Structural biology is solved — now what?

The splendid computational success of AlphaFold and RoseTTAFold in solving the 60-year-old problem of protein folding raises an obvious question: what new avenues should structural biology explore? We propose a strong pivot toward the goal of reading mechanism and function directly from the amino acid sequence. This ambitious goal will require new data analytical tools and an extensive database of the atomic-level structural trajectories traced out on energy landscapes as proteins perform their function.

Abbas Ourmazd, Keith Moffat and Eaton Edward Lattman

We have a confession to make. Not long ago, we were skeptical that the accurate structure of a fully folded protein could be deduced computationally from its amino acid sequence. The spectacular success of AlphaFold and RoseTTAFold algorithms[1,2] in determining the fully folded structure of proteins from their amino acid sequence, often to high accuracy, has eliminated any doubt.

This delightful success is the culmination of four decades-long efforts: (1) deposition of more than 170,000 experimentally determined protein structures in the openly accessible Protein Databank[3]; (2) deposition of a large number of amino acid sequences of entire families of proteins and their evolutionary relationships in public repositories; (3) elucidation of multiple sequence alignments; and (4) the resurgence of neural-inspired machine-learning algorithms[4]. This resurgence constitutes an impressive demonstration of the power of sophisticated deep learning. In brief, AlphaFold 2 consists of a module for extracting information from so-called multisequence alignments to gain insight into segments of the studied protein. This module operates in tandem with a second that is able to build a model of the protein structure, including the side-chains.

Does this success mean that structural biology, as an experimental discipline, is 'solved'? Can we, in good conscience, continue to ask our students and young collaborators to spend months, if not years, determining protein structures? Or is the heyday of protein structure determination finally over? As with any success, it is important to ask what is next.

We venture to believe that the full impact of structural biology is yet to come. For, as impressive as these new algorithms are, they cannot predict both protein function and mechanism directly from the amino acid sequence. Structural biology has rested on the credo that knowledge of structure is key to understanding function and mechanism.

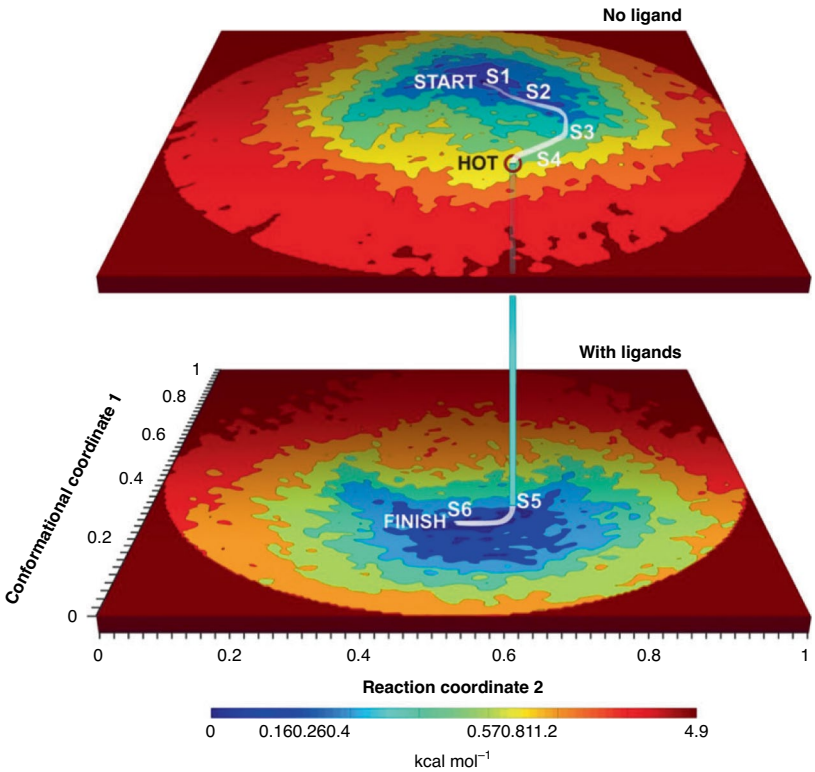

Fig. 1 | Experimentally determined energy landscapes for the protein ryanodine receptor 1 (RYR1) with and without ligands. The upper and lower landscapes represent the energy landscape without ligand (upper surface), and with ligands (Ca^{2+}, ATP and caffeine) (lower surface). The landscapes are described in terms of the most important two mutually orthogonal conformational coordinates. The curved path represents the minimum-energy functional route to the binding of ligands. This path starts at the minimum-energy conformation of RYR1 without ligands (START), follows the conduit of lowest energy to a point with a high probability of transition to the with-ligands energy landscape (HOT) and terminates at the minimum-energy conformation with ligands (FINISH)[7].

In practice, structure provides only hints about these elements.

As an example, take the ribosome — a molecular machine that uses the energy released by guanosine triphosphate (GTP) hydrolysis to synthesize a polypeptide chain though the serial addition of amino acid residues, as encoded by the cognate mRNA. This basic function was uncovered largely by the methods of biochemistry and molecular biology. Structural analysis, however, greatly enriched our understanding of this function and stimulated searches for antibiotics that target the ribosome. Structure provided insights into mechanism, but did not reveal function as biologists understand the term.

Determining what proteins do and how they do it currently requires an extensive library of methods developed to infer function and mechanism. Within this library, structure determination represents but a single book. To make

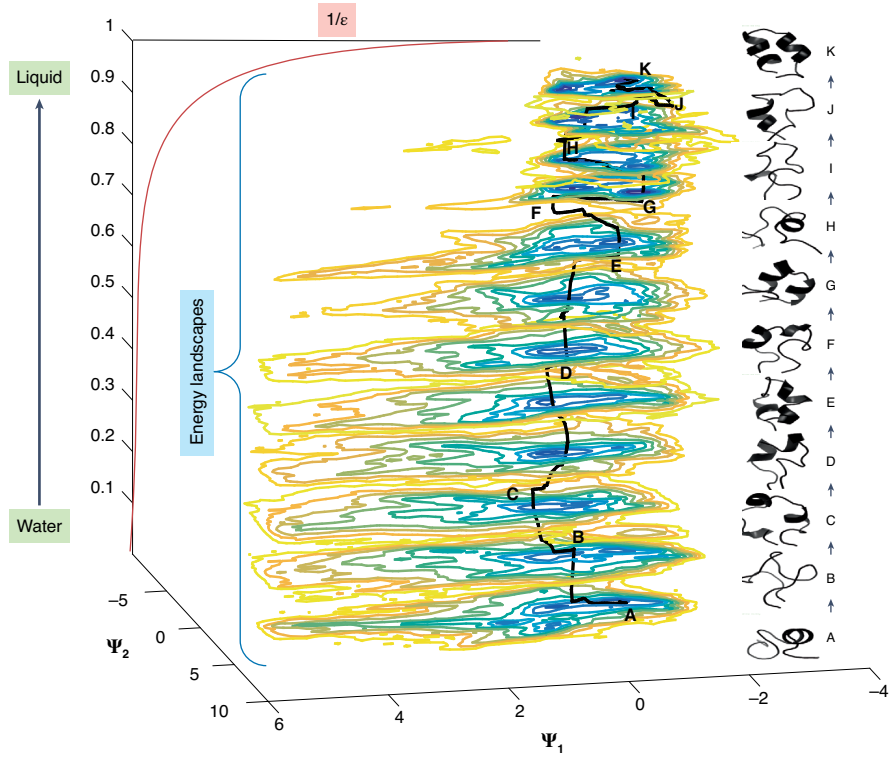

Fig. 2 | Series of energy landscapes for the hemagglutinin fusion peptide membrane insertion simulated by molecular dynamics. Each landscape is in equilibrium with a reservoir of a slightly different pH, with ε representing the permittivity. The two most important conformational coordinates are represented by Ψ_1 and Ψ_2, respectively. The schematic shows how a non-equilibrium process (insertion into a membrane) can be approximated by a trajectory involving a series of energy landscapes. Horizontal trajectory segments represent structural or conformational motions at a constant pH, and vertical transitions the effect of changing pH. The structural evolution is shown in the column on the right (J. Copperman, P. Schwander and A.O., unpublished observations).

matters worse, the mechanism by which function is performed almost invariably involves a complex sequence of concerted changes in a protein's structure and/or conformation. Until the recent advent of new data analytical techniques able to deal with continuous conformational spectra[5–10], structural biology was largely confined to determining one — or at best a few discrete — static structures.

Of course, the grand challenge of determining both function and mechanism directly from the amino acid sequence presents major hurdles. First, we need to abandon the notion of a single structure for each protein; as a protein performs a function, it can adopt a continuum of different structures. Second, we must develop means for mapping the continuous conformational motions of proteins from random snapshots of the protein of interest. Third, we must develop means for identifying the functionally relevant conformational motions. Fourth, we need to develop a conceptual basis for determining and codifying functional trajectories — the pathways populated by the functionally relevant structures. Fifth, we must establish a sufficiently large experimental library of functionally relevant trajectories to exploit the power of machine learning. Finally, we need to develop algorithmic means for linking functional trajectories to amino acid sequences.

Overcoming these hurdles entails significant difficulties. For example, function often involves a protein complex, in which the structures of individual proteins depend on those of their neighbors. Some motions within protein complexes can be as large as tens of angstroms. It is an open question whether we can amass a sufficiently large and diverse library of function to train powerful machine-learning algorithms. We also do not yet know how well AlphaFold and RoseTTAFold can predict the structures of complexes, although early indications are surprisingly positive. Following the success of machine learning in beating humans when playing the boardgame 'Go', perhaps our library could be extended by computer simulation, just as the artificial intelligence algorithm learned from the games it played with itself.

We must learn to walk before we can run. Thirty years ago, Frauenfelder and colleagues[11] pointed out that the concept in enzymology of one intermediate structure (or a few discrete intermediate structures) is tantamount to the conformational energy landscape[12] of a protein consisting of one or few deep energy minima surrounded by high barriers. In fact, the barriers often turn out to be comparable with the thermal energy available under physiological conditions. This means that the notion of one or only a few distinct structures is inadequate. New data analytical algorithms capable of handling continuous conformational motions, first demonstrated a decade ago[5,13,14], are now developing at a healthy clip[6–9,15].

The next challenge involves identifying the conformational motions relevant to function. Here, the concept of an energy landscape[11] is key, where each point corresponds to a structure of a particular energy. More than a century ago, Boltzmann pointed out that at (and near) thermal equilibrium, only the energetically lowest-lying conformational states of a protein are significantly occupied, and that the occupation probability of a conformation decreases exponentially with the energetic cost of assuming that conformation. This means that near equilibrium, function proceeds primarily along heavily occupied minimum-energy conduits, just as water flows along the rivers in a hilly landscape. The functionally relevant conformational motions thus correspond to, and can be deduced from, minimum-energy trajectories on energy landscapes. In this picture, reactions such as ligand binding represent 'vertical' transitions between different energy landscapes (Fig. 1)[7].

Experimental mapping of energy landscapes is best carried out by structural studies of individual particles by, for example, cryo-electron microscopy or single-particle X-ray scattering. This is because in crystallography, the inevitable averaging over the many particles in a crystal reveals the growth and decay of populations, rather than the dynamics of individual particles.

A period of intellectual recognition but little action is being rapidly replaced by the development of tools for mapping minimum-energy (that is, functionally relevant) continuous conformational trajectories[7,16]. As in the early days of the automobile, a plethora of algorithmic tools is being proposed and investigated. Just

as the key point of the automobile was locomotion rather than the placement of the clutch, the key point to recognize is that near equilibrium, function entails (near-)minimum-energy conformational trajectories, because (near-)equilibrium states are the only ones with a significant occupation probability.

Accumulation of a sufficiently large and diverse database of single-particle functional trajectories should enable us to harness the power of modern machine learning, to 'read' function (and mechanism) from the amino acid sequence.

Of course, this is only part of the story. Adenosine triphosphate (ATP) hydrolysis releases about 12 k_BT of energy, and the absorption of a visible-light photon by a signaling photoreceptor deposits even more energy. These energies drive the system far from equilibrium, and are too high to describe all biologically relevant reactions as quasi-equilibrium trajectories on a single energy landscape.

There are at least two ways around this problem. The first approximates a non-equilibrium process as a succession of quasi-equilibrium processes on a series of landscapes (Fig. 2), just as one can think of a big vertical step as a series of small ones. In this picture, a functional trajectory consists of a succession of small vertical steps from one quasi-equilibrium landscape to the next, interspersed with horizontal segments for conformational relaxation on each step. An alternative approach would directly determine the non-equilibrium conformational trajectories using methods developed to study ultrafast processes[17].

As Ludwig Wittgenstein famously noted in the preface to his monumental work *Tractatus Logico-Philosophicus*, "…the problems…have in essentials been finally solved. And if I am not mistaken in this, then the value of this work…consists in the fact that it shows how little has been achieved when these problems have been solved".

It has taken decades of effort by structural and computational biologists and data scientists to read protein structure from the amino acid sequence. This represents a major achievement. It is therefore right that credit should go to the AlphaFold and RoseTTAFold teams, and to the thousands of contributors who, over decades, selflessly made the results of their labors available to the scientific community. It is exciting to think that this achievement is but a prelude to 'solving' protein mechanism and function. ❐

Abbas Ourmazd[1]✉, Keith Moffat[2] and Eaton Edward Lattman[3]

[1]*University of Wisconsin Milwaukee, Milwaukee, WI, USA.* [2]*Department of Biochemistry & Molecular Biology and the Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL, USA.* [3]*Baltimore, MD, USA.*

✉*e-mail: Ourmazd@uwm.edu*

Published online: 11 January 2022
https://doi.org/10.1038/s41592-021-01357-3

References

1. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
2. Baek, M. et al. *Science* **373**, 871–876 (2021).
3. Bernstein, F. C. et al. *J. Mol. Biol.* **112**, 535–542 (1977).
4. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
5. Dashti, A. et al. *Proc. Natl Acad. Sci. USA* **111**, 17492–17497 (2014).
6. Hosseinizadeh, A. et al. *Nat. Methods* **14**, 877–881 (2017).
7. Dashti, A. et al. *Nat. Commun.* **11**, 4734 (2020).
8. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. *Nat. Methods* **18**, 176–185 (2021).
9. Punjani, A. & Fleet, D. J. *J. Struct. Biol.* **213**, 107702 (2021).
10. Giraldo-Barreto, J. et al. *Sci. Rep.* **11**, 13657 (2021).
11. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. *Science* **254**, 1598–1603 (1991).
12. Ourmazd, A. *Nat. Methods* **16**, 941–944 (2019).
13. Schwander, P., Fung, R., Phillips, G. N. & Ourmazd, A. *New J. Phys.* **12**, 035007 (2010).
14. Schwander, P., Fung, R. & Ourmazd, A. *Phil. Trans. R. Soc. Lond. B* **369**, 20130567 (2014).
15. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. *eLife* **7**, e36861 (2018).
16. Giraldo-Barreto, J. *Sci. Rep.* **11**, 13657 (2021).
17. Hosseinizadeh, A. et al. *Nature* **599**, 697–701 (2021).

Acknowledgements

We acknowledge valuable discussions with past and present members of the University of Wisconsin Milwaukee data science group. The development of underlying techniques was supported by the US Department of Energy, Office of Science, Basic Energy Sciences, under award DE-SC0002164 (underlying dynamical techniques) and by the US National Science Foundation under awards STC-1231306 (underlying data analytical techniques) and DBI-2029533 (underlying analytical models).

Competing interests

The authors declare no competing interests.