KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

9 OCTOBER 2024

Scientific Background to the Nobel Prize in Chemistry 2024

COMPUTATIONAL PROTEIN DESIGN AND PROTEIN STRUCTURE PREDICTION

The Nobel Committee for Chemistry

BOX 50005 (LILLA FRESCATIVÄGEN 4 A), SE-104 05 STOCKHOLM, SWEDEN
TEL +46 8 673 95 00 ⋆ WWW.KVA.SE

Computational Protein Design and Protein Structure Prediction

The Royal Swedish Academy of Sciences has decided to award ***David Baker****,* ***Demis Hassabis*** *and* ***John Jumper*** *the Nobel Prize in Chemistry 2024, for computational protein design and protein structure prediction.*

Introduction

The first three-dimensional (3D) structures of proteins were determined by X-ray crystallography about 65 years ago.[1,2] Ever since, scientists have been fascinated by how the polypeptide chains fold themselves up into well-defined and complex 3D patterns. It is also precisely these specific structures that confer proteins their function. It thus became clear that the ability to predict the 3D structure of a protein would enable prediction also of its function and biochemical properties.

In 1972, Christian Anfinsen was awarded the Nobel Prize in Chemistry for the remarkable finding that protein 3D structures were basically encoded by the sequence of amino acids in the polypeptide chain. That is, he found that if a protein is reversibly denatured, it would always refold into the same 3D conformation.[3] This finding led to the long scientific quest of predicting 3D structures directly from the primary amino acid sequence. The problem is considered so important that Nobel Prize Laureate Venki Ramakrishnan has described it as a "50-year-old grand challenge in biology".[4]

Determining protein structures by experimental means is labour intensive. It is noteworthy that while the number of DNA sequences in public databases is now close to 3 billion, and the number of protein sequences that have so far been identified in organisms is over 200 million, the Protein Data Bank[5] still contains only a small fraction of the corresponding 3D protein structures (~200 000). To be able to predict protein structures directly from the amino acid sequence would thus be a major achievement.

The problem here is that the number of theoretically possible conformations of a protein is truly astronomical. Cyrus Levinthal estimated this number and gave name to what is called "Levinthal's paradox".[6] It is often stated in terms of the number of possible conformations for a 100-amino-acid residue protein, which would be on the order of 10^{47}. Hence, the inevitable conclusion is that proteins do not fold by means of a random search of all these conformations, but by biased folding pathways.[6,7]

Study of the actual protein folding process is a very large scientific subject area in itself and has made considerable progress over the years, both by experimental work and theoretical calculations. However, predicting protein structures from sequence is a different problem where the final stable structure of the folding process is the ultimate goal.

The structure prediction problem can also be formulated in another way, where one instead asks what amino acid sequences would yield a certain folding pattern. This question is at the heart of protein design, a field where a target structure is envisaged and then sequences that would yield this structure are identified by computational means.

This year's Nobel Prize in Chemistry recognizes decisive breakthroughs in solving both of these problems – structure prediction from sequence and sequence prediction from structure – and the implications are truly profound. Most monomeric protein structures can now be predicted with high fidelity, and large databases of hundreds of millions of structures have thus been created, with huge impact on biochemical and biological research. Likewise, completely new protein structures, not found in nature, can now be created by computational design and used in various biotechnological and biomedical applications.

Background

After determination of the first protein structures, it was immediately recognized that protein tertiary (3D) structures contain recurring regular so-called secondary structure elements, such as α-helices and β-sheets, the orientation and packing of which, together with connecting loop regions, define the actual tertiary topology (Figure 1). In fact, the α-helical polypeptide pattern was predicted by Linus Pauling as early as 1951.[8] The earliest attempts to predict protein structure from amino acid sequence therefore focused on secondary structure prediction, rather than tertiary.

Hence, in 1974, Chou and Fasman used a dataset of 15 proteins with known conformation to determine propensities for all 20 natural amino acids to be found in either α-helical or β-sheet regions of proteins. By calculating the average α- and β-probabilities of different chain fragments, they could thus predict the secondary structure of a polypeptide chain.[9]

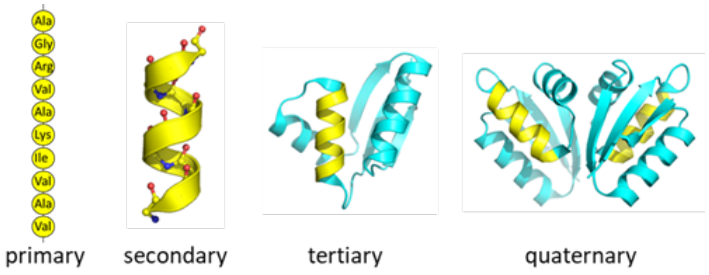

Figure 1. Hierarchy of protein structure. Primary: the amino acid sequence that is determined by the corresponding sequence of DNA base triplets. Secondary: formation of regular geometric patterns of α-helices and β-sheets. Tertiary: the detailed 3D shape of the polypeptide chain. Quaternary: the association of several polypeptide chains or subunits.

The predictions based on this approach, however, turned out not to be very accurate. This is mainly due to the fact that 3D tertiary interactions also are important for establishing the secondary structure, and not just the one-dimensional (1D) sequence of amino acids (primary structure). At that time, the total number of experimentally determined structures in the Protein Data Bank was also very modest (a few hundred), and it remained at this level until the 1990s, when the database really started to grow due to several advances in protein crystallography (Figure 2).[10]

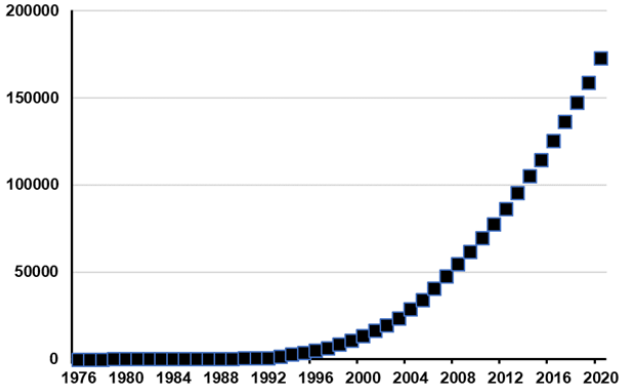

Figure 2. Time evolution of the number of experimentally determined protein structures deposited in the Protein Data Bank.

Despite the limited available protein structures during these earlier years, some profound physical-chemical rules could be gleaned. The hydrophobic effect was clearly manifested in these structures, where nonpolar amino acid sidechains were predominantly found to be packed in the protein interior, well shielded from the surrounding water solvent. Conversely, polar and, particularly, charged sidechains were instead found on the protein surface, where they could interact with the polar solvent. Polar sidechains were also seen in the interior, but then they tended to form networks of hydrogen bonds to compensate for the loss of solvation energy upon folding of the protein. Such seemingly simple rules gave rise to the first attempts of protein design, where researchers tried to design polypeptide sequences that essentially would obey the simple rule – nonpolar inwards, polar outwards.[11]

Based on this information, a perfect case for protein design would be amphiphilic helical structures, where one face of each α-helix is hydrophilic, exposed to the solvent, and the other faces hydrophobic, which would allow packing with similarly hydrophobic surfaces in the interior of the designed structure (Figure 3). This was realized by Regan and DeGrado, who, in 1988, constructed a four-helix bundle protein obeying these principles, where three loops were needed to connect the four α-helices. The resulting structure was characterized by biochemical means and found to be a highly helical and stable monomeric protein.[12]

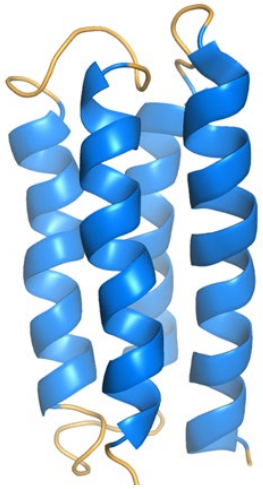

Figure 3. Example of a four-helix bundle structure with a hydrophobic interior and a hydrophilic exterior surface. Only the protein backbone is shown.

Four-helix bundles thus became common targets for protein design in the early years of this field, and the concept of a “binary code” with hydrophobic and hydrophilic amino acid residues was further

elaborated by Hecht and coworkers. These researchers constructed a large library of synthetic genes coding for the same pattern of polar and nonpolar residues and showed that most of the designed protein sequences folded into compact α-helical structures.[13]

The first steps towards protein design were thus taken in the late 1980s, but simple biophysical principles did not suffice for construction of more complex structures, such as mixed α/β-topologies, and with atomic detail. The breadth of the problem clearly called for an automated computational approach.

Computational protein design

The first successful design of a small protein via computation was published by Dahiyat and Mayo in 1997.[14] As a target, they chose the so-called zinc-finger motif that coordinates one or two Zn^{2+} ions. This structure is largely stabilized by interactions with the Zn^{2+} ions, and the goal of the design was to find a new sequence that would adopt the same structure but without any metal ions. The generation of new proteins with sequences unrelated to those in nature is usually termed *de novo* protein design.

The Zn-finger structure is relatively small, made up of ~30 amino acids, and contains one α-helical segment and two β-strands, plus connecting short loops (ββα motif). The same protein had been chosen as a target the year before by Imperiali and coworkers, who iteratively designed a 23-residue variant of the Zn finger, with about one-third of the residues in common with the original sequence; they showed by NMR spectroscopy that it acquired the desired structure.[15]

Dahiyat and Mayo, however, took an entirely computational approach to the problem.[14] They kept the polypeptide backbone fixed, removed the Zn^{2+} ions and computationally searched for amino acid sequences that would yield the desired 3D structure. This involved not only searching among a huge number of possible sequences (~10^{27}), but also optimizing the orientation of amino acid sidechains in terms of their rotamers (torsional angles). For this combinatorial search, they used a dead-end-elimination (DEE) algorithm together with Monte Carlo simulations, with an empirical scoring function to evaluate the conformational energies. The resulting optimal design had 6 out of 28 residues in common with the original Zn-finger sequence (21% identity) and its 3D structure was determined by NMR spectroscopy, demonstrating a close resemblance to the computationally predicted structure (Figure 4).

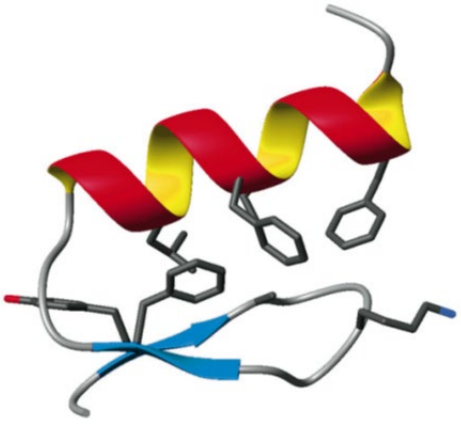

Figure 4. Schematic structure of the designed zinc-finger protein by Dahiyat and Mayo.[14]

This study represents an important milestone. However, the algorithm was still limited to relatively short sequences, and it therefore did not yet provide a general solution to the design problem. Moreover, a general computational design approach would also need to optimize the protein backbone conformation.[16]

The breakthrough in computational *de novo* protein design came in 2003, when **David Baker** and coworkers published the design and crystallographic validation of a 93-residue α/β-protein

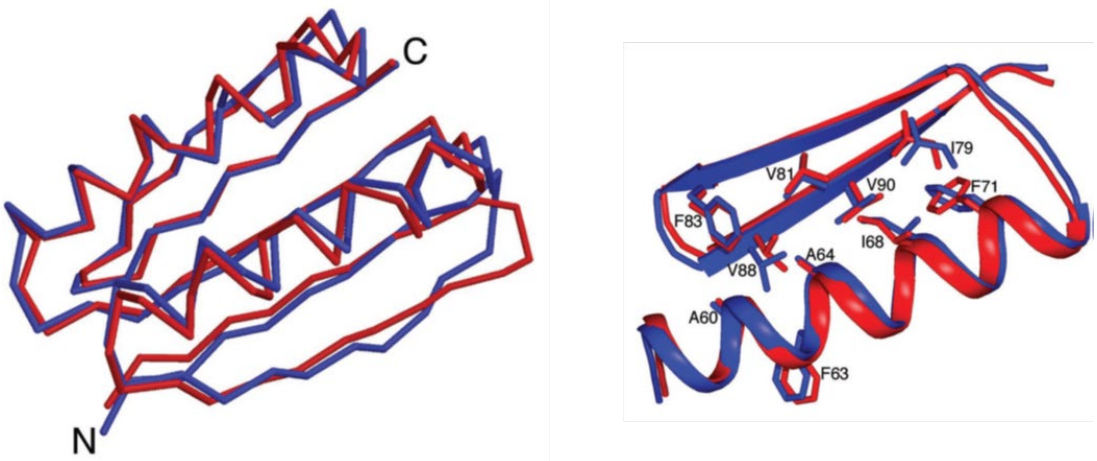

Figure 5. Left: comparison of the predicted backbone structure of Top7 (blue) with the determined X-ray structure (red). Right: view of superimposed sidechains in the cores of the designed model and the solved structure.[17]

named Top7.[17] This was a truly remarkable achievement for several reasons. First, it was a relatively large protein with two α-helices and a β-sheet made up from five β-strands, and its predicted structure was in agreement with the experimental one, including the detailed sidechain positions (Figure 5). Second, the authors sought to design a folding pattern not found for any globular protein in the Protein Data Bank, and third, the final Top7 design had no significant sequence similarity to any naturally occurring protein in the sequence databases. Hence, Top7 was an entirely new protein, both structurally and sequence-wise, designed by automated computation with full optimization of both backbone and sidechains.

How did **Baker** and his coworkers solve the problem? The key to success here was their initial development a few years earlier (1999) of the Rosetta computer program.[18] It assembles short structural fragments from unrelated protein structures with similar local sequences in the Protein Data Bank and simultaneously optimizes sequence and structure with respect to the target backbone conformation. Monte Carlo optimization was used in the calculations with an energy function that treated van der Waals interactions (6-12 Lennard-Jones potential), hydrogen bonding and solvation effects; sidechain orientations were sampled from a large library of rotamers. The program generates many putative solutions and ranks them in terms of energies.

Most important, Rosetta was designed to be a general program both for protein structure prediction and design, and it has continuously been developed since its inception, with a large cadre of users and co-developers. The key idea to build proteins from short fragments can be traced back to the work of Jones and Thirup, who showed that in the context of automated protein model building into crystallographic electron density maps, assembling proteins from known substructures is an effective strategy.[19]

Baker and colleagues went on to show that a wide range of protein structures could be designed using the Rosetta software.[20] While protein design was initially just focused on designing structures, more recent work has aimed to also design advanced protein functions. This still poses major challenges in terms of understanding protein dynamics, structural transitions, allostery, catalytic effects, and so forth, and is thus an area of active research.

In 2008, **Baker** and coworkers reported the first attempts at *de novo* enzyme design, or the design of novel enzymes that can catalyse reactions for which no naturally occurring enzymes exist.[21,22] Although the *de novo* designed enzymes showed increased catalytic rates compared to the non-catalytic background reactions, their overall rates were relatively low compared to those of natural enzymes. However, the designer enzymes could be markedly improved by rounds of experimental directed evolution (for which Frances Arnold was awarded the Nobel Prize in Chemistry 2018).

An area where computational design of function immediately delivered impressive results was for ligand-binding proteins.[23] Here, **Baker** and coworkers could design protein structures that bind steroids with high affinity and selectivity. Already the initial designs showed binding affinities in the micromolar range, and they could be improved by laboratory evolution to reach the nano- to picomolar range.[23] They further demonstrated how new protein nanomaterials could be designed and could create self-assembling icosahedral virus-like particles on the megadalton scale.[24] Another area of great interest is the design of protein switches and sensors for different analytes, with promising applications.[25,26]

Protein structure prediction from sequence – slow progress for many years

A very important initiative within the field of protein 3D structure prediction was the so-called CASP experiments (Critical Assessment of protein Structure Prediction) founded by John Moult and Krzysztof Fidelis in 1994.[27] These biannual challenges allowed truly blind predictions to be made and evaluated by comparison to new experimental structures determined by X-ray crystallographers and NMR spectroscopists, who would withhold their data until after the submission deadline of the CASP entries. This made it possible to assess the progress in structure prediction in an unbiased way, and the most difficult category was termed *ab initio* to indicate that there was no relationship to already known complete structures in such cases.[28]

In the early days of CASP, progress was definitely slow, but several new and important ideas of how to go about the problem of structure prediction saw the light of day. The strategies used by the participants varied substantially, with different knowledge-based approaches and search techniques such as genetic algorithms, Monte Carlo methods, and so on. Here, it may also be noted that artificial neural networks had entered into secondary structure prediction as early as 1988.[29,30]

Another approach that gained some momentum over the years has been "brute-force" molecular dynamics (MD) simulations of protein folding in solution, starting from an arbitrary unfolded state. In 1998, Duan and Kollman reported a microseconds-long simulation of a 36-residue miniprotein and observed folding to a native-like state,[31] and in 2010, Shaw and coworkers showed that several small proteins could be folded to the correct structure during milliseconds-long MD simulations.[32] The computational effort involved in these all-atom simulations was, however, prohibitive and made it clear that plain MD simulation would not be scalable to proteins of larger size in the foreseeable future. Nevertheless, these studies demonstrated that the empirical force fields used were good enough to yield folding to the experimentally observed structures.

With the advances in DNA sequencing in the early 1990s, the number of available protein sequences also grew rapidly. This meant that larger numbers of sequences for a given protein family could be aligned and compared (multiple sequence alignment, or MSA), and researchers realized that correlated mutations in such alignments would contain information about pairs of amino acids in contact with each other in 3D.[33] This was considered such an important concept that a contact prediction category was introduced in CASP2 as early as 1996. However, the reliability of contact prediction remained low for many years, until CASP12 in 2016, when the accuracy suddenly increased dramatically. It turned out that the methods used to analyse correlated mutations had been oversimplified and could not distinguish between directly and indirectly correlated residues. The latter are non-causally correlated by intervening directly correlated residues[34,35], and significant improvements could be achieved by disentangling direct from indirect statistical dependencies. In CASP12, methods involving machine learning and neural networks were also common, and together with the methods improvements for correlated mutations, a contact prediction precision of over 45% was reached.

The accuracy of 3D structure predictions in CASP is measured in terms of a global distance test (GDT) score that reports the largest percentage of α-carbons falling within a certain distance cutoff from the experimental structure, after iteratively superimposing the two structures.[28] An average of several such cutoffs (typically between 1 and 8 Å) is taken as the final GDT score. This is a more robust measure than the common root-mean-square coordinate deviation, which is more sensitive to outliers. Despite the progress outlined above, the average GDT score for the best *ab initio* predictions was stuck below 40% up until and including CASP12 in 2016.

Protein structure prediction and AlphaFold

Following the leap in contact prediction accuracy seen in 2016, the next round of CASP13 in 2018 witnessed another major improvement in contact accuracy which had now reached 70%.[36] Now the accuracy had become high enough to allow translation of the contact or distance maps into 3D structures. The main reason for this improvement was the impact of deep learning methods using convolutional neural networks.

It had become clear that the problem of training a network to predict 2D distance maps had distinct similarities with image recognition tasks, where convolutional networks are widely used. Such networks apply filter optimization to extract features in a hierarchical manner, thereby simplifying the number of neural network connections needed. Hence, for the CASP13 experiment in 2018, the company DeepMind, founded and led by **Demis Hassabis**, constructed a computer program based on a convolutional neural network, which they called AlphaFold (now known as AlphaFold1 or AF1).[37] The program was trained on Protein Data Bank structures to produce a distance map between residues, or rather a map of probability distributions for the distances, based on multiple sequence alignments. From this map, a potential of mean force could be constructed and optimized by a gradient descent algorithm to generate structures.[37]

With deep learning entering the structure prediction field, the performance had now risen to a GDT score of about 60% (Figure 6), and the AlphaFold team was clearly ahead of other participants. **Hassabis** and his team had already taken the community of Go and Chess players by storm in 2018, when they published the AlphaZero program, also based on deep learning, that

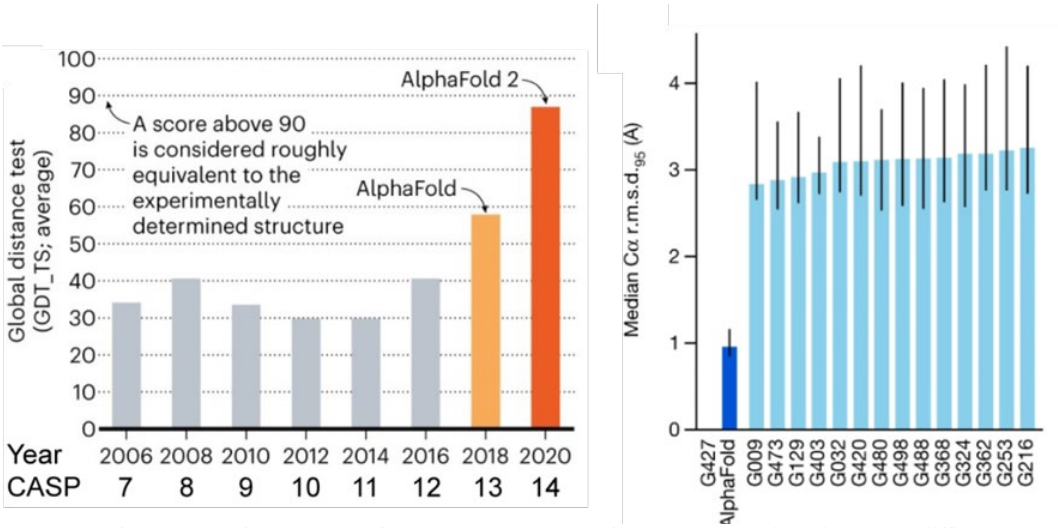

Figure 6. Left: progress of the CASP performance over the years for the best models and the most difficult targets.[38] Right: performance of AlphaFold2 relative to the top 15 entries by other groups in CASP14. Data are the median coordinate error and the 95% confidence interval of the median, estimated from 10 000 bootstrap samples.[41]

showed unsurpassed performance in these and other two-party games.[39] However, there was still some way to go for protein structure prediction to reach near-experimental accuracy.

AlphaFold2 – the real breakthrough

In the next round of CASP in 2020 (CASP14), the group from DeepMind had again not only achieved another leap in accuracy but could now actually present an accuracy competitive with experimental structures for a majority of targets. Hence, while the contact and distance prediction performance remained at around 70% in CASP14, the GDT score for the best predictions on difficult targets[40] had now reached about 90%, and this was due to the new AlphaFold2 (AF2) program[41] (Figure 6). A GDT score of about 90% is generally considered on par with experimental accuracy, since experimental structure determination is, of course, also associated with some errors. The AF2 team led by **John Jumper** and **Hassabis** had thus finally succeeded in solving the protein structure prediction problem for monomeric proteins to within a backbone accuracy of about 1 Å.

Importantly, the neural network model of AF2 had been entirely redesigned compared to that of AF1.[41] The convolution approach was abandoned, and instead a transformer architecture was used, with the essential attention mechanism for learning which parts of the input are more important for the objective of network.[42] The network is also of the end-to-end type, where atomic coordinates are directly produced as output rather than contact information, which had to be post-processed separately in AF1.

The AF2 network has two main blocks called the Evoformer and the Structure module (Figure 7). The Evoformer works simultaneously with a multiple sequence alignment representation (a 2D matrix of aligned sequences from different species) and a pair representation (a 2D matrix of pair distances). The multiple sequence alignment and pair representations exchange information during the learning process and update each other, thus allowing both to evolve. The Structure module then directly operates on a 3D backbone structure using the pair representation and the target sequence, where the backbone geometry is defined in terms of triangles formed by the N-Cα-C atoms of each residue (Figure 7). These triangles float around freely as rigid bodies and are moved by the network to form the structure. This “residue gas” representation is updated iteratively using affine matrices that rotate and translate the residues in space, where an “invariant point attention” mechanism is used.[41] Finally, the sidechain rotation χ-angles, which determine the detailed sidechain conformations, are predicted and 3D atomic coordinates can be computed. The output structure is then recycled back to the Evoformer a number of times to improve the final result. For further technical details, the reader is referred to Ref. 41 and its supplementary information.

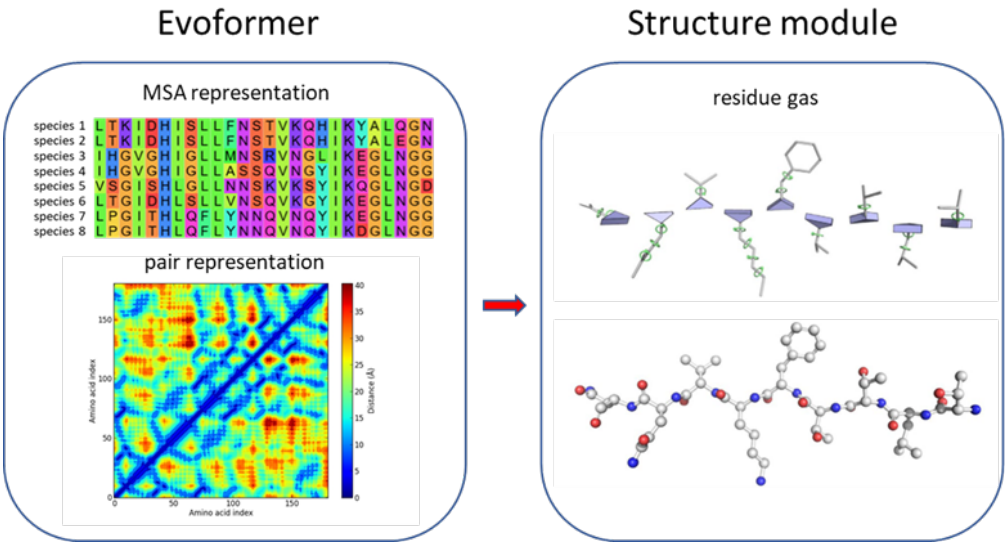

Figure 7. Schematic description of the two main modules of AF2. An input sequence together with data from sequence and structure databases serve as input to the Evoformer. The Structure module produces as output a 3D model of the protein structure corresponding to the input sequence.

Overall, the AF2 architecture can been described as an ingenious piece of neural network engineering by **Jumper**, **Hassabis** and their coworkers, with a multitude of new inventions, and it can be viewed as the first real scientific breakthrough of artificial intelligence. The fact that the AF2 source code was made public also decisively contributed to its impact, as it could be extensively tested and validated. A deep learning architecture similar to that of AF2 was also rapidly adopted by **Baker** and colleagues in the RoseTTAFold program.[43]

One might ask whether the deep learning methodology of AF2 is more or less equivalent to pattern recognition, but this is not really the case, since the number of possible conformations of a typical protein is astronomically larger than the number of atomic constellations found in the Protein Data Bank. Instead, the probable explanation for the program's performance is that it has effectively learned a potential of mean force (free energy surface), i.e., probability distributions for interatomic distances between pairs of atom types. There is thus a direct connection to the physical principles of protein structure, where the "knowledge" acquired by the program can be used to accurately determine structures.

One may also wonder why the protein design field appeared to be ahead of structure prediction for many years. The likely explanation for this is that the target structures for protein design are usually highly regular and idealized. Once a sequence for the target structure has been optimized, it is therefore more probable that its predicted structure will be rather accurate. In *ab initio* structure prediction, on the other hand, one just starts from a sequence with no information about what its 3D structure may be, and it may, in fact, be more or less irregular.

Summary and outlook

We are now at a stage where both the structural design and prediction problems are largely solved. The implications of this are far-reaching.

The AlphaFold2 team immediately created large databases of predicted protein structures, first for the human proteome[44] and then for the majority of sequences (> 200 million) available in the UniProt (Universal Protein Resource) database.[45] This means that almost overnight, we got access to orders of magnitude more structural information. Likewise, the protein design field has reached a stage where some of the most exciting areas of research are biomedical applications, such as vaccines and protein-based inhibitors,[46] and applications in synthetic biology.[47]

The progress described above would not have been possible, of course, without the efforts from structural biologists in providing all the experimentally determined structures that have gone into the Protein Data Bank. A number of these spectacular protein structures have also been recognized with Nobel Prizes in Chemistry over the years. These data, resulting from decades of research in protein structure determination, have laid the foundation for the decisive breakthroughs in protein design and structure prediction by this year's Laureates.

In summary, the achievements of **David Baker**, **Demis Hassabis** and **John Jumper** in the fields of computational protein design and protein structure prediction are truly profound. Their work has opened up a new era of biochemical and biological research, where we can now predict and design protein structures in ways that had not been possible before. Hence, a long-standing goal has finally been met, and the impact of this will have far-reaching consequences.

Johan Åqvist
Professor of Theoretical Chemistry
Member of the Royal Swedish Academy of Sciences
Member of the Nobel Committee for Chemistry

References

1. Kendrew, J.C.; Bodo, G.; Dintzis, H.M.; Parrish, R.G.; Wyckoff, H.; Phillips, D.C. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **1958**, *181*, 662-666.
2. Perutz, M.F.; Rossmann, M.G.; Cullis, A.F.; Muirhead, H.; Will, G.; North, A.C.T. Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* **1960**, *185*, 416-422.
3. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **1973**, *181*, 223-230.

4. Bouatta, N.; Sorger, P.; AlQuraishi, M. Protein structure prediction by AlphaFold2: are attention and symmetries all you need? *Acta Cryst.* **2021**, *D77*, 982-991.
5. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer Jr., E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535-542.
6. Levinthal, C. How to fold graciously. *Mössbauer Spectroscopy in Biological Systems Proceedings* **1969**, *67*, 22-26.
7. Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's paradox. *Proc. Natl. Acad. Sci. USA* **1991**, *89*, 20-22.
8. Pauling, L.; Corey, R.B.; Branson, H.R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **1951**, *37*, 205-211.
9. Chou, P.Y.; Fasman, G.D. Prediction of protein conformation. *Biochemistry* **1974**, *13*, 222-245.
10. https://www.rcsb.org/stats/growth/growth-released-structures. Retrieved October 9, 2024.
11. Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **1981**, *34*, 167-339.
12. Regan, L.; DeGrado, W.F. Characterization of a helical protein designed from first principles. *Science* **1988**, *241*, 976-978.
13. Kamtekar, S.; Schiffer, J.M.; Xiong, H.; Babik, J.M.; Hecht, M.H. Protein design by binary pattering of polar and nonpolar amino acids. *Science* **1993**, *262*, 1680-1685.
14. Dahiyat, B.I; Mayo, S.L. De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82-87.
15. Struthers, M.D.; Cheng, R.P.; Imperiali, B. Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science* **1996**, *271*, 342-345.
16. Harbury, P.B.; Plecs, J.J.; Tidor, B.; Alber, T.; Kim, P.S. High-resolution protein design with backbone freedom. *Science* **1998**, *282*, 1462-1467.
17. Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364-1368.
18. Simons, K.T.; Bonneau, R.; Ruczinski, I.; Baker, D. Ab initio protein structure prediction of CASPIII targets using ROSETTA. *Proteins: Struct. Funct. Genet. Suppl.* **1999**, *3*, 171-176.
19. Jones, T.A.; Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* **1986**, *5*, 819-822.
20. Huang, P.S.; Boyken, S.E.; Baker, D. The coming of age of *de novo* protein design. *Nature* **2016**, *537*, 320-327.
21. Jiang, L.; Althoff, E.A.; Clemente, F.R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J.L.; Betker, J.L.; Tanaka, F.; Barbas III, C.F.; Hilvert, D.; Houk, K.N.; Stoddard, B.L.; Baker, D. De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387-1391.

22. Röthlisberger, D.; Khersonsky, O.; Wollacott, A.M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J.L.; Althoff, E.A.; Zanghellini, A.; Dym, O.; Houk, K.N.; Tawfik, D.S., Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190-195.
23. Tinberg, C.E.; Khare, S.D.; Dou, J.; Doyle, L.; Nelson, J.W.; Schena, A.; Janowski, W.; Kalodimos, C.G.; Johnsson, K.; Stoddard, B.L.; Baker, D. *Nature* **2013**, *501*, 212-216.
24. Bale, J.B.; Gonen, S.; Liu, Y.; Sheffler, W.; Ellis, D.; Thomas, C.; Cascio, D.; Yeates, T.O.; Gonen, T.; King, N.P.; Baker. D. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **2016**, *353*, 389-394.
25. Langan, R.A.; Boyken, S.E.; Ng, A.H.; Samson, J.A.; Dods, G.; Westbrook, A.M.; Nguyen, T.H.; Lajoie, M.J.; Chen, Z.; Berger, S.; Khipple Mulligan, V.; Dueber, J.E.; Novak, W.R.P.; El-Samad, H.; Baker, D. De novo design of bioactive protein switches. *Nature* **2019**, *572*, 205-210.
26. Bick, M.J.; Greisen, P.J.; Morey, K.J.; Antunes, M.S.; La, D.; Sankaran, B.; Reymond, L.; Johnsson, K.; Medford, J.I.; Baker, D. Computational design of environmental sensors for the potent opioid fentanyl. *eLife* **2017**, *6*, e28909.
27. Moult, J.; Pedersen, J.T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Struct. Funct. Genet.* **1995**, *23*, ii-iv.
28. Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins: Struct. Funct. Genet. Suppl.* **2002**, *5*, 2-7.
29. Qian, N.; Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **1988**, *202*, 865-884.
30. Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R.M.J.; Lautrup, B.; Norskov, L.; Olsen, O.H.; Petersen, S.B. Protein secondary structure and homology by neural networks. The α-helices in rhodopsin. *FEBS Lett.* **1988**, *241*, 223-228.
31. Duan, Y.; Kollman, P.A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740-744.
32. Shaw, D.E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S., Dror, R.O.; Eastwood, M.P.; Bank, J.A.; Jumper, J.M.; Salmon, J.K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341-346.
33. Göbel, U.; Sander, C.; Schneide, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **1994**, *18*, 209-317.
34. Giraud, B.G.; Heumann, J.M.; Lapedes, A.S. Superadditive correlation. *Phys. Rev. E* **1999**, *59*, 4983-4991.
35. Weigt, M.; White, R.A.; Szurmant, H.; Hoch, J.A.; Hava, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 67-72
36. Shrestha, R.; Fajardo, E.; Gill, N.; Fidelis, K.; Kryshtafovych, A.; Monastyrskyy, B.; Fiser, A. Assessing the accuracy of contact predictions in CASP13. *Proteins* **2019**, *87*, 1058-1068.
37. Senior, A.W.; Evan, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A.W.R.; Bridgland, A.; Penedones, H.; Pedersen, S.; Simonyan, K.; Crossan, S.;

Kohli, P.; Jones, D.T.; Solver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potential from deep learning. *Nature* **2020**, *577*, 706-710.

38. Callaway, E. 'It will change everything': AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588*, 203-204-
39. Silver, S.; Hubert, T., Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi and Go through self-play. *Science* **2018**, *262*, 1140-1144.
40. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of protein structure prediction (CASP) – round XIV. *Proteins* **2021**, *89*, 1607-1617.
41. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S.A.A.; Ballard, A.J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S, Reiman, D.; Clancy, E., Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A.W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-589.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inform. Process. Syst.* **2017**, *30* (NIPS 2017).
43. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Rie Lee, G.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; Millan, C.; Park, H.; Adams, C.; Glassman, C.R.; DeGiovanni, A.; Pereira, J.H.; Rodrigues, A.V.; van Dijk, A.A.; Ebrecht, A.C.; Opperman, D.J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M.K.; Dalwadi, U.; Yip, C.K.; Burke, J.E.; Garcia, K.C.; Grishin, N.V.; Adams, P.D.; Read, R.J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871-876.
44. Tunyasuvunakool, K.; Adler, J; Wu, Z.; Green, T.; Zielinski, M.; Zidek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G.J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S.A.A.; Potapenko, A.; Ballard, A.J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reima .; Petersen, S.; Senior, A.W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590-596.
45. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Zidek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R., Vora, A.; Lufti, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Databse: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucl. Acids Res.* **2022**, *50*, D439-D444.

46. Cao, L.; Goreshnik, I.; Coventry, B.; Case, J.B., Miller, L.; Kozodoy, L.; Chen, R.E.; Carter, L.; Walls, A.C.; Park, Y.J.; Strauch, E.M.; Stewart, L.; Diamond, M.S.; Veesler, D.; Baker, D. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **2020**, *370*, 426-431.
47. Butterfield, G.L; Lajoie, M.J.; Gustafson, H.H.; Sellers, D.L.; Natterman, U.; Ellis, D.; Bale, J.B.; Ke, S.; Garreck, G.H.; Yehdego, A.; Ravichandran, R.; Pun, S.H.; King, N.P.; Baker, D. Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **2017**, *552*, 415-420.