TYPE Mini Review
PUBLISHED 28 February 2023
DOI 10.3389/fbinf.2023.1120370

Check for updates

OPEN ACCESS

EDITED BY
Petras Kundrotas,
University of Kansas, United States

REVIEWED BY
Andrzej Kolinski,
University of Warsaw, Poland
Arun Prasad Pandurangan,
Department of Biochemistry, University of Cambridge, United Kingdom
Lim Heo,
Michigan State University, United States

*CORRESPONDENCE
Samantha K. Teixeira,
samantha.teixeira@hc.fm.usp.br

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Protein Bioinformatics, a section of the journal Frontiers in Bioinformatics

RECEIVED 09 December 2022
ACCEPTED 17 February 2023
PUBLISHED 28 February 2023

CITATION
Bertoline LMF, Lima AN, Krieger JE and Teixeira SK (2023), Before and after AlphaFold2: An overview of protein structure prediction.
Front. Bioinform. 3:1120370.
doi: 10.3389/fbinf.2023.1120370

Before and after AlphaFold2: An overview of protein structure prediction

Letícia M. F. Bertoline†, Angélica N. Lima†, Jose E. Krieger and Samantha K. Teixeira*†

Laboratory of Genetics and Molecular Cardiology, Heart Institute, University of São Paulo Medical School, São Paulo, Brazil

Three-dimensional protein structure is directly correlated with its function and its determination is critical to understanding biological processes and addressing human health and life science problems in general. Although new protein structures are experimentally obtained over time, there is still a large difference between the number of protein sequences placed in Uniprot and those with resolved tertiary structure. In this context, studies have emerged to predict protein structures by methods based on a template or free modeling. In the last years, different methods have been combined to overcome their individual limitations, until the emergence of AlphaFold2, which demonstrated that predicting protein structure with high accuracy at unprecedented scale is possible. Despite its current impact in the field, AlphaFold2 has limitations. Recently, new methods based on protein language models have promised to revolutionize the protein structural biology allowing the discovery of protein structure and function only from evolutionary patterns present on protein sequence. Even though these methods do not reach AlphaFold2 accuracy, they already covered some of its limitations, being able to predict with high accuracy more than 200 million proteins from metagenomic databases. In this mini-review, we provide an overview of the breakthroughs in protein structure prediction before and after AlphaFold2 emergence.

KEYWORDS
protein structure prediction, AlphaFold, template-based modeling, free modeling, protein language model

1 Introduction

Protein is the term proposed by the Swedish chemist Jacob Berzelius to compounds containing nitrogen and constituted by a combination of amino acids linked by a bond, called peptide bond (Fruton, 1985; Wisniak, 2000; Lehninger et al., 2005; Voet et al., 2014). They are responsible for housekeeping and specific functions essential to life, such as cell structural support, immune protection, enzymatic catalysis, cell signal transduction to transcription and translation regulation (Pearce and Zhang, 2021). The biological function of a protein depends on its tertiary/quaternary structures that derives ultimately from the folding of a polypeptide sequence(s) considering physical chemistry principles and the lowest free energy level, being the understanding of protein folding one of the most important goals in structural biology (Lehninger et al., 2005; Voet et al., 2014).

Experimentally, tertiary protein structures are resolved by X-ray crystallography, nuclear magnetic resonance and electron cryomicroscopy (cryo-EM). Nevertheless, those techniques are complex, time consuming, expensive, and often the structure is not in its native form. Under these constrains, it is no surprise that the number of proteins with resolved tertiary structures is small (200,988 PDB entries) compared to the large number of proteins sequence known (229,580,745 entries on UniProtKB as of January 2023). This is an open challenge waiting for innovative ways to develop novel protein structure prediction approaches such as the ones using computationally models to predict the three dimensional protein structures starting from a polypeptide sequence (Dill et al., 2007; Nassar et al., 2021).

Several algorithms and web servers have been developed with the aim to improve the protein structure prediction. The relevance of these efforts are underscored by the different application that shall be affected including rational drug design, mutational studies and structural comparison, evolutionary analysis, folding studies. Here, we provide an overview of the methods developed to perform protein structure prediction to compare with AlphaFold2, which combines neural networks and homology modeling to generate models that may have experimental accuracy in a growing number of examples. We also discuss its applicability, limitations and efforts to improve the overall algorithm's performance.

2 Structure prediction methods

The prediction methods are usually divided into template-based modeling (TBM) and free modeling (FM), considering the use or not of templates (Gromiha et al., 2018; Bongirwar and Mokhade, 2022; Paiva et al., 2022), even though, recently, some TBM methods use energy-guided model refinement, and part of FM uses fragment-based sampling approaches, extracting information from Protein Data Bank (PDB) through machine learning. Within these two groups, the algorithms developed are usually classified into three different groups: *ab initio* (a FM methodology), threading/fold recognition (a TBM methodology), and homology (a TBM methodology) (Gromiha et al., 2018; Agnihotry et al., 2022). Despite controversial, this classification corresponds to the categories from the Critical Assessment of Structure Prediction (CASP), a biennial competition with the aim to establish the current state of the art in protein structure prediction, which contains three categories: TBM, FM, and an intermediate category, FM/TBM. In this contest, participants submit their models for proteins whose experimental structure has not been published yet (Kryshtafovych et al., 2019).

The *ab initio* approaches are based on the thermodynamics hypothesis that the native protein structure presents the lowest free energy possible (Hardin et al., 2002; Yuan et al., 2003; Gromiha et al., 2018; Agnihotry et al., 2022). The idea of these methods are to predict new folds considering physicochemical properties from the protein fold process, such as hydrogen bonding, contact potential energy, PDB-derived secondary structure propensities, and folding involving both bonded and non-bonded interactions. The *ab initio* method may or may not take into account motif identification in databases using small fragments. Some reviews even divide this category into two distinct methodologies, one dependent and another independent on database information. In this mini-review, we will use the term *ab initio* for methods that model protein structure without a template and that use the laws of thermodynamics as a basis, even though FM methods commonly exploit the information from known structures.

The main advantage of *ab initio* methods is the capacity to obtain novel and unknown protein folds (Dorn et al., 2014). Nevertheless, the complexity of the problem and the high number of conformational possibilities is computationally demanding, limiting the use for long protein sequences. Methods that use fragments may help to reduce the conformational space, but again avoid the prediction of new protein folds. One example of *ab initio* method that considers fragments is QUARK, a template-free novel program developed by Xu and Zhang (2012) in 2012. Briefly, this program breaks the protein sequence into fragments of 20 amino acids, searches for structure in a database, and then, using replica-exchange Monte Carlo simulations, unites the fragments taking into account the force field and energy terms, constituting a complete model.

The threading/fold recognition methodologies are based on the idea that structure is more conserved than the amino acid sequence and has a limited number of protein structure folds in nature (Rost et al., 1997; Dorn et al., 2014; Gromiha et al., 2018; Agnihotry et al., 2022). They consist in choosing the best 3D template of known foldings that fit well in the target sequence considering a scoring function built on pairwise potential, second structure comparison, as well as solvent properties. Thus, the target sequence is aligned with the structure model with the optimal scoring function, reorganizing the atoms of the target sequence in the aligned backbone. Finally, the affinity of the sequence with three-dimensional fold is verified followed by a manual verification. GenTHREADER is a program that uses threading techniques to evaluate the alignments, made using a sequence profile method, and then generates models that will be evaluated by a neural network to give a confidence measure (Jones, 1999).

The homology models derive from the fact that two amino acid sequences that are highly similar have similar structures (Dorn et al., 2014; Gromiha et al., 2018; Agnihotry et al., 2022; Sanjeevi et al., 2022). For this, the target sequence is aligned to a sequence in which the structure is known and an atomic model for the target protein is generated taking into account its similarities with the template backbone, followed by modeling loop regions and sidechains. The final step is to submit the model to energy minimization and evaluate it using the Ramachandran plot. Usually, homology methods achieve protein structures with higher accuracy than other methodologies. But, as other TBM methods, they are limited due to their inability to predict structures for new proteins, as they are dependent on templates. SWISS-MODEL is an automated system that uses homology modeling to predict a three dimensional structure of a protein (Guex et al., 2009; Kiefer et al., 2009). It was the first web server available to generate a 3D protein structure. This program integrates and automates all the processes involved in a homology modeling method, creating a fully automated workflow using a PERL based framework (Kiefer et al., 2009). The program presents an interface friendly for non-bioinformatician users and information, such as PFAM domain annotation and other tools from SWISS-MODEL (Kiefer et al., 2009).

Recently, new hybrid techniques have been published combining tools or improving known methods with artificial intelligence approaches. This was possible in part due to the

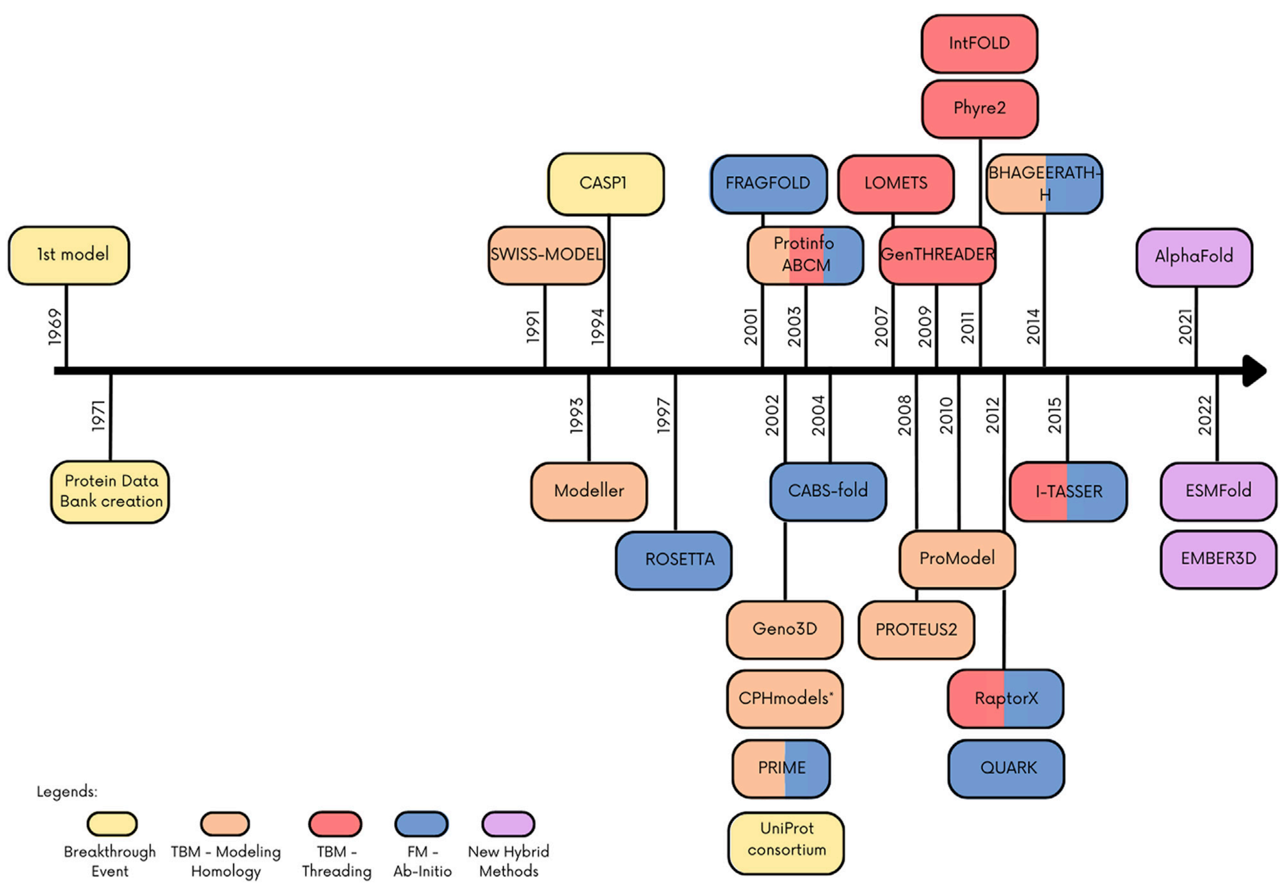

FIGURE 1
Timeline with main events and programs/webserver in the protein structure prediction. Colored boxes indicate the method or important event in the field.

development and improvement of computer processing. One of the breakthrough methods that not only combines methodology but also uses artificial intelligence is AlphaFold, an algorithm that beats the other tools in CASP13 and currently represents the state of art in protein structure prediction. Figure 1 depicts a timeline of the emergence of protein structure prediction programs/web servers and their classification considering the groups of methods adopted in this review, as well as other important dates for this field. More comprehensive and detailed reviews about protein structure prediction methods/tools can be found in Paiva et al. (2022), Bongirwar and Mokhade (2022).

3 AlphaFold

In 2018, DeepMind, a startup of Google, presented a new software that best performed in the 13th edition of CASP, named AlphaFold. In this competition, AlphaFold achieved the best position in the FM (best-of-five), reaching a summed z-score of 52.8 *versus* 36.6 from the second place and, combining FM and FM/TBM categories, achieved 68.3 z-score *versus* 48.2 (Senior et al., 2019). Even without using a template, AlphaFold also scored well in the TBM category (Senior et al., 2019).

The first version of AlphaFold used deep learning to predict the protein structure, demonstrating that it is possible to learn protein specific potential by training a neural network giving only the protein sequence. It contains a convolutional neural network that is trained by PDB structures to predict the distances between residues, creating distograms. From the amino acid sequence of the target protein, the neural network predicts a distogram based on multiple sequence alignment (MSA) features, in which a separate output from prediction network predicts the probability of backbone torsion distribution. The combined potential obtained by both ends is then optimized by gradient descent, predicting the protein structure itself (Senior et al., 2020).

Presented at CASP14 between May and July 2020, AlphaFold2 predicted protein structures with more accuracy than other competing methods, demonstrating a root-mean-square deviation (RMSD) among prediction and experimental backbone structures of 0.8Å *versus* the 2.8Å from the next best performing method. Moreover, AlphaFold2 scored 244.0 in summed z-scores compared with 90.8 for the next closest group (Jumper et al., 2021a; Jumper et al., 2021b). The great performance of AlphaFold2 in all Casp14 categories is depicted in Figure 2.

It is important to emphasize that AlphaFold2 contrasts considerably from the first version of AlphaFold. The authors

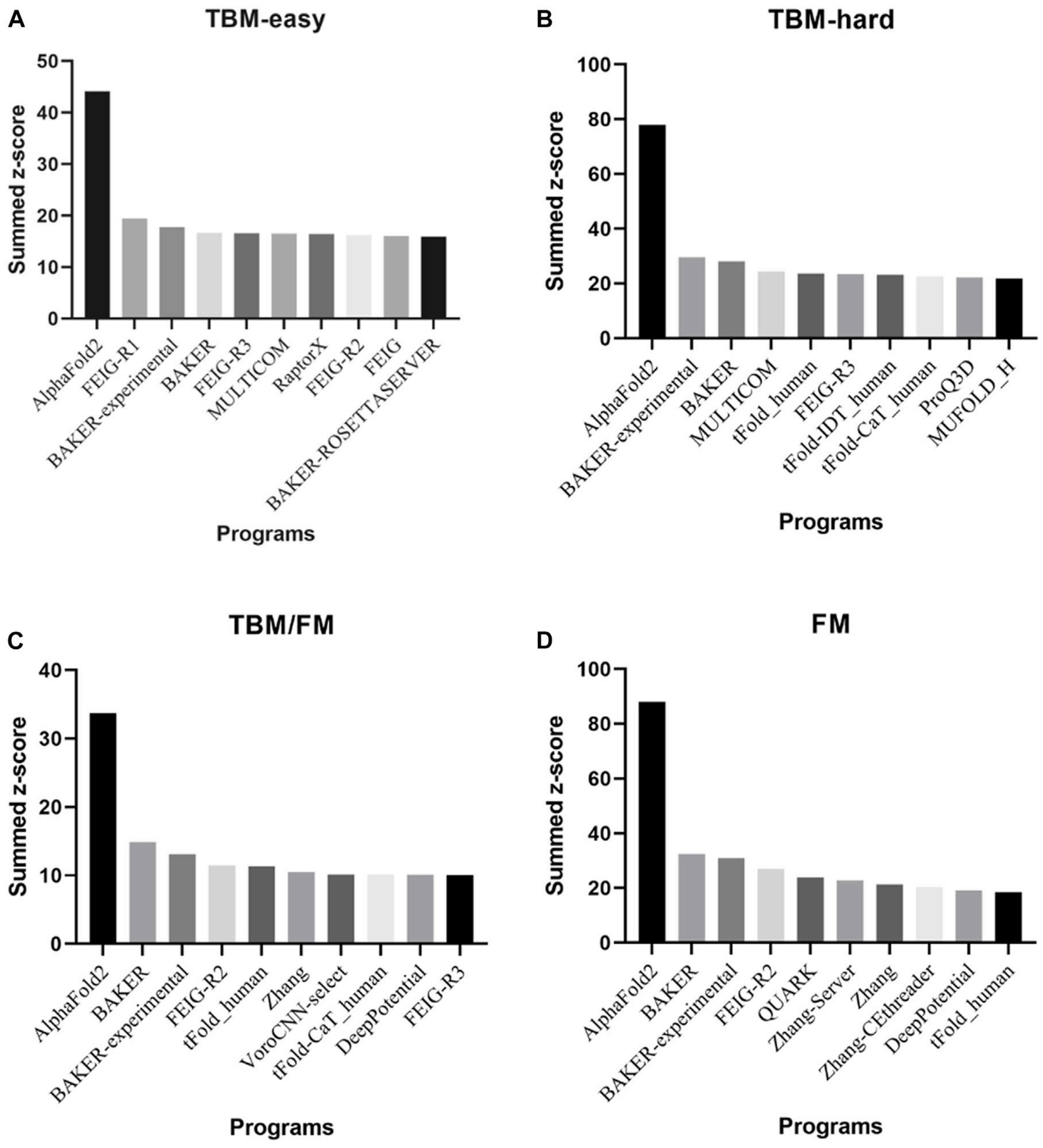

FIGURE 2
The top 10 programs and/or web servers in CASP14 in **(A)** TBM-easy, **(B)** TBM-hard, **(C)** TBM/FM, and **(D)** FM categories considering summed z-score. Data extracted by CASP official website.

attribute the high performance of AlphaFold2 by "incorporating novel neural networks architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures" (Jumper et al., 2021b; Bouatta et al., 2021; Callaway, 2022). AlphaFold2 uses as an input amino acid sequence to construct a MSA based on several databases of protein sequences to determine which parts of the sequence are mutation prone, detecting correlation between them. It also identifies proteins with similar structure with the input that will be used to build an initial representation of the target sequence (template), named as pair representation. Both strategies are not new and are shared by other algorithms in CASP14. Nevertheless, the breakthrough of AlphaFold2 is due to its neural network architectures, more specifically, the two neural network modules, evoformer and the structure module (Jumper et al., 2021b; Oxford Protein Informatics Group, 2021; Skolnick et al., 2021).

The evoformer extracts information from MSA and templates, exchanging information between them in flows back and forth throughout the network, improving the assessment of the MSA, that in turns modifies the protein structures hypothesized by the templates, allowing the MSA and templates in the correct "embedding space". It consists of two transformers, networks that use attention to boost the speed with which a model can be trained, each of them specialized for a type of data, MSA or pair representations, with a clear communication channel between them. This allows the MSA transformer attention mechanism to incorporate information from the pair representation, adding a bias term from it, augmenting the attention mechanism and allowing it to pinpoint interacting pairs of residues. The pair representation transformer also works in a similar way, but includes an attention to terms of triangles of residues. The structure module, that also contains an attention architecture, uses both representations to prioritize the orientation of the protein backbone, considering the residue rotations and translations, localizing each side chain of each residue in highly constrained within a frame, followed by local refinement and minimization by gradient descent (Jumper et al., 2021b; Oxford Protein Informatics Group, 2021; Skolnick et al., 2021).

Protein-protein interactions are the basis of the biological process, and high-resolution structural characterization of these interactions give rise to insights of their molecular mechanisms and function, as well as direct the design of new drugs that are able to modulate these molecular pathways. Alphafold2 has been used to predict protein-protein interaction, using flexible linkers or artificial gaps and, in general, it predicted heterodimeric protein complexes accurately, exceeding docking approaches usually used in these analysis (Bryant et al., 2022; Yin et al., 2022). Nevertheless, it was limited to predict complexes, such as antigen-antibody and AlphaFold2 accuracy was also limited to predict complexes with protein from different species (Yin et al., 2022). In October 2021, DeepMind extended Alphafold2 to multiple chains - called AlphaFold-Multimer (Evans et al., 2022). For this, AlphaFold-Multimer was trained with protein complexes, and a series of changes in the code were made. The developers observed that performance was better in homomeric than heteromeric interfaces (Jumper et al., 2021b) and it did not predict binding of antigen to antibodies (Yin et al., 2022). Both, AlphaFold2 and AlphaFold-Multimer, are open-source codes and are available on github (https://github.com/deepmind/alphafold).

In the same year, in a partnership between DeepMind and the EMBL-European Bioinformatics Institute (EMBL-EBI), the AlphaFold Protein Structure Database (AlphaFold DB—Available in https://alphafold.ebi.ac.uk) was created, making available over 360,000 predicted structures from 21 organism proteomes (Varadi et al., 2022). Today, AlphaFold DB has over 200 million entries from the human and 47 other organism proteomes, with the structure predictions and their respective analyses freely available to the scientific community. Porta-Pardo et al. (2022) demonstrated that AlphaFold2 increases the structural coverage from 48% to 76% of all human protein residues, dropping the number of human protein without structural coverage from 5027 to 29. Moreover, they quantified that, among the 5027 of the proteins without structure previously, 4459 had structure prediction for over 50% of the protein's length (88,7%), in which 1408 with high-accuracy (28%). Despite the large amount of data, protein sequences containing non-standard amino acids, like selenocysteine, have been excluded, as well as multiple isoforms codifying by the same gene (Varadi et al., 2022). The database usability is easy, as input, protein name, gene, Uniprot accession number or organism can be used. As output, the AlphaFold DB provides the atomic coordinates in PDB and mmCIF formats and Predicted Aligned Error (PAEs) in JSON format. It is also possible to give feedback about the prediction structure through "Looks great" or "Could be improved" buttons.

Considered as the ground-breaking application of AI in science, AlphaFold has promised to revolutionize structure biology. Its application has been considered to design better protein expression experiment; To solve experimental structures faster, overcoming tedious model building, especially for X-ray crystallography, and to facilitate the interpretation of low-resolution cryo-EM; To protein design and drug development; To examine the effect of mutation in protein function, elucidating their potential impact on human diseases; To provide novel insights in poorly known molecular mechanisms (Perrakis and Sixma, 2021; Porta-Pardo et al., 2022). In this context, Noone et al. (2022) has demonstrated that indeed AlphaFold offers shortcuts to solve protein structures experimentally, predicting the remaining N-terminal region of PTX3 complex, which was, then, validated with cryo-EM class averaging. The hybrid cryoEM/AlphaFold structure allowed the mapping of the putative sites and regions of interaction, giving insights of the functions of PTX3.

However, despite its breakthrough accuracy and performance to predict protein structures, AlphaFold2 models have important limitations. First, AlphaFold2 has difficulty to predict intrinsically disordered proteins/regions (Ruff and Pappu, 2021) and loops (Stevens and He, 2022), especially considering the importance of the latter for drug screening and design, since they are exposed in protein surface and readily available to solvent and other proteins. Ruff and Pappu (2021) demonstrated that residues and regions predicted with low accuracy by AlphaFold2 overlaps intrinsically disordered regions, while Stevens and He (2022) showed that only short loops (<20 amino acids) are predicted with high accuracy by AlphaFold and it has the tendency to over-predict secondary structures in loop regions, usually alpha helix. Both regions are known to be hypervariable and flexible across orthologies, making it difficult to uncover evolutionary constraints from MSA.

Second, AlphaFold2 predicts only a single conformer, not identifying the apo and holo forms. In 67% of a dataset tested, AlphaFold prediction resembled holo form and the proteins were less predictable when the conformational differences between apo and holo forms increased (Saldaño et al., 2022). Moreover, Azzaz et al. (2022) demonstrated that structure prediction of membrane proteins by AlphaFold is not reliable, mainly because it presents inconsistencies in the location of the transmembrane domains. They stress that the protein environment influences the amino acid sequence, imposing folding constraints. These evidences together with the AlphaFold's inability to predict structures with metal ions, cofactors and other ligands, complexes with DNA or RNA, or post-translational modifications, such as glycosylation, methylation and phosphorylation (Perrakis and Sixma, 2021) highlight the steps to be overcome to improve AphaFold models in drug screening and design. Indeed, Scardino et al. (2023) demonstrated that AlphaFold models showed worse performance in high-throughput docking when compared to their corresponding experimental PDB structures, while Wong et al. (2022) showed that AlphaFold2 protein structure prediction exhibits weak performance on reverse docking in a search for binding targets of bacterial compounds, emphasizing that, even though AlphaFold2 provides rich structural information, more accurate models of protein-ligand interactions are needed to improve use of AlphaFold2 for drug discovery.

Third, AlphaFold fails to predict defects in protein folding due to point mutations. As demonstrated by Buel and Walters (2022), the differences between mutated and wild-type models predicted by AlphaFold are very small, represented by backbones RMSD lower than 1Å. Moreover, Pak et al. (2021) also demonstrated that there is no correlation between AlphaFold accuracy metrics (pLDDT) and the impact of mutations on protein stability change (ΔΔG), neither with the side chain size change.

Finally, of AlphaFold2 cannot predict novel structures, since its algorithm is based on MSA and requires known structures databases. Another important aspect is that the use of evolutionary information from larger MSAs, requiring environmental systems and storage to detect the homology between known and target sequences, demands a powerful computing processors and its structure prediction is time consuming as protein length increases. To overcome this limitation Google offers the Google Colaboratory, which enables access to powerful GPUs. One of these solutions is ColabFold, a fast and easy-to-use software that replaces the AlphaFold2's homology by MMseq2 (Mirdita et al., 2022), making the computational demands less relevant.

4 New methods of protein structure prediction using protein language model

Recently, new free modeling methods have been published to overcome some of the limitations of AlphaFold2, such as the inability to predict novel structures and the necessity of high time and computing processes. These methods are based on protein language models, derived from natural language processing (NLP), which uses the amino acids sequence only and is able to learn evolutionary, structural and functional patterns derived from sequences available in databases, predicting a structural conformation. The idea behind those methods is that the amino acids correspond to words/tokens and proteins to sentences in NLP, assuming that similar semantics come from amino acids that occur in similar contexts.

There are three approaches used in language models: autoregressive, bidirectional, and masked. The first takes into count the previous tokens (amino acids) to predict the probability of a token, the second considers the previous and following tokens independently to estimate the probability of a token, and the last model considers all tokens in a sequence and replaces each token with a mask token. A synthesis of the recent advances in protein language modeling and their applications to protein prediction problems can be found elsewhere (Bepler and Berger, 2021).

Two methods have gained attention this year. ESMfold, developed by Lin et al. (2022) uses a masked transformer protein language model trained in deep information about biological properties, using 15 billion parameters. Compared with AlphaFold2, it did not present the same performance, achieving lower TM-scores (0.68 *versus* 0.85 using AlphaFold2 on CASP14). However, when evaluating AlphaFold2 without MSA, using only the amino acid sequence, ESMFold performed better (0.68 *versus* 0.37 using AlphaFold2 on Casp14). Moreover, it presented an accuracy comparable to AlphaFold2 for structures predicted with high confidence, demonstrating a median all-atom RMSD of 1.91Å and a backbone RMSD of 1.33Å, reaching similar experimental-level accuracy. Finally, this approach also demonstrated a significant improvement in prediction speed, since it does not require the construction of MSA. Using this approach, the authors presented the ESM Metagenomic Atlas, where they predicted more than 617 million structures from metagenomic databases, in which 225 million structures were predicted with high confidence, including those that are novel (Lin et al., 2022).

EMBER2 is a protein language-model developed by Weissenow et al. (2022a), that uses embeddings to predict inter-residue distance (2D structure) introducing attention heads derived from a pre-trained protein language model instead of MSA. Using EMBER2 in *trRosetta* to predict 3D structures is less accurate than AlphaFold2 in predicting a native structure and presents an inferior TM score (0.5 *versus* 0.79 in ColabFold) (Yang et al., 2020). However, it is faster than ColabFold by about 35 fold, similarly to ESMfold. As the comparison made in this work was not fair, since EMBER2 predicts 2D structures and AlphaFold, 3D structures, the authors developed a new approach that uses EMBER2 model, but now applied in three-dimensional structure prediction, named EMBER3D (Weissenow et al., 2022b). Again, Ember3D did not outperform AlphaFold, but is much faster than it and ESMFold. Whereas AlphaFold2 do not perform efficiently in the study of the impact of single amino acids variants into protein structure, Weissenow et al. (2022b) demonstrated that the differences in predicted distances maps generated by EMBER3D correlated well with native and mutant 3D structures from deep mutational scanning, having a better result than ESMfold. Furthermore, they developed a tool that presents the difference between native and mutant predicted structures by all possible amino acid exchanges in each position of a protein sequence. The similarity between de native amino acid and the mutated one helps the identification of exchanges that may cause a high impact on the protein structure.

These new approaches highlight the powerful capacity of language models to identify evolutionary, structural and functional patterns from massive protein sequence databases to solve protein prediction problem, improving prediction speed and requiring less computational power. It is expected that those approaches will develop and gain accuracy with the inclusion of biological knowledge and multi task learning.

5 Conclusion

Three-dimensional protein structure determination is important to elucidate the protein function, being critical to understanding biological processes and addressing human health and life science problems in general. Due to the difficulty of determining protein structures by experimental methods, their predictions have been one of the central problems of the scientific community. The advent of AlphaFold2 and the release of millions of protein structures predicted with high accuracy and available in AlphaFold DB allowed an unprecedented expansion of different research fields in life science, impacting the most biological sciences, followed by biochemistry and cell biology, genetics, medical and health sciences and chemical sciences (Varadi and Velankar, 2022).

Protein structure prediction can be applied from understanding the interaction between pathogen and host, how pathogens survive and reproduce, and why they are resistant to certain drugs used, to the development of new and more efficient drugs, reducing the cost in drug discovery and development, as well as the development of new and improved vaccines (Duran-Frigola et al., 2013; Hazra and Patra, 2021). The development of a new vaccine strategy used against COVID-19 by Pfizer, Moderna and Johnson & Johnson, in which the prediction of the mutations and its potential effects in spike protein domain allowed the generation of an immunogen, highlights the window of opportunity that AlphaFold2 offers in the structure-guided vaccine design (Higgins, 2021).

The recent advances in protein structure prediction have contributed to improve protein folding issues. The emergence of AlphaFold2 has taken the problem of protein structure prediction to another level, reaching similar experimental-level accuracy in some cases. Nevertheless, improvements are needed to overcome the limitation to predict novel structures, intrinsically disordered regions and loops, the ability to predict only a single conformer without ligands and still present inconsistencies in its models, and the inability to predict the impact of missense mutation on protein structure. These limitations, ultimately, are important drawbacks to expand the use of AlphaFold in life science.

Efforts are in progress to improve AlphaFold performance and models. Johansson-Åkhe and Wallner (2022) demonstrated that randomly perturbing the neural network weights, forcing it to sample more conformational spaces can improve AlphaFold Multimer performance. Terwilliger et al. (2022) suggested that inclusion of new experimental information can improve parts of the models, showing that application of experimental density maps used iteratively allows the rebuilding of models that can be used as templates by AlphaFold new prediction. Finally, Hekkelman et al. (2022) enriched the models in the AlphaFold DB through transplantation of small molecules and ions based on homologous protein structures. They presented a new resource, the AlphaFill databank, to overcome the limitation presented by AlphaFold models that do not present ligands and co-factors, to help life scientists test new hypotheses and design target experiments.

Simultaneously, new strategies using protein language models are arising to compete with AlphaFold2 in terms of performance and accuracy and to overcome some of its limitations. These strategies, more specific ESMfold, offers an opportunity to identify new proteins and novel functions, allowing the identification of new species, including microorganisms and viruses that endanger human health, as well as those that offer solution to mitigate environmental problems, such as the degradation of polluting and the development of transgenic microorganisms for more efficient product production. For this, ESM Metagenomic Atlas is an important 3D protein structures resource to be deeply investigated by the scientific community (Lin et al., 2022). Finally, the efforts to better predict how mutations affect protein structure using these new approaches are essential to gain insights in human genetic disease and further improve disease management and prediction. Today, these methods do not outperform AlphaFold2, but with the improvement of deep language models and the enrichment of these models with biological information through multi-task learning, they promise to revolutionize the protein structural biology.

Author contributions

LB: Conception and design, literature revision, drafting the manuscript and revising it critically for important intellectual content. AL: Conception and design, literature revision, drafting the manuscript and revising it critically for important intellectual content. JK: Revising it critically for important intellectual content, and final approval of the version to be published ST: Conception and design, literature revision, drafting the manuscript and revising it critically for important intellectual content, and final approval of the version to be published.

Funding

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [INCT—2014/50889-7 to JK], the Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq [INCT—465586/2014-7 to JK], the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES [PROEX—169192/2018-0 to LB] and the Zerbini Foundation and Foxconn Brazil as part of a research grant "Machine Learning in Cardiovascular Medicine".

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agnihotry, S., Pathak, R. K., Singh, D. B., Tiwari, A., and Hussain, I. (2022). "Protein structure prediction," in *Bioinformatics* (Amsterdam, Netherlands: Elsevier), 177–188. doi:10.1016/B978-0-323-89775-4.00023-7

Azzaz, F., Yahi, N., Chahinian, H., and Fantini, J. (2022). The epigenetic dimension of protein structure is an intrinsic weakness of the AlphaFold program. *Biomolecules* 12, 1527. doi:10.3390/biom12101527

Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell. Syst.* 12, 654–669.e3. doi:10.1016/j.cels.2021.05.017

Bongirwar, V., and Mokhade, A. S. (2022). Different methods, techniques and their limitations in protein structure prediction: A review. *Prog. Biophysics Mol. Biol.* 173, 72–82. doi:10.1016/j.pbiomolbio.2022.05.002

Bouatta, N., Sorger, P., and AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallogr. Sect. D. Struct. Biol.* 77, 982–991. doi:10.1107/S2059798321007531

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* 13, 1265. doi:10.1038/s41467-022-28865-w

Buel, G., and Walters, K. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* 29, 1–2. doi:10.1038/s41594-021-00714-2

Callaway, E. (2022). What's next for AlphaFold and the AI protein-folding revolution. *Nature* 604, 234–238. doi:10.1038/d41586-022-00997-5

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A. (2007). The protein folding problem: When will it be solved? *Curr. Opin. Struct. Biol.* 17, 342–346. doi:10.1016/j.sbi.2007.06.001

Dorn, M., E Silva, M. B., Buriol, L. S., and Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Comput. Biol. Chem.* 53, 251–276. doi:10.1016/j.compbiolchem.2014.10.001

Duran-Frigola, M., Mosca, R., and Aloy, P. (2013). Structural systems pharmacology: The role of 3D structures in next-generation drug development. *Chem. Biol.* 20, 674–684. doi:10.1016/j.chembiol.2013.03.004

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2.

Fruton, J. S. (1985). Contrasts in scientific style. Emil fischer and franz hofmeister: Their research groups and their theory of protein structure. *Proc. Am. Philos. Soc.* 129, 313–370.

Gromiha, M. M., Nagarajan, R., and Selvaraj, S. (2018). Protein structural bioinformatics: An overview. *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.* 1–3, 445–459. doi:10.1016/B978-0-12-809633-8.20278-1

Guex, N., Peitsch, M. C., and Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-model and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 30, S162–S173. doi:10.1002/elps.200900140

Hardin, C., Pogorelov, T. V., and Luthey-Schulten, Z. (2002). *Ab initio* protein structure prediction. *Curr. Opin. Struct. Biol.* 12, 176–181. doi:10.1016/S0959-440X(02)00306-8

Hazra, S., and Patra, S. (2021). *Importance of protein structure and function in pathogenesis: Highlights on the multifaceted organism* Mycobacterium tuberculosis. Amsterdam, Netherlands: Elsevier. doi:10.1016/b978-0-12-820084-1.00030-2

Hekkelman, M., de Vries, I., Joosten, R., and Perrakis, A. (2022). AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat. Methods* 20, 205–213. doi:10.1038/s41592-022-01685-y

Higgins, M. K. (2021). Can we AlphaFold our way out of the next pandemic? *J. Mol. Biol.* 433, 167093. doi:10.1016/j.jmb.2021.167093

Johansson-Åkhe, I., and Wallner, B. (2022). Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Front. Bioinforma.* 2, 959160. doi:10.3389/fbinf.2022.959160

Jones, D. T. (1999). GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 797–815. doi:10.1006/jmbi.1999.2583

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021a). Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinforma.* 89, 1711–1721. doi:10.1002/prot.26257

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021b). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* 37, 387–392. doi:10.1093/nar/gkn750

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—round XIII. *Proteins Struct. Funct. Bioinforma.* 87, 1011–1020. doi:10.1002/prot.25823

Lehninger, A., Nelson, D., and Cox, M. (2005). *Principles of biochemistry*. 4th editio. New York.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2022.07.20.500902v2.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi:10.1038/s41592-022-01488-1

Nassar, R., Dignon, G. L., Razban, R. M., and Dill, K. A. (2021). The protein folding problem: The role of theory. *J. Mol. Biol.* 433, 167126. doi:10.1016/j.jmb.2021.167126

Noone, D. P., Dijkstra, D. J., van der Klugt, T. T., van Veelen, P. A., de Ru, A. H., Hensbergen, P. J., et al. (2022). PTX3 structure determination using a hybrid cryoelectron microscopy and AlphaFold approach offers insights into ligand binding and complement activation. *PNAS* 33, e2208144119. doi:10.1073/pnas.2208144119

Oxford Protein Informatics Group (2021). *AlphaFold 2 is here: what's behind the structure prediction miracle*. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction- miracle/.

Paiva, V. A., Gomes, I. S., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., et al. (2022). Protein structural bioinformatics: An overview. *Comput. Biol. Med.* 147, 105695. doi:10.1016/j.compbiomed.2022.105695

Pak, M., Markhieva, K., Novikova, M., Petrov, D., Vorobyev, I., Maksimova, E., et al. (2021). *Using AlphaFold to predict the impact of single mutations on protein stability and function*. bioRxiv. doi:10.1101/2021.09.19.460937v1

Parto-Pardo, E., Ruiz-Serra, V., Valentini, S., and Valencia, A. (2022). The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput. Biol.* 18, 1–17. doi:10.1371/journal.pcbi.1009818

Pearce, R., and Zhang, Y. (2021). Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* 297 (1), 100870. doi:10.1016/j.jbc.2021.100870

Perrakis, A., and Sixma, T. K. (2021). AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Rep.* 22, e54046–6. doi:10.15252/embr.202154046

Rost, B., Schneider, R., and Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* 270, 471–480. doi:10.1006/jmbi.1997.1101

Ruff, K. M., and Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* 433, 167208. doi:10.1016/j.jmb.2021.167208

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., et al. (2022). Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* 38, 2742–2748. doi:10.1093/bioinformatics/btac202

Sanjeevi, M., Hebbar, P. N., Aiswarya, N., Rashmi, S., Rahul, C. N., Mohan, A., et al. (2022). *Methods and applications of machine learning in structure-based drug discovery*. Amsterdam, Netherlands: Elsevier. doi:10.1016/B978-0-323-90264-9.00025-8

Scardino, V., Di Filippo, J. I., and Cavasotto, C. N. (2023). How good are AlphaFold models for docking-based virtual screening? *iScience* 26, 105920. doi:10.1016/j.isci.2022.105920

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinforma.* 87, 1141–1148. doi:10.1002/prot.25834

Skolnick, J., Gao, M., Zhou, H., and Singh, S. (2021). AlphaFold 2: Why it works and its implications for understanding the relationships of protein sequence, structure, and function. *J. Chem. Inf. Model.* 61, 4827–4831. doi:10.1021/acs.jcim.1c01114

Stevens, A. O., and He, Y. (2022). Benchmarking the accuracy of AlphaFold 2 in loop structure prediction. *Biomolecules* 12, 985. doi:10.3390/biom12070985

Terwilliger, T., Poon, B., Afonine, P., Schlicksup, C., Croll, T., Millan, C., et al. (2022). Improved AlphaFold modeling with implicit experimental information. *Nat. Methods* 19, 1376–1382. doi:10.1038/s41592-022-01645-6

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50, D439–D444. doi:10.1093/nar/gkab1061

Varadi, M., and Velankar, S. (2022). The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*, e2200128. doi:10.1002/pmic.202200128

Voet, D., Voet, J. G., and Charlotte, W. P. (2014). *Fundamental of biochemistry: Life at the molecular level*. 4th editio. Porto Alegre: wiley.

Weissenow, K., Heinzinger, M., and Rost, B. (2022a). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* 30, 1169–1177.e4. doi:10.1016/j.str.2022.05.001

Weissenow, K., Heinzinger, M., Steinegger, M., and Rost, B. (2022b). Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. 1–16. doi:10.1101/2022.11.14.516473

Wisniak, J. (2000). Jons Jacob Berzelius A guide to the perplexed chemist. *Chem. Educ.* 5, 343–350. doi:10.1007/s00897000430a

Wong, F., Krishnan, A., Zheng, E., Stark, H., Manson, A., Earl, A. M., et al. (2022). Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* 18, e11081. doi:10.15252/msb.202211081

Xu, D., and Zhang, Y. (2012). *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinforma.* 80, 1715–1735. doi:10.1002/prot.24065

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1496–1503. doi:10.1073/pnas.1914677117

Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* 31, e4379. doi:10.1002/pro.4379

Yuan, X., Shao, Y., and Bystroff, C. (2003). *Ab initio* protein structure prediction using pathway models. *Comp. Funct. Genomics* 4, 397–401. doi:10.1002/cfg.305