attribute the change in optical reflectivity to a pressure-induced phase transition in which electrons in the sample become free to move like those in a metal. Hydrogen remains as a molecular solid up to the transition pressure; it possibly stays in this state above 425 GPa, but it is difficult to confirm this by spectroscopy because there is a reduced coupling between light and matter in these extreme conditions.

It can certainly be argued that a definite proof for metallic hydrogen would come only from a measurement of the sample's electrical conductivity at high pressure as a function of temperature. Solid hydrogen should exhibit a high level of electrical conduction that should then decrease as the sample temperature is raised. However, even with experimental techniques developed in the past few decades to study condensed matter in extreme conditions, electrical-transport measurements of hydrogen remain a huge challenge[9,10].

Nevertheless, Loubeyre and co-workers' findings should be considered as a close-to-definite proof of dense hydrogen reaching a metallic state in extreme-pressure conditions. Computational predictions of the pressure at which molecular hydrogen enters a metallic state still lack accuracy, because they require many different quantum-mechanical corrections that are difficult to address. However, the experimental value of 425 GPa agrees with calculations[11] that predict a transition in hydrogen to a different solid phase at a similar pressure.

Loubeyre and colleagues' study has combined innovative techniques for ultra-high-pressure generation with advanced experimental methods using synchrotron radiation. In doing so, it has raised expectations for the discovery of other remarkable properties of solid hydrogen at extreme density. For the time being, many questions remain. For instance, could electrical resistivity be measured across the metallic transition? Could superconductivity at a record-high temperature be achieved in hydrogen? And could the molecular order be disrupted under ultrahigh pressure and lead to an atomic phase in the solid state?

Competition is still strong between different research groups seeking to answer these questions, and to further unveil and understand the characteristics of hydrogen at extreme density. More exciting findings are sure to come at every stage of the race.

Serge Desgreniers is in the Department of Physics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.
e-mail: serge.desgreniers@uottawa.ca

1. Wigner, E. & Huntington, H. B. *J. Chem. Phys.* **3**, 764–770 (1935).
2. Mao, H. K. & Hemley, R. J. *Science* **244**, 1462–1465 (1989).
3. Eremets, M. I., Troyan, I. A. & Drozdov, P. Preprint at https://arxiv.org/abs/1601.04479 (2016).
4. Dias, R. P. & Silvera, I. F. *Science* **355**, 715–718 (2017).
5. Loubeyre, P., Occelli, F. & Dumas, P. *Nature* **577**, 631–635 (2020).
6. Dewaele, A., Loubeyre, P., Occelli, F., Marie, O. & Mezouar, M. *Nature Commun*. **9**, 2913–2922 (2018).
7. Jenei, Zs. *et al. Nature Commun*. **9**, 3563 (2018).
8. Loubeyre, P., Occelli, F. & LeToullec, R. *Nature* **416**, 613–617 (2002).
9. McMinis, J., Clay, R. C. III, Lee, D. & Morales, M. A. *Phys. Rev. Lett.* **114**, 105305 (2015).
10. Azadi, S., Drummond, N. D. & Foulkes, W. M. C. *Phys. Rev. B* **95**, 035142 (2017).
11. Eremets, M. I. & Troyan, I. A. *Nature Mater*. **10**, 927–931 (2011).

Computational biology

Protein-structure prediction gets real

Mohammed AlQuraishi

Two threads of research in the quest for methods that predict the 3D structures of proteins from their amino-acid sequences have become fully intertwined. The result is a leap forward in the accuracy of predictions. **See p.706**

Proteins perform or catalyse nearly all chemical and mechanical processes in cells. Synthesized as linear chains of amino-acid residues, most proteins spontaneously fold into one or a small number of favoured three-dimensional structures. The sequence of amino acids specifies a protein's structure and range of motion, which in turn determine its function. Over decades, structural biologists have experimentally determined thousands of protein structures, but the difficulty of these studies has made the promise of a computational approach for predicting protein structure from sequence alluring. On page 706, Senior *et al.*[1] describe an algorithm, AlphaFold, that takes a leap forward in solving this classic problem by bringing to bear modern machine-learning techniques.

The diversity of protein structures precludes the possibility of obtaining simple folding rules, making structure prediction difficult. Protein folding is ultimately driven by quantum mechanics. Were it possible to compute the exact energy of protein molecules from quantum theory, and to do so for every possible conformation, then predicting a protein's most energetically favoured structure would be easy. Unfortunately, a quantum treatment of proteins is computationally intractable (quantum computers might change this), and the total set of possible conformations that any protein can take is astronomical, prohibiting such a brute-force approach.

This has not stopped scientists from attempting a direct attack on the problem. Physical chemists have devised tractable, but approximate, energy models for proteins[2], and computer scientists have developed ways to explore protein conformations[3]. Much progress has been made on the first problem but the second has proved more recalcitrant.

The set of shapes that a protein might take can be likened to a landscape: different locations in the landscape correspond to different shapes, with nearby locations having similar shapes. The height of a location corresponds to how energetically favourable the associated shape is, with the lowest point being the most favoured. Natural proteins evolved to have funnel-shaped landscapes that enable newly synthesized proteins, jostled by the thermal fluctuations of the cell, to cross the landscape and find their way to a favoured conformation in physiologically relevant timescales (milliseconds to minutes)[4]. Algorithms can search the landscape to find favoured conformations by following the landscape's inclination, but the ruggedness of the terrain causes them to get stuck in troughs and valleys far from the lowest basin.

"The algorithm outperformed all entrants at the most recent blind assessment of methods used to predict protein structures."

The course of the structure-prediction field changed nearly a decade ago with the publication of a series of seminal papers[5–7] exploring the idea that the evolutionary record contains clues about how proteins fold. The idea is predicated on the following premise: if two amino-acid residues in a protein are close together in 3D space, then a mutation that replaces one of them with a different residue (for example, large for small) will probably induce, at a later time, a mutation that alters the other residue in a compensatory direction (in our example, swapping small for large). The set of co-evolving residues therefore encodes valuable spatial information, and can be found by analysing the sequences of evolutionarily related proteins.

By transforming this co-evolutionary information into a matrix known as a binary contact map, which encodes which residues are proximal, the set of conformations that merit consideration by algorithmic searches can be restricted. This in turn makes it possible to accurately predict the most favourable protein conformation, especially for proteins for which many evolutionarily related sequences are known. The idea was not new[8], but the rapid growth in available sequence data in the early 2010s, coupled with crucial algorithmic breakthroughs, meant that its time had finally come.

Co-evolutionary analysis has been responsible for most progress in protein-structure prediction in the past few years, but it has not obviated the need for algorithms to search the energy landscapes of proteins: binary contact maps constrain the search space, but do not pin down a single 3D structure. Furthermore, the mathematics underpinning the conversion of co-evolutionary data into contact maps is restricted by the types of input used and the output generated. The initial injection of deep learning (a type of machine learning) into co-evolutionary analyses improved matters by incorporating richer inputs[9]. AlphaFold takes things a step further by changing the outputs.

In lieu of binary contact data, AlphaFold predicts the probabilities of residues being separated by different distances. Because probabilities and energies are interconvertible, AlphaFold predicts an energy landscape – one that overlaps in its lowest basin with the true landscape, but is much smoother. In fact, AlphaFold's landscape is so smooth that it nearly eliminates the need for searching. This makes it possible to use a simple procedure to find the most favourable conformation, rather than the complex search algorithms employed by other methods.

The idea that a complex search could be unnecessary for structure prediction is, in hindsight, unsurprising. Mathematically, the distances between points determine their relative locations. Predictions of distances can therefore predict structure. Moreover, relatively simple models of protein energy landscapes known as Gō potentials, in which experimentally determined distances between residues are favoured, can lead to protein-folding pathways that resemble ones experienced by real proteins[10]. This suggests that proteins fold more like simple origami than like an intricate knot – all parts can come together at once. My own work has shown that folding can be predicted implicitly using a deep-learning model without searching[11], and minimal search procedures have also been embedded within another deep-learning model to predict protein structures[12].

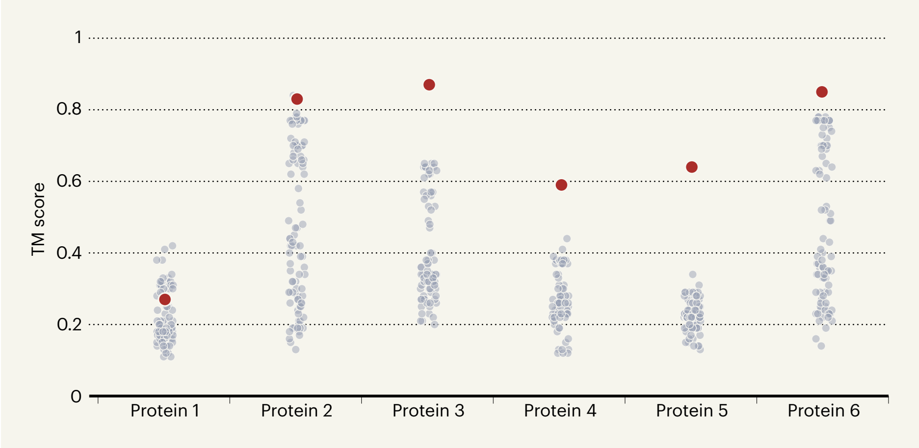

Figure 1 | Predictions of protein structures. Senior *et al.*[1] report a machine-learning system called AlphaFold, which predicts the 3D structures of proteins from their amino-acid sequences. Template modelling (TM) scores measure how well a predicted structure matches the overall shape of the actual structure, on a scale from 0 to 1. TM scores for AlphaFold were better than those of other prediction systems for 25 out of 43 proteins in a blind test. Here, the TM scores for AlphaFold (red) are compared with those of other prediction systems (grey) in the blind test for six proteins whose 3D structures could be modelled only on the basis of their amino-acid sequences – no 3D structures of proteins that have similar amino-acid sequences were available to use as a starting point for modelling. AlphaFold made the most accurate predictions for five of these six proteins. (Adapted from Fig. 1b of ref. 1.)

What is notable about AlphaFold is that it predicts distances with sufficient accuracy to outperform state-of-the-art search methods (Fig. 1). Senior *et al.* used advances in deep learning to extract as much structural information as possible from protein sequences. The resulting algorithm outperformed all entrants at the most recent blind assessment of methods used to predict protein structures (the CASP13 event), generating the best structure for 25 out of 43 proteins, compared with 3 out of 43 for the next-best method. AlphaFold's predictions had a median accuracy of 6.6 ångströms on this set of proteins – that is, for the middle-ranked protein in this set, the atoms in the proposed structures were on average 6.6 Å away from their actual positions.

Challenges remain. AlphaFold is not yet accurate enough for most applications, such as working out the catalytic mechanisms of enzymes or how drugs bind to proteins (which both typically require 2–3 Å resolution). And although AlphaFold's search procedure is much simpler than most modern methods, it can still be slow, taking tens to hundreds of hours to make a single prediction. For applications such as protein design, which require the structures of many different protein sequences to be modelled, the lack of speed is an impediment.

Nevertheless, this is a watershed moment for the field. Given continued growth in the number of available protein sequences, it is possible that the coarse structures (about 4 Å resolution) of most proteins that consist of a single folded domain will become available in the next five years from structure predictions. Such broad availability of structural information might transform the life sciences, just as sequence information did in the preceding decades. This could mean that, combined with the rapid advances in protein–structure determination enabled by cryo-electron microscopy, we are entering a golden age of structural biology – one that makes possible a quantitative and mechanistic basis for the life sciences, broadly grounded in firm structural hypotheses.

Mohammed AlQuraishi is in the Laboratory of Systems Pharmacology, Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mail: alquraishi@hms.harvard.edu

1. Senior, A. W. *et al. Nature* **577**, 706–710 (2020).
2. Guvench, O. & MacKerell, A. D. Jr *Methods Mol. Biol.* **443**, 63–88 (2008).
3. Maximova, T., Moffatt, R., Ma, B., Nussinov, R. & Shehu, A. *PLoS Comput. Biol.* **12**, e1004619 (2016).
4. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. *Proteins* **21**, 167–195 (1995).
5. Marks, D. S. *et al. PLoS ONE* **6**, e28766 (2011).
6. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. *Bioinformatics* **28**, 184–190 (2012).
7. Kamisetty, H., Ovchinnikov, S. & Baker, D. *Proc. Natl Acad. Sci. USA* **110**, 15674–15679 (2013).
8. Lapedes, A. S., Giraud, B. G., Liu, L. & Stormo, G. D. *IMS Lecture Notes Monogr. Ser.* **33**, 236–256 (1999).
9. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. *PLoS Comput. Biol.* **13**, e1005324 (2017).
10. Hills, R. D. & Brooks, C. L. *Int. J. Mol. Sci.* **10**, 889–905 (2009).
11. AlQuraishi, M. *Cell Syst.* **8**, 292–301 (2019).
12. Ingraham, J., Riesselman, A., Sander, C. & Marks, D. in *7th Int. Conf. Learn. Represent.* https://openreview.net/forum?id=Byg3y3C9Km (2019).

This article was published online on 15 January 2020.