# Cryo-EM, XFELs and the structure conundrum in structural biology

Single-particle techniques offer an unprecedented opportunity to understand the role of structural variability in biological function. They also call into question the meaning of 'a structure' and its relevance to function.

Abbas Ourmazd

Ludwig Wittgenstein famously said, the limits of our language define the limits of our world. If so, a key limit to our understanding of structural biology is this: what do we mean by 'structure'?

We are interested in biomolecular structure, because it sheds light on biological function. But function involves changes in structure.

Structural changes are often thought of as 'jumps' between a small number of discrete structures separated by high energy barriers. But there is longstanding evidence[1,2] that the hierarchy of barriers includes surprisingly low ones — at most, a few times the thermal energy available under physiological conditions[3,4]. This means that many structures coexist, each with a sizable probability. So many, in fact, that the concept of one or even a few discrete structures seems inadequate, if not outright misleading[5,6].

The coexistence of many structures means that 'a structure' is but a point in a continuum of conformational states. More specifically, the possible conformations of a system span a continuous (hyper)surface, with a dimensionality determined by the number of conformational degrees of freedom available to the system[6] (Fig. 1). If the object of interest were my five-fingered (and stiff-jointed) hand, the hypersurface describing the conformations of my hand would be five-dimensional. In Fig. 1, each one-dimensional trajectory (line) on the hypersurface represents a specific sequence of conformational changes. Just as my fingers can move in many ways, there are many possible trajectories on the conformational hypersurface. But only a few of the trajectories correspond to the action of, say, threading a needle. This means only a small subset of all possible conformational trajectories is functionally relevant. If we are interested in function, we need to identify these functional trajectories.

Prevailing approaches in structural biology attempt to infer the functional trajectory by careful study of a few discrete structures. This is plainly inadvisable. You cannot learn to swim from photos of a swimmer before she has dived in, and after she has got out of the pool, even if you include a few more snapshots along the way. Swimming corresponds to a specific sequence of continuous conformational motions — a specific trajectory on the conformational hypersurface. We must discover this specific conformational trajectory if we are to learn what conformational changes lead to a pleasurable swim, rather than death by drowning.

The key lies in recognizing that different conformations have different free energies. This means the energy landscape of biomolecular conformations is not flat like a pancake, but a hilly landscape with peaks and valleys[2,3,7]. For theoretical and experimental reasons, one expects the rivulets in this energy landscape to represent functional paths, at least near equilibrium[4,8] (Fig. 2).

As outlined in Box 1, such energy landscapes can be compiled by counting the frequency with which each conformation occurs in an ensemble of molecules in equilibrium[3,9]. This rules out powerful techniques such as crystallography and nuclear magnetic resonance (NMR), which average over all the conformations present in the sample. One must therefore utilize so-called single-particle methods to obtain snapshots of the individual members of an ensemble. At present, cryogenic electron microscopy (cryo-EM) is the primary means of obtaining structural snapshots of single-particles. Having labored for decades under the moniker 'cryo-blobology' for its relatively poor spatial resolution, modern cameras have enabled cryo-EM to deliver structural snapshots of individual particles at near-atomic resolution. But the 'cryogenic' part remains; the biological entities examined are frozen in vitreous ice. Their conformational spectrum may thus be different from that present under physiological conditions. Equally important, the largest datasets currently available comprise, at most, a few million single-particle snapshots. This, as will shortly become clear, is another significant limitation.

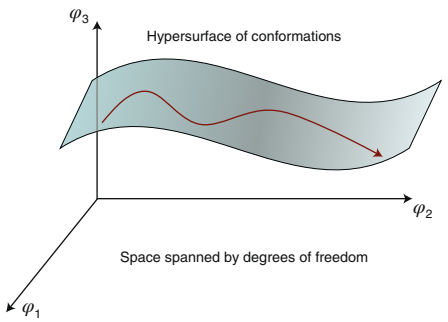


**Fig. 1 | Hypersurface representing all possible conformations of a molecule.** The hypersurface is described in terms of conformational reaction coordinates, $\varphi$, learned from the data. Each point represents a specific conformation, with similar conformations lying close to each other. The dimensionality of the hypersurface is determined by the number of conformational degrees of freedom available to the molecule. The red line represents a specific conformational trajectory. If the hypersurface represents all possible motions of my body, then swimming corresponds to a specific trajectory on the hypersurface. Figure adapted from ref. [27], Springer Nature Ltd.

The examination of functionally relevant ('functional') conformational trajectories from experimentally determined energy landscapes offers substantial information beyond static structures[10], revealing, for example, the propagation of allosteric signals in complex biological molecules, and important clues to the functional sensitivity of sites associated with disease. In fact, functional information derived from energy landscapes can be very different from that inferred from discrete rigid structures. The differences include not only the sequence and the extent of conformational motions, but even the nature of the conformationally active structural blocks[10]. This calls into question the wisdom of deriving functional information from static structures.

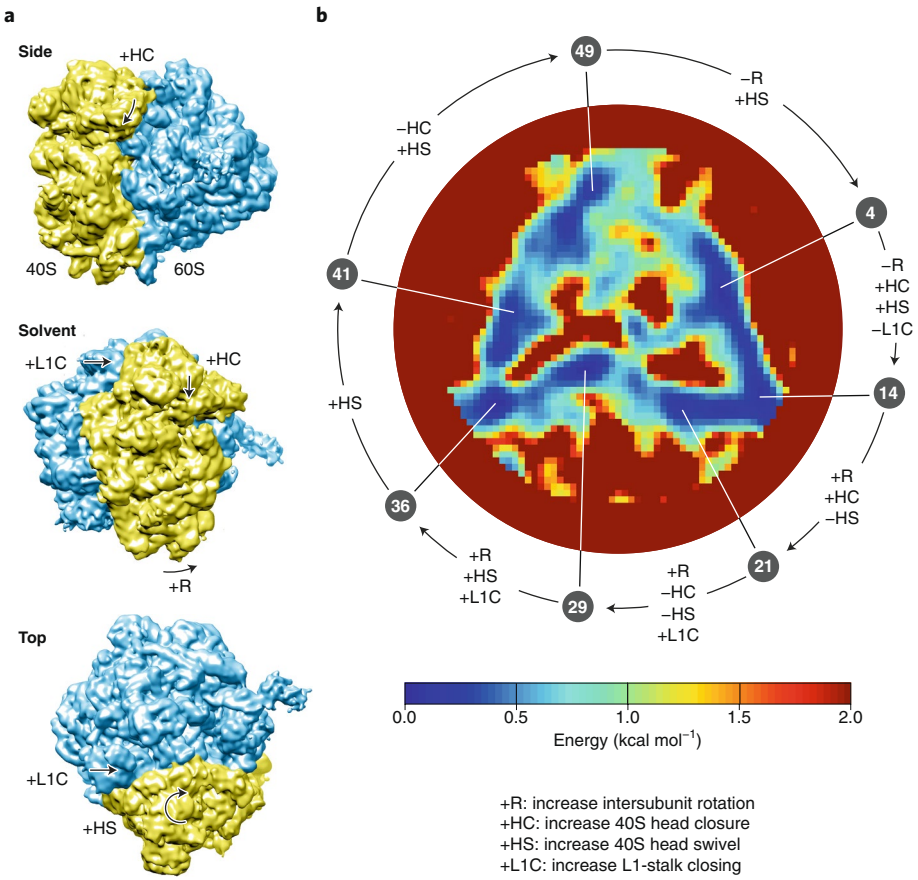


**Fig. 2 | Experimentally determined conformational motions and energy landscape of a ribosome.** **a**, Three views of a cryo-EM map of the 80S ribosome from yeast, with arrows indicating four key conformational changes associated with the elongation work-cycle of the ribosome. **b**, The experimentally determined energy landscape traversed by the ribosome. The color bar shows the energy scale. The error in energy determination along the closed triangle is 0.05 kcal mol$^{-1}$. The roughly triangular minimum-free-energy trajectory is divided into 50 states. The arrows indicate the structural changes between seven selected states, each identified by its place in the sequence of 50 states. Figure reproduced from ref. [3], PNAS.

So far, we have discussed near-equilibrium processes. Of course, one would also like to study the functional motions associated with exothermic processes, such as those driven by ATP hydrolysis or the absorption of one or more photons[11,12]. As outlined in Box 1, a collection of single-particle snapshots samples all states of the system, including those at high energies, albeit with exponentially diminishing probability[6,10]. A 'sufficiently large' dataset of snapshots will thus include high-energy conformations. States at energies comparable with that released by ATP hydrolysis, for example, begin to appear in datasets with ~$10^9$ single-particle snapshots obtained from an equilibrium ensemble of molecules[6].

This offers the possibility to investigate biological processes based on their functional trajectories, without having to 'track' each process in time. A particularly exciting possibility is gaining direct access to the structure of so-called transition states, which critically determine the course of biochemical reactions. The key is the ability to collect and analyze billion-strong collections of single-particle snapshots. Until recently, this was beyond the horizon.

Not long ago, X-ray free-electron lasers (XFELs) revolutionized crystallographic techniques[13]. The recent advent of high-repetition-rate XFELs now promises the collection of very large, single-particle datasets, and emerging data-analytical techniques offer the possibility to analyze them[6,14]. To obtain single-particle snapshots by XFEL, one intersects single, hydrated molecules with intense pulses of X-rays, each a few femtoseconds ($10^{-15}$ s) long[15]. The molecule is, of course, blown to smithereens by the pulse, but only after it has diffracted a few photons. So, one keeps repeating the procedure until a sufficiently large number of snapshots has been collected[16,17]. As the object of interest is encased in water (rather than ice, as in cryo-EM), it is in the biologically relevant state.

Much like cryo-EM before its so-called resolution revolution, the spatial resolution of XFEL single-particle methods is currently in the nanometer range[18]. This modest spatial resolution stems primarily from difficult-to-control instrumental factors, including shot-to-shot variations in the position and intensity of the incident X-ray beam, stray scattering from up-stream apertures, the thickness of the hydrating water jacket surrounding each particle injected into the X-ray beam, and nonlinear and history-dependent detector response[14]. An international collaboration initiated four years ago by the linear collider light source at SLAC National Laboratory in California has made substantial progress in mitigating these issues by a combination of instrumental measures and post-facto algorithmic corrections[19]. In the latest experiments, the achieved spatial resolution was limited to 9 nm by the placement of the detector[18].

Importantly, it seems the intensity of the incident pulse is not the limiting factor. Indeed, recent experimental and theoretical advances indicate that it is

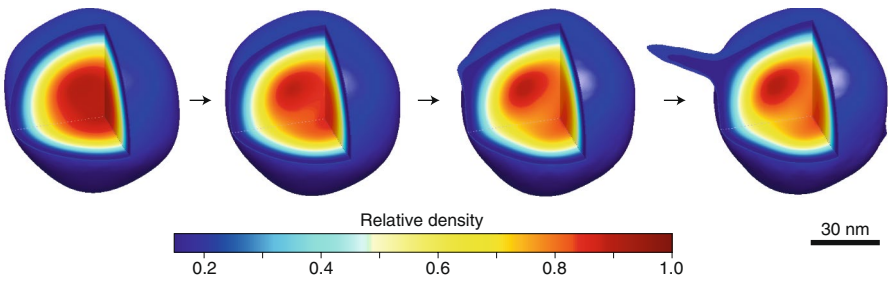


**Fig. 3 | Four frames of a 50-frame movie showing the conformational changes in the PR772 virus.** The movie was compiled from experimental single-particle XFEL snapshots. Note the accumulation of viral content near the fivefold portal, from which a tubular structure emerges. Figure reproduced from ref. [18], Springer Nature America, Inc.

### Box 1 | What is an energy landscape?

In an ensemble of molecules in equilibrium, the probability of finding a specific conformation drops exponentially with the free energy of the conformation. One can, therefore, use the frequency with which a particular conformation is encountered in a random sample of snapshots to determine its energy. The larger the dataset, the greater the probability that high-energy conformations will be sighted. An energy landscape thus includes conformations up to the highest energy determined by the size of the dataset[6,10].

In practice, one sorts a collection of snapshots from individual members of the ensemble into conformationally homogeneous bins approximating a continuum. How many bins are needed for a complete description is determined by the Shannon–Nyquist sampling theorem and the signal-to-noise ratio[3].

An energy landscape is, in general, multi-dimensional. The dimensionality of an energy landscape depends on the number of degrees of freedom exercised by the system during the experiment. The degrees of freedom give rise to so-called conformational reaction coordinates (RCs). Each reaction coordinate controls a concerted set of conformational motions.

The number and suitable set of RCs must be 'learned' from the data. It is inadvisable to adopt arbitrary parameters (for example, an angle or a distance between two parts of a molecule) as RCs. These may not capture the functionally relevant changes, and are unlikely to be orthogonal. Orthogonality ensures that each RC represents a distinct set of concerted changes.

Many important biological processes, such as those triggered by the binding of a ligand molecule, involve more than one energy landscape[10]. In such cases, the functional path crosses from one landscape to another, with the crossing probability varying from point to point.

possible to recover three-dimensional structure by XFEL techniques at incident X-ray intensities substantially lower than previously thought necessary[18,20,21]. Further progress requires improved control of imaging conditions, a higher success rate in intersecting each injected particle with an X-ray pulse, and data-analytical algorithms able to correct instrumental artefacts post facto to extract conformational information from large (and noisy) datasets. These constitute difficult, but not impossible challenges. A concerted 'pre-competitive' international effort is likely the most efficient route to solving them.

XFEL-based approaches are, of course, still in their infancy. Fairly or unfairly, let us discount the possibility of further progress and ask whether very large, noisy datasets with modest spatial resolution can be of scientific value. The answer is very likely yes. A recent proof-of-principle study demonstrated the possibility of mapping the concerted conformational changes associated with the extrusion of the genome from a virus (Fig. 3, Box 2), without the need for tracking the process in time[18]. It is reasonable to expect that many important functional trajectories entail large, high-energy, and thus rarely sighted, conformational motions. In such cases, the ability to compile and data-analytically extract specific functional information from large, modest-resolution datasets promises unprecedented experimental access to fleetingly occupied, rate-limiting states in biological and chemical processes. Thus, when combined with molecular simulations, one may be able to gain insight even into the atomic-level details of cooperative conformational motions. Reminiscent of the powerful synergy between quantum mechanical calculations and infra-red spectroscopy[22], advances in cryo-EM and XFEL single-particle techniques may allow one to compare experimental and simulated molecular movies, and thus gain a deep understanding of biological function at the atomic level[10].

Not all biological processes, however, are amenable to the energy-landscape approach. Examples include strongly driven processes far from equilibrium, such as those induced by the absorption of a photon. Under these conditions, concepts borrowed from (near-) equilibrium thermodynamics, such as free energy, are no longer valid. Indeed, the outcome of a non-equilibrium process is route- and rate-dependent. How much fuel you need to reach Chicago from Milwaukee depends on the route you take, and how fast you drive. This means one cannot directly compare different non-equilibrium experiments, even if they both involve the same conformational trajectory. Bridging the gap between equilibrium concepts (such as energy landscapes) and the intrinsically non-equilibrium nature of life is thus an important research goal.

One way to circumvent this problem is to 'transform' the non-equilibrium process of interest to what would have been observed in a series of quasi-equilibrium processes. In this picture, a non-equilibrium process corresponds to a thread through a stack of quasi-equilibrium energy landscapes. Alternatively, one might develop a recipe for extracting equilibrium parameters from non-equilibrium measurements, as demonstrated in the celebrated approach pioneered, for example, by Jarzinski[23,24], Crooks[25] and others[26].

A rigorous treatment of non-equilibrium biological processes constitutes what we in the United States call 'a challenge'. Personally, I prefer the German equivalent: 'it is music of the future' — a happy thought. ❐

### Box 2 | Mapping the conformations of a virus by single-particle XFEL

Recent experimental results have demonstrated that complex conformational changes can be detected and accurately quantified by single-particle X-ray scattering[18]. The movie frames shown in Fig. 3 pertain to the PR772 virus, a member of the Tectiviridae family. The data-analytical approach directly revealed the reaction coordinate controlling concerted conformational changes in the virus. For the viral particles investigated, a single conformational-coordinate couples and controls the reorganization of the internal membrane and genome, the growth of a tubular structure from a portal vertex, and the release of the genome. These observations clearly demonstrate the capability of determining the landscape associated with complex conformational changes in biological entities by XFEL single-particle techniques. This opens the way to systematic investigation of the influence of biologically important effects, such as pH, temperature, ligands and cofactors, on function.

Abbas Ourmazd
*Department of Physics, University of Wisconsin Milwaukee, Milwaukee, WI, USA.*
*e-mail: ourmazd@uwm.edu*

Published online: 18 September 2019
https://doi.org/10.1038/s41592-019-0587-4

References

1. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. *J. Mol. Biol.* 7, 95–99 (1963).

2. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. *Science* **254**, 1598–1603 (1991).
3. Dashti, A. et al. *Proc. Natl. Acad. Sci. USA* **111**, 17492–17497 (2014).
4. Henzler-Wildman, K. & Kern, D. *Nature* **450**, 964–972 (2007).
5. Frank, J. *Biochemistry* **57**, 888 (2018).
6. Ourmazd, A. in *X-Ray Free Electron Lasers: Applications in Materials, Chemistry and Biology Energy and Environment Series* (eds Uwe Bergmann, V. Yachandra & J. Yano) 418–433 (Royal Society of Chemistry, 2017).
7. Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V. & Stark, H. *Nature* **466**, 329–333 (2010).
8. Neu, J. C., Ghanta, A. & Teitsworth, S. in *Coupled Mathematical Models for Physical and Biological Nanoscale Systems and Their Applications* **Vol. 232** Springer Proceedings in Mathematics and Statistics (eds Bonilla, L. L. et al.) 153–167 (Springer 2018).
9. Frank, J. & Ourmazd, A. *Methods* **100**, 61–67 (2016).
10. Dashti, A. et al. Preprint at https://doi.org/10.1101/291922 (2019).
11. Pande, K. et al. *Science* **352**, 725–729 (2016).
12. Kupitz, C. et al. *Nature* **513**, 261–265 (2014).
13. Boutet, S. et al. *Science* **337**, 362–364 (2012).
14. Hosseinizadeh, A., Dashti, A., Schwander, P., Fung, R. & Ourmazd, A. *Struc. Dyn.* **2**, 041601 (2015).
15. Gaffney, K. J. & Chapman, H. N. *Science* **316**, 1444–1448 (2007).
16. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. *Nature* **406**, 752–757 (2000).
17. Ekeberg, T. et al. *Phys. Rev. Lett.* **114**, 098102 (2015).
18. Hosseinizadeh, A. et al. *Nat. Meth.* **14**, 877–881 (2017).
19. Munke, A. et al. *Sci. Data* **3**, 160064 (2016).
20. Ayyer, K. et al. Preprint at https://arxiv.org/abs/1905.05008 (2019).
21. von Ardenne, B., Mechelke, M. & Grubmuller, H. *Nat. Commun.* **9**, 2375 (2018).
22. Neugebauer, J., Reiher, M., Kind, C. & Hess, B. A. *J. Comput. Chem.* **23**, 895–910 (2002).
23. Jarzynski, C. *Phys. Rev. Lett.* **78**, 2690–2693 (1997).
24. Jarzynski, C. *Phys. Rev.* **56**, 5018–5035 (1997).
25. Crooks, G. E. *Phys. Rev.* **61**, 2361 (2000).
26. Pohorille, A., Jarzynski, C. & Chipot, C. *J. Phys. Chem.* **114**, 10235–10253 (2010).
27. Fung, R., Shneerson, V., Saldin, D. K. & Ourmazd, A. *Nat. Phys.* **5**, 64–67 (2009).



## Acknowledgements

I am indebted to my colleagues at the University of Wisconsin Milwaukee for many discussions, and to J. Frank, A. Singharoy for valuable comments on the manuscript. The research conducted at the University of Wisconsin Milwaukee was supported by the US Department of Energy, Office of Science, Basic Energy Sciences under award DE-SC0002164 (algorithm design and development), and by the US National Science Foundation under awards STC 1231306 (numerical trial models and data analysis) and 1551489 (underlying analytical models).


## Competing interests

The author declares no competing interests.