Bayesian Inference of Initial Models in Cryo-Electron Microscopy Using Pseudo-atoms

Group Meeting

Yu-Hsiang Lien 連昱翔

Reference

Biophysical Journal Volume 108 March 2015 1165–1175

Article

Bayesian Inference of Initial Models in Cryo-Electron Microscopy Using Pseudo-atoms

Paul Joubert[1,*] and Michael Habeck[1,2,*]

[1]Felix-Bernstein Institute for Mathematical Statistics, Georg-August-Universität Göttingen, Göttingen, Germany; and [2]Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

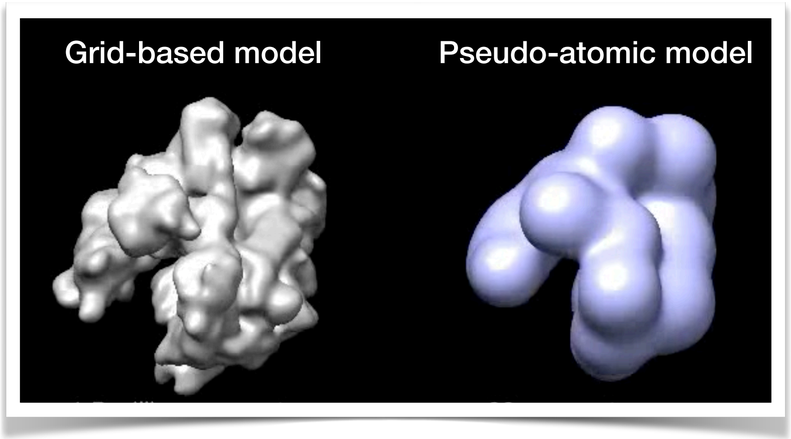
Outline

- Pseudo atomic model
- Data generation model
- Ab initio algorithm
- Tests of algorithm

Pseudo-Atomic Model

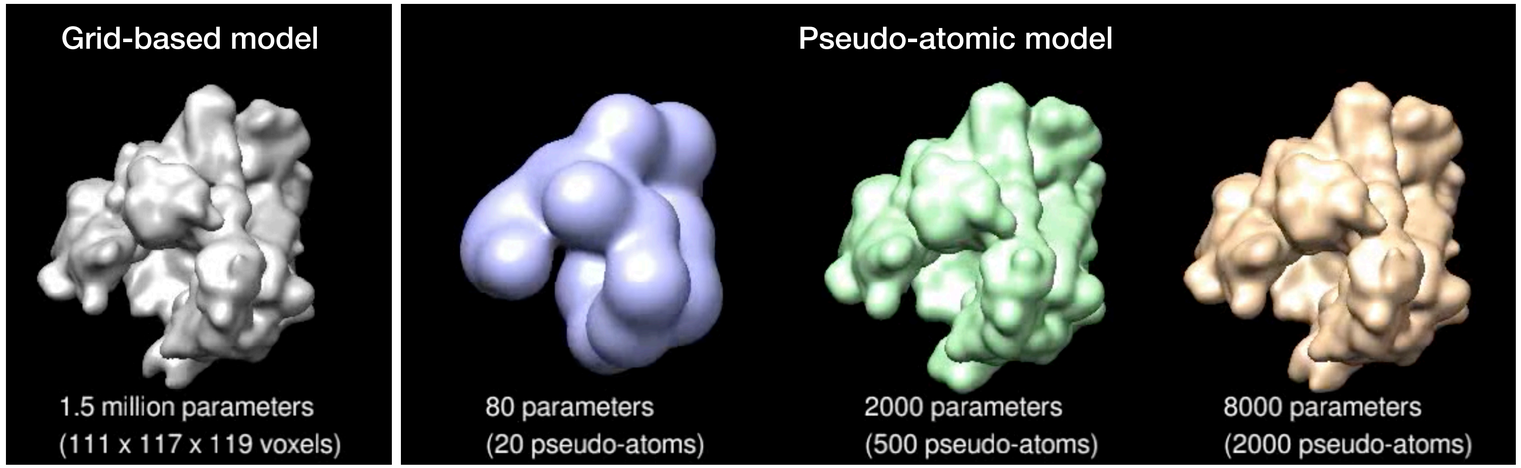
- Use a coarse-grained representation of the 3D structure as **a cloud of K pseudo-atoms**.
- Choose pseudo-atoms to have a Gaussian shape ➡ **Gaussian Mixture Model (GMM)**.

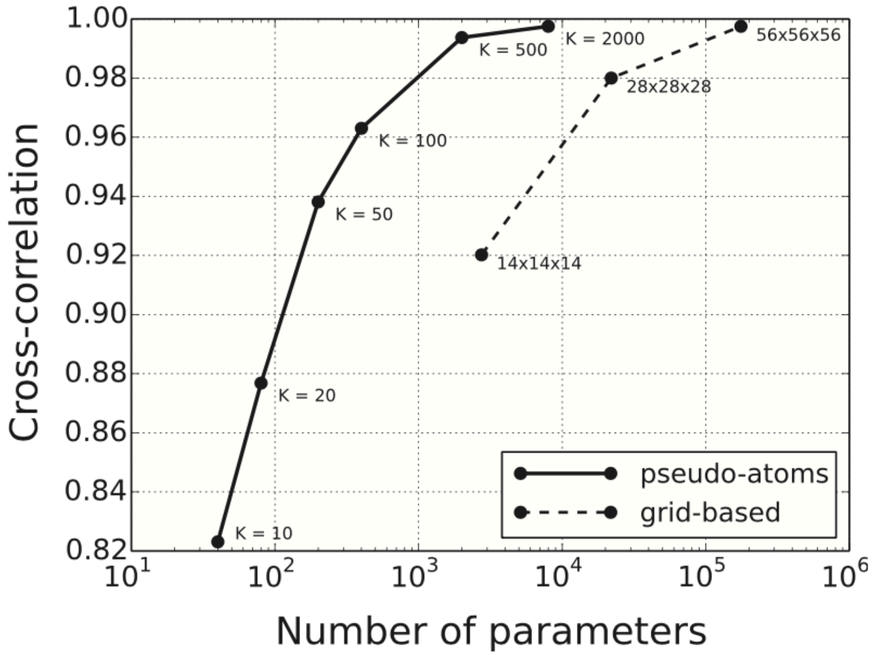
$$\rho(x) = \sum_{k=1}^{K} w_k \, G_{3\mathrm{D}}(x; \mu_k, \sigma)$$
$$= \sum_{k=1}^{K} \frac{w_k}{(2\pi)^{3/2}\sigma^3} \exp\left\{ -\frac{||x - \mu_k||^2}{2\sigma^2} \right\}$$



- Each pseudo-atom is represented by a Gaussian function $G_{3\mathrm{D}}(x; \mu_k, \sigma)$ describing the density at the 3D point x of a pseudo-atom centered at μ_k with radius σ.
- The density map ρ representing the entire 3D structure is a weighted sum of K such pseudo-atoms.
- $||x - \mu_k||^2$ denotes the Euclidean distance between any 3D point x and the position of the pseudo-atom μ_k .

Advantage of Pseudo-Atomic Model — I





- The pseudo-atomic model requires far fewer parameters to describe a 3D structure. (Each pseudo-atom needs only four parameters to describe its position and weight)
- Computing 3D rotations and 2D projections is simple and fast in the pseudo-atomic model w.r.t grid-based model.

Advantage of Pseudo-Atomic Model — II

- The 3D structure is projected along the corresponding direction by first rotating it by R, and then integrating along the z axis to obtain an image in the x, y plane.
- Apply the rotation by transforming each pseudo-atom position μ_k to $R\mu_k$. (R : 3D rotation matrix.)
- Project to the x, y plane by just discarding the z coordinate $PR\mu_k$, where $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$
- Translate the projection by t : $PR\rho(x) = \sum_{k=1}^{K} w_k \, G_{2\mathrm{D}}(x; PR\mu_k + t, \sigma)$, where $G_{2\mathrm{D}}(x; \mu, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{||x-\mu||^2}{2\sigma^2} \right\}$
- The parameters of the pseudo-atomic model, together with a rotation $R_i$and translation t_i for each image:

$$\theta = \left\{ \{\mu_k\},\ \sigma,\ \{w_k\},\ \{R_i\},\ \{t_i\} \right\}$$

  - The size σ of the pseudo-atoms is estimated by the algorithm.
  - Should specify the number of pseudo-atoms, K, which implicitly determines the optimal size σ.

Simulated Data — Data Generation Model

- Generating the i-th image :
  - **Step 1**: Generate a 3D point cloud with C points covering the same region as the pseudo-atomic model. Each point in the point cloud is created by :
    - 1. Randomly selecting a pseudo-atom according to its weight w_k .
    - 2. Randomly placing the point near the center of the pseudo-atom.
  - **Step 2**: The 3D point cloud is then rotated and translated by R_i and t_i , and projected to a 2D point cloud by discarding the z coordinate.
  - **Step 3**: A 2D histogram is formed by using the 2D pixels as bins. (A quantized image = 2D histogram)
- The randomness in generating the 3D point cloud translates into randomness in the generated data $\mathscr{D}$.

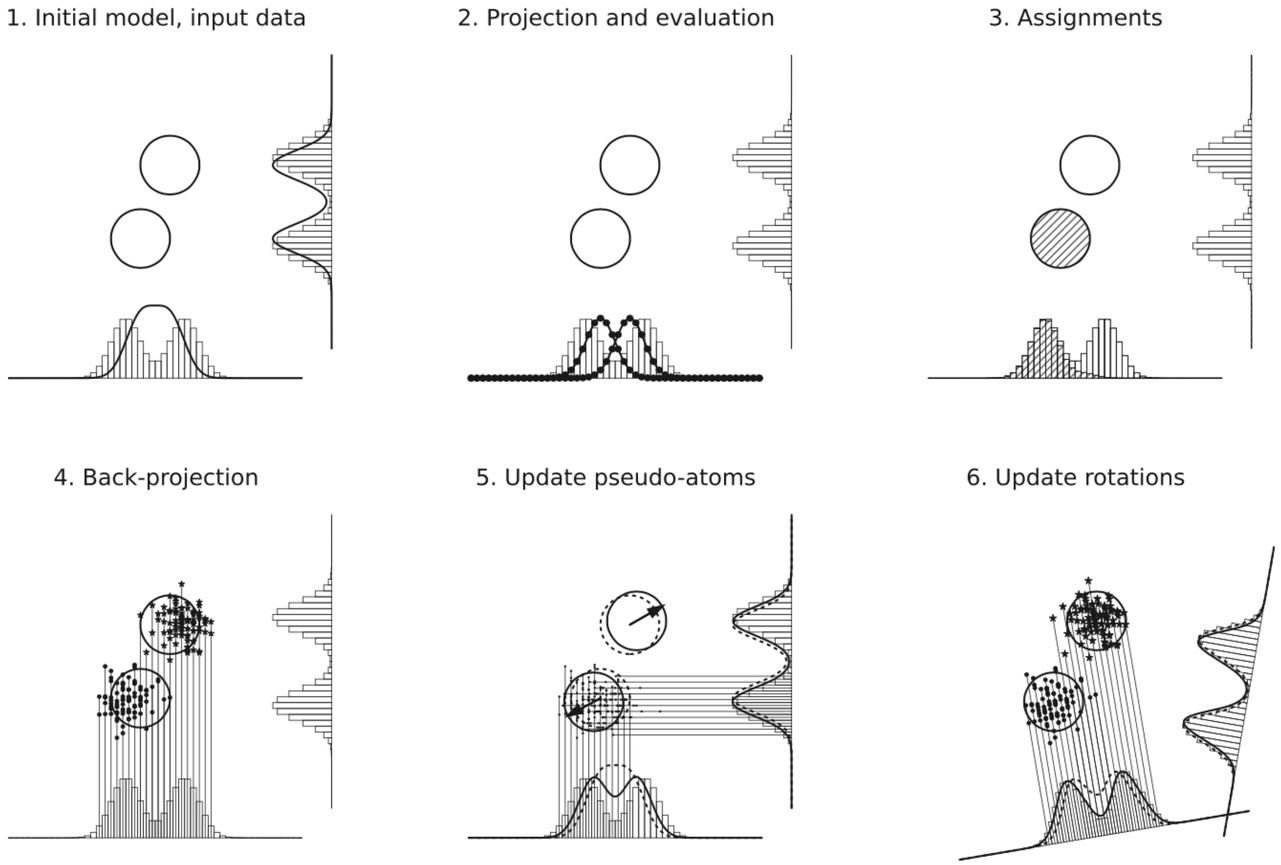
Ab initio Algorithm

- The idea of the ab initio algorithm —> **reverse the data-generation process**.
  - **Step 1**: Starting with 2D points from the quantized image, one first back-project them to 3D points by estimating their missing z coordinates.
  - **Step 2**: Assign each 3D point to a pseudo-atom that was likely to have generated it.
  - **Step 3**: Move the pseudo-atom to align it to its assigned 3D points —> **Gibbs sampling** (similar to expectation maximization)

Gibbs Sampling

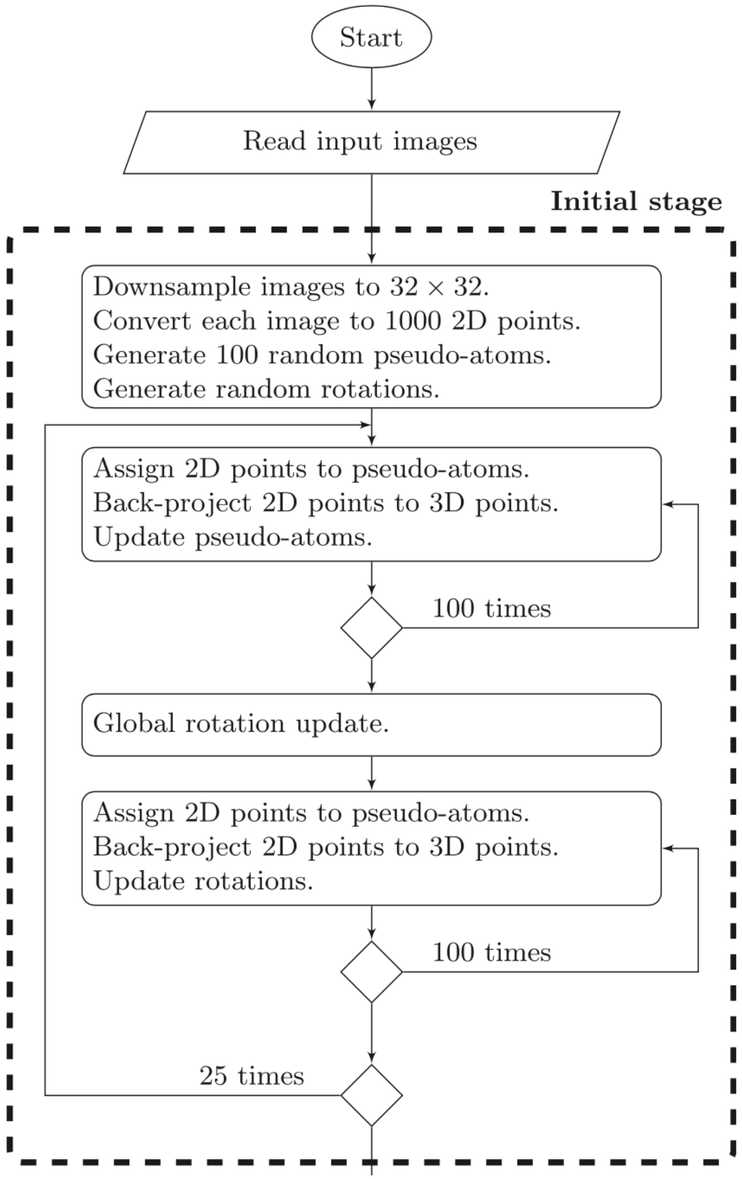
- Gibbs sampling is used for **generating model realizations that follow the posterior distribution** and are therefore consistent with both the data and the prior information.
- Generate a random initial model by sampling the model parameters from the prior distribution.
- The parameters are then updated in turn:
  - The assignments of points to pseudo-atoms and missing z coordinates (back-projection).
  - The pseudo-atom parameters (positions, weights, and size).
  - The rotations and translations.
- Each parameter update depends on the current values of the other parameters.

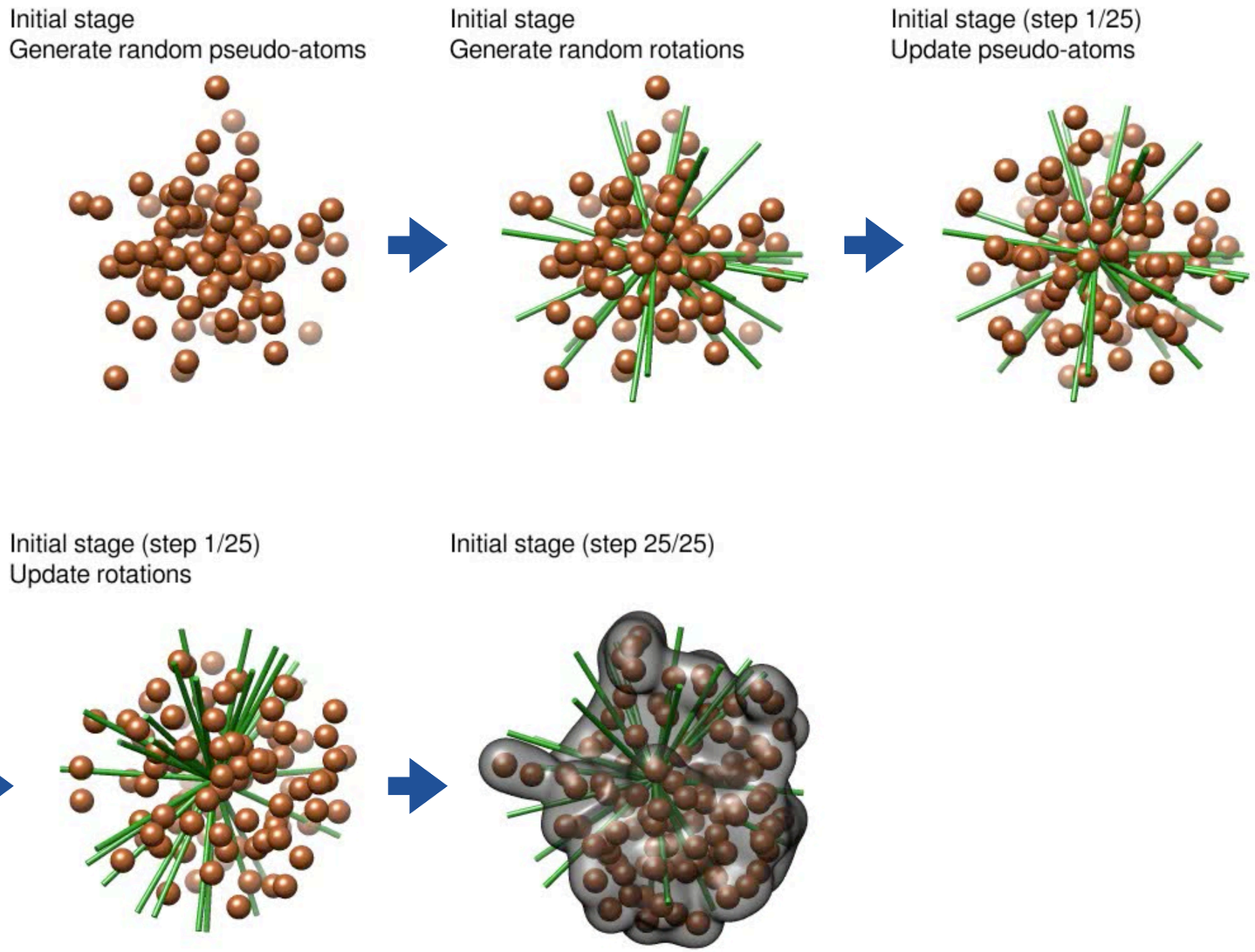
Example of a Single Gibbs Sampling



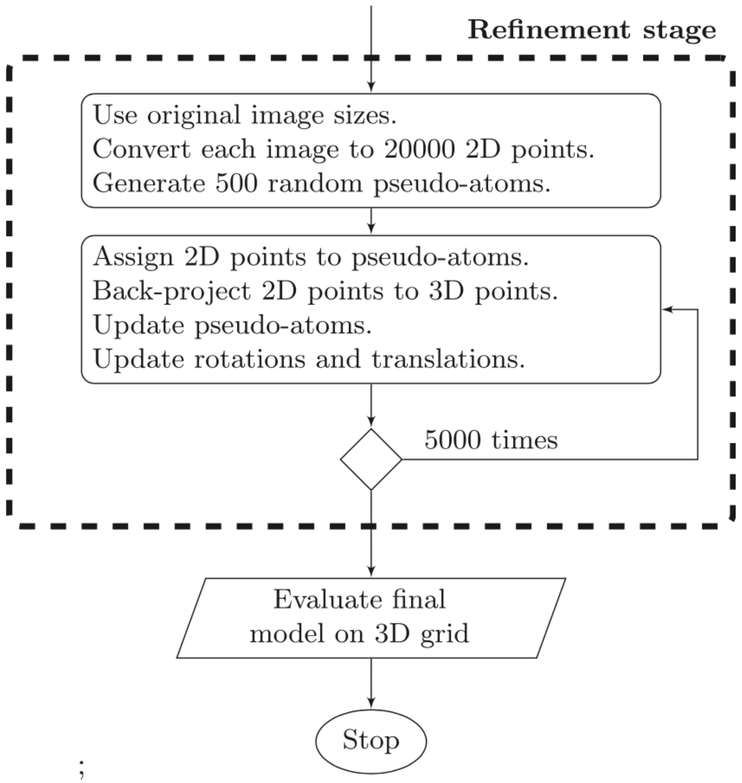
- **Step 1**: Evaluate the 1D projection of each pseudo-atom at all the 1D pixels, and assign points to pseudo-atoms.
- **Step 2**: At each pixel, the relative value of the two pseudo-atoms determines the proportion of points to assign to each.
- **Step 3**: Estimate the missing y coordinates. For each 1D point, its y coordinate is chosen randomly near the y coordinate of the pseudo-atom to which it was assigned.
- **Step 4**: Update the pseudo-atoms, i.e., their weights, positions, and size.
- **Step 5**: Update the image orientation/rotation.

Algorithm — Initial Stage

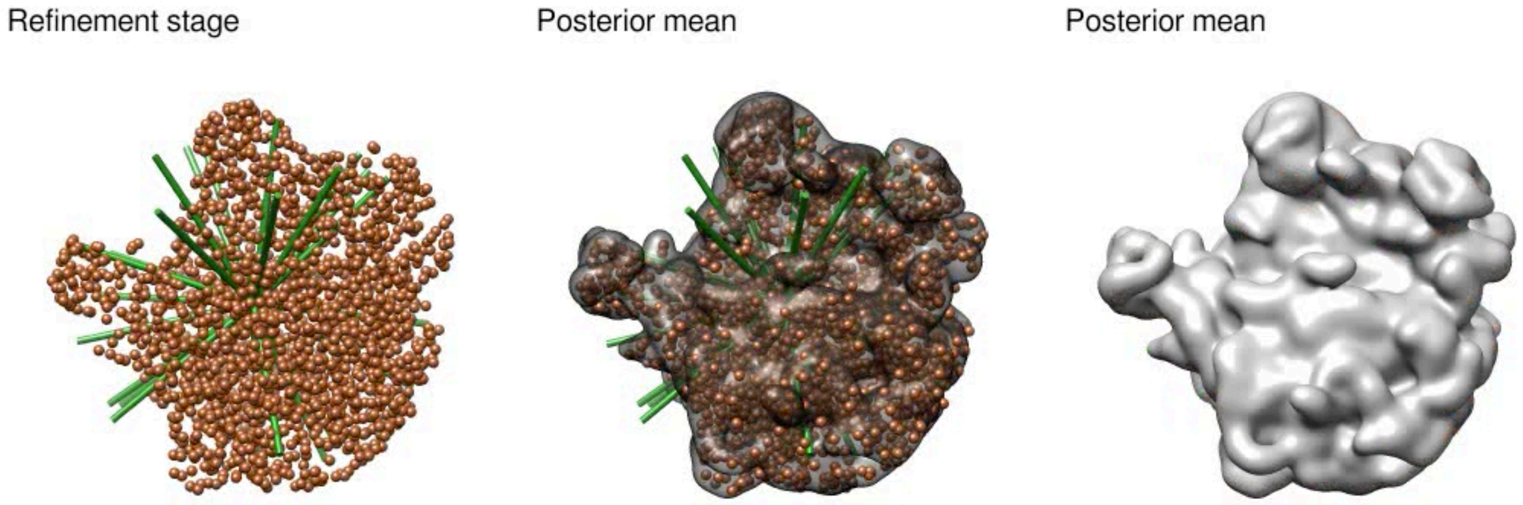




Algorithm — Refinement Stage

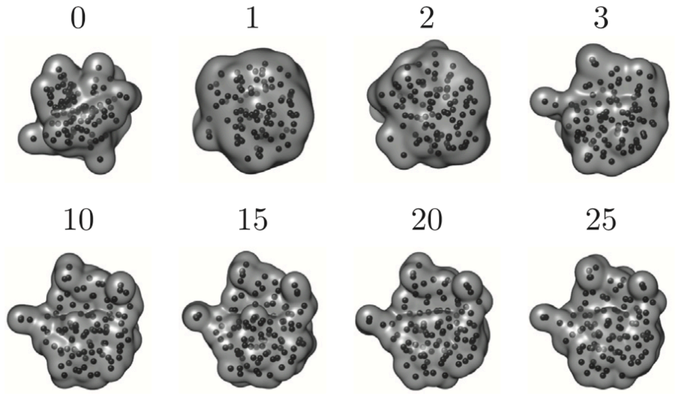


;

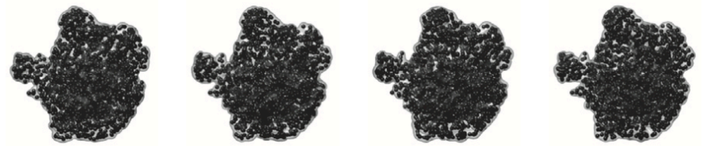


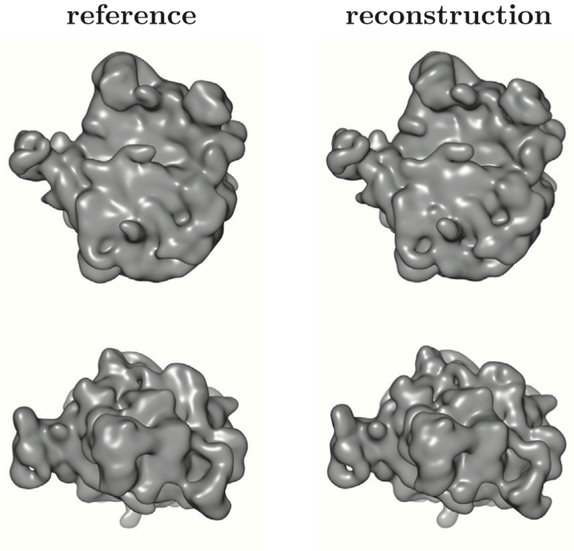
Test of Algorithm 1 — Simulated Class Averages

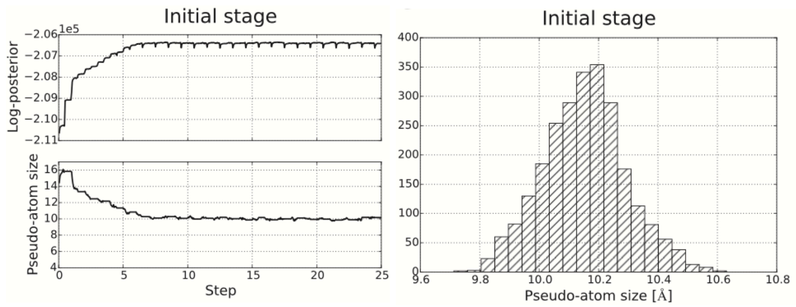
Initial Stage (100 pseudo-atoms)

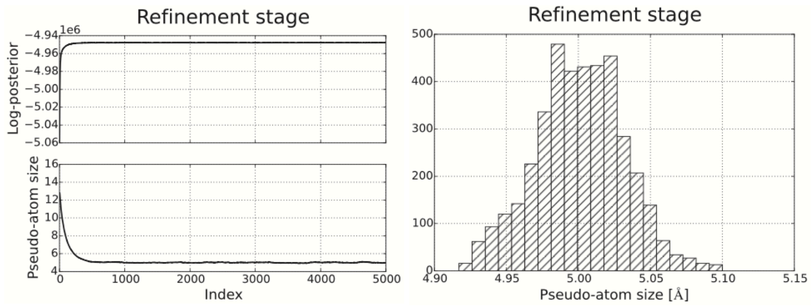


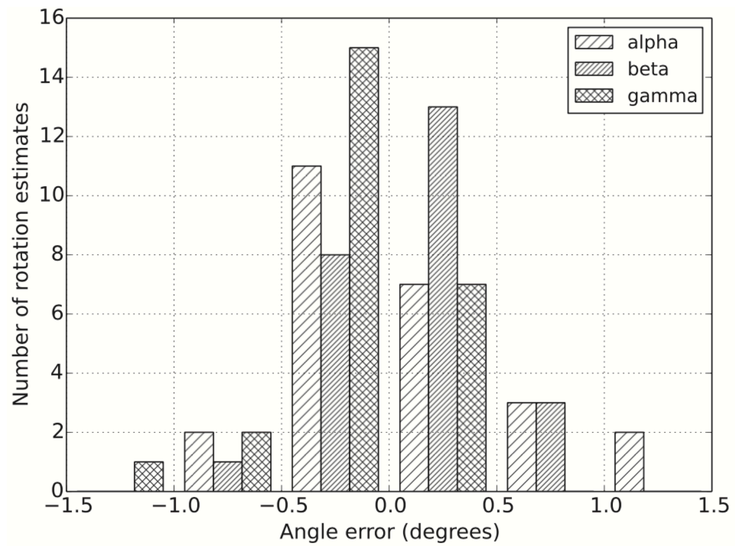
Refinement Stage (2000 pseudo-atoms)





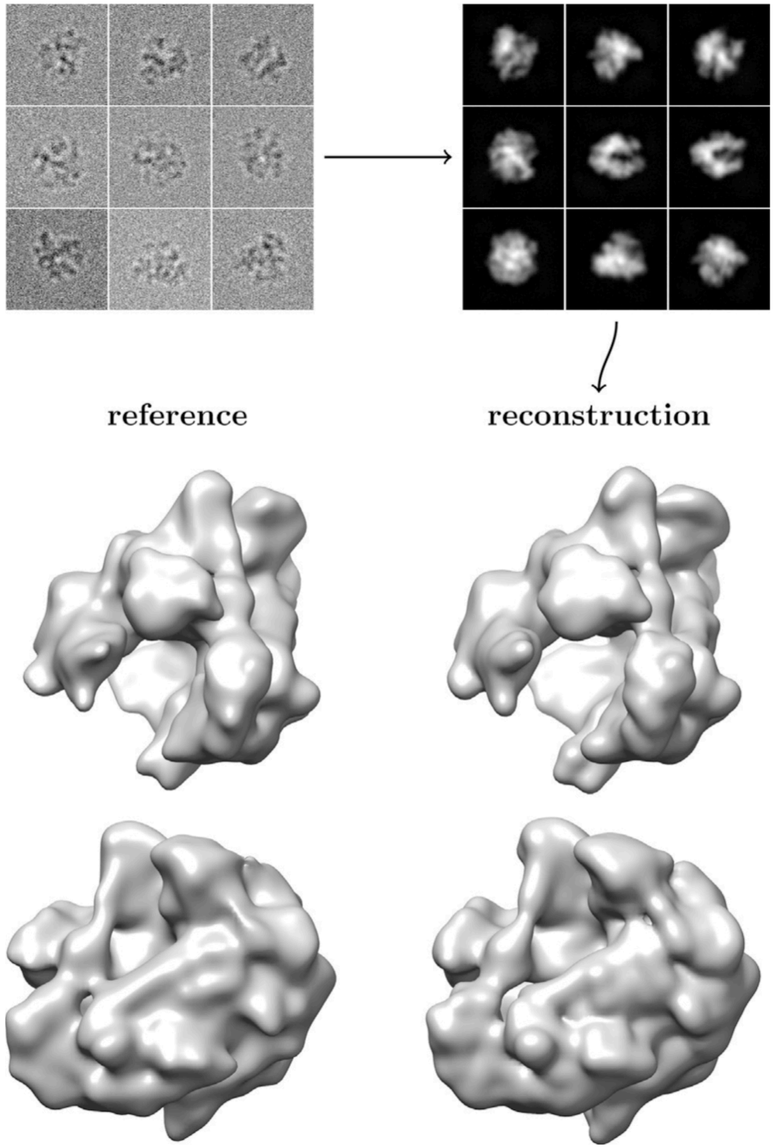






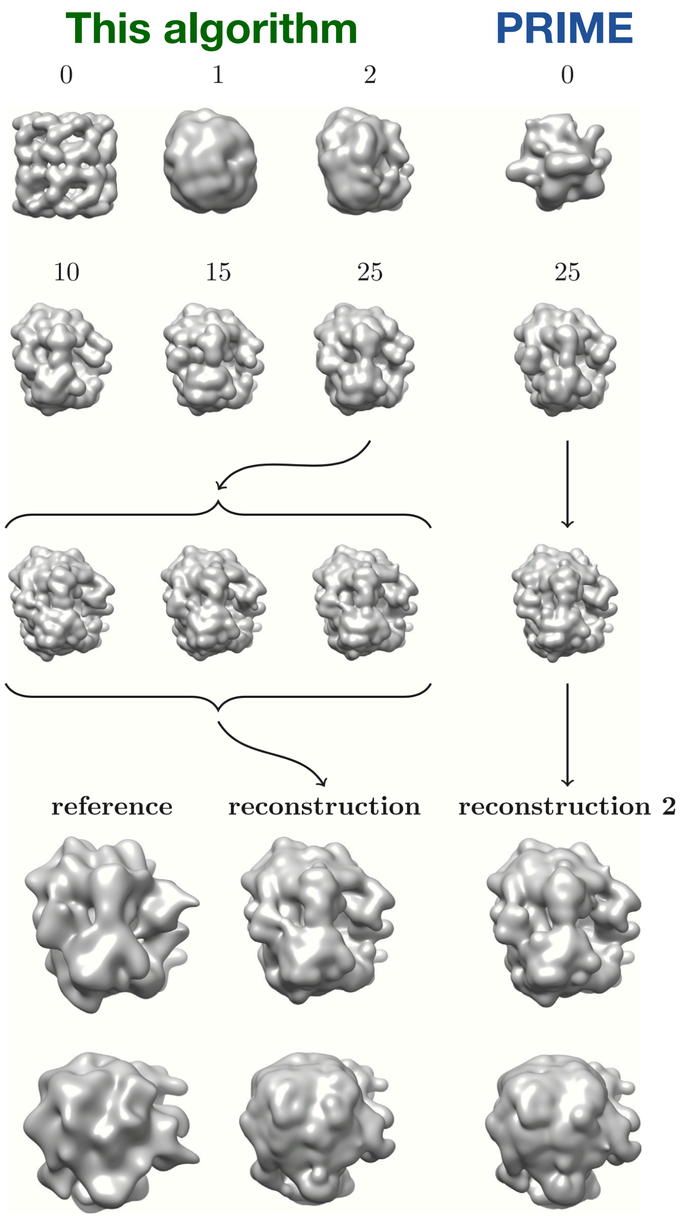
- 25 class averages of the ribosome 50S. Image size = 50x50 pixels w/ sampling rate = 6 Å/pixel
- The computation took 22 min on a single core (7 and 15 min for the initial and refinement stages, respectively).
- The normalized cross-correlation between the true/reconstructed structures is 0.990 at 16 Å.

Test of Algorithm 2 — Realistically Simulated Particles



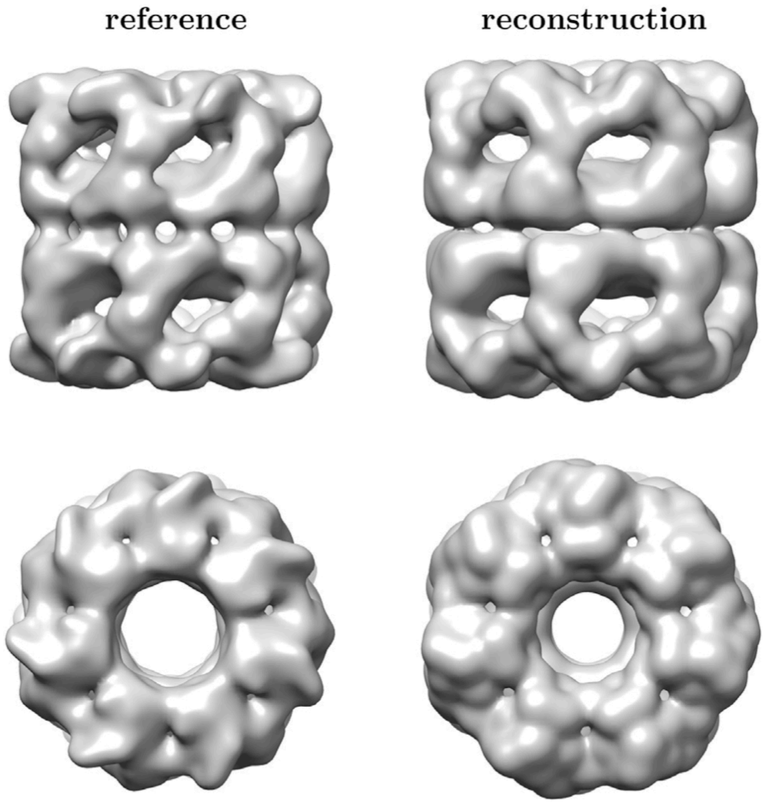
- 5000 realistically simulated RNA polymerase II particles.
- Image size = 100x100 pixels w/ sampling rate = 2.5 Å/pixel
- The CTF was applied with a random defocus value for each image, followed by Gaussian noise with SNR = 0.2.
- Used EMAN2 to compute 41 class averages in 98 min, followed by deconvolution, which took 73 min. Applying this ab initio algorithm to the images took another 38 min, a total of 209 min.
- The final reconstruction has a cross-correlation of 0.966 compared to the reference model at 20 Å.

Test of Algorithm 3 — Experimental 70S Ribosome Data



- The dataset consists of 5000 images with size 130x130 at a sampling rate of 2.82 Å/pixel.
- The normalized cross-correlation between the two structures is 0.900.
- Time cost comparison:
  - The computation of the first reconstruction (starting with GroEL) took 102 min in total (50 min for class averages, 24 min for deconvolution, and 28 min for the initial and refinement stages).
  - The PRIME reconstruction took ~10 h on a cluster with 40 cores.

Test of Algorithm 4 — Experimental GroEL Dataset



- ~5000 images with size 128x128 at a sampling rate of 2.12 Å/pixel.
- EMAN2 was used to obtain 13 class averages in 19 min, followed by deconvolution, which took 2 min. Applying the initial and refinement stages took another 13 min, for a total of 34 min.
- Final result has a cross-correlation of 0.927 with the reference model at 20 Å.

Summary

- With pseudo-atoms method, the number of parameters needed to describe a structure are significantly reduced, which bring some advantages:
  - The algorithm is very fast.
  - Reducing the model complexity.
  - Smooth position description instead of limited by grid.

- During the initial stage when there are only a few large pseudo-atoms it is impossible to represent high-frequency information in the 3D structure, however, this excludes a large number of undesired models from the search space —> **Same as a low-pass filter, which is the property of this model itself**.