Deep learning opens up protein science's next frontiers

Computer models can now provide stunningly accurate predictions of proteins' three-dimensional structures. But what about their biological functions?

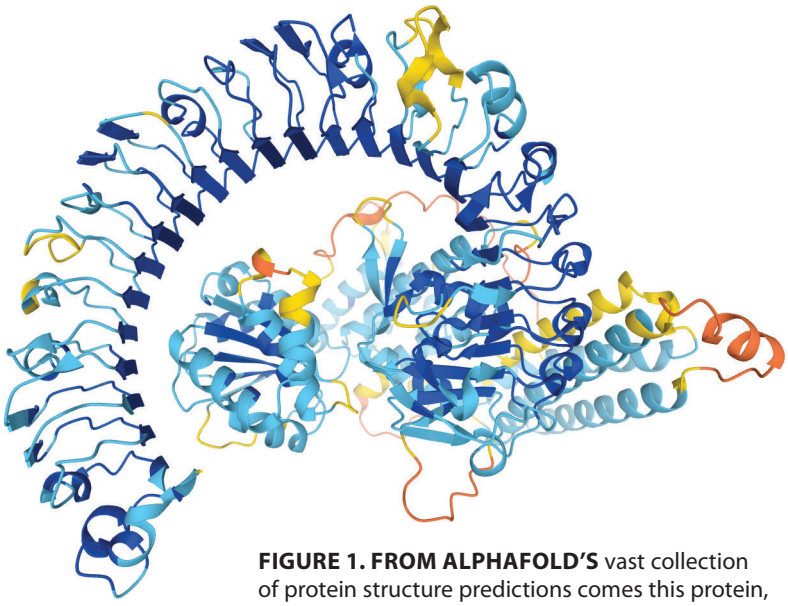

FIGURE 1. FROM ALPHAFOLD'S vast collection of protein structure predictions comes this protein, which may promote disease resistance in the Eurasian wildflower *Arabidopsis thaliana*. The colors represent regions of the protein predicted with high (blue) through low (yellow and orange) confidence. The structure has yet to be observed experimentally. (Courtesy of the AlphaFold Protein Structure Database, CC BY 4.0.)

At their heart, proteins are much like any other polymers: flexible linear chains of amino-acid monomers drawn from a library of just 20 or so building blocks. But unlike synthetic polymers, which tend to flop around stochastically, proteins reliably fold into characteristic three-dimensional shapes. The diversity of those shapes gives rise to the complexity of the biological world.

Uncovering the relationship between amino-acid sequence and folded structure has been a grand challenge of the past half century, with connections to cell biology, chemistry, biophysics, and medicine. To date, more than 180 000 protein structures have been made available to the world in the Protein Data Bank (PDB). But even that enormous resource barely makes a dent in the tens of millions of proteins known to be encoded by genes across all living species.

Last November, as part of the Critical Assessment of Structure Prediction (CASP) project, researchers at DeepMind in London showed that their AlphaFold2 model had made astonishing progress. Given a protein's amino-acid sequence, AlphaFold2 could often predict its structure with most atomic positions correct to within an angstrom—less than the length of a chemical bond.[1] The team has now released its own database of predicted protein structures, including the complete human proteome and many nonhuman proteins whose structures, such as the one in figure 1, experimenters have yet to resolve.

Inspired by AlphaFold2 but with only a rough idea of the model's architecture, Minkyung Baek, in David Baker's group at the University of Washington, and her colleagues developed a similarly capable model, called RoseTTAFold, in time to publish their results concurrently with AlphaFold2's this summer.[2]

Both AlphaFold2 and RoseTTAFold use deep learning—a type of artificial intelligence—which means that their inner workings are largely a black box. But their guiding principles are some of the same ones that have been guiding structural biologists for years. And their success has researchers thinking about how to paint an even more complete picture of proteins and their biological environments.

A hard problem

Compared with the rest of organic chemistry, understanding of proteins came late. The first known protein structures, of myoglobin and hemoglobin, weren't discovered until 1958 and 1959, respectively—half a decade after the structure of DNA.

And unlike DNA's elegant double helix, protein structures were a mess. Linus Pauling, among others, had predicted years earlier that amino-acid chains could organize into orderly alpha helices and beta sheets. Indeed, those motifs do show up in protein structures, but they're interspersed with wild twists and turns that hadn't been anticipated. "There was a sense of, 'Holy cow, there's no symmetries here,'" says Ken Dill of Stony Brook University. But the structure wasn't disordered either: For any given protein, it was always the same.

The hemoglobin and myoglobin structures had been found through x-ray crystallography, the long-time gold standard for probing the atomic structure of any material, not just biomolecules. (See the article by Wayne Hendrickson, PHYSICS TODAY, November 1995, page 42.) But the powerful technique is beleaguered by a pair of challenges. First, it requires a crystalline sample—an unnatural form of matter for most proteins. Second, the x-ray diffraction pattern retains only half the crystal's structural information: The x rays' intensities are easily measured, but their phases are lost. Max Perutz, discoverer of the hemoglobin structure, solved the so-called phase problem by inserting various heavy-metal atoms into the protein to scramble the phases. (See Perutz's obituary in PHYSICS TODAY, August 2002, page 62.) But that trick doesn't always work.

In recent years, cryoelectron microscopy has started to rival x-ray crystallography in its ability to image proteins with atomic resolution. (See PHYSICS TODAY, December 2017, page 22.) It has the benefit of not needing a crystal—instead, the molecules are embedded in a thin sheet of vitreous ice—but it's still challenging. The folded proteins might unravel under the effects of surface tension, and one needs to computationally align many 2D images at random angles to convert them into a composite 3D structure. Finding protein structures experimentally by any method remains difficult and laborious.

What about theoretically? Although some proteins require chaperone molecules to fold correctly, most can solve their own folding problem using nothing but the laws of physics. It's tempting to try to reproduce their solutions in a computer simulation (see PHYSICS TODAY, December 2013, page 13), but the complexity of the system quickly runs up against the limits of computer power for all but the smallest proteins.

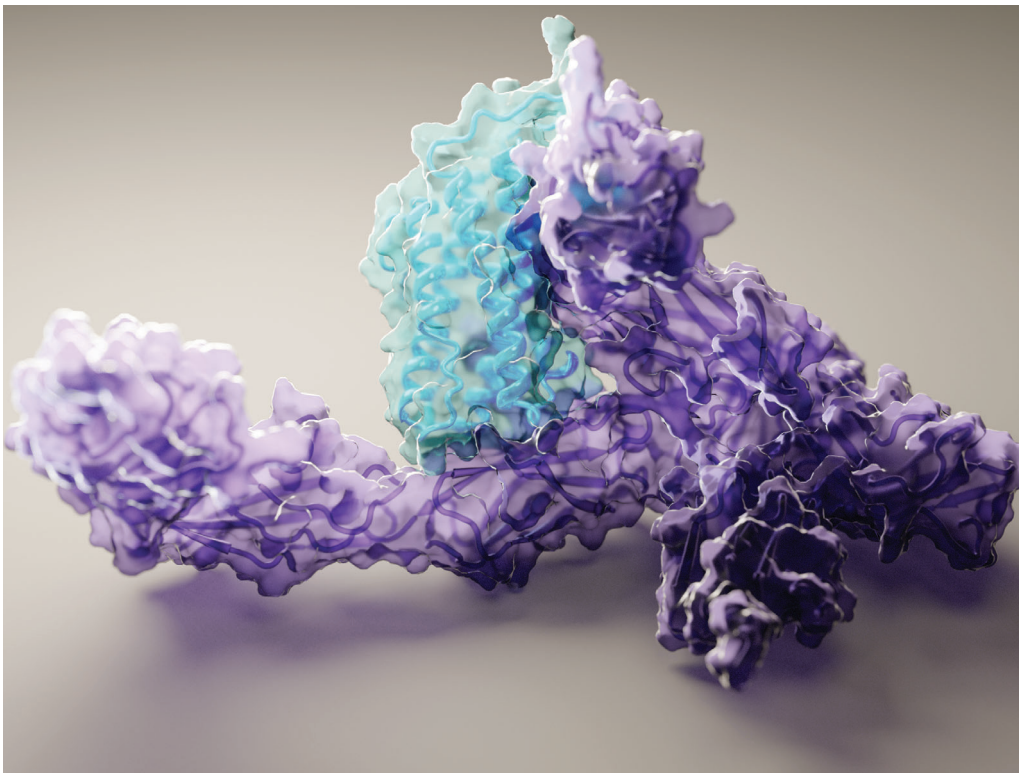

FIGURE 2. ROSETTAFOLD generated this structure of the human signaling protein interleukin-12 (purple) bound to its receptor (blue). Although the structures of molecular complexes are tougher to predict than those of single proteins, the structure here agrees well with one found experimentally through cryoelectron microscopy. (Courtesy of Ian Haydon, Institute for Protein Design, University of Washington.)

And the physics of folding is subtle. The folded structure is generally the thermodynamically favored one, but not by much, and the details of interatomic forces are extremely important. "If you don't know the forces, then a bigger computer doesn't help you get the right answers," says Dill. "It just gives the wrong answers a lot faster."

Biological shortcuts

Any successful approach to protein-structure prediction, including the new deep-learning models, needs to draw on insights from biology, not just chemistry and physics. "Protein sequences aren't random. They've been distilled by natural selection," says Temple University's Vincenzo Carnevale. Any given protein has thousands of evolutionary cousins from across the tree of life; that evolutionary context can provide hints about structure.

Proteins with similar sequences probably have similar structures. If the structure of a related, homologous protein is already known, it can be used as a template: The new sequence is fitted into the old structure, then adjusted accordingly. "That works in a stupendous way," says Carnevale, "but only because the community realized that to succeed with this approach, we needed not just to increase the size of the PDB but to explore the right regions of protein-sequence space."

From 2000 until 2015, the Protein Structure Initiative guided the discovery of thousands of new protein structures, chosen not haphazardly but with the goal of systematically exploring the possible structures proteins can adopt. By the end of the project, says Carnevale, "there was no longer anything novel being discovered in protein structural space. It had all been exhaustively mapped out." Although not every protein sequence has a sufficiently similar known structure to use as a template, the days of wholly unanticipated new protein structures were past.

Meanwhile, theorists were working on ways to glean information about a protein structure from its evolutionary context even when none of its relatives' structures are known. That surprising feat is possible because each protein in a family is the product of its own evolutionary optimization. When one amino acid in a protein randomly mutates, the mutation usually isn't enough to ruin the entire structure, but it does destabilize it. Evolutionary pressure therefore builds on the amino acid's neighbors in 3D space to mutate also and thus restore the structure's stability. If, in a list of many related protein sequences, two amino-acid positions show a tendency to mutate in tandem, they likely sit next to each other in the folded protein.

Structural revolution

Those methods and others were already being tried and tested at CASP before DeepMind entered the fray. Part computational experiment, part competition, CASP challenges hundreds of research groups every two years to reproduce protein structures that have recently been found experimentally but not yet published. Target structures are classified by difficulty, based in part on whether a homologous structure exists to use as a template. Structure predictions are graded on a scale from 0 to 100: A random guess might score below 20;

an atomically precise structure, above 90.

From the early days of CASP, models have been scoring above 80 for the easiest template-based predictions, while scores for the most difficult targets have been stuck around 40. So DeepMind's first CASP entry, the original AlphaFold model in 2018, shook things up by scoring above 70 for more than half of the most difficult targets.[3]

The power of deep learning is that it can recognize patterns, interpolate between known structures, and identify mutation correlations more keenly than human observers or more straightforward algorithms can. AlphaFold wasn't the first machine-learning model to be entered into CASP; its superior performance came, in part, from using the structure and correlation data to predict not just which pairs of amino acids are in contact but the full matrix of all their pairwise distances.

For the 2020 CASP assessment, the DeepMind team had revamped its model into AlphaFold2, whose predictions scored near 90 even for the most difficult targets—scores so high that they were probably limited by the imprecision of the experimental structures the predictions were graded against.

The AlphaFold2 code and method weren't made public at first; all that was released to the world was a 30-minute presentation that described a model that processed data on two parallel tracks. One carried the list of protein sequences thought to be related to the target protein; the other, a pairwise amino-acid distance matrix. By exchanging information between the two tracks, the model repeatedly updated both sets of data until it converged on a final prediction, from which a 3D structure was extracted.

Building on the ideas in the presentation, Baek and colleagues developed RoseTTAFold: a three-track model that iteratively updates the sequence data, distance matrix, and 3D structure itself. If it had been entered into CASP in 2020, its scores for the hardest targets would have averaged about 80.

Now that the AlphaFold2 details are published, Baek concedes that it's a better engineered method. "Almost every component is based on some physical insight," she says. For example, AlphaFold2 requires its amino-acid distances to satisfy the triangle inequality—two points can't be farther from each other than the sum of their distances to a third point—so it saves time by maintaining a degree of physicality even at intermediate steps.

Furthermore, while RoseTTAFold was trained on all the PDB structures, AlphaFold2's training data included additional structures predicted by the model itself. "Training data is very critical," says Baek, "so I think that more complete coverage of protein space helped them a lot."

Molecular interactions

Has deep learning solved the notorious protein-folding problem? That depends on how the problem is defined. Dill draws a distinction between predicting protein structures—what AlphaFold2 and RoseTTAFold do—and understanding protein folding. He considers the latter, which involves mapping the funnel-shaped energy landscapes that guide amino-acid chains into their folded structures, to be largely solved already by statistical physics.[4]

As far as structure prediction is concerned, the deep-learning models have reached a milestone, but they're far from the finish line. Proteins in nature aren't isolated structures. They interact with surrounding molecules, including water, and they combine with other proteins to build large molecular machines—and, ultimately, multicellular living organisms.

Deep-learning methods have made some headway toward solving the structures of multimolecular complexes: The structure in figure 2, found by RoseTTAFold, shows the signaling protein interleukin-12 (purple) bound to its receptor (blue). Multiprotein structures are much more challenging to predict than single-protein ones. The models rely heavily on structural clues from evolutionary context and mutation correlations. But amino acids don't always mutate in tandem if they're in different molecules—especially if those molecules come from different species, such as a pathogen and its host.

"Experimental methods are by no means obsolete," says DeepMind scientist Kathryn Tunyasuvunakool. "They can provide information that AlphaFold currently can't." The model's big advantage, she says, is that it produces structural starting points quickly—in minutes, rather than months or years—and in large numbers. "That's useful, for example, for generating hypotheses and planning experiments." The deep-learning

models are already helping experimenters to fill in the missing structural data from their x-ray crystallography and cryoelectron microscopy experiments and to tackle ever more challenging structure problems.

New drugs

One of the most important uses of protein structures is in drug development. To stop a protein from performing some harmful action in the body, pharmaceutical scientists study the protein's structure, identify a nook or cranny that may correspond to the protein's active site, and design a molecule to plug it up like a cork in a wine bottle.

"But there's really no such thing as 'the' structure," says Carnevale, because proteins flex and contort. Focusing on fitting a molecule to one static structure ignores all the other conformations a protein can adopt or the transitions between them, any one of which might offer a more effective way to disrupt the protein's function.

In some cases, the dynamic approach to drug development might be the only viable one. In neurodegenerative conditions like Alzheimer's and Parkinson's diseases, amino-acid chains get tangled up and fold into the wrong structure, called an amyloid fibril. The fibril structure is known (see PHYSICS TODAY, June 2013, page 16), but the structure alone doesn't say much about how the fibril forms—or how to stop it from forming.

It would take a far more sophisticated model than the ones available today to predict a protein's entire conformational ensemble and range of motion. But as Carnevale points out, "Surely the sequence must encode that information, because nature knows what it is."

Another ambitious goal that's on Baek and colleagues' minds is to free their model from the need to consider evolutionary relationships at all and predict the folded structure based only on the amino-acid sequence. Evolution has produced a wondrous array of proteins and functions, but it hasn't come close to exploring every possible protein. The Baker lab's specialty is in designing proteins from scratch to do things that natural ones can't. (See PHYSICS TODAY, June 2020, page 17.) But those bespoke proteins don't come with millions of years of evolutionary relatives to analyze.

Says Dill, "The whole field is headed toward bigger, better, faster": bigger proteins, more complex actions, and more detailed information than has ever been possible before. Lately, modelers and experimenters alike have been working on understanding the spike protein of SARS-CoV-2—the virus that causes COVID-19—whose binding to a host cell involves a cascade of large conformational changes.[5] As Dill explains, "It's a huge protein that's part of an even huger complex, the virus, with all kinds of moving parts like a big Rube Goldberg machine."

Johanna Miller

References

1. J. Jumper et al., *Nature* **596**, 583 (2021); K. Tunyasuvunakool et al., *Nature* **596**, 590 (2021).
2. M. Baek et al., *Science* **373**, 871 (2021).
3. A. W. Senior et al., *Nature* **577**, 706 (2020).
4. R. Nassar et al., *J. Mol. Biol.* (2021), doi: 10.1016/j.jmb.2021.167126.
5. See, for example, E. Brini, C. Simmerling, K. Dill, *Science* **370**, 1056 (2020); T. Sztain et al., *Nat. Chem.* (2021), doi:10.1038/s41557-021-00758-3.

A seismometer maps Mars's anatomy

NASA's *InSight* is the first mission to explore seismic waves in a planetary body since *Apollo 17* in 1972.

On 26 November 2018, the *InSight* lander—whose acronym stands for Interior Exploration Using Seismic Investigations, Geodesy and Heat Transport—touched down on Mars's Elysium Planitia. Within two months on that flat, volcanic plain, the lander's robotic arm removed a seismometer from the lander deck and placed it on the ground (figure 1), where it started listening for vibrational signals. Eight orbiters currently survey the gravitational fields, magnetism, and atmosphere of Mars, and six rovers have explored its surface chemistry and geology. *InSight*'s seismometer is the only current direct probe of the planet's interior.

To date, the instrument has picked up more than 1000 distinct seismic events. Of the several hundred marsquakes it's recorded, the vast majority were small—none exceeded a moment magnitude of 4. A low level of seismic activity was not unexpected. Unlike Earth, whose sharply defined tectonic plates intersect at boundaries that wind around the planet like the seam of a baseball, Mars has a single, thick plate.

The Martian activity, however, is even lower than what some planetologists expected for the thousands of faults that populate the surface. Most may have formed from stresses on the planet as it shrinks while slowly cooling. Some could have arisen from internal dynamics—mantle convection and volcanism.

The outer part of Mars solidified from a magma ocean produced by accretion early in solar-system history. An iron-rich core formed as heavy, molten metal sank into the planet's center and lighter, silicate-rich material rose; part of that lighter material melted and refroze into a brittle crust. Orbital measurements of the planet's gravity, tidal response, and moment of inertia provided early hints of that differentiation.

An international collaboration of 65 seismologists and planetary scientists from 12 countries has now published three papers that describe the first direct observations of those distinct layers.[1–3] The teams' quantitative measurements of the structure set the stage for understanding how the planet evolved into its current thermochemical state.

Single seismometer

InSight isn't the first spacecraft to bring a seismometer to Mars. The two *Viking* landers each carried one when they landed on Mars in 1976. But uncaging mishaps and the seismometers' onboard installation prevented either from definitively detecting anything but the wind.

Working out planetary structure is largely a matter of interpreting shear (*S*) and compressional (*P*) seismic waves, which travel through the planet at different speeds and refract and reflect from the boundaries of the planet's layers. Those speeds vary with stiffness (or shear and bulk moduli, in geological parlance), density, and temperature. The difference in the waves' arrival times at the seismometer provides the distance to