Review Article

OPEN ACCESS CHECK FOR UPDATES

molecular systems biology

Deep learning for protein structure prediction and design—progress and applications

Jürgen Jänes[1,2] & Pedro Beltrao[1,2]✉

Abstract

Proteins are the key molecular machines that orchestrate all biological processes of the cell. Most proteins fold into three-dimensional shapes that are critical for their function. Studying the 3D shape of proteins can inform us of the mechanisms that underlie biological processes in living cells and can have practical applications in the study of disease mutations or the discovery of novel drug treatments. Here, we review the progress made in sequence-based prediction of protein structures with a focus on applications that go beyond the prediction of single monomer structures. This includes the application of deep learning methods for the prediction of structures of protein complexes, different conformations, the evolution of protein structures and the application of these methods to protein design. These developments create new opportunities for research that will have impact across many areas of biomedical research.

Keywords AlphaFold2; Structural Bioinformatics; Protein Design; Protein Conformations; Structural Systems Biology
Subject Categories Computational Biology; Structural Biology
https://doi.org/10.1038/s44320-024-00016-x
Received 26 June 2023; Revised 21 December 2023; Accepted 11 January 2024
Published online: 30 January 2024

Introduction

Predicting protein structure from sequence information has been a long-standing challenge in the field of molecular biology. The ability to accurately predict protein structure from sequence information alone would have far-reaching implications for our understanding of biological processes as well as disease and for developing new drugs and therapies.

Historically, one approach to predicting protein structure from sequence information has been homology modeling (Browne et al, 1969). This method relies on the assumption that proteins with similar sequences will have similar structures. By identifying a known protein structure that shares sequence similarity with the target protein, a model of the target protein's structure can be built. In addition to homology modeling, researchers have also explored the use of co-evolutionary information to predict protein structure. This approach is based on the observation that residues in a protein that are in close spatial proximity often co-evolve (Göbel et al, 1994; Benner and Gerloff, 1991). By analyzing patterns of co-evolution in multiple sequence alignments, it is possible to infer residue-residue contacts and use this information to predict protein structure. The development of prediction methods has progressed steadily over the years including improvements in obtaining residue distance constraints from multiple sequence alignments (Thomas et al, 2008; Dunn et al, 2008; Bartlett and Taylor, 2008; Wang et al, 2017) and in using this information for the prediction of 3D structures (Senior et al, 2020; Xu, 2019; AlQuraishi, 2019). These advances and their historical perspective have been reviewed elsewhere (AlQuraishi, 2021; Laine et al, 2021; Elofsson, 2023) and can be summarized by an increase in usage of neural network models along key parts of the protein structure prediction problem. These developments have led to the notable advance demonstrated by AlphaFold2 that has achieved very high accuracy in sequence-based structure prediction.

In this Review, we will discuss the recent developments and applications of deep learning-based methods for protein structure prediction and design.

Artificial Intelligence for sequence-based structure prediction

DeepMind showcased the results of AlphaFold2 in the 14th CASP conference in December of 2020. This led to a flurry of activity from different research groups resulting in several end-to-end deep learning models for sequence-based protein structure predictions. These are split into two main groups: alignment-based predictors—e.g., AlphaFold2 (Jumper et al, 2021), RoseTTAFold (Baek et al, 2021), and OpenFold (Ahdritz et al, 2022)—and protein language model-based predictors—including RGN2 (Chowdhury et al, 2022), ESMfold (Lin et al, 2023), OmegaFold (Wu et al, 2022), and EMBER2 (Weissenow et al, 2022). AlphaFold2 takes as inputs a multiple sequence alignment (MSA) and an initial set of pairwise distance measurements that could be optionally initialized via a structural template from a homologous sequence. The architecture is composed of two stages. The first stage processes the MSA and pairwise distances through repeated layers of a transformer-based neural network block dubbed Evoformer. The second stage is a so-called structure module that represents the rotation and translation for each protein residue. Each residue is represented as a triangle of the 3 backbone atoms (nitrogen, alpha-carbon, carbon) and the neural network has learned to move these triangles to the correct place in 3D space to form the predicted structures. The

[1]Institute of Molecular Systems Biology, ETH Zürich, 8093 Zürich, Switzerland. [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland. ✉E-mail: pbeltrao@ethz.ch

improvements derived from this and other architectures have been reviewed elsewhere (AlQuraishi, 2021) but one critical point is that these models are able to learn how best to extract co-evolution information from a multiple sequence alignment in order to predict distances between residues and the final structure. Removing the possibility of using structural templates has a minimal impact on model performance (Jumper et al, 2021).

RoseTTAFold (Baek et al, 2021) was an explicit attempt to design a model inspired by DeepMind's presentation of AlphaFold2 at CASP14, at a point when it was unclear if the details of the model were going to be publicly released. The model had a three-track neural network that processes MSA, pairwise distance, and 3D coordinate information simultaneously to produce structure predictions with accuracies that were approaching those of AlphaFold2. A more recent implementation of RoseTTaFold brings its accuracy closer to AlphaFold2 and extends its capabilities to the prediction of RNA and DNA structures (Baek et al, 2024). Finally, OpenFold (Ahdritz et al, 2022) is a reimplementation of AlphaFold2 that has the same architecture but having the code available as well as the data required for re-training. The OpenFold implementation also contains some technical improvements that improve speed and memory usage efficiency. The possibility to retrain this model has already been important to gain insight into what the model has learned and to adapt it for specific applications (as discussed in further detail below). Even though most of these approaches work by integrating co-evolutionary information and structure, there are still differences in their performance. Additional research will determine which specific design decisions are critical for performance.

The major input signal for the models described above is the multiple sequence alignment with the depth of the alignment impacting on the accuracy of the models. However, there are several proteins for which an MSA will not be available. Among others, these include proteins that: have been recently evolved; are rapidly evolving, are designed, or those with rearrangements. Protein language model-based predictors have been developed that aim to replace the need for an MSA with high-dimensional representations of protein sequences that are learned from the protein sequence universe. Protein language models (e.g., epiBERTope (Park et al, 2022), ESM (Rives et al, 2021), ProtTrans (Elnaggar et al, 2022), or ProteinBERT (Brandes et al, 2022)) are neural network models that are trained on predicting masked amino-acids from a very large number of protein sequences. As observed with large language models that have been popularized by chat bots, the simple task of learning masked words has led to models that read and write the language. Similarly, these protein language models capture a representation of protein sequence space that can be passed on to neural network models capable of using this representation to predict protein structure. RGN2, ESMfold, OmegaFold, and EMBER2 are examples of such models that vary in the protein language model used and how these protein sequence representations feed into the structure prediction. All of these models have in common that, in comparison to AlphaFold2, they have a simplified architecture, can run much faster but do not reach the same level of performance when an MSA is available (Elofsson, 2023). These different models have yet to be compared directly on extensive benchmarks but given their speed, they offer the possibility of being applied on a larger scale. As an example, the ESMfold has been used to make predictions for over 700 million proteins (Lin et al, 2023).

Choosing which model to use will depend on the user's application. For predicting individual structures, the user is better off using AlphaFold2 and putting a larger effort on improving the multiple sequence alignment. For most cases, ColabFold (Mirdita et al, 2022) has modified AlphaFold2 and other methods to run at reduced computational cost with minimal loss in accuracy. Further, ColabFold can fold sequences of up to 1000 residues on Google Colaboratory, without any computational requirements for the user. For individual examples where other approaches are unsuccessful, aggressive sampling as implemented in AFSample (Wallner, 2023) would have the highest chance of success at the expense of significantly more computational resources. Finally, in the absence of an MSA or for very large scale applications, one of the protein language models is likely better suited to the task.

Opening the black box: what have these models learned

Deep learning methods are complex models with a very large number of parameters which are often described as "black box" models given the difficulty in studying how the models make their predictions and what they have learned. During the training process, the OpenFold team has studied what information their model captured at intermediate steps of the training process (Ahdritz et al, 2022). Independent training runs tended to follow a similar progression where, within the initial steps, the model learned a 1D representation of the structure, followed by 2D and 3D phases of learning that reached reasonably accurate backbone representations. Only then are the representations of the secondary structural elements fully learned, despite the fact that local secondary structure can be predicted even from sequence alone. One of the questions initially raised by the release of AlphaFold2 was the degree of generalization to unseen parts of the protein structural universe. OpenFold addressed this question by training the model on very distinct types of structures, leaving out different protein families or even training on structures composed of single secondary structure elements. Encouragingly, training on these subsets revealed that this architecture is quite robust and can generalize to structures of unseen protein families.

Classical protein structure prediction methods have relied on an energy function to rank possible solutions by considering different energy terms such as the contribution of steric clashes, the formation of hydrogen bonds or electrostatic interactions, etc. To explore if AlphaFold2 may have learned an energy function, Roney and Ovchinnikov used it to rank different related template structures without providing an MSA (Roney and Ovchinnikov, 2022). This analysis indicates that, in the absence of co-evolution signals, AlphaFold2 can rank which structural templates are a better fit to a sequence, suggesting that this model has also learned something akin to an energy function. Additional explorations of these models will be needed to better understand what aspects of biophysics may have been incorporated into them and at what stage of the training process these are acquired.

Protein structure comparisons empowering evolutionary studies

The development of high-confidence sequence-based structure predictions opens the door for the prediction of structures for large

parts of the protein universe. While the unique proteins with solved structures represented in PDB are on the order 100,000, the protein sequences available for analysis are on the order of billions. Currently, the AlphaFold database contains 217 million structures predicted by AlphaFold2 and the ESMfold Metagenomic Atlas contains predicted structures for 772 million proteins. This dramatic increase in the available predicted structures should empower many studies, including the study of diversity of the protein structural universe, the evolution of protein sequences, structures and function, and the potential discovery of novel enzymes. However, analyzing such large numbers of structures also requires the development of highly efficient computational approaches. Even accounting for some high level of redundancy in the sequence databases, such methods would need to be applicable to the scale of tens to hundreds of millions of structures. Examples in this context include the development of efficient methods for pocket comparison (Simonovsky and Meyers, 2020), comparison of protein structures (van Kempen et al, 2023; Durairaj et al, 2020), clustering of protein structures (Barrio-Hernandez et al, 2023) and compression of structural data files (Kim et al, 2023).

The 365,000 structures that were first released in the AlphaFold2 database led to initial attempts of clustering and characterization. A proof-of-principle analysis showed that clustering of these structures based on the similarity of their structural elements could be used to recover groups of known protein families, supporting the use of such approaches for evolutionary studies (Akdel et al, 2022). A protein family analysis of these same structures suggested that around 92% of the predicted domains within this set matched already known superfamilies (Bordin et al, 2023). Recently, an efficient clustering method was used to cluster the 217 million structures in AlphaFold DB, leading to the identification of 2.27 M non-singleton clusters (Barrio-Hernandez et al, 2023). While 31% of these clusters were deemed to represent likely novel structures, these clusters lacking annotations only cover 4% of all proteins in the database. These observations would suggest that the majority of protein structures have, at least partial matches to known protein families. However, the diversity of shapes and functions within each cluster can still be of interest. In addition, these clusters were used for evolutionary studies, identifying cases of remote structural similarity between eukaryotic and prokaryotic structures where sequence-based methods would not easily identify a link.

Protein complexes and integrative structural modeling

While AlphaFold2 was trained to predict the structures of individual proteins, co-evolutionary information has been used to predict protein-protein interactions since the development of direct coupling analysis algorithms (Weigt et al, 2009). Earlier deep learning methods like Raptor-X (Jing et al, 2020) were making use of protein-interaction contact site predictions for predicting complex structures and it was therefore not unexpected that roseTTAFold and AlphaFold2 could also be applied to this challenge (Mirdita et al, 2022; Akdel et al, 2022; Evans et al, 2021; Ko and Lee, 2021; Bryant et al, 2022a). Multiple independent reports have benchmarked the capacity of AlphaFold2 to predict the structures of complexes and attempted to improve its performance at this task either by improving the correct species pairing of sequences for both proteins in the MSA (Bryant et al, 2022a) or by explicitly training the models on structures of protein complexes (Evans et al, 2021). When comparing against experimental structures, these methods have reported correct predictions of the interface running between 50 to 70% of the cases. In addition, the size of the predicted interface and the confidence estimate for the residues at the interface can be used to rank the predicted models according to an estimated accuracy (Bryant et al, 2022a). It is important to note that this estimated accuracy only relates to cases when it is known that two proteins interact via a direct interface. The prediction accuracy for protein pairs that may or may not include direct interfaces will be lower than the accuracy described above. This can be somewhat circumvented by using the estimated confidence values from the predicted models but should be considered even when deciding what protein pairs to attempt to model. These methods have been applied on a larger scale to predict the structures for previously known protein interactions for *S. cerevisiae*, *B. subtilis*, and human (Humphreys et al, 2021; O'Reilly et al, 2023; Burke et al, 2023), showcasing how these can be applied for example to predict the impact of mutations at protein interfaces. Given the combinatorial nature of protein interactions, using these methods on all known or predicted protein interactions in a species remains resource intensive. The faster protein language-based models, once adapted to this problem, may serve as a useful screening method that could be applied on a very large scale.

The prediction of complexes containing larger numbers of proteins remains challenging due to computer memory limitations in these methods. Given these limitations, an approach (see Fig. 1A) has been to predict models for pairs or triplets of proteins in the same complex and develop a strategy to assemble them by superimposition (Burke et al, 2023; Bryant et al, 2022b). These studies have pointed out some limitations, including the need of prior knowledge on the stoichiometry of subunits within the complex and the higher error rates for correct placement of paralogous subunits within the same complex.

The availability of predicted structures for monomers and complexes can also be used as part of pipelines for integrative structural modeling. In integrative modeling of protein complexes, different data modalities are used as restraints in order to score the possible conformations of represented molecules of interest (see review (Ziemianowicz and Kosinski, 2022)). In this application, even the predicted monomer structures are of high interest since they can be used to fit predicted atomic structures for specific proteins in lower-resolution experimental data of larger assemblies. A notable recent example of this was the combination of AlphaFold2-based predictions with cryo-electron tomography data to solve a 70-megadalton model of the human Nuclear Pore Complex (NPC) (Mosalaganti et al, 2022).

From single structures to ensembles

Proteins are dynamic and can exist in different conformations. Despite this, AlphaFold2 and related methods have been trained to produce a single structural representation for a given protein sequence. Early studies suggested that AlphaFold2 could not predict the structural changes of mutated sequences (Buel and

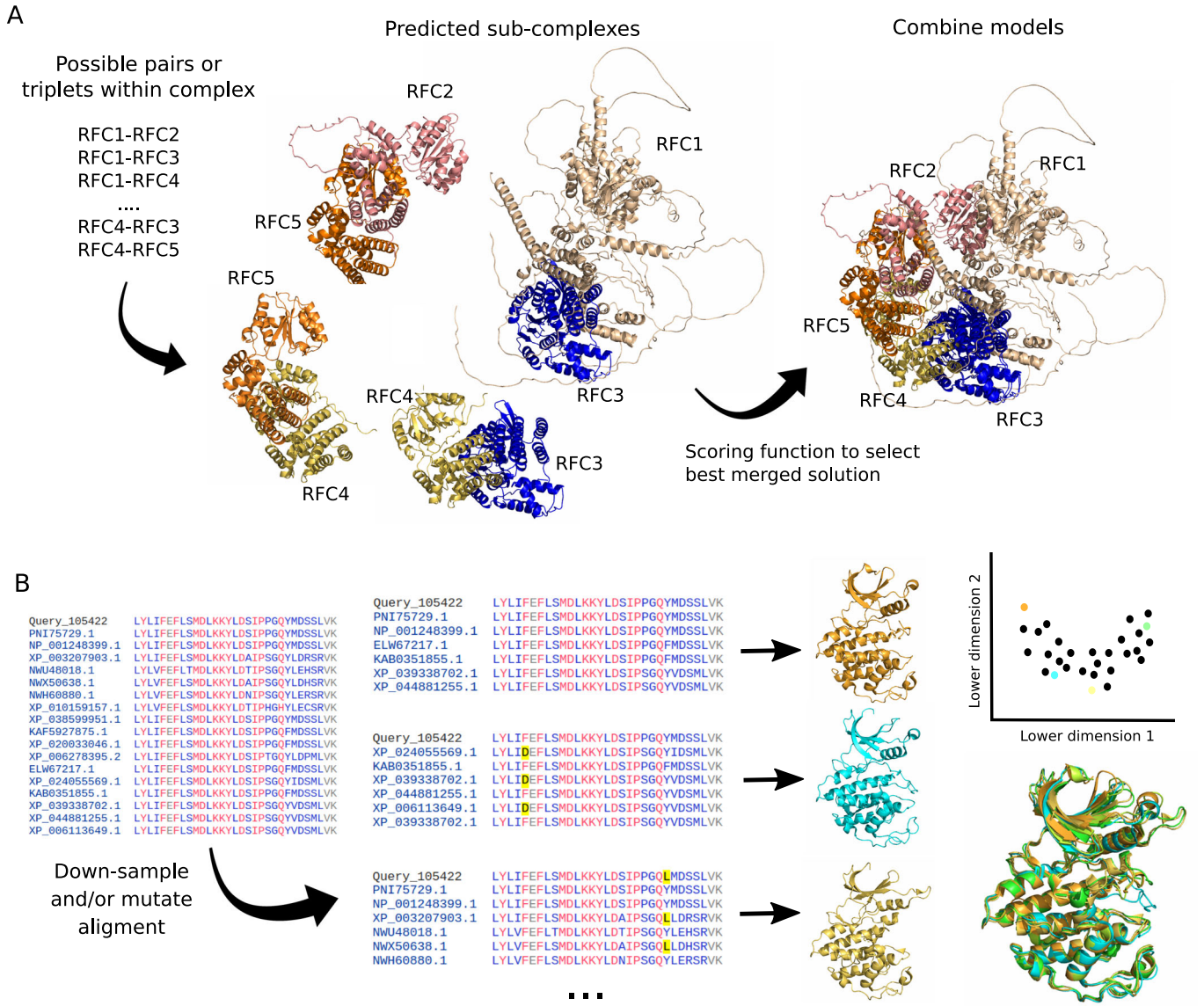

Figure 1. Example applications of AlphaFold2 beyond single protein structure prediction.

(A) Alphafold2 has shown to be capable of predicting structures for binary protein complexes but predicting structures for larger assemblies remains challenging. A suggested procedure has been to predict the structures for possible sub-complexes and then combine them using superimposition of common subunits (see main text). **(B)** While AlphaFold2 is trained to predict a single conformation, it has been shown that subsampling of the alignment that serves as the main input, can result in the prediction of different conformations that sometimes resemble known conformations.

Walters, 2022) or different conformation of proteins that are known to change in structure when bound to a small molecule (Saldaño et al, 2022). However, this was based on multiple runs with the same parameters. Several independent groups then reported success in predicting different conformations by providing, in different runs, an MSA with a smaller set of random sequences from the full alignment (Del Alamo et al, 2022b) (see Fig. 1B). Presumably, such random down-sampling of the alignment may expose co-evolution signals that predispose the prediction towards different conformations. In addition to uniform down-sampling of the alignment, other strategies include the mutation of residues in the alignment that correspond to positions of contact within the structure (Stein and Mchaourab, 2022) or down-sampling of the alignment after clustering the sequences (Wayment-Steele et al, 2024). Down-sampling the MSA by selecting them from a sequence clustering method was shown to substantially improve the prediction of known alternative conformations when compared to uniform random down-sampling. In all cases, the methodology follows a similar strategy: generating a large number of predictions with different perturbations of the alignment; grouping predictions by structural similarity to identify high-confidence predictions that are different from each other; comparing with existing structures or external sources of data.

The prediction of different conformations has now been successfully applied in a number of different systems (see review (Sala et al, 2023)), including transporters and GPCRs (Del Alamo et al, 2022b), different "metamorphic" proteins (KaiB, RfaH, Mad2) (Wayment-Steele et al, 2024), small molecule binding proteins (adenylate cyclase, ribose binding protein, tryptophan synthase) (Casadevall et al, 2022; Stein and Mchaourab, 2022), and proteins with pockets that open in specific conformations (i.e cryptic pockets) (Meller et al, 2023). In one case, the analysis and perturbations of the alignments were used to predict specific mutations that could result in a change of the preferred conformation (Wayment-Steele et al, 2024). Not all attempts at predicting known alternative conformations were successful and it is unclear how the success rate depends on the representation of the different conformation in the training data. Nevertheless, these results indicate that AlphaFold2 and related methods might have the capacity to predict different conformations and can be combined with orthogonal sources data (del Alamo et al, 2022a). However, it is likely that better methodology may be developed, using deep learning models that are specifically trained for the purpose of predicting alternative conformations.

In related efforts, the distributions of inter-residue confidence estimates predicted by AlphaFold2 have been used to construct structural ensembles of intrinsically disordered proteins (Faidon Brotzakis et al, 2023). Alternatively, AlphaFold2 confidence scores have been combined with elastic networks to generate structural ensembles (Jussupow and Kaila, 2023).

Advances in deep learning methods for protein design

Protein design aims to generate proteins with a pre-determined shape and/or function with great potential for the rational design of enzymes, scaffolds, high-affinity binders, and other functions of biotechnological or therapeutic value. Rational protein design can be seen as the inverse problem of sequence-based protein structure prediction where the objective is to predict a sequence that will have a pre-determined structure or function. Traditionally, this has been achieved by computational protocols that can search through favorable sequences that are ranked according to a physics inspired energy function (review in (Kuhlman and Bradley, 2019)). As for protein structure prediction, deep learning neural network models have been recently applied to dramatically improve on the capacity to design proteins with diverse characteristics (Anand et al, 2022; Strokach et al, 2020; Huang et al, 2022; Anishchenko et al, 2021; Madani et al, 2023; Verkuil et al, 2022; Watson et al, 2023). Recent approaches have adapted similar architectures used for protein structure predictions for the generative task of protein design leading to order-of-magnitude increases in success rates. High experimentally confirmed rates have been reported on the design of proteins with pre-defined backbones (67% success rate measured as solubility and monomeric state) (Verkuil et al, 2022), novel sequences for an existing enzyme family (73%) (Madani et al, 2023), pre-defined oligomerization states with novel proteins (11.5%), novel ion binding proteins (40%) and binders to specific target proteins (18%) (Watson et al, 2023). Importantly, the controllability of the designs has also improved substantially (Watson et al, 2023; Hie et al, 2022) whereby the target protein can be steered by easy to implement user defined constraints. These can, in principle, be tuned to any function for which the target sequence and/or structure can be measured against. It is important to note that these success rates are not strictly comparable due to the differences in defining success for different design tasks and that it remains to be seen if these success rates can be easily replicated in different labs.

Perspective

AlphaFold2 and related methods have made great progress at predicting structures for well folded single-domains, and made significant advances in other areas such as protein-protein interactions. One major aspect of AlphaFold2's success lies in combining PDB structures with co-evolutionary signals derived from large protein sequence databases in an end-to-end differentiable way. We think these approaches have the potential to be extended in several directions.

Several mass-spectrometry (MS) assays, such as cross-linking (XL-MS), hydrogen deuterium exchange (HDX-MS), limited proteolysis-coupled mass spectrometry (LiP-MS), can capture information on regions that are in close contact, freely accessible to the solvent or that change in accessibility under some conditions. AlphaFold2 has already been adapted to use in-cell cross-linking information for improved protein modeling (Stahl et al, 2023) and this could be further generalized to consider other sources of constraints that could include the above mentioned MS methods and also constraints from structural methods such as NMR, x-ray crystallography and cryo-EM.

In addition to proteins, it is likely that methods related to AlphaFold may be extendable to other types of molecules. RoseTTAFold has been adapted to predict protein-nucleic acid complexes (Baek et al, 2024). RoseTTAFoldNA has already improved the state of the art despite the low number of available nucleic acid structures. Here, further improvements could be obtained by integrating data from high-throughput protein-nucleic acid profiling experiments. AlphaFold2-related methods cannot yet predict protein-small molecule interactions and docking small ligands into the structure is challenging (Holcomb et al, 2023). A recent method, DiffDock, shows better performance on computationally folded structures (Corso et al, 2022), although overall success rates remain low. Databases such as ChEMBL contain binding information on a non-overlapping set of ligands and targets (Liu et al, 2015), although without structural information. Nevertheless, similarly to the co-evolutionary signal from Uniprot, an end-to-end differentiable pipeline combining PDBbind—a collection of measured binding affinity data for complexes deposited in PDB—with small-molecule binding data could improve protein-small molecule complex predictions.

In addition, the current approaches are limited to predicting a single structure per input sequence (Lane, 2023). Prediction of multiple discrete confirmations is possible in some cases, but complex dynamics such as predicting the exact folding pathways is beyond the reach of current methods (Outeiral et al, 2022). Here, trajectories from molecular dynamics simulations could be used as complementary training input. Early results suggest that this is possible, and can generalize to systems beyond the training set (Janson et al, 2023).

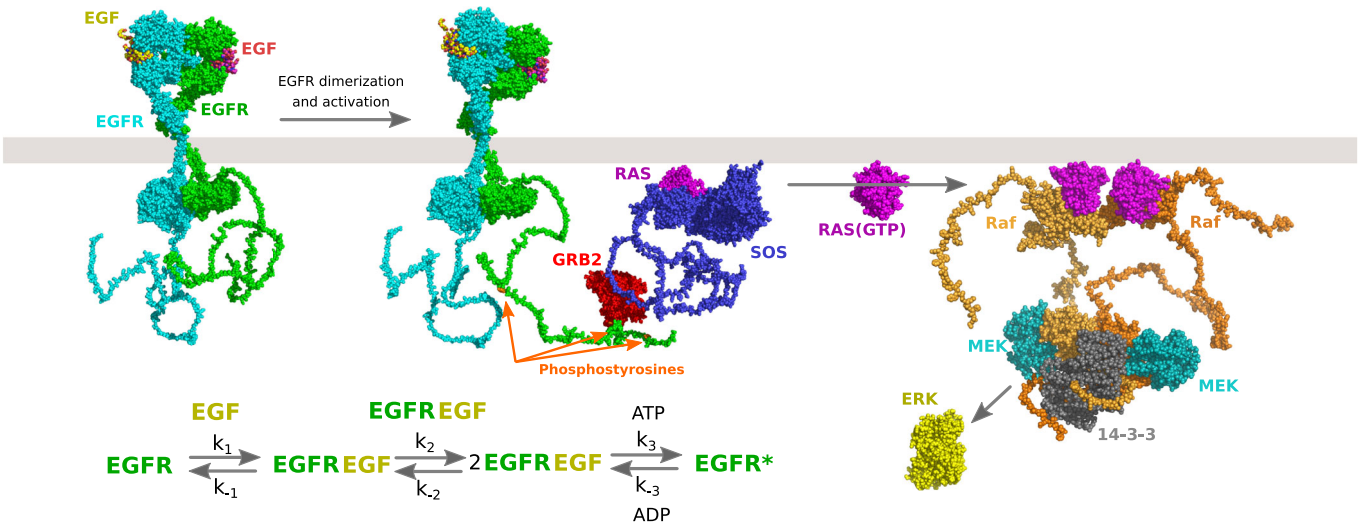

Figure 2. Proteome-wide structural systems biology.

Structural details for the initial steps of EFG pathway activation. For representation, the AlphaFold2 predicted structures of pathway components were combined with experimental structures from years of study of this pathway, including PDB ids: 1egf, 1nql, 1m17, 2jwa, 3njp, 2gs6, 1gri, 1xd2, 3ksy, 5p21, 6xi7, 6q0j, 2y4i, 1pme. The AlphaFold2 models help complete the missing protein sequence information not represented in the experimental results, in particular for the long unstructured regions. The example is inspired by similar visualization in PDB-101 (https://pdb101.rcsb.org/learn/exploring-the-structural-biology-of-cancer). It may become possible to use the protein sequences and structures to derive reaction parameters that would allow us to better understand the mechanisms underlying a system of interest.

The combination of improvements in experimental and computational approaches is leading to a revolution in structural biology whereby structural information is expected to cover the full proteomes of key species of interest. In Fig. 2, we combine AlphaFold2 models with experimental structures for proteins in the early steps of EGF pathway activation. In this example, the AlphaFold2 models helped in particular in visualizing the long unstructured regions which give context to those missing sequences. While this model is likely to contain many errors it challenges us to think about the complete atomic details of multi-component cellular processes. While it is clear to us that this expanded structural view of the cell should open many possible research questions it is not yet obvious exactly what the most promising future directions might be. As an example, future developments in this area may include the ability to derive reaction parameters directly from protein sequences/structures in order to model a system of interest. Structural models have always been a means towards better understanding the mechanisms of life. It is up to the research community now to take these advances in bold new directions.

References

Ahdritz G, Bouatta N, Kadyan S, Xia Q, Gerecke W, O'Donnell TJ, Berenberg D, Fisk I, Zanichelli N, Zhang B et al (2022) OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Preprint at bioRxiv https://doi.org/10.1101/2022.11.20.517210

Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G et al (2022) A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol 29:1056–1067

AlQuraishi M (2019) End-to-end differentiable learning of protein structure. Cell Syst 8:292–301.e3

AlQuraishi M (2021) Machine learning in protein structure prediction. Curr Opin Chem Biol 65:1–8

Anand N, Eguchi R, Mathews II, Perez CP, Derry A, Altman RB, Huang P-S (2022) Protein sequence design with a learned potential. Nat Commun 13:746

Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK et al (2021) De novo protein design by deep network hallucination. Nature 600:547–552

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876

Baek M, McHugh R, Anishchenko I, Jiang H, Baker D, DiMaio F (2024) Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. Nat Methods 21(1):117–121. https://doi.org/10.1038/s41592-023-02086-5

Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M (2023) Clustering predicted structures at the scale of the known protein universe. Nature 622(7983):637–645. https://doi.org/10.1038/s41586-023-06510-w

Bartlett GJ, Taylor WR (2008) Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction. Proteins 71:950–959

Benner SA, Gerloff D (1991) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv Enzyme Regul 31:121–181. https://doi.org/10.1016/0065-2571(91)90012-B

Bordin N, Sillitoe I, Nallapareddy V, Rauer C, Lam SD, Waman VP, Sen N, Heinzinger M, Littmann M, Kim S et al (2023) AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. Commun Biol 6:160

Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38:2102–2110

Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL (1969) A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 42(1):65–86 https://doi.org/10.1016/0022-2836(69)90487-2

Bryant P, Pozzati G, Elofsson A (2022a) Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13:1265

Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P, Elofsson A (2022b) Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Nat Commun 13:6028

Buel GR, Walters KJ (2022) Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol 29:1–2

Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A et al (2023) Towards a structurally resolved human protein interaction network. Nat Struct Mol Biol 30:216–225

Casadevall G, Duran C, Estévez-Gay M, Osuna S (2022) Estimating conformational heterogeneity of tryptophan synthase with a template-based Alphafold2 approach. Protein Sci 31:e4426

Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, Rochereau C, Ahdritz G, Zhang J, Church GM et al (2022) Single-sequence protein structure prediction using a language model and deep learning. Nat Biotechnol 40:1617–1623

Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T (2022) DiffDock: diffusion steps, twists, and turns for molecular docking. Preprint at https://arxiv.org/abs/2210.01776

Del Alamo D, DeSousa L, Nair RM, Rahman S, Meiler J, Mchaourab HS (2022a) Integrated AlphaFold2 and DEER investigation of the conformational dynamics of a pH-dependent APC antiporter. Proc Natl Acad Sci USA 119:e2206129119

Del Alamo D, Sala D, Mchaourab HS, Meiler J (2022b) Sampling alternative conformational states of transporters and receptors with AlphaFold2. Elife 11:e75751

Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340

Durairaj J, Akdel M, de Ridder D, van Dijk ADJ (2020) Geometricus represents protein structures as shape-mers derived from moment invariants. Bioinformatics 36:i718–i725

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M et al (2022) ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44:7112–7127

Elofsson A (2023) Progress at protein structure prediction, as seen in CASP15. Curr Opin Struct Biol 80:102594

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J et al (2021) Protein complex prediction with AlphaFold-Multimer. Preprint at bioRxiv https://doi.org/10.1101/2021.10.04.463034

Faidon Brotzakis Z, Zhang S, Vendruscolo M (2023) AlphaFold prediction of structural ensembles of disordered proteins. Preprint at bioRxiv https://doi.org/10.1101/2023.01.19.524720

Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins: Structure Function and Bioinformatics 18(4):309–317. https://doi.org/10.1002/prot.340180402

Hie B, Candido S, Lin Z, Kabeli O, Rao R, Smetanin N, Sercu T, Rives A (2022) A high-level programming language for generative protein design. Preprint at bioRxiv https://doi.org/10.1101/2022.12.21.521526

Holcomb M, Chang Y-T, Goodsell DS, Forli S (2023) Evaluation of AlphaFold2 structures as docking targets. Protein Sci 32:e4530

Huang B, Xu Y, Hu X, Liu Y, Liao S, Zhang J, Huang C, Hong J, Chen Q, Liu H (2022) A backbone-centred energy function of neural networks for protein design. Nature 602:523–528

Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR et al (2021) Computed structures of core eukaryotic protein complexes. Science 374:eabm4805

Janson G, Valdes-Garcia G, Heo L, Feig M (2023) Direct generation of protein conformational ensembles via machine learning. Nat Commun 14:774

Jing X, Zeng H, Wang S, Xu J (2020) A web-based protocol for interprotein contact prediction by deep learning. Methods Mol Biol 2074:67–80

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589

Jussupow A, Kaila VRI (2023) Effective molecular dynamics from neural network-based structure prediction Models. J Chem Theory Comput 19:1965–1975

Kim H, Mirdita M, Steinegger M (2023) Foldcomp: a library and format for compressing and indexing large protein structure sets. Bioinformatics 39(4):btad153. https://doi.org/10.1093/bioinformatics/btad153

Ko J, Lee J (2021) Can AlphaFold2 predict protein-peptide complex structures accurately? Preprint at bioRxiv https://doi.org/10.1101/2021.07.27.453972

Kuhlman B, Bradley P (2019) Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 20:681–697

Laine E, Eismann S, Elofsson A, Grudinin S (2021) Protein sequence-to-structure learning: is this the end(-to-end revolution)? Proteins 89:1770–1786

Lane TJ (2023) Protein structure prediction has reached the single-structure frontier. Nat Methods 20(2):170–173

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y et al (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123–1130

Liu Z, Li J, Liu J, Liu Y, Nie W, Han L, Li Y, Wang R (2015) Cross-mapping of protein - ligand binding data between ChEMBL and PDBbind. Mol Inform 34:568–576

Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos Jr JL, Xiong C, Sun ZZ, Socher R et al (2023) Large language models generate functional protein sequences across diverse families. Nat Biotechnol 41(8):1099–1106

Meller A, Bhakat S, Solieva S, Bowman GR (2023) Accelerating cryptic pocket discovery using AlphaFold. J Chem Theory Comput 19(14):4355–4363

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022) ColabFold: making protein folding accessible to all. Nat Methods 19:679–682

Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, Zimmerli CE, Buczak K, Schmidt FH, Margiotta E, Mackmull MT, Hagen WJH, Hummer G, Kosinski J, Beck M (2022) AI-based structure prediction empowers integrative structural analysis of human nuclear pores. Science 376(6598):eabm9506. https://doi.org/10.1126/science.abm9506

O'Reilly FJ, Graziadei A, Forbrig C, Bremenkamp R, Charles K, Lenz S, Elfmann C, Fischer L, Stülke J, Rappsilber J (2023) Protein complexes in cells by AI-assisted structural proteomics. Mol Syst Biol 19:e11544

Outeiral C, Nissley DA, Deane CM (2022) Current structure predictors are not learning the physics of protein folding. Bioinformatics 38(7):1881–1887

Park M, Seo S-W, Park E, Kim J (2022) EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. Preprint at bioRxiv https://doi.org/10.1101/2022.02.27.481241

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J et al (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118:e2016239118

Roney JP, Ovchinnikov S (2022) State-of-the-art estimation of protein model accuracy using AlphaFold. Phys Rev Lett 129:238101

Sala D, Engelberger F, Mchaourab HS, Meiler J (2023) Modeling conformational states of proteins with AlphaFold. Curr Opin Struct Biol 81:102645. https://doi.org/10.1016/j.sbi.2023.102645

Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, Gonik E, García Melani A, Novomisky Nechcoff J, Salas MN et al (2022) Impact of protein conformational diversity on AlphaFold predictions. Bioinformatics 38:2742–2748

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710

Simonovsky M, Meyers J (2020) DeeplyTough: learning structural comparison of protein binding sites. J Chem Inf Model 60:2356–2366

Stahl K, Graziadei A, Dau T, Brock O, Rappsilber J (2023) Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. Nat Biotechnol 41(12):1810–1819. https://doi.org/10.1038/s41587-023-01704-z

Stein RA, Mchaourab HS (2022) SPEACH_AF: sampling protein ensembles and conformational heterogeneity with Alphafold2. PLoS Comput Biol 18:e1010483

Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM (2020) Fast and flexible protein design using deep graph neural networks. Cell Syst 11:402–411.e4

Thomas J, Ramakrishnan N, Bailey-Kellogg C (2008) Graphical models of residue coupling in protein families. IEEE/ACM Trans Comput Biol Bioinform 5:183–197

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M (2023) Fast and accurate protein structure search with Foldseek. Nat Biotechnol https://doi.org/10.1038/s41587-023-01773-0. Online ahead of print

Verkuil R, Kabeli O, Du Y, Wicky BIM, Milles LF, Dauparas J, Baker D, Ovchinnikov S, Sercu T, Rives A (2022) Language models generalize beyond natural proteins. Preprint at bioRxiv https://doi.org/10.1101/2022.12.21.521521

Wallner B (2023) AFsample: improving multimer prediction with AlphaFold using massive sampling. Bioinformatics 39(9):btad573. https://doi.org/10.1093/bioinformatics/btad573

Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 13:e1005324

Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF et al (2023) De novo design of protein structure and function with RFdiffusion. Nature 620(7976):1089–1100. https://doi.org/10.1038/s41586-023-06415-8

Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, Ovchinnikov S, Colwell L, Kern D (2024) Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625(7996):832–839. https://doi.org/10.1038/s41586-023-06832-9

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106:67–72

Weissenow K, Heinzinger M, Rost B (2022) Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. Structure 30:1169–1177.e4

Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B et al (2022) High-resolution *de novo* structure prediction from primary sequence. Preprint at bioRxiv https://doi.org/10.1101/2022.07.21.500999

Xu J (2019) Distance-based protein folding powered by deep learning. Proc Natl Acad Sci USA 116:16856–16865

Ziemianowicz DS, Kosinski J (2022) New opportunities in integrative structural modeling. Curr Opin Struct Biol 77:102488 https://doi.org/10.1016/j.sbi.2022.102488

Acknowledgements

PB is supported by the Helmut Horten Stiftung and the ETH Zurich Foundation.

Author contributions

Jürgen Jänes: Conceptualization; Visualization; Writing—original draft; Writing—review and editing. **Pedro Beltrao**: Conceptualization; Visualization; Writing—original draft; Writing—review and editing.

Disclosure and competing interests statement

The authors declare no competing interests.