De novo structural ensemble determination from single molecule X-ray scattering

CRC 1456 MATHEMATICS OF EXPERIMENT

Steffen Schultze and Helmut Grubmüller

Department of Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Sciences

MAX PLANCK INSTITUTE FOR MULTIDISCIPLINARY SCIENCES

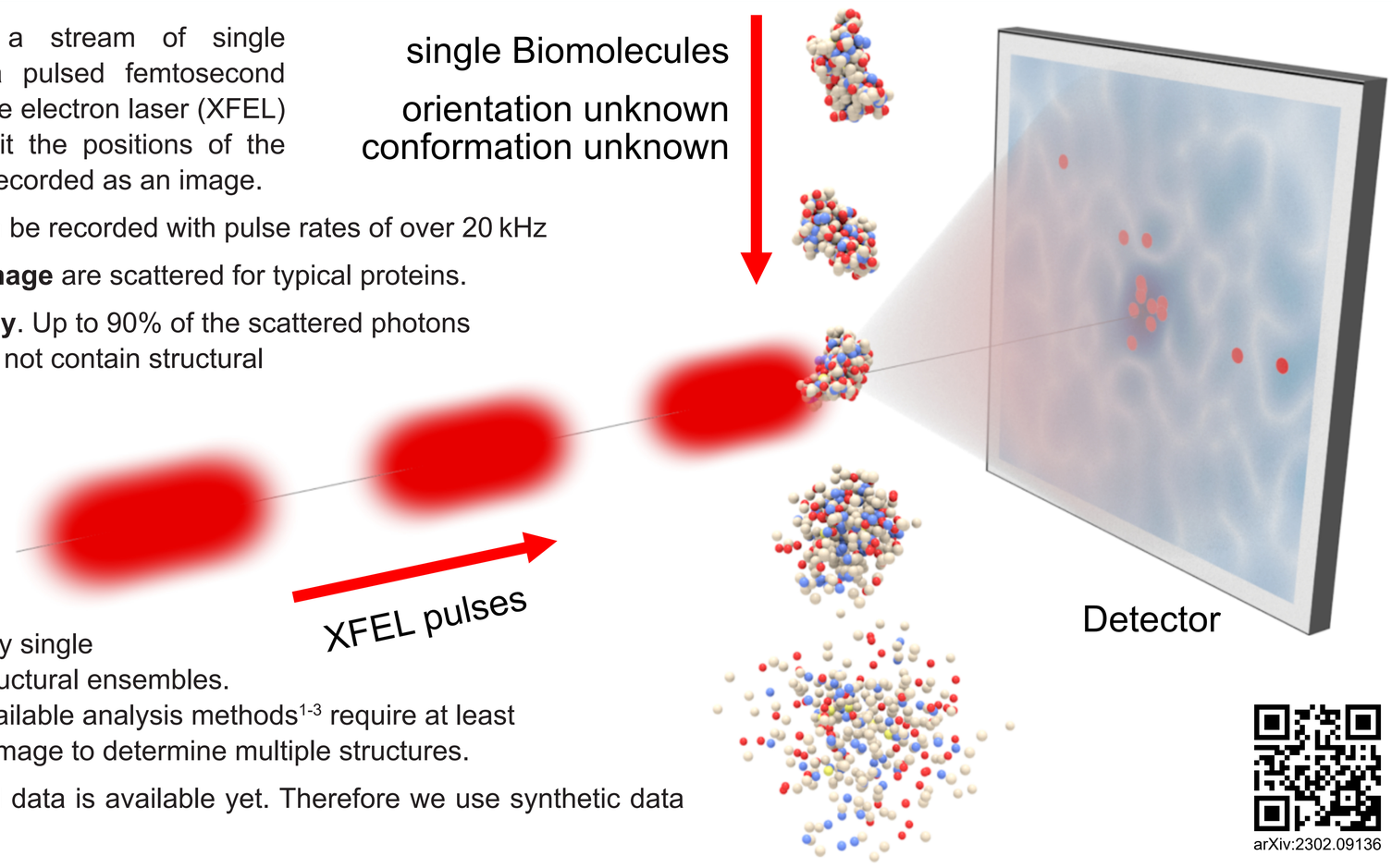
Abstract

Single molecule X-ray scattering experiments with free electron lasers have opened a new route to the structure determination of biomolecules. Because typically only very few photons per scattering image are recorded, structure refinement is quite challenging. In addition, in each scattering event the orientation of the biomolecule is random and unknown. Further, many biomolecules show structural heterogeneity and conformational transitions between different distinct structures; these structural dynamics are averaged out by existing refinement methods.

To overcome these limitations, here we developed and tested a rigorous Bayesian approach and demonstrate that it should be possible to determine not only a single structure, but an entire structural ensemble from these experiments. Using 10^7 synthetic scattering images generated from molecular dynamics trajectories, our approach was able to resolve the unfolded ensemble of the mini-protein chignolin at 4 − 7 Å resolution. These findings show that X-ray scattering experiments using state-of-the-art free electron lasers should allow one to determine not only biomolecular structures, but whole structure ensembles and, ultimately, 'molecular movies'.

Introduction

In the experiments, a stream of single biomolecules enters a pulsed femtosecond high-intensity X-Ray free electron laser (XFEL) beam, and for each hit the positions of the scattered photons are recorded as an image.



- **10^6 to 10^9** images can be recorded with pulse rates of over 20 kHz
- **10-50 photons per image** are scattered for typical proteins.
- **Images are very noisy**. Up to 90% of the scattered photons are incoherent and do not contain structural information.
- **Orientations are unknown.**
- **Conformations are unknown.**

Potentially, this allows determination of not only single structures but whole structural ensembles.
However, previously available analysis methods[1-3] require at least 100-1000 photons per image to determine multiple structures.

Almost no experimental data is available yet. Therefore we use synthetic data to test our approach.

arXiv:2302.09136

Method

A

$$P\left(\text{50\% 40\% 10\%} \,\middle|\, \text{images}\right) \sim P\left(\text{images} \,\middle|\, \text{50\% 40\% 10\%}\right) \cdot P\left(\text{50\% 40\% 10\%}\right)$$

B

$$P(\mathcal{I} \mid \boldsymbol{\rho}, \mathbf{w}) \propto \underbrace{\prod_{(\mathbf{k}_1,\ldots,\mathbf{k}_l)\in\mathcal{I}}}_{\text{product over images}} \underbrace{\sum_{i=1}^{n} w_i}_{\text{weighted ensemble average}} \underbrace{\int_{\mathrm{SO}(3)}}_{\text{average over orienations}} \underbrace{\exp\left(-N \int_D \left|\mathcal{F}\{\rho_i\}(\mathbf{Rk})\right|^2 \mathrm{d}\mathbf{k}\right)}_{\text{Poisson distribution for number of photons}} \underbrace{\prod_{j=1}^{l} \left|\mathcal{F}\{\rho_i\}(\mathbf{Rk}_j)\right|^2 \mathrm{d}\mathbf{R}}_{\text{product over single photons}}$$

C

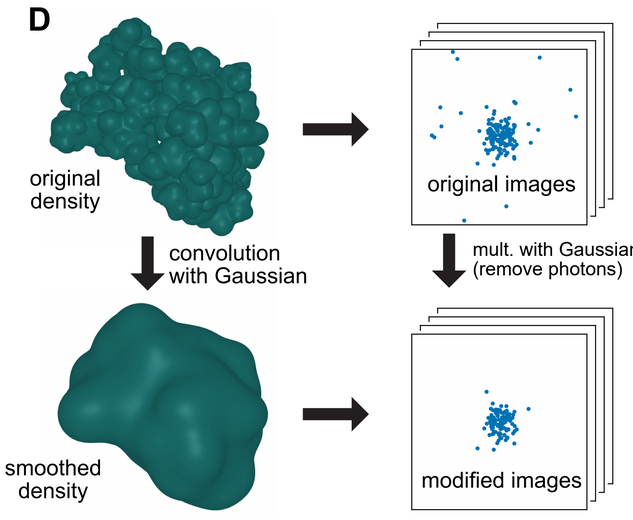
$$\rho(\mathbf{r}) = \sum_{i=1}^{m} \frac{h_i}{\left(\sigma_i\sqrt{2\pi}\right)^3} \exp\left(\frac{1}{2\sigma_i^2}\|\mathbf{r}-\mathbf{y}_i\|^2\right)$$

A. Our Bayesian approach determines a weighted ensemble of structures by maximizing or sampling from the Bayesian posterior probability, computed using Bayes' theorem.

B. The likelihood decomposes into a product of the probabilities of the independent images, and for each image, includes a weighted ensemble average and an integral over all possible orientations of the molecule.

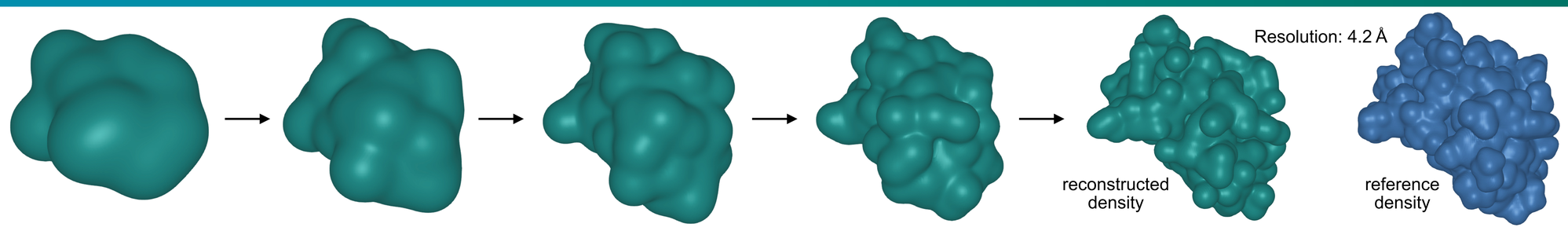
C. The structures are represented as a sum of Gaussian functions. The optimal positions of these Gaussians are determined using simulated annealing.

D. To enhance convergence, sampling is performed in multiple hierarchical stages of increasing resolution. For this, images corresponding to smoothed structures are generated by rejection sampling using the convolution theorem.



Results

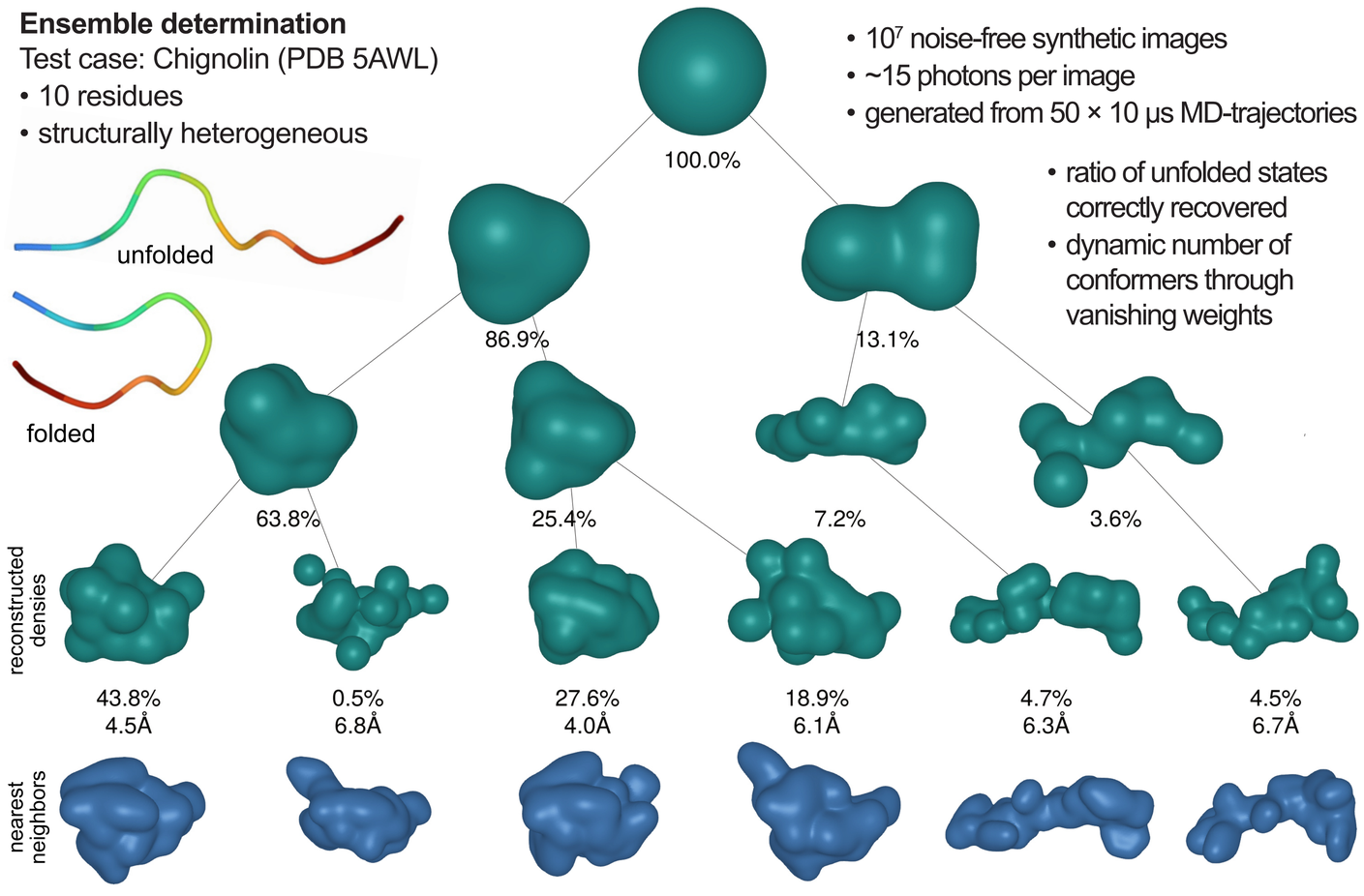
Single structure determination
Test case: Crambin (PDB 1EJG)
- 46 residues
- 10^8 noise-free synthetic images
- ~15 photons per image
- five hierarchical stages



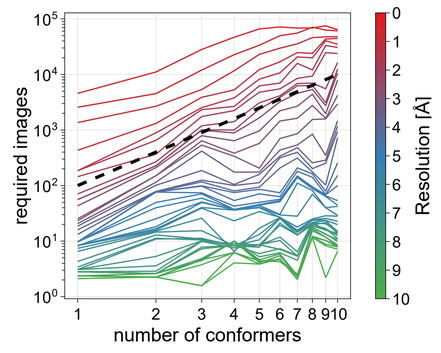
Ensemble determination
Test case: Chignolin (PDB 5AWL)
- 10 residues
- structurally heterogeneous

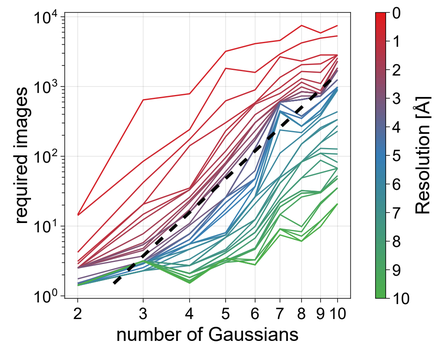
- 10^7 noise-free synthetic images
- ~15 photons per image
- generated from 50 × 10 µs MD-trajectories

- ratio of unfolded states correctly recovered
- dynamic number of conformers through vanishing weights

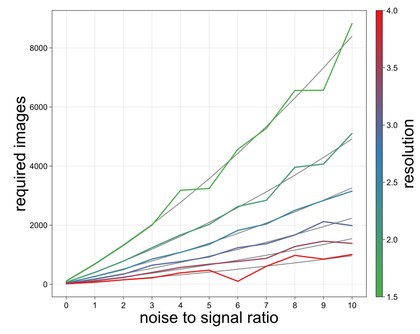


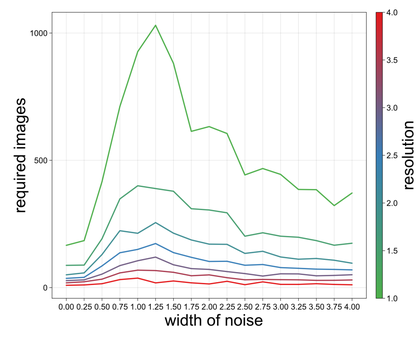
How many images are needed?
- determining an ensemble of n conformers requires $O(n^2)$ images
- determining a single structure with m atoms requires $O(n^{5\pm1})$ images





Robustness to noise (preliminary)
- in the presence of noise more images are required
- strongly dependent on distribution of noisy photons





References

1. Zhuang, Y. et al. 2022. IUCrJ 9, 204–214.
2. Hosseinizadeh, A. et al. 2012. Nature Methods 14, 877–881.
3. von Ardenne, B., Mechelke, M. & Grubmüller, H. 2018. Nat Commun 9, 2375.

Acknowledgements

Funded by Federal Ministry of Education and Research, joint research project 05K20EGA Fluctuation XFEL and Deutsche Forschungsgemeinschaft, CRC 1456/1 Mathematics of Experiments.
MD-trajectories kindly provided by Nicolai Kozlowski.