.*, 2007) which has been used in various applications ranging from face recognition, visual tracking, and data fusion.

The Kpax algorithm by Ritchie (2016) uses a related match criterion for protein structure alignment. The major difference is that Kpax's objective function is not the total sum over all elements in the kernel matrix as in Eq. (1), but reduced to the aligned pairs of points in $\boldsymbol{X}$ and $\boldsymbol{Y}$. For general point cloud alignment, establishing an alignment is no longer suitable (think of circularly permuted protein structures or cryo-EM maps).

An important parameter is the kernel bandwidth σ: The smaller σ the rougher will be the kernel correlation. However, a large value of σ will result in very ambiguous registrations. Larger values of σ are more suitable for global registration, whereas a small σ value allows us to identify locally similar subsets of points in both clouds. In principle, σ is a free parameter that can be chosen by the user or by methods used in kernel density estimation (e.g. Silverman's rule or crossvalidation). In our applications, we typically set σ based on the fact that we are dealing with biomolecular structures. For example, when working with cryo-EM maps, the resolution of the map gives an estimate for the appropriate σ. For the comparison of alpha carbon clouds derived from atomic resolution structures, we typically use $\sigma = 5$ Å based on the following reasoning: We have $\sigma = 5 \approx \sqrt{2} \times 3.5$ Å where 3.5 Å is roughly the average distance between alpha carbons. In contrast, Ritchie (2016) uses a smaller bandwidth, $\sigma = \sqrt{2} \times 1.4$ Å for aligning protein structures.

2.3 Local optimization of the kernel correlation by iterative majorization-minimization (MM)

The negative logarithm of the kernel correlation, $-\log \mathrm{KC}(\boldsymbol{R},\boldsymbol{t})$, can be used as an objective function for finding the best rigid registration of two point clouds by minimization. Minimization of $-\log \mathrm{KC}$ is a six-dimensional non-convex optimization problem with many local minima. Global optimization of $-\log \mathrm{KC}$ therefore involves heuristics such as simulated annealing or branch-and-bound (Straub *et al.*, 2017).

A simple *local* optimization of $-\log \mathrm{KC}$ is achieved by constructing an upper bound which can be optimized easily. By iterating over updating the upper bound and subsequent upper bound minimization, we arrive at a Majorization-Minimization (MM) algorithm (Hunter and Lange, 2004). With the help of Jensen's inequality we have

$$\begin{aligned}
-\log \mathrm{KC}(\boldsymbol{R},\boldsymbol{t}) &= -\log \sum_{ij} q_i\, p_j\, \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|) \\
&= -\log \sum_{ij} q_i\, p_j\, w_{ij}\, \frac{\phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)}{w_{ij}} \\
&\le -\sum_{ij} q_i\, p_j\, w_{ij} \log \frac{\phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)}{w_{ij}} \\
&= \frac{1}{2\sigma^2} \sum_{ij} q_i p_j w_{ij} \|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|^2 + \mathrm{const}
\end{aligned}$$

where the constant, $\sum_{ij} q_i p_j w_{ij} \log w_{ij}$, does not depend on the parameters of the rigid transformation. The inequality is valid for all weights w_{ij} satisfying $\sum_{ij} q_i p_j w_{ij} = 1$. If we choose the weights to be proportional to the entries of the kernel matrix, $w_{ij} \propto \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)$, the upper bound touches $-\log \mathrm{KC}$ at $\boldsymbol{R}, \boldsymbol{t}$, and the inequality becomes an identity.

This suggests an iterative scheme to locally optimize the kernel correlation by cycling between updates of w_{ij} for given $\boldsymbol{R}$ and $\boldsymbol{t}$ followed by updates of $\boldsymbol{R}$ and $\boldsymbol{t}$ by minimizing the upper bound

$$U(\boldsymbol{R},\boldsymbol{t}) = \frac{1}{2\sigma^2} \sum_{ij} q_i p_j w_{ij} \|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|^2 \qquad (3)$$

where the weights w_{ij} are proportional to the entries of the kernel matrix evaluated at the current transformation and normalized such that $\sum_{ij} q_i p_j w_{ij} = 1$.

The solution of $\arg\min_{\boldsymbol{R},\boldsymbol{t}} U(\boldsymbol{R},\boldsymbol{t})$ is available in closed form. The optimal translation is:

$$\hat{\boldsymbol{t}} = \overline{\boldsymbol{x}} - \boldsymbol{R}\,\overline{\boldsymbol{y}} \qquad (4)$$

with centers of mass

$$\overline{\boldsymbol{x}} = \sum_{ij} q_i p_j w_{ij} \boldsymbol{x}_i, \quad \overline{\boldsymbol{y}} = \sum_{ij} q_i p_j w_{ij} \boldsymbol{y}_j\,.$$

The optimal rotation can be computed by solving the matrix nearness problem

$$\hat{\boldsymbol{R}} = \underset{\boldsymbol{R}\in SO(3)}{\arg\min} \|\boldsymbol{R} - (\boldsymbol{X} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T \boldsymbol{Q}\boldsymbol{K}\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{1}\overline{\boldsymbol{y}}^T)\|_F^2 \qquad (5)$$

where $\|\cdot\|_F$ indicates the Frobenius norm and $\boldsymbol{Q}$, $\boldsymbol{P}$ are diagonal matrices with diagonal elements q_i and p_j, respectively. Problem (5) can be solved by singular value decomposition of the 3×3 matrix $(\boldsymbol{X} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T \boldsymbol{Q}\boldsymbol{K}\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{1}\overline{\boldsymbol{y}}^T)$ (Higham, 1989).

The following iterative MM procedure minimizes the negative logarithm of the kernel correlation locally:

- Initialization: generate a random rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$ (alternatively, we can try to find good initial values by some heuristic)
- Iterate until convergence (e.g. when changes in $-\log \mathrm{KC}$ are no longer significant) or unit a maximum number of iterations has been reached:
  1. Evaluate the kernel matrix $\Phi_{ij} = \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)$ at the current pose $(\boldsymbol{R},\boldsymbol{t})$ and compute the normalized weights $w_{ij} = \Phi_{ij} / \sum_{i'j'} q_{i'} p_{j'} \Phi_{i'j'}$.
  2. Minimize the upper bound $U(\boldsymbol{R},\boldsymbol{t})$ by calculating the optimal rotation $\hat{\boldsymbol{R}}$ and translation $\hat{\boldsymbol{t}}$ according to equations (5) and (4).

This MM algorithm is analogous to Expectation Minimization where updating w_{ij} corresponds to the E-step and calculation of $(\hat{\boldsymbol{R}},\hat{\boldsymbol{t}})$ corresponds to an M-step. Supplementary Figure S1 illustrates the MM iterations for a specific example.

2.4 Deterministic annealing

Choosing the kernel width σ should not be seen as a burden, but as a means to incorporate prior knowledge and control the shape of the objective function $\mathrm{KC}(\boldsymbol{R},\boldsymbol{t})$. One of the most widely used methods for rigid point cloud registration is the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992; Chen and Medioni, 1992). Like our upper bound minimization approach, ICP is also an iterative algorithm, but lacks the bandwidth parameter. In ICP, the procedure analogous to our first step establishes a correspondence between points i in $\boldsymbol{X}$ and points j in $\boldsymbol{Y}$ by matching those pairs (i,j) whose distance $\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|$ is smallest. ICP's analog of our second step solves the least-squares fitting problem for all pairs of corresponding points. The only algorithmic parameter in ICP is the number of iterations.

In our approach, we can use the kernel bandwidth σ to gradually change the objective function. Because the iterative MM algorithm is only a local search strategy, it will strongly depend on the initial pose (the same is true

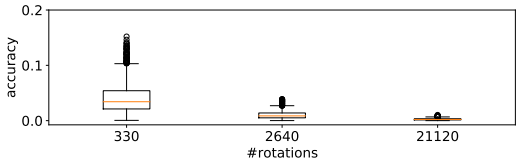


Fig. 1. Box-plot showing the accuracy achieved with tessellations of $SO(3)$ based on the 600-cell and its refinements. The accuracy is measured in terms of the matrix distances between 1000 random rotations and the closest rotation in the partition of rotation space. At the coarsest level of discretization, there are 330 rotations. At each higher level, the number of rotations covering $SO(3)$ increases eight-fold.

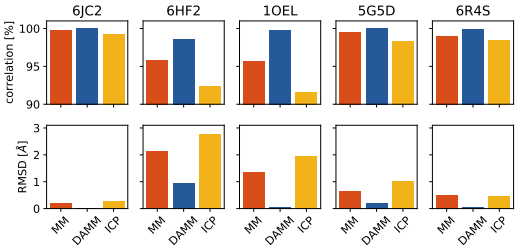


Fig. 2. Testing local optimization strategies on various self-matching problems. A PDB structure (indicated in panel titles) is fitted against a permuted and randomly transformed version of itself. The top row shows the correlation coefficient (ratio of actual kernel correlation and maximum achievable kernel correlation) obtained when starting local optimization runs from 10 random initial poses. The bottom row shows the RMSD (defined in equation 7) of corresponding points after striking the estimated pose.

for ICP). To avoid getting trapped in the nearest pose, we propose a simple modification of the algorithm reminiscent of deterministic annealing (Rose *et al.*, 1990). During the iterative minimization of the upper bound (3), we decrease the kernel bandwidth gradually until we reach the desired σ value. The bandwidth acts as a parameter similar to a temperature: Large σ values (high temperatures) result in a flat cost function with shallow minima, annealing (i.e. reduction of σ) makes the cost function rougher, but also more selective. In our tests, we chose a simple linear annealing schedule starting at a large $\sigma_{\max}$ (typically 15 Å or more generally 3σ) and decrease the kernel bandwidth by a constant increment in each iteration. There are many more options for the initial bandwidth and the progression of the annealing iterations. We found that the simple linear annealing schedule is sufficient for the registration problems considered in this article. We will use the abbreviation DAMM to denote the combination of iterative majorization-minimization and deterministic annealing.

2.5 Global optimization by a six-dimensional grid search

Maximization of the kernel correlation (1) over all six rotational and translational degrees of freedom is a challenging non-concave optimization problem. Also, the annealing strategy is not guaranteed to always find the best registration. But thanks to the low dimensionality of the search space, we can attempt to find the best pose in a systematic fashion by discretizing the space of all rotations and translations.

A simple approach to discretize the space of all translations is to restrict them to a cubic lattice of finite size. In principle, the search space for translations is the entire Euclidean space $\mathbb{R}^3$. However, if translations are larger than the sum of the maximum extensions of both point clouds, the kernel correlation is close to zero (provided that σ is not exceedingly large). Moreover, if both point clouds are centered, we expect the optimal translation to lie within in a cubic box centered at the origin. Therefore, we typically restrict the grid search to a box $[-L, L]^3$ where $L = 10\,\text{Å}$. In general, a reasonably accurate spacing of the grid cells should be chosen to depend on the bandwidth; in our tests, we use a spacing of $\sigma/2$.

For a fixed rotation $\boldsymbol{R}$, we can then scan all kernel correlations in a systematic fashion by observing that we have to evaluate a mixture of MN spherical Gaussians

$$f(\boldsymbol{t}) = \sum_{ij} q_i p_j \, \phi_\sigma(\|\boldsymbol{t} - \boldsymbol{z}_{ij}\|) \tag{6}$$

with component means $\boldsymbol{z}_{ij} = \boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j$, weights $q_i p_j$ and isotropic covariance matrix $\sigma^2 \boldsymbol{I}$. Since translations are restricted to a regular cubic grid, $f(\boldsymbol{t})$ can be evaluated very rapidly by limiting the support of each Gaussian to 5σ, say, in all three spatial directions and/or by using Fourier accelerations based on the Wiener-Khinchin theorem.

To discretize rotation space, we adopt a strategy proposed by Straub *et al.* (2017). We parameterize rotations by unit quaternions (Horn, 1987) and discretize 3-sphere, the unit sphere embedded in a four-dimensional space, by using the 600-cell, a 4D analog of the icosahedron. The 600-cell is composed of even sized tetrahedra whose corners lie on the unit sphere. By projecting the center of a tetrahedron onto the unit sphere, we obtain a unit quaternion that parameterizes a valid rotation matrix. Due to the degeneracy of the quaternions, we only have to consider the upper half of 3-sphere that is covered by 330 tetrahedra at the coarsest level of discretization. To obtain finer tessellations of the rotation group, we can split each tetrahedron into eight tetrahedra whose corners again lie on the unit sphere. By repeating this process, we can obtain finer and finer tessellations of the 3D rotation group.

Figure 1 shows the accuracy that different tessellations of the 600-cell can achieve. We generated 1000 random rotation matrices and measured the distance between each random rotation and the closest rotation matrix within the tessellation. As a distance measure, we used a multiple of the squared Frobenius distance: $1 - \text{tr}[\boldsymbol{R}_1^T \boldsymbol{R}_2]/3$ for $\boldsymbol{R}_1, \boldsymbol{R}_2 \in SO(3)$. This distance is restricted to values ranging from 0 to 4/3 for 3D rotation matrices. As indicated by this test, a discretization based on 2640 rotations and more achieves a very good coverage of the space of all rotations.

Our six-dimensional grid search proceeds as follows. For each rotation $\boldsymbol{R}_i$ obtained by a tessellation of the unit sphere, we scan all translations by evaluating the mixture model (6) and pick the translation $\boldsymbol{t}_i$ achieving the largest value. This generates a list of poses $(\boldsymbol{R}_i, \boldsymbol{t}_i)$ where i runs over all tetrahedra that are part of the 600-cell or finer tessellations. Finally, the solutions can be refined running the MM algorithm or ICP starting from the pose that was found by the systematic search.

3 Results

In the following, we first report tests on the local and global optimization strategies outlined in Methods. We then proceed by applying point cloud registration techniques to various structure comparison and fitting tasks.

3.1 Matching a permuted and randomly transformed structure against itself

To investigate the strengths and shortcomings of the local and global optimization techniques detailed in Methods, we applied them to the following test case: For a given crystal structure from the PDB, we first extracted all alpha carbon positions and assigned a weight of one to each position. This point cloud served as target against which a modified version of the same point cloud was matched. Since the kernel correlation is invariant under permutations of the point clouds that are compared against each other, we should achieve the same kernel correlation with a permuted version. Therefore, to modify the target point cloud (CA atom positions

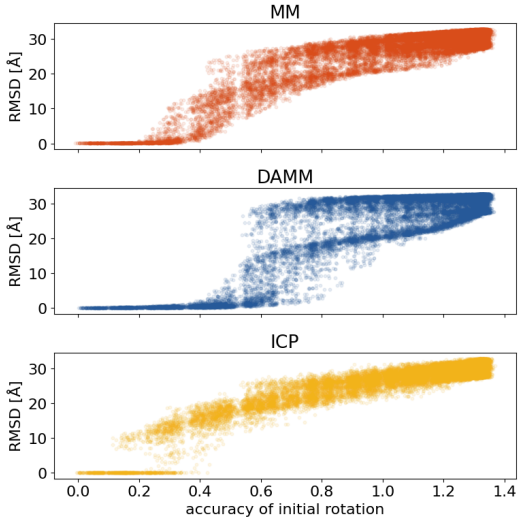


Fig. 3. Radius of convergence of the local registration methods. The TIM barrel structure 6HF2 was matched against itself starting from initial rotations that form a tessellation of rotation space at a very fine level of discretization. The distance between the initial and the correct rotation is plotted against the RMSD achieved by each registration method.

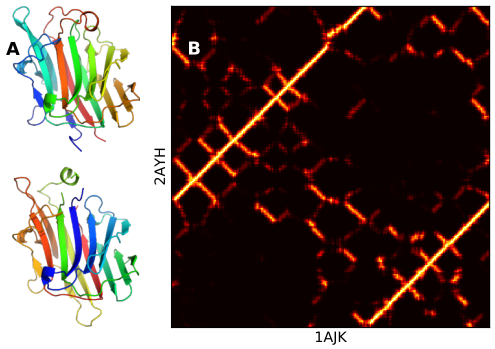


Fig. 4. KC-based matching of circularly permuted structures 2AHY and 1AJK. (A) The left panel shows both structures after registration by maximizing the kernel correlation (Pymol's chainbow coloring indicates the order of amino acids in each structure). (B) The right panel shows the kernel matrix has a heatmap. The row indices correspond to the amino acid sequence of 2AYH running from bottom (N-terminus) to top (C-terminus). The column indices correspond to the sequence of 1AJK and run from left (N-terminus) to right (C-terminus). The brightness of the heatmap is directly proportional to the entries in the kernel matrix.

from PDB file), we applied a random permutation and random rotation. If the rigid registration procedure is successful, it should then produce the same kernel correlation that is achieved when matching the target point cloud against itself without rotating or translating it, i.e. the maximum achievable kernel correlation is $\sum_{ij} \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$. In the following tests, we set $\sigma = 5$ Å. We ran self-matching tests on crystal structures of varying size (PDB codes: 6JC2 (212), 6HF2 (325), 1OEL (524), 5G5D (160), 6R4S (382) where the numbers in brackets indicate the number of CA atom positions in each structure).

3.1.1 Testing local registration algorithms

We first tested the local registration techniques, i.e. the MM algorithm outlined in subsection (2.3) and its annealed version DAMM (see subsection 2.4), and compared these with ICP. For each test structure, we generated 100 random matching problems by randomly shuffling the positions of the target and transforming it by a random rotation and translation. For each registration problem, 10 random initial poses were generated from which each of the local optimization methods (MM, DAMM, and ICP) was started. The maximum number of iterations for each registration approach was set to 50.

For each algorithm, we started from the same initial poses and measured the success of the registration method by comparing the kernel correlation that was obtained by the optimization procedure with the maximum possible value. Moreover, we assessed the quality of the registration by computing the root mean square deviation (RMSD) between corresponding points in both point clouds. For a given pose $(\boldsymbol{R}, \boldsymbol{t})$, we defined the RMSD in the following way:

$$\mathrm{RMSD} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} \min_{1\le j\le N}\{\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|^2\}} \qquad (7)$$

For the self-matching tasks, the optimal RMSD is zero.

Figure 2 shows a summary of the self-matching benchmark of the local registration methods. Both in terms of the correlation (which is propotional to the objective function of MM and DAMM) as well as the RMSD (which is the objective function of ICP), the deterministic annealing approach performs best, reaching correlations near 100% and RMSDs close to zero for most test cases. Also the plain MM approach without annealing outperforms ICP, but not as clearly as DAMM. Nevertheless, all approaches face difficulties with target 6HF2, a member of the TIM barrel fold family. The likely reason for the problems with self-matching 6HF2 is to be found in the quasi-symmetry of the structure. Rotations about the barrel axis achieve similar kernel correlations and RMSDs, which adds to the severeness of the registration problem, because the chance of getting trapped in a local optimum is increased. However, these difficulties can be overcome by using a larger number of initial poses. For example, when we increased the number of random initial poses from 10 to 50, the average correlations improved to 99.9% (MM), 100.0% (DAMM), and 98.5% (ICP), as do the average RMSDs: 0.2 Å (MM), 0.1 Å (DAMM), and 0.6 Å (ICP).

3.1.2 Testing global registration algorithms

Both the MM algorithms as well as ICP are local optimization methods and potentially suffer from getting trapped in a local minimum. Therefore, we also tested the global optimization strategies outlined in subsection 2.5.

The first modification of the setup used to benchmark the local search strategies was to generate the initial rotation randomly as before, but to choose the initial translation systematically rather than randomly. The initial translation was obtained by evaluating the spherical mixture model (6) over a cubic grid of cell size 1 Å. This adds negligible computational cost, because a mixture of spherical Gaussians can be evaluated very rapidly over a rectilinear grid. Again, we used 10 initial poses for each of the 100 test cases per target.

The optimized choice of the initial translation has only a minor impact on the success of the registration methods tested here. For most of the five targets, we see a slight but negligible improvement in the correlation and RMSD with all registration algorithms (see Supplementary Figure S2).

Next, we also scanned the rotations systematically by using the tessellation provided by the 600-cell. At the coarsest level, there are 330 rotations that need to be probed. For each rotation, we systematically searched all translations by evaluating the spherical Gaussian mixture model over a rectilinear grid of cell size 1 Å. Because the number of

rotations increased by a factor of 33, the computational costs are scaled accordingly. Again, our test calculations did not show an improvement in the performance of the registration algorithms that would justify the increase in computation time.

Finally, we used the discretization of rotation space to estimate the radius of convergence of the three registration algorithms. To do so, we matched the TIM barrel structure 6HF2 against itself such that the correct pose is $(\boldsymbol{I}, \boldsymbol{0})$. Using a very fine tessellation of $SO(3)$ based on 21120 rotations $\boldsymbol{R}_i$, we launched all three algorithms from initial poses $(\boldsymbol{R}_i, \boldsymbol{0})$. The success of the registration algorithms was then measured by computing the standard RMSD (not the RMSD defined in Eq. 7) between CA positions in the target and its transformed state. As is evident from Figure 3, the radius of convergence is highest for the MM methods, confirming our findings from the previous tests. The MM algorithms find the correct pose even, if the initial rotation has a distance smaller than 0.2 (MM) or 0.4 (DAMM) to the correct rotation. On the other hand, ICP starts to produce suboptimal fits already at rotations as close as 0.1. Similar results were obtained for the other target structures (see Supplementary Figures S3 and S4).

3.2 Comparison of circularly permuted structures

Circular permutation breaks the sequential order of the amino acids and thereby poses a challenge to standard protein alignment methods. The kernel correlation does not require a position-to-position correspondence between both structures and is invariant under shuffling the order of points in each cloud. Therefore, circularly permuted structures can directly be superimposed and compared with KC-based registration methods.

To illustrate this point, we matched the two circularly permuted structures 2AYH and 1AJK using the deterministic annealing approach with $\sigma = 5$ Å. The aligned structures are shown in Figure 4A. The correct alignment can be found very rapidly with DAMM. Out of 10 random initial poses, 3 produced the correct structural fit. At the right (Fig. 4B), we show the kernel matrix with elements $\phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)$ as a heatmap. The kernel matrix clearly delineates corresponding structural regions, which are indicated by "hot" matrix elements that run parallel to the diagonal and are "folded" due to circular permutation.

3.3 Rigid fitting of bead models from small-angle scattering

Our registration algorithms can also be used to dock structures into bead models derived from small-angle scattering (SAS) curves. The bead models were obtained from the Small Angle Scattering Biological Data Bank (SASBDB) (Valentini *et al.*, 2014). The first target is a bead model of exportin CRM1 derived from a SAS curve (SASBDB code SASDAJ4). We fitted the crystal structure 4HZK into the bead model by maximizing the kernel correlation with $\sigma = 5$ Å. Again, we used the deterministic annealing approach to find the pose that maximizes KC. Figure 5 shows the SAS bead model and a CA trace of the superimposed crystal structure. The correlation between both point clouds is 75 %. More examples can be found in Supplementary Figures S5 and S6.

3.4 Rigid fitting of cryo-EM maps

Next, we used the point-cloud registration methods to align two cryo-EM maps. We superimposed three pairs of low-resolution maps after converting them to weighted point clouds. All density maps were downloaded from the EMDataBank (Lawson *et al.*, 2010) and converted to weighted point clouds by running the DP-means algorithm (Kulis and Jordan, 2012). We set the desired bead radius to 10 Å for the first pair (two 80S ribosome maps) and to 5 Å for the second and third pairs (characterizing a motor protein and RNA polymerase II). The first pair of medium-resolution maps shows the 80S ribosome at 11.7 Å (EMD-1067) and 9.7 Å resolution

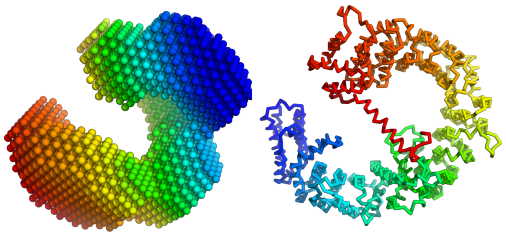


Fig. 5. Docking a crystal structure of exporting CRM1 into a bead model derived from a SAS curve. The bead model obtained from SASBDB is shown on the left. The right panel shows the high-resolution structure 4HZK that was docked into the bead model by maximizing the kernel correlation.

(EMD-1343). DP-means generated weighted point clouds with 425/666 beads representing EMD-1067/EMD-1343. The second docking task is to superimpose two low-resolution maps of axonemal dynein-c without and with nucleotide bound. The low-resolution maps EMD-2155 (apo dynein-c at 19 Å resolution) and EMD-2156 (dynein-c with bound nucleotide at 22 Å resolution) are represented by 433 and 405 beads, respectively. The third task is to superimpose bead models derived from EM maps EMD-2189 and EMD-2190 showing human RNA polymerase II at 25 Å resolution in complex with different RNAs. The bead models are composed of 889 and 951 particles. In each of the docking tasks, the kernel bandwidth σ was chosen to be twice as large as the bead radius.

Figure 6 shows the superpositions obtained with the deterministic annealing approach. Visual inspection reveals that the 3D superpositions found by DAMM are meaningful. To quantify this further, we also investigated the correspondence between the negative log kernel correlation (which is the target function of the MM algorithms) and more traditional measures for assessing the overlap of two structures or 3D density maps. The kernel correlation can serve as a surrogate of the cross-correlation coefficient (CCC), which is typically maximized to superimpose two cryo-EM maps using a voxel representation. Supplementary Figure S7 shows that there is high correlation between both metrics. An advantage of the kernel correlation compared to the CCC is that it can be evaluated very efficiently and does not require the interpolation of the moving cryo-EM structure over a voxel grid. Similarly, we also see a high agreement between $-\log \mathrm{KC}$ and the RMSD as defined in Eq. (7).

3.5 Rigid fitting of subunits into cryo-EM density maps of symmetric assemblies

Finally, we used KC-based point cloud registration to dock a high-resolution structure of a single subunit into a cryo-EM map of a symmetric assembly. As is shown in the Supplementary Material, the kernel correlation as well as the majorization-minimization strategy for finding the best superposition can readily be adapted to the symmetric case.

We demonstrate the ability to dock structures into symmetric assemblies for the high-resolution map of the capsaicin receptor TRPV1 (EMD-5778). This assembly exhibits a four-fold cyclic symmetry and served as a target for rigidly fitting a single subunit into the map. The map and the symmetry operators were downloaded from the EMDataBank. We converted the map to a weighted point cloud by applying DP-means using a bead radius of 5 Å. The modeled structure of a single subunit (PDB code 3J5P) was docked into the assembly by maximizing the symmetrized kernel correlation.

Figure 7 shows the cryo-EM map of the assembly and its point cloud representation next to the assembly predicted by fitting the structure of the subunit into the assembly. Visual inspection confirms that the subunit has been docked correctly into the point cloud representing the assembly. The

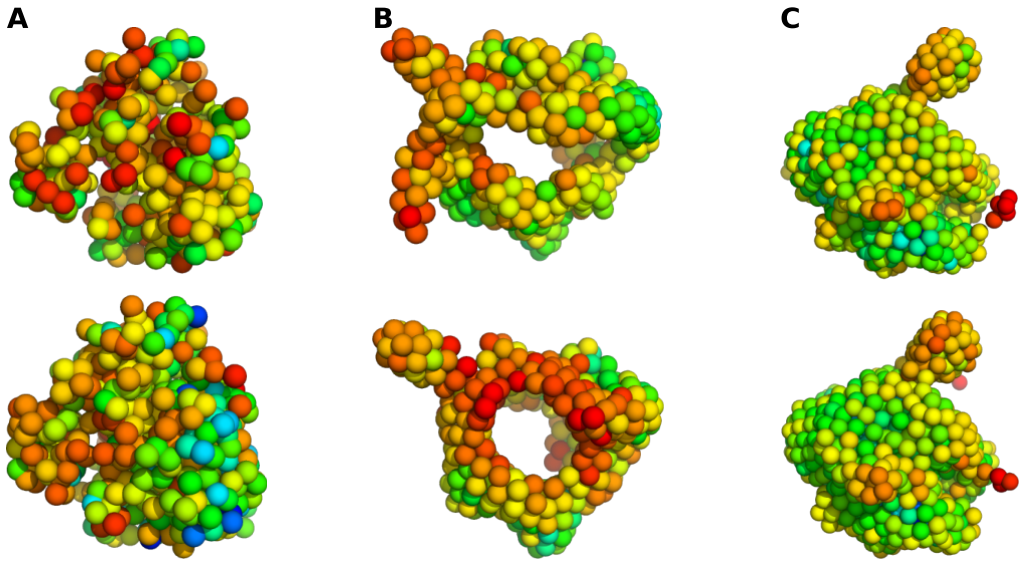


Fig. 6. Superposition of low-resolution cryo-EM maps by rigid point cloud registration. The colors indicate the weight of the particles ranging from blue (high weight) to red (low weight). (A) Superposition of two bead models of intermediate-resolution maps of the 80S ribosome (EMD-1067, EMD-1343). (B) Two bead models of the free and nucleotide-bound structure of axonemal dynein-c (EMD-2155, EMD-2156). (C) Two bead models of human RNA polymerase II (EMD-2189, EMD-2190) in complex with different RNAs.

correlation between both point clouds is 65 %. More examples of rigid fits into symmetric assemblies are presented in Supplementary Figure S8.

4 Conclusion

3D superposition of biomolecular structure is a common task in structural biology that is typically solved by specialized algorithms and software that depend on the representation of a 3D structure. In this article, we aim to address 3D fitting problems within a common framework based on the registration of weighted point clouds. As a measure of similarity, we use the kernel correlation, and introduce iterative algorithms for optimizing it so as to superimpose two point clouds.

An advantage of the kernel correlation over RMSD-based approaches is that KC does not require a point-to-point correspondence. This advantage comes at the cost of having to deal with an objective function that exhibits multiple optima and is therefore more difficult to optimize than the standard RMSD or its modified versions. Local point-cloud registration algorithms such as ICP therefore risk to get trapped in local optima. To overcome these challenges, we introduced an iterative MM algorithm and its annealed version, which both have a larger radius of convergence and thereby a lower chance of getting trapped in suboptima than the commonly used ICP method. Due to the generality of the representation, our rigid registration approach can be applied to various 3D fitting problems including the comparison of circularly permuted structures or the superposition of bead models and density maps from cryo-EM.

There is still room to improve the efficiency of the registration algorithm. Its current implementation is not fast enough for large-scale similarity searches based on point clouds. Algorithmic improvements that we envision include the use of data structures and algorithms for fast neighborhood queries such as KD-trees or neighbor lists.

The computation time to evaluate the kernel correlation scales with the size of both point clouds. Therefore, to improve the search over all rigid transformation, we plan to pursue a multiscale approach based on a hierarchical representation of point clouds. Coarser representations would involve a smaller number of points thereby allowing us to evaluate the kernel correlation more rapidly. Combined with a global grid search a multiscale approach should enable us to exhaustively scan all rigid transformations and further reduce the chance to miss the global optimum.

Another interesting direction is to fit multiple subunits into a point cloud representing an assembly, and to fit 3D point clouds against 2D point clouds derived from projection images obtained with electron microscopy or tomography as well as other imaging methods.

Funding

The authors acknowledge funding from the German Research Foundation (DFG) under project SFB 860, TP B09 as well as funding from the Carl Zeiss Foundation.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Besl, P. and McKay, N. D. (1992). A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **14**(2), 239–256.

Chen, Y. and Medioni, G. (1992). Object modelling by registration of multiple range images. *Image and vision computing*, **10**(3), 145–155.

Damm, K. L. and Carlson, H. A. (2006). Gaussian-weighted rmsd superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys J*, **90**(12), 4558–4573.

Higham, N. J. (1989). Matrix nearness problems and applications. In M. J. C. Gover and S. Barnett, editors, *Applications of Matrix Theory*, pages 1–27. Oxford University Press.

Hirsch, M. and Habeck, M. (2008). Mixture models for protein structure ensembles. *Bioinformatics*, **24**, 2184–2192.

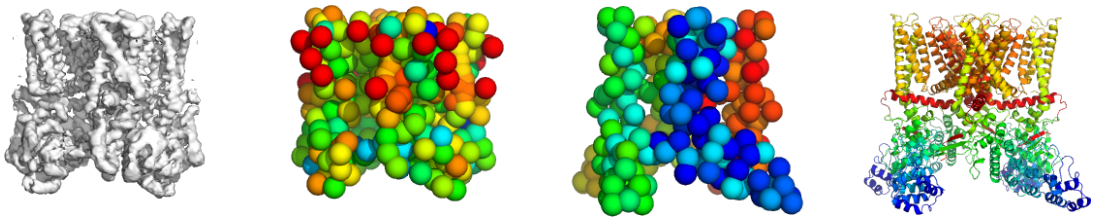


Fig. 7. Rigid fitting of an atomic structure of a subunit into a high-resolution map of the symmetric channel TRPV1. Left: Cryo-EM map EMD-5778. Middle left: Bead model representing the cryo-EM map where the colors indicate the weight of particles. Middle right: Docked structure of the subunit and symmetry mates shown as bead models. Right: Docked high-resolution structure.

Horn, B. K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, **4**, 629–642.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, **58**(1), 30–37.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, **A32**, 922–923.

Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophys. J.*, **95**, 4643–4658.

Kulis, B. and Jordan, M. I. (2012). Revisiting k-means: New algorithms via bayesian nonparametrics. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 513–520, New York, NY, USA. ACM.

Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S. J., Newman, R. H., Oldfield, T. J., Rees, I., Sahni, G., Sala, R., Velankar, S., Warren, J., Westbrook, J. D., Henrick, K., Kleywegt, G. J., Berman, H. M., and Chiu, W. (2010). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Research*, **39**(suppl_1), D456–D464.

Liu, W., Pokharel, P., and Principe, J. (2007). Correntropy: Properties and Applications in Non-Gaussian Signal Processing. *Signal Processing, IEEE Transactions on*, **55**(11), 5286 –5298.

Mechelke, M. and Habeck, M. (2010). Robust probabilistic superposition and comparison of protein structures. *BMC Bioinformatics*, **11**, 363.

Ritchie, D. W. (2016). Calculating and scoring high quality multiple flexible protein structure alignments. *Bioinformatics*, **32**(17), 2650–2658.

Rose, K., Gurewitz, E., and Fox, G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, **11**(9), 589–594.

Rusu, R. B., Blodow, N., and Beetz, M. (2009). Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE.

Straub, J., Campbell, T., How, J. P., and Fisher, J. W. (2017). Efficient Global Point Cloud Alignment Using Bayesian Nonparametric Mixtures. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 2403–2412.

Suzuki, H., Kawabata, T., and Nakamura, H. (2016). Omokage search: shape similarity search service for biomolecular structures in both the pdb and emdb. *Bioinformatics*, **32**(4), 619–620.

Theobald, D. L. and Wuttke, D. S. (2006). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proc. Nat. Acad. Sci.*, **103**, 18521–18527.

Tsin, Y. and Kanade, T. (2004). A correlation-based approach to robust point set registration. In *European conference on computer vision*, pages 558–569. Springer.

Vakili, N. and Habeck, M. (2021). Bayesian random tomography of particle systems. *Frontiers in Molecular Biosciences*, **8**, 399.

Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M., and Svergun, D. I. (2014). SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Research*, **43**(D1), D357–D363.

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.

Supplementary Material for *Matching biomolecular structures by registration of point clouds*

Nima Vakili [1,2], Michael Habeck[1,2,*]

September 27, 2021

[1]Microscopic Image Analysis Group, Jena University Hospital, 07743 Jena, Germany.
[2]Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany.

*E-Mail: michael.habeck@uni-jena.de

1 Iterative majorization-minimization (MM)

The figure below illustrates the majorization-minimization (MM) approach used in this paper to minimize the negative log kernel correlation $-\log \mathrm{KC}(\boldsymbol{R}, \boldsymbol{t})$. Instead of minimizing $-\log \mathrm{KC}(\boldsymbol{R}, \boldsymbol{t})$ directly over all rotational and translational degrees of freedom, our MM algorithms minimize an upper bound $U(\boldsymbol{R}, \boldsymbol{t})$ detailed in equation (3) of the main article. The progression of the MM iterations is shown for a self-matching task where PDB structure 6JC2 is fitted against itself. As is evidenced by the plot, the upper bound $U(\boldsymbol{R}, \boldsymbol{t})$ is rather tight and becomes even tighter in the course of the MM iterations.

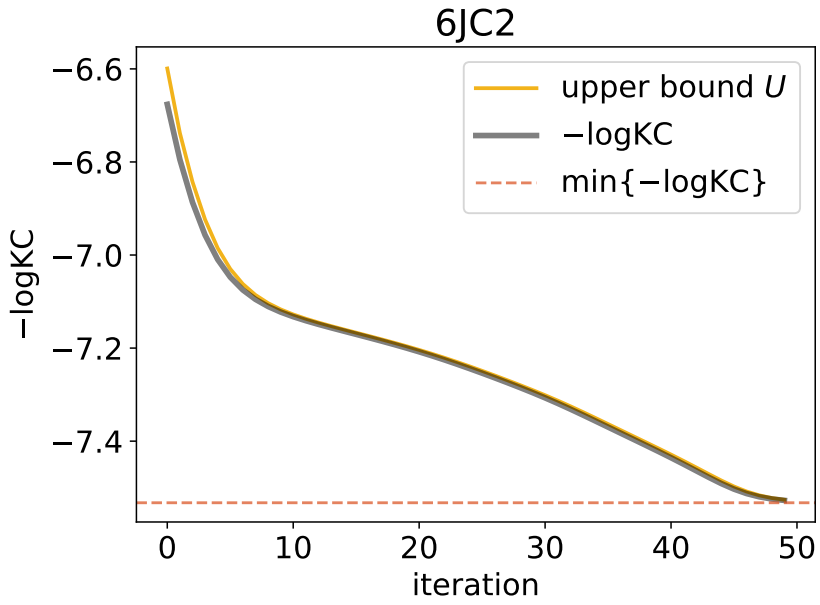


Figure S1: Evolution of the negative log kernel correlation $-\log \mathrm{KC}$, our objective function to superimpose two weighted point clouds, during the MM iterations. Instead of $-\log \mathrm{KC}(\boldsymbol{R}, \boldsymbol{t})$ itself, each iteration minimizes a tight upper bound $U(\boldsymbol{R}, \boldsymbol{t})$ shown in yellow.

2 Self-match benchmark with random initial rotation and grid search of the translation

The self-matching benchmark (subsection 3.1.1 of the main article) was modified as follows. Instead of choosing the initial translation randomly, it was optimized by maximizing the kernel correlation over a regular cubic grid. The resulting average correlation values and RMSDs can found in the following figure:

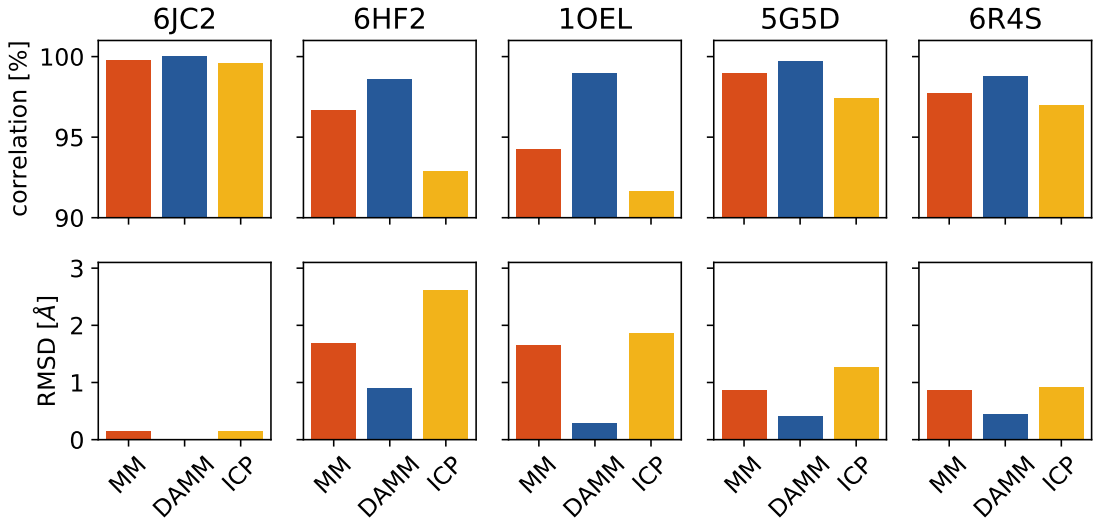


Figure S2: Testing the impact of using an initial translation that is optimized over a cubic grid of cell size 1 Å. A PDB structure (indicated in panel titles) is fitted against a permuted and randomly transformed version of itself. The top row shows the correlation coefficient (ratio of actual kernel correlation and maximum achievable kernel correlation) obtained when starting local optimization runs from 10 random initial rotations (and optimized translations). The bottom row shows the RMSD (defined in equation 7 of the main paper) of corresponding points after striking the estimated pose.

3 Radius of convergence of iterative registration methods

In addition to the example shown in Figure 3 of the main manuscript, we also assessed the radius of convergence of three registration methods (ICP: iterative closest point; MM: iterative majorization-minimization; DAMM: MM with deterministic annealing) using four other examples. The setup is the same as described in the main paper.

In all tests, the MM methods have a larger radius of convergence than ICP (see Fig. **??**).

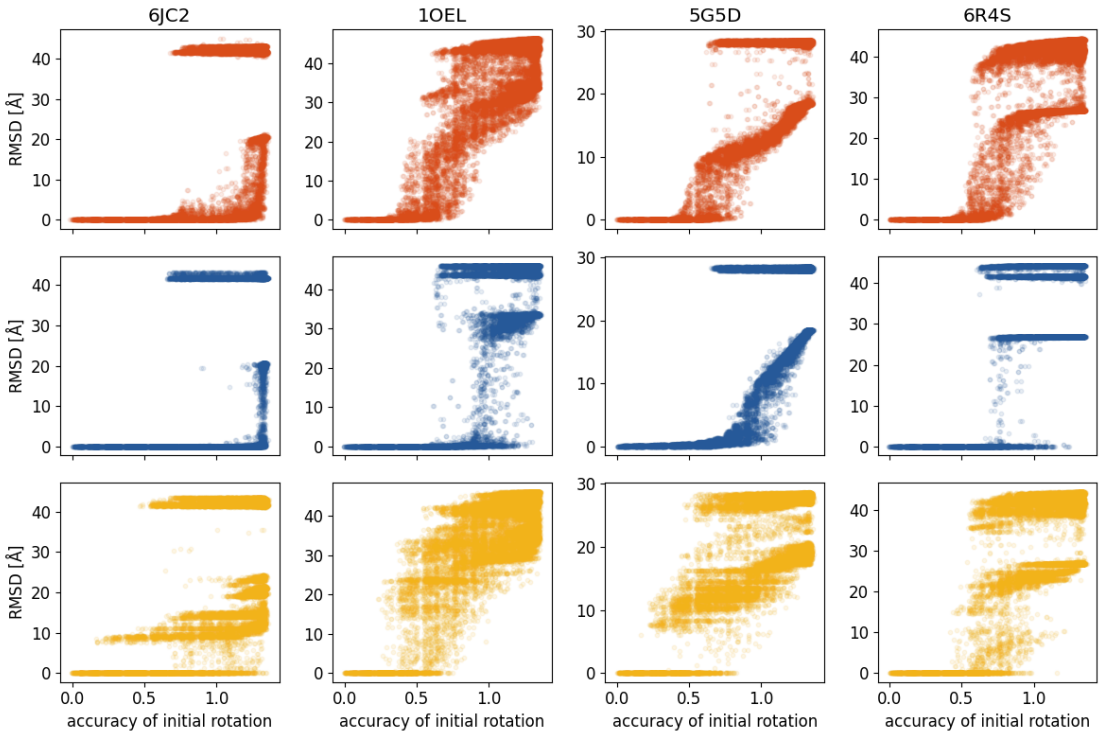
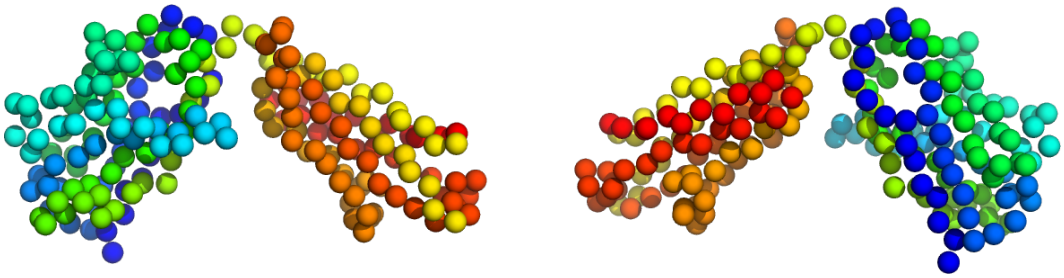


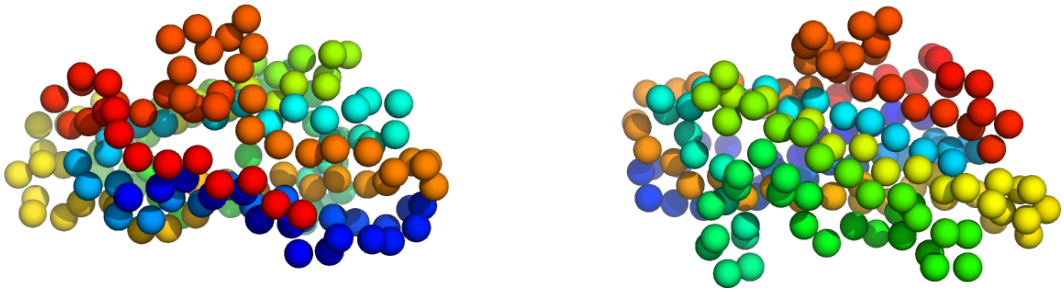
Figure S3: Radius of convergence. Distance between initial and true rotation versus the RMSD of the final pose generated by the registration methods: MM (red), DAMM (blue), ICP (yellow).

In most examples, we see the emergence of a second dominant cluster of solutions with high RMSD values. These registrations correspond to an alternative 3D superpositions due to a quasi-symmetry of the point cloud. For example, in case of 6JC2 we are dealing with a heterodimer where the two monomers are very similar to each other. The bad fit with an RMSD of 41.6 Å achieves a correlation 91% and is shown in Fig.

??(a). In case of 5G5D, we are dealing with a member of the Carbohydrate-binding domain superfamily, which also exhibits a quasi-symmetry. If we ignore the sequence information and only look at the spatial arrangement of CA atoms, the bad fit with an RMSD of 28.2 Å achieves a correlation of 97% and is shown in Fig. **??**(b).



(a) 6JC2



(b) 5G5D

Figure S4: Left: target point cloud. Right: bad pose with a good correlation but high RMSD.

4 Fitting of atomic structures into bead models derived from small-angle scattering curves

The main article illustrates the ability to fit high-resolution structures into bead models from small-angle scattering (SAS) curves for an exportin structure. Here, we demonstrate this for two additional examples. The first example also involves exportin CRM1. We fitted a bead model of free CRM1 (SASDAJ4) against CRM1 RanGTP (SASDAK4). The figure **??** shows the final superposition found by maximizing the kernel correlation with σ = 5 Å. The correlation of the final fit is 80.3%.

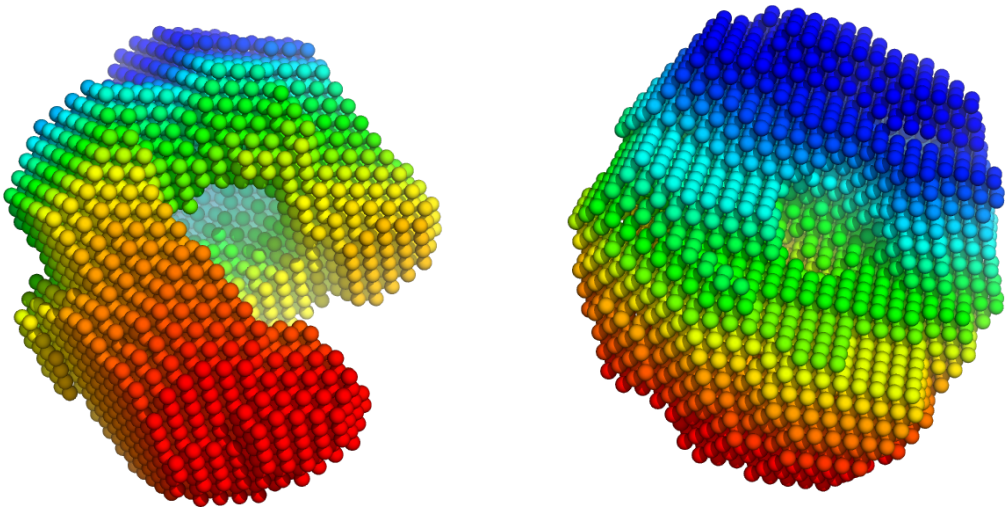


Figure S5: 3D superposition of SASDAK4 (CRM1 RanGTP, shown on the right) onto SASDAJ4 (free CRM1, shown on the left).

The second example involves fitting a bead model of Human Chromatin Remodeler CHD4 (SASBDB code SASDAA5) against two other SASBDB structures: the apo form of full length ObgE (SASDBS6) and mitochondrial heat shock protein 70 (SASDBY6). The superposition is shown in figure **??**.

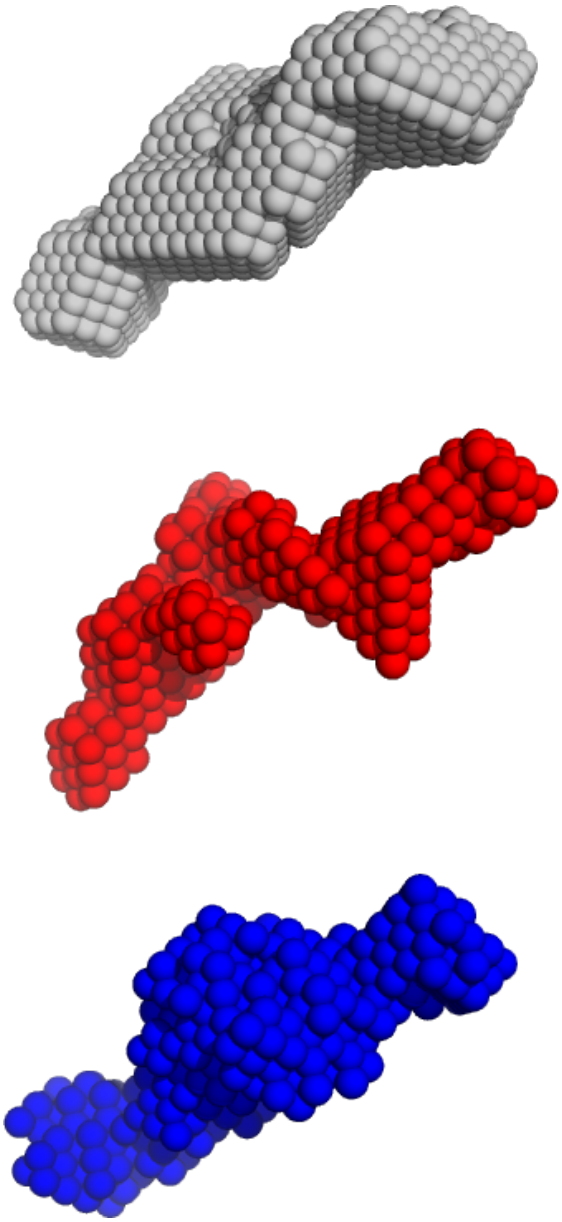


Figure S6: Point cloud registration onto SASDAA5 (grey top). Bead models SASDBS6 (red) and SASDBY6 (blue) were fitted onto SASDAA5 by maximizing the kernel correlation using DAMM.

5 Kernel correlation as a proxy for the cross-correlation coefficient and the RMSD

By construction, the kernel correlation is highly related to the cross-correlation coefficient (CCC) that is often used to measure the overlap of two cryo-EM density maps represented on voxel grids. This is demonstrated in the figure below for one of the cryo-EM docking targets discussed in the main text (superposition of two RNA polymerase II structures). We observe a high correlation between the negative log kernel correlation (which is optimized by the MM algorithms introduced in Methods) and CCC between the target map and the transformed map. Therefore, by minimizing $-\log \mathrm{KC}$ we effectively maximize the CCC between the two density maps.

A high correlation is also observed between the RMSD (as defined in equation 7 of the main article) and the negative log kernel correlation. The RMSD is optimized by ICP.

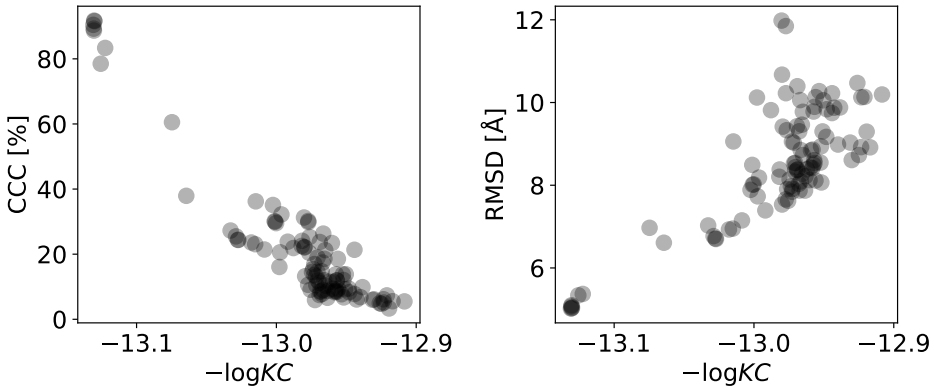


Figure S7: Kernel correlation as a proxy for RMSD and cross-correlation coefficient (CCC). Left: High correlation between $-\log \mathrm{KC}$ and CCC. Right: High correlation between $-\log \mathrm{KC}$ and the RMSD (see equation 7 in main manuscript.)

6 Docking the high-resolution structure of a subunit into a cryo-EM map of a symmetric assembly

The first point cloud, $(\boldsymbol{X}, \boldsymbol{q})$, represents the full assembly and is derived, for example, from a cryo-EM map. The second point cloud, $(\boldsymbol{Y}, \boldsymbol{p})$, represents the subunit that will be docked into the full assembly. We assume that the assembly is symmetric and the symmetry mates can be generated from a single subunit by the action of C rigid transformations $\{(\boldsymbol{R}_k, \boldsymbol{t}_k)\}_{k=1}^{C}$. The subunit needs to be transformed by an unknown rigid transformation $(\boldsymbol{R}, \boldsymbol{t})$ such that the overlap between the target and the full model structure is as large as possible. The coordinates of the k-th subunit after rigid transformation are:

$$\boldsymbol{y}'_{jk} = \boldsymbol{R}_k(\boldsymbol{R}\boldsymbol{y}_j + \boldsymbol{t}) + \boldsymbol{t}_k$$

The kernel correlation between the point cloud representing the assembly and the structure of the assembly built by applying the symmetry operators is:

$$\mathrm{KC}_{\mathrm{sym}}(\boldsymbol{R}, \boldsymbol{t}) = \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{C} q_i p_j \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}_k(\boldsymbol{R}\boldsymbol{y}_j + \boldsymbol{t}) - \boldsymbol{t}_k\|)$$

We can rewrite the kernel correlation of a symmetric assembly:

$$\mathrm{KC}_{\mathrm{sym}}(\boldsymbol{R}, \boldsymbol{t}) = \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{C} q_i p_j \phi_\sigma(\|\boldsymbol{R}_k^T(\boldsymbol{x}_i - \boldsymbol{t}_k) - \boldsymbol{R}\boldsymbol{y}_j - \boldsymbol{t}\|)$$

Upper bound:

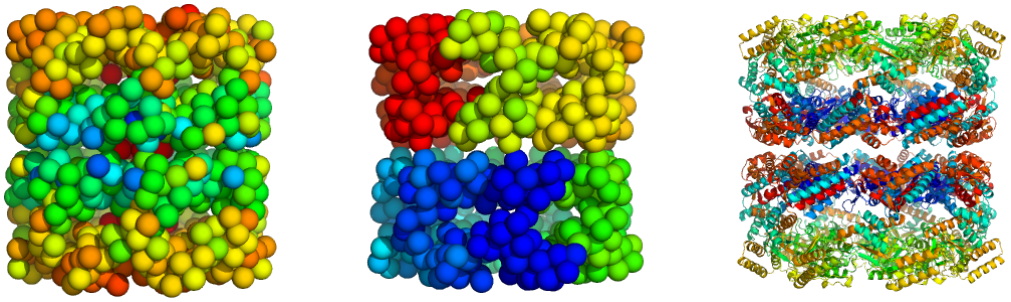
$$\begin{aligned}
-\log \mathrm{KC}_{\mathrm{sym}}(\boldsymbol{R}, \boldsymbol{t}) &= -\log\left\{\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{C} q_i p_j \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}_k(\boldsymbol{R}\boldsymbol{y}_j + \boldsymbol{t}) - \boldsymbol{t}_k\|)\right\} \\
&= -\log\left\{\sum_{ijk} q_i p_j w_{ijk} \frac{\phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}_k(\boldsymbol{R}\boldsymbol{y}_j + \boldsymbol{t}) - \boldsymbol{t}_k\|)}{w_{ijk}}\right\} \\
&\leq \frac{1}{2\sigma^2}\sum_{ijk} q_i p_j w_{ijk} \|\boldsymbol{x}_i - \boldsymbol{R}_k(\boldsymbol{R}\boldsymbol{y}_j + \boldsymbol{t}) - \boldsymbol{t}_k\|^2 + \text{const.}
\end{aligned}$$

where the weights

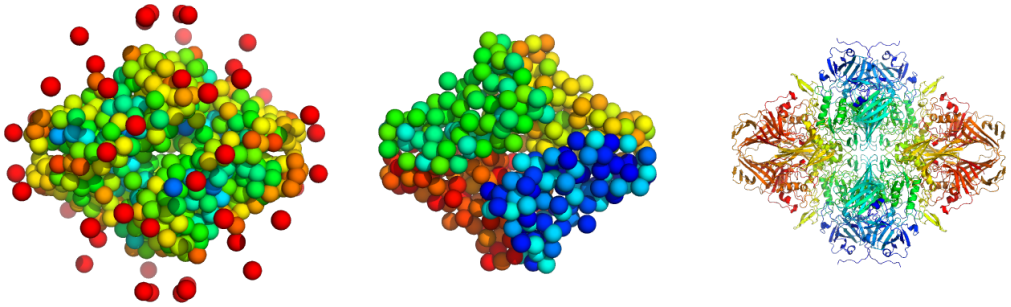
$$w_{ijk} \propto \phi_\sigma(\|\boldsymbol{x}_i - \boldsymbol{R}_k(\boldsymbol{R}\boldsymbol{y}_j + \boldsymbol{t}) - \boldsymbol{t}_k\|)$$

are normalized such that $\sum_{ijk} q_i p_j w_{ijk} = 1$.

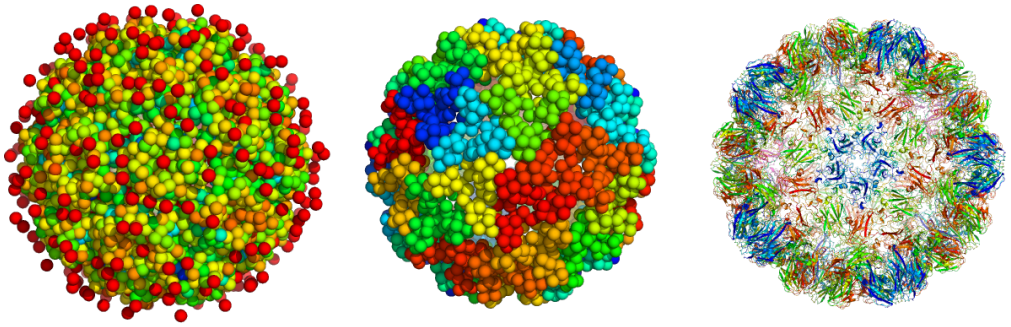
More examples of 3D fitting into symmetric assemblies by maximizing $\mathrm{KC}_{\mathrm{sym}}$ are shown in Fig. **??**.



(a) EMD-6422



(b) EMD-5995



(c) EMD-6000

Figure S8: Rigid docking of a subunit into a symmetric assembly. EMD-6422: GroEL, D7 symmetry. EMD-5995: beta-galactosidase, D2 symmetry. EMD-6000: Brome mosaic virus. In each row, the point cloud on the left shows the particle representation of the cryo-EM map of the assembly obtained by running DP-means with a particle radius of 5 Å. The colors indicate the weight of the particles. The middle panel shows the fitted structure of the subunit that was also coarse-grained by running DP-means. The structure in blue is the subunit; all other particles were generated on the fly by applying the symmetry operators. The right panel is the high-resolution structure of the full assembly obtained by transforming the high-resolution structure of the subunit rigidly by using the estimated pose and by generating the symmetry mates by applying the symmetry operators.

Chapter 5

Conclusion and future works

5.1 Summary

A common feature of all algorithms proposed in this thesis is the use of an RBF network to represent molecular densities than their atomic models in real space, a representation that allows all the particles to be located off the grid. We model the molecular densities by using the Kernel representation, the isotropic Gaussian radial basis function (RBF). This thesis comprises three manuscripts. In chapter 2, the goal was to develop a probabilistic model in a Bayesian framework to infer a volumetric shape from 2D projection images acquired by cryo-EM. We represent the density using a radial basis function kernel (Gaussian mixture model). Moreover, to ensure that estimated particle configurations show a physically plausible packing, we adopt an excluded volume prior. We fit our model to the datasets (images or point-clouds) by using MCMC algorithms, such as HMC and global sampling. In chapter 3, we focus on the tomographic reconstruction problem which is simpler because the relative orientations of the particles are assumed to be known. We propose three kernel-based approaches for tomographic reconstruction problem in real space. The first approach considers the particle positions to be located on a regular grid (hexagonal grid) while estimating the particle weights by using an iterative non-negative least square method. The second approach is a mesh-free method and employs Expectation Maximization technique, and it involves a subsampling step in which a chosen number of pixels is randomly selected from the projection images. Our third proposal is also a grid-free approach that proposes a fully Bayesian method which incorporates a prior information such as excluded volume and fits the particle positions with HMC. In chapter 4, we

develop a mathematical expression to compute the assessing match between two volume densities. We represent volumes by kernel representation and then superposition of two particle clouds can be reached by minimizing our Kernel correlation equation. We introduce an Upper Bound strategy for finding the minimum mode. Since this is a local search method and there is a possibility of getting trapped in a local minima, we augment our proposal by using either a deterministic annealing or a global rotational/translational search methods.

5.2 Outlook

Our current version of the Bayesian random tomography uses only class averages with high SNR. In the future, we plan to do the reconstruction from raw cryo-EM images by adding the CTF term to our model.

Unfortunately, our current Bayesian implementation that underlines the reconstruction task is computationally demanding. One direction for extending the research is to use hardware accelerators like graphics processing units (GPUs) (Kimanius *et al.*, 2016; Zhang, 2016). New implementation of our algorithm with GPU support can address the most computationally intensive steps in the reconstruction workflow.

Another direction of research would be to increase the number of our pseudo atoms to achieve higher resolution. It's clear that increasing the number of particles is at the cost of high computation expenses, but we can adopt some approaches to address this issue. Like the one that we did in chapter 2, which we start with very few particles and while the log-posterior converged, we increase the number of particles by fixing the rotational parameters.

We have already applied in chapter 2 our proposal method on marine plankton species as a different application. One of the great advantage of our method is its versatility to other tomographic reconstruction areas. The method is not limited to application in cryo-EM and it would be interesting to assess its efficiency in reconstruction problems arising in optical tomography or thermoacoustic imaging.

Another route for extending the approach is to take conformational heterogeneity into account, which is one of the major bottlenecks in cryo-EM data processing. The algorithm would then produce multiple structures instead of only one and then their relation should be considered as well.

Another example is studying the dynamics of a macromolecular assembly and identify the possible conformational changes in protein. This can be done by adopting Normal Mode Analysis (NMA) which is a useful technique to analyze dynamics of large macromolecular assemblies around an equilibrium structural conformation. (Nogales-Cadenas *et al.*, 2013; Jin *et al.*, 2014) use pseudo-atomic representation of the atomic structures to explore the conformational changes and this brought us curiosity how easily our point-cloud representation can be adopted in this strategy.

In our RBF kernel representation, the number of particles K was fixed. Another direction of research would be to estimate the number of particles based on the projection data, which can be considered as an alternative way to measure the resolution of the input data. Alternatively, we can estimate the number of particles by recruiting an extension of MCMC algorithm i.e., Reversible Jump Markov Chain Monte Carlo (RJMCMC) (Green, 1995). The method provides moves between subsapaces of varying dimensionality and can explore multi dimensional parameter spaces efficiently. Therefore, when the number of model parameters is unknown, simulating the posterior distribution is still possible and we can simultaneously update model indicator K as well as model parameters.

Many of the heuristic approaches that are available in cryo-EM for the particle picking are designed based on template matching, computing the cross-correlation of the micrographs against pre-calculated of the particle of interest (Nicholson and Glaeser, 2001; Roseman, 2004). In case of non-ideal datasets, when we face particles overlap, conformationally heterogeneous, these methods are usually prone to failure. Adopting the deep convolutional neural network (CNN) augmented with GPU support can be one of the interesting extension for our per-processing reconstruction algorithm (Schmidhuber, 2015; Wang *et al.*, 2016; Rawat and Wang, 2017; Wagner *et al.*, 2019).

The other outlook would generalize our proposed kernel correlation approach (chapter 4) for the cases that someone needs to superimpose several sub-units simultaneously into atomic structures. Besides, our method currently works for the rigid-transformation, it would also very interesting to consider non-rigid transformation for the future or even the cases when the target and the source structures lie in different spaces, i.e., one in R^n and the other on lies in R^{n-1}. In this scenario, we need to use a regularizer term which should be added to the kernel equation.

Nowadays, point clouds are used in many areas of technical practice and one of the most interesting area is autonomous driving. Light Detection And Ranging (LiDAR) sensors (as one of the most important sensors in autonomous cars) collect 3D point clouds that

precisely record the external surfaces of objects and scenes. By computing point clouds registration problem, we can estimate the motion-planning or control decisions of the car. This area would be a great opportunity to extend our research (chapter 4) in the industrial field.

Bibliography

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., *et al.* (2007). *Nature*, **450**(7170), 683–694.

Anishinraj, M., Venkataraman, B., and Vaithiyanathan, V. (2012). *Eur. J. Sci. Res*, **4**(68), 584–590.

Besl, P. J. and McKay, N. D. (1992). Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics.

Bishop, C. M. (2006). *Machine learning*, **128**(9).

Ceulemans, H. and Russell, R. B. (2004). *Journal of molecular biology*, **338**(4), 783–793.

Chacón, P. and Wriggers, W. (2002). *Journal of molecular biology*, **317**(3), 375–384.

Cheng, Y. and Walz, T. (2009). *Annual review of biochemistry*, **78**, 723–742.

Elmlund, D. and Elmlund, H. (2012). *Journal of structural biology*, **180**(3), 420–427.

Elmlund, H., Lundqvist, J., Al-Karadaghi, S., Hansson, M., Hebert, H., and Lindahl, M. (2008). *Journal of molecular biology*, **375**(4), 934–947.

Elmlund, H., Elmlund, D., and Bengio, S. (2013). *Structure*, **21**(8), 1299–1306.

Frank, J. (2006). *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press.

Frank, J. (2008). *Electron tomography: methods for three-dimensional visualization of structures in the cell*. Springer.

Garzón, J. I., Kovacs, J., Abagyan, R., and Chacón, P. (2007). *Bioinformatics*, **23**(4), 427–433.

Gordon, R., Bender, R., and Herman, G. T. (1970). *Journal of theoretical Biology*, **29**(3), 471–481.

Green, P. J. (1995). *Biometrika*, **82**(4), 711–732.

Herman, G. T. (2009). *Fundamentals of computerized tomography: image reconstruction from projections*. Springer Science & Business Media.

Jaitly, N., Brubaker, M. A., Rubinstein, J. L., and Lilien, R. H. (2010). *Bioinformatics*, **26**(19), 2406–2415.

Jin, Q., Sorzano, C. O. S., De La Rosa-Trevín, J. M., Bilbao-Castro, J. R., Núñez-Ramírez, R., Llorca, O., Tama, F., and Jonić, S. (2014). *Structure*, **22**(3), 496–506.

Jonić, S. and Sorzano, C. Ó. S. (2015). *IEEE Journal of Selected Topics in Signal Processing*, **10**(1), 161–173.

Joubert, P. and Habeck, M. (2015). *Biophysical Journal*, **108**(5), 1165–1175.

Kak, A. C. and Slaney, M. (2001). *Principles of computerized tomographic imaging*. SIAM.

Kawabata, T. (2008). *Biophys. J.*, **95**, 4643–4658.

Kimanius, D., Forsberg, B. O., Scheres, S. H., and Lindahl, E. (2016). *elife*, **5**, e18722.

Kulis, B. and Jordan, M. I. (2012). Revisiting k-means: New algorithms via bayesian nonparametrics. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 513–520, New York, NY, USA.

Leary, R., Midgley, P. A., and Thomas, J. M. (2012). *Accounts of chemical research*, **45**(10), 1782–1791.

Levin, B. D., Padgett, E., Chen, C.-C., Scott, M., Xu, R., Theis, W., Jiang, Y., Yang, Y., Ophus, C., Zhang, H., *et al.* (2016). *Scientific data*, **3**(1), 1–11.

Miao, J., Förster, F., and Levi, O. (2005). *Physical Review B*, **72**(5), 052103.

Miao, J., Ercius, P., and Billinge, S. J. (2016). *Science*, **353**(6306).

Midgley, P. A. and Dunin-Borkowski, R. E. (2009). *Nature materials*, **8**(4), 271–280.

Neal, R. M. (2011). *Handbook of Markov Chain Monte Carlo*, pages 113–162.

Neal, R. M. *et al.* (2011). *Handbook of markov chain monte carlo*, **2**(11), 2.

Nicholson, W. V. and Glaeser, R. M. (2001). *Journal of Structural Biology*, **133**(2-3), 90–101.

Nogales-Cadenas, R., Jonic, S., Tama, F., Arteni, A., Tabas-Madrid, D., Vázquez, M., Pascual-Montano, A., and Sorzano, C. O. S. (2013). *Nucleic acids research*, **41**(W1), W363–W367.

Penczek, P. A., Zhu, J., and Frank, J. (1996). *Ultramicroscopy*, **63**(3-4), 205–218.

Pham, M., Yuan, Y., Rana, A., Miao, J., and Osher, S. (2020). *arXiv preprint arXiv:2004.10445*.

Pryor, A., Yang, Y., Rana, A., Gallagher-Jones, M., Zhou, J., Lo, Y. H., Melinte, G., Chiu, W., Rodriguez, J. A., and Miao, J. (2017). *Scientific reports*, **7**(1), 1–12.

Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. (2017). *Nat. Methods*, **14**(3), 290–296.

Rabbani, T., Van Den Heuvel, F., and Vosselmann, G. (2006). *International archives of photogrammetry, remote sensing and spatial information sciences*, **36**(5), 248–253.

Radermacher, M. (2007). Weighted back-projection methods. In *Electron tomography*, pages 245–273. Springer.

Rawat, W. and Wang, Z. (2017). *Neural computation*, **29**(9), 2352–2449.

Rohou, A. and Grigorieff, N. (2015). *Journal of structural biology*, **192**(2), 216–221.

Roseman, A. (2004). *Journal of structural biology*, **145**(1-2), 91–99.

Roseman, A. M. (2000). *Acta Crystallographica Section D: Biological Crystallography*, **56**(10), 1332–1340.

Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M. (2008). *Robotics and Autonomous Systems*, **56**(11), 927–941.

Saghi, Z. and Midgley, P. A. (2012). *Annual Review of Materials Research*, **42**, 59–79.

Scheres, S. H. (2010). *Methods Enzymol.*, **482**, 295–320.

Scheres, S. H. (2012). *J. Struct. Biol.*, **180**(3), 519–530.

Scheres, S. H., Valle, M., and Carazo, J.-M. (2005). *Bioinformatics*, **21**(suppl_2), ii243–ii244.

Scheres, S. H., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P., Frank, J., and Carazo, J. M. (2007). *Nat. Methods*, **4**(1), 27–29.

Schmidhuber, J. (2015). *Neural networks*, **61**, 85–117.

Schölkopf, B., Smola, A. J., Bach, F., *et al.* (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Singer, A. and Shkolnisky, Y. (2011). *SIAM journal on imaging sciences*, **4**(2), 543–572.

Straub, J., Campbell, T., How, J. P., and Fisher, J. W. (2017). Efficient global point cloud alignment using bayesian nonparametric mixtures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2941–2950.

Takeda, H., Farsiu, S., and Milanfar, P. (2007). *IEEE Transactions on image processing*, **16**(2), 349–366.

Thompson, R. F., Walker, M., Siebert, C. A., Muench, S. P., and Ranson, N. A. (2016). *Methods*, **100**, 3–15.

Tian, X., Kim, D. S., Yang, S., Ciccarino, C. J., Gong, Y., Yang, Y., Yang, Y., Duschatko, B., Yuan, Y., Ajayan, P. M., *et al.* (2020). *Nature materials*, **19**(8), 867–873.

Topf, M. and Sali, A. (2005). *Current opinion in structural biology*, **15**(5), 578–585.

Topf, M., Baker, M. L., John, B., Chiu, W., and Sali, A. (2005). *Journal of structural biology*, **149**(2), 191–203.

Trampert, J. and Leveque, J.-J. (1990). *Journal of Geophysical Research: Solid Earth*, **95**(B8), 12553–12559.

Tsin, Y. and Kanade, T. (2004). A correlation-based approach to robust point set registration. In *European conference on computer vision*, pages 558–569. Springer.

Van Heel, M. (1987). *Ultramicroscopy*, **21**(2), 111–123.

Vasishtan, D. and Topf, M. (2011). *Journal of structural biology*, **174**(2), 333–343.

Villa, E. and Lasker, K. (2014). *Current opinion in structural biology*, **25**, 118–125.

Volkmann, N. and Hanein, D. (1999). *Journal of structural biology*, **125**(2-3), 176–184.

Volkmann, N. and Hanein, D. (2003). *Methods in enzymology*, **374**, 204–225.

Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., *et al.* (2019). *Communications biology*, **2**(1), 1–13.

Wang, F., Gong, H., Liu, G., Li, M., Yan, C., Xia, T., Li, X., and Zeng, J. (2016). *Journal of structural biology*, **195**(3), 325–336.

Wriggers, W. and Chacon, P. (2001). *Structure*, **9**(9), 779–788.

Wriggers, W., Milligan, R. A., Schulten, K., and McCammon, J. A. (1998). *Journal of molecular biology*, **284**(5), 1247–1254.

Yang, Y., Zhou, J., Zhu, F., Yuan, Y., Chang, D. J., Kim, D. S., Pham, M., Rana, A., Tian, X., Yao, Y., *et al.* (2021). *Nature*, **592**(7852), 60–64.

Zhang, K. (2016). *Journal of structural biology*, **193**(1), 1–12.

Zhang, K., Gao, X., Tao, D., and Li, X. (2012). *IEEE Transactions on Image Processing*, **21**(11), 4544–4556.

Zhu, Y., Carragher, B., Glaeser, R. M., Fellmann, D., Bajaj, C., Bern, M., Mouche, F., De Haas, F., Hall, R. J., Kriegman, D. J., *et al.* (2004). *Journal of structural biology*, **145**(1-2), 3–14.