


## REVIEW

# An outlook on structural biology after AlphaFold: tools, limits and perspectives

 Serena Rosignoli, Maddalena Pacelli, Francesca Manganiello and Alessandro Paiardini 

Department of Biochemical sciences "A. Rossi Fanelli", Sapienza Università di Roma, Italy

## Keywords

AlphaFold; machine learning; structural bioinformatics; structure prediction

## Correspondence

A. Paiardini and M. Pacelli, Department of Biochemical sciences "A. Rossi Fanelli", Sapienza Università di Roma, Rome 00185, Italy

 E-mail: [alessandro.paiardini@uniroma1.it](mailto:alessandro.paiardini@uniroma1.it); [maddalena.pacelli@uniroma1.it](mailto:maddalena.pacelli@uniroma1.it)

Serena Rosignoli and Maddalena Pacelli contributed equally to this article.

(Received 13 March 2024, revised 19 August 2024, accepted 13 September 2024)

doi:10.1002/2211-5463.13902

Edited by Claudio Soares

AlphaFold and similar groundbreaking, AI-based tools, have revolutionized the field of structural bioinformatics, with their remarkable accuracy in *ab-initio* protein structure prediction. This success has catalyzed the development of new software and pipelines aimed at incorporating AlphaFold's predictions, often focusing on addressing the algorithm's remaining challenges. Here, we present the current landscape of structural bioinformatics shaped by AlphaFold, and discuss how the field is dynamically responding to this revolution, with new software, methods, and pipelines. While the excitement around AI-based tools led to their widespread application, it is essential to acknowledge that their practical success hinges on their integration into established protocols within structural bioinformatics, often neglected in the context of AI-driven advancements. Indeed, user-driven intervention is still as pivotal in the structure prediction process as in complementing state-of-the-art algorithms with functional and biological knowledge.

The advent of AlphaFold2 (AF) and its landmark performance at the 14th edition of the Critical Assessment of Protein Structure Prediction (CASP) marked a substantial shift in biomedical research, with the newfound ability to easily access millions of 3D-structures of proteins, for which only their sequence was previously known [1,2]. Within a year from AF's debut, a collaborative effort with EMBL-EBI led to the creation of the UniProt-indexed AF database (AFDB) [3,4]. By releasing more than 200 million AF-predicted structures, AFDB significantly enhanced the accessibility to this groundbreaking tool for the global research community. Overall, these advancements have significantly enhanced

the structural coverage of the human proteome. Initially limited to just 10% when relying solely on experimental structures, this coverage has now expanded to 58% with the incorporation of high-accuracy AF models (predicted Local Distance Difference Test, pLDDT, scores above 70) [5–7]. However, this "Big Bang" of the protein structures universe, ignited by AF, also prompted critical examination regarding the accessibility, accuracy, reliability, and potential biases inherent in the data produced. Therefore, to avoid the potential risk of relying too much on readily available 3D-structures without critical thinking and judgment, it is important to ask: "How do people from various

## Abbreviations

AF, AlphaFold2; AF3, AlphaFold 3; AFDB, AF database; AI, artificial intelligence; CASP, critical assessment of protein structure prediction; EDAM, ontology of bioscientific data analysis and data management; ESM, evolutionary scale modeling; GDTTS, global distance test total score; MD, molecular dynamics; MSA, multiple sequence alignment; PDB, Protein Data Bank; pLDDT, predicted local distance difference test; pMHC, peptide-major histocompatibility complex; SCOP, structural classification of proteins; TCR, T-cell receptor; TM, transmembrane proteins; UX/UI, user experience/user interface; XAI, explainable artificial intelligence.

academic backgrounds or research settings make use of this wealth of new structural information?”

### CASP—CAlyzing a shift in the paradigm

From its foundation in 1994, to the groundbreaking achievements at CASP14, the evolution of CASP mirrors the progress in Computational Biology and the growing intersection with Artificial Intelligence (AI) (Fig. 1) [9,10]. In the first decade of the CASP competitions, the division into categories—Comparative Modeling, Fold Recognition, and *Ab-Initio* Prediction—accompanied the birth of the first template-based modeling algorithms [11–13], such as MODELLER [14], and algorithms for *ab-Initio* folding, primarily governed by fragment-based algorithms, notably Rosetta [15,16] and I-Tasser [17,18].

The concept of coevolution, which is utilized in the AF algorithm, has been proposed since the early 1990s [19]. It suggests that detecting co-evolutionary signals within multiple sequence alignments (MSAs) could indicate potential physical interactions between molecules. Clearly, even before the advancements of CASP14, leading-edge methods already utilized the analysis of co-evolutionary patterns through MSAs. Over the years, these methods have been refined to effectively distinguish between direct correlations, signifying actual physical contact, and indirect correlations. The latter may not denote direct interaction but still reveal a relationship, possibly due to other influencing factors [20,21]. However, at that time, identifying accurate signals required extensive MSAs and computational resources, which were not readily available, leading to a diminished focus on this methodology.

The decade from 2010 to 2020 witnessed a significant shift as traditional bioinformatics began to integrate more effectively with advanced AI techniques. This period saw the resurgence of interest in the concept of co-evolutionary signals [22–24], which were then combined with deep learning models to better predict protein contact patterns [25,26]. This approach was based on the understanding that protein contacts

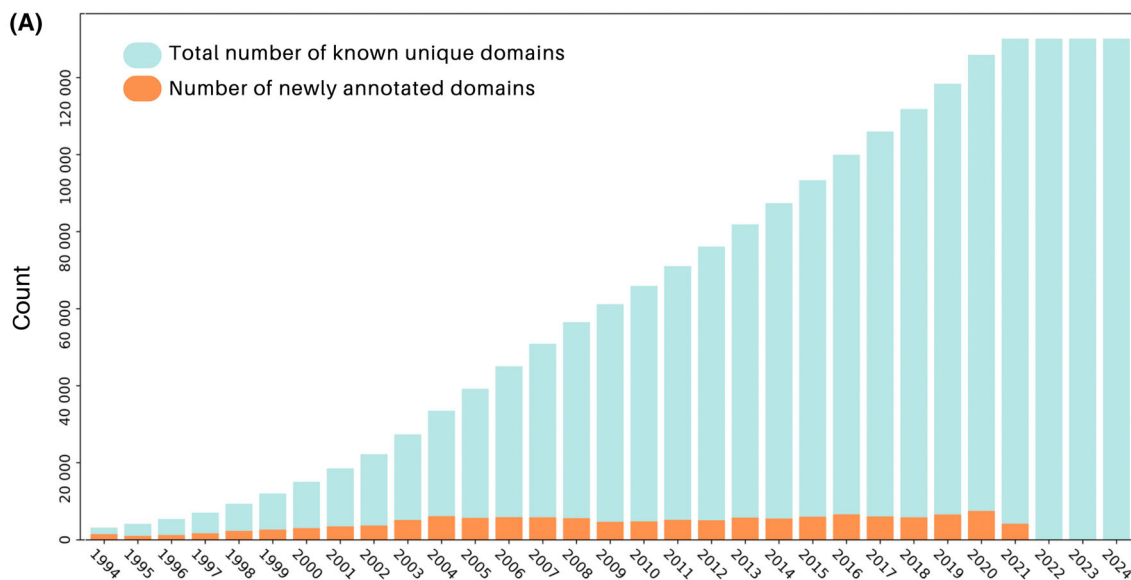
are not randomly distributed, but they have a biological sense, given by domains and structural motifs of the category of proteins. Subsequently, in CASP 13, the use of the contact predictions concept was replaced in the first version of AlphaFold by neural-network based distance probabilities [27], employing the steepest descent method as an optimization algorithm—a strategy that, although not yet fully successful, will later be crucial for enhancing accuracy. However, it was the revised neural network model proposed in CASP14 [1,2], utilizing MSA transformers to extract co-evolutionary signals directly from raw MSAs [28], which proved to be pivotal for enhancing accuracy.

After the milestone of CASP14, predicting single protein domains with accuracy has become significantly less challenging. Interestingly, this milestone coincides with an important accomplishment: the convergence of the domain structural knowledge (Fig. 1A). Throughout its history, CASP has dynamically evolved with a balance between retiring and modulating specific categories, while maintaining the core principles underlying the assessment process of the CASP competition largely unchanged (Fig. 1B). This evolution persisted up to CASP14, which saw significant changes, including the discontinuation of categories such as contact prediction and refinement for single protein models, alongside the introduction of novel challenges toward universal modeling [29,30].

### Building on the outcomes of AlphaFold

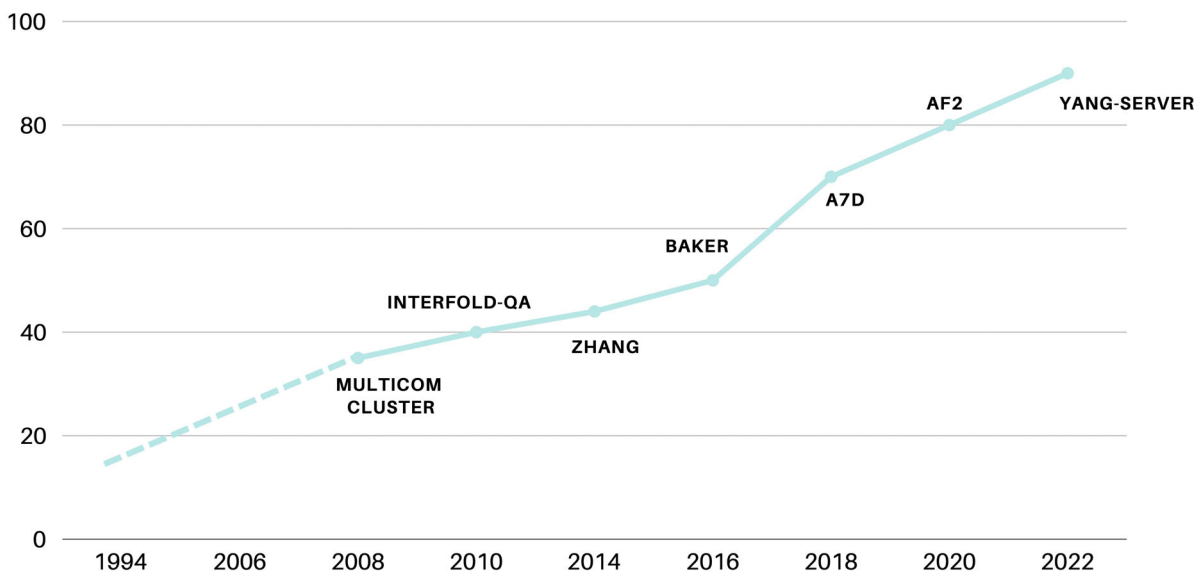
Immediately after AF's performance at CASP14, and further propelled by the open-source availability of the AF code, researchers in the field became aware of the vast potential inherent to the huge amount of available structural information and moved accordingly to ensure an optimal integration of AF into their specific areas of research. In this sense, a dual approach has been undertaken by the scientific community, mainly with the aim of addressing some of the limitations of AF, for example, the lack of physico-chemical

**Fig. 1.** CASP: CAlyzing a Shift in the Paradigm. In analyzing the events in the field of protein structure prediction, a shifting trend emerges in recent years according to various metrics. (A) The data from domain annotation in SCOP (Structural Classification of Proteins), as reported by the Protein Data Bank Statistics (RCSB PDB – Growth in Domain SCOP [8]), is plotted to show the number of known domains available each year from 1994 to 2024 (light blue). The orange highlights indicate the newly annotated domains each year. This visualization reveals, for the first time in 2022, an absence of newly annotated domains. This suggests that as experimental information on domain structures has increased, methods for accurately predicting structural domains have become more prominent. The table (B) reports on the prediction categories independently assessed at the Critical Assessment of protein Structure Prediction (CASP) over the years, mirroring the progress in the field and highlighting the most significant conceptual changes observed at CASP15. The plot illustrates the progression of the median GDTS (Global Distance Test Total Score) values across different CASP editions from 2008 till now.



**(B)**

	1994–2004	2006–2010	2012–2016	2018	2020	2022
Prediction categories	<ul style="list-style-type: none"> <li>Comparative Modeling</li> <li>Fold Recognition (Threading)</li> <li><i>Ab initio</i> (New Fold)</li> </ul>	<ul style="list-style-type: none"> <li>Template-based modeling</li> <li>Refinement and physics-based modeling</li> <li>Template-free modeling</li> <li>Prediction of function, domains and contacts</li> </ul>	<ul style="list-style-type: none"> <li>Template-based modeling</li> <li>Refinement</li> <li>Template-free modeling</li> <li>Contact-assisted prediction</li> </ul>	<ul style="list-style-type: none"> <li>Assembly</li> <li>Template-based modeling; Refinement</li> <li>Template-free modeling</li> <li>Contacts; Xlink, SAXS, NMR-assisted</li> <li>Accuracy assessment</li> </ul>	<ul style="list-style-type: none"> <li>Assembly</li> <li>High accuracy modeling</li> <li>Template-free modeling</li> <li>Contacts</li> <li>Refinement</li> <li>Accuracy assessment</li> <li>Biological Relevance</li> </ul>	<ul style="list-style-type: none"> <li>Assembly</li> <li>Single protein /domain</li> <li>Accuracy assessment</li> <li>Ligands</li> <li>RNA</li> </ul>
CASP	1–6	7–9	10–12	13	14	15



interpretation of proteins and their folding process. On the one hand, some efforts led to the development of platforms and tools that not only incorporate the algorithm's capabilities but also ensure they complement and enhance traditional physics-based protocols (Table 1). On the other hand, a different set of initiatives has taken a more exploratory approach, capitalizing on the novel capabilities and insights provided by AF, and using them as a catalyst for expanding the application of AI to the biochemical and biological fields.

### AlphaFold integration in new databases

The establishment of a database derived from AF, that is AFDB, has significantly enhanced the efficiency of

information sharing, emerging as an essential component for methodologies that depend on structural databases [51]. The AFDB contains predictions for all protein sequences annotated in the UniProt reference proteome, of length between 6 and 2700 amino acids, with the maximum limit decreasing to 1280 for noncurated sequences. A minimum length is required for having an informative MSA, while the maximum limits are set because of computational capacity. The integration of AFDB into pre-existing databases represents a natural progression in response to AFDB's utility and relevance in the field. Indeed, key protein family databases such as InterPro and Pfam, as many others in the field [52–55], have introduced dedicated pages for visualizing AF models. Additionally, the 2023 version of Ensembl [56] has incorporated AF models' visualization to map

**Table 1.** Databases development and integrations after AlphaFold. The table organizes into categories the development of new databases and the integration of existing ones, following the release of the AlphaFold Database (AFDB).

Category	Tool	Description	Ref.
Development of new databases	AlphaFold Database	Initially comprising 360 000 predicted structures, the AFDB has expanded to over 214 million structures, providing a comprehensive resource for structural biologists	[3,4]
	ESM Metagenomic Atlas	Offers structural predictions for 600 million metagenomic sequences, complementing the AFDB by extending coverage to environmental and microbiome samples	[31]
	AlphaFill	Enhances AF predictions by adding ligands from similar PDB structures, providing functional context to the predicted models	[32]
	TmAlphaFold	Incorporate predicted membrane planes into AF models, aiding in the study of membrane proteins	[33]
	AFTM	Leverages AF models to identify candidate human TMPs	[34]
Integration into software packages and existing data-resources	CCP4 Suite	In crystallography, automatically fetch predicted structures from the AFDB to solve crystal structures by molecular replacement without user intervention	[35]
	MrBUMP and MrPARSE	Can automatically fetch AF predictions, integrating them seamlessly into existing crystallographic analysis pipelines	[36,37]
	ISOLDE	It leverages AF predictions to refine models based on experimental data, such as cryo-EM or X-ray crystallography density maps	[38]
	CCP-EM	Imports structures directly from the AFDB for electron microscopy applications	[39]
	ChimeraX	Uses ColabFold for modeling, retrieves structures from the AFDB, and provides interactive visualization of predicted aligned error (PAE) plots	[40]
	COOT	Imports AF models for detailed molecular modeling and refinement	[41]
	DALI Server and Foldseek Search Server	Perform structure-based searches over the AFDB, enabling researchers to find structurally similar proteins	[42,43]
	Jalview and Mol* Viewer	These tools import AF structures for sequence alignment and interactive 3D visualization, respectively	[44,45]
	PHENIX	Integrates AF into molecular replacement pipelines, facilitating the incorporation of predicted structures into crystallographic workflows	[38,46,47]
	DeepTracer-ID	Combines DeepTracer and AF to identify proteins in cryo-EM maps by searching the AF library and iteratively refining the atomic model	[48]
DeepProLigand	Uses DeepTracer and AF to predict protein-ligand interactions by leveraging known structures available in the AlphaFold library or the RCSB PDB	[49]	
EMBUILD	Integrates U-Net and AF to construct main chain maps and fit AlphaFold2 predicted chains into the maps for cryo-EM applications	[50]	

variant predictor effects onto the structure. Platforms such as PDB [8] and UniProt [57] are revolutionizing the management of predicted protein models within widely acknowledged databases, now incorporating AF predictions alongside experimental ones. The escalating volume of structures managed by these platforms has intensified the demand for scalability, a challenge efficiently met by renowned conformational search protocols, FoldSeek [42] and DALI [43], as well as newly developed ones [58]. Finally, new databases have been developed conveying newly available information to specific topics, for example, transmembrane proteins [33,34]. It is important to note that this list is not exhaustive, and the landscape of AF integration is constantly evolving. Researchers are continuously finding new ways to incorporate AF into existing databases and cater to specific needs. These may include specialized databases for protein families, disease-related proteins, or other specific fields. For instance, specialized protocols have emerged for analyzing kinases [59] and intrinsically disordered proteins [60]. The availability of such comprehensive data has significantly advanced high-throughput and -omics research and has facilitated benchmarks of existing protocols, providing insights into their performance with predicted models [61,62].

### AlphaFold for experimental structure determination

The advent of AF, with its remarkable accuracy in predicting protein structures, has ushered in a transformative era for experimental structural biology, opening new avenues for investigating complex biological systems. One of the most significant contributions of AF lies in its application to Molecular Replacement (MR) in X-ray crystallography. Traditionally, MR has relied on experimentally determined structures from the PDB, posing limitations when suitable homologs were not available. However, AF has revolutionized MR by offering high-quality predicted protein structures as alternative search models [63,64]. This breakthrough has significantly expanded the scope of MR, making it applicable to a broader range of proteins, including those with no known homologs in the PDB [65]. The integration of AF predictions into established software [36–38,46,47] underscores its rapid adoption and widespread impact across various macromolecular structure determination methodologies. In PHENIX, AF models can be utilized through a dedicated set of functions [38,46], including a PHENIX-AF webservice to run predictions remotely from the GUI [47], their import from ColabFold [66], their trimming and splitting into single domains, and finally their positioning in unit cells. The resulting models

can be examined with PHENIX validation tools to identify and manually fix any problematic areas. Similarly, CCP4 provides seamless integration of AF models for MR [35,39,41], interacting with the AFDB.

AF has been also integrated into various Cryo-electron microscopy (cryo-EM) pipelines, streamlining the workflow and improving both speed and accuracy [67]. Traditionally, building atomic models from cryo-EM density maps has been a laborious and error-prone process. AF predictions provide researchers with a high-quality starting point. Software tools such as MrParse [36] and UCSF ChimeraX [40] can seamlessly access the AFDB, allowing researchers to achieve precise protein positioning within the cryo-EM map by superimposing the AF model, which significantly improves the final model's quality, while substantially reducing manual building time. ISOLDE is another tool incorporating AF models during the refinement process [38]. This allows ISOLDE to utilize the predicted information alongside the cryo-EM data, potentially leading to a more refined and accurate final structure in agreement with experimental data.

In particular, AF predictions can significantly improve the quality of cryo-EM reconstructions, especially when dealing with data with low resolution and can be used as accurate starting models to fit components into cryo-EM densities [67–71]. This is particularly helpful for determining the structures of large protein assemblies, such as the nucleopore complex [72]. Here, the authors utilized AF to enhance the structural determination of the nuclear pore complex's (NPC) cytoplasmic ring using integrative cryo-EM. The high-accuracy predictions of AF were crucial in providing detailed atomic models, accurately positioning proteins within the cryo-EM density maps, and bridging gaps in incomplete experimental data. This integrative approach led to a more comprehensive and accurate model, revealing intricate protein interactions and conformations.

In another recent study, researchers working to solve the structure of the mycobacterial lipid transporter MceI were able to assign density to a previously unknown subunit of the complex, LucB protein. They were able to perform a structural search of a density-derived poly-Ala model against a large number of predictions in AFDB, which returned LucB as a hit. The assignment was subsequently experimentally validated [73].

### Unleashing the potential of AI in protein structure prediction

As the creation of AFDB streamlined the integration of AF into commonly utilized databases and tools [44,45,48–50], in a similar vein, releasing the source

code of AF fostered the development of other AI-based tools for protein structure prediction (Table 2). Consequently, AF promoted the idea that, as in other fields, AI was progressing exponentially, and promising results could also be achieved in structural biology. Although neural networks have long been applied to structural prediction, the point at which AI-based predictions began to significantly outperform traditional methods coincided with the introduction of Transformer models [88], exemplified by the model used in AF.

Almost in parallel with AF, the RosettaFold [84] neural network for protein structure prediction has emerged. The “two-track” version of the network was outperforming trRosetta [2]. However, a performance improvement, approaching that of AF, has been observed with the development of a “three-track” neural network. This latter was inspired by the features that contributed to the performance of AF, reworking them to operate in 3D coordinate space in order to establish a closer relationship between sequence, residue–residue distances and orientations, and atomic coordinates.

Methods such as AF, which rely on co-evolution information extracted from MSA, inevitably hinder the possibility of prediction when an accurate MSA is lacking, as is the case of orphan proteins. The optimization and broader application of language models in different areas have paved a new direction in protein modeling too. These advancements have led to the development of MSA-free approaches that are both computationally efficient and highly accurate. By

leveraging the contextual understanding of protein sequences provided by language models, these tools can generate accurate predictions even in the absence of extensive sequence homology. Notable examples of these advancements include ESMFold [31], OmegaFold [83], and AminoBERT [89]. A significant achievement has been made with the large-scale application of ESMFold, whose training data retrieval has been inspired by AF, culminating in the creation of their Metagenomic Atlas, comprising over 700 million predicted structures.

## Building on the limits of AlphaFold

With the AF exploit at CASP14 as a turning point, focus also shifted to exploring other essential aspects of structural prediction. This trajectory of progress transitioned into CASP15, which embraced “universal modeling” by expanding into RNA structure and protein–ligand complex prediction (Fig. 1B) [29]. CASP15 aimed to refine the evaluation metrics for RNA and protein–ligand complexes, underscoring the complexities in accurately predicting these structures.

Unfortunately, the outcomes of CASP15 fell short of expectations. In the case of RNA, none of the models presented managed to surpass the performance of methods evaluated in other competitions. For both RNA [90] and protein–ligand complexes [91], adaptations of algorithms typically applied to proteins were explored, but these adaptations achieved only limited success, potentially due to a lack of comprehensive training data. This suggests that, unlike with protein

**Table 2.** Deep-learning-based tools for protein structure and complex prediction. Summary of the deep-learning-based tools developed and/or improved after CASP14. The ‘Method’ column highlights the key distinguishing feature or unique aspect of each model. The aim is categorized as follows: MDM, multidomain modeling; PLC, protein–ligand complexes; PNC, protein–nucleic-acids complexes; PPC, protein–protein complex; PSP, protein structure prediction.

Tool	Method	Release year	Aim	Ref.
AFSample	Stochastic perturbation of AF	2023	PPC	[74]
AlphaFold 3	Adapted AF + diffusion	2024	PPC; PNC; PLC	[75]
CombFold	AF + deterministic combinatorial assembly algorithm	2024	PPC	[76]
DeepAssembly	Population-based evolutionary algorithm	2023	MDM	[77]
DMFold-Multimer	DeepMSA	2024	PPC	[78]
EMBER3D	Protein language model	2022	PSP	[79]
EquiFold	SE(3)-equivariant	2022	PSP	[80]
ESMFold	Protein language model	2023	PSP	[31]
HelixFold	Large-scale protein language model	2023	PSP	[81]
MoLPC	AF + Monte Carlo tree search	2022	PPC	[82]
OmegaFold	Deep transformer-based protein language model	2022	PSP	[83]
RosettaFold	Three-track neural Network	2021	PSP	[84]
RosettaFold-All-Atom	Adapted RosettaFold + diffusion	2024	PPC; PNC; PLC	[85]
RosettaFoldNA	Adapted RosettaFold	2023	PNC	[86]
Umol	Evoformer + Structural module	2024	PLC	[87]

structure prediction, deep learning methods may find it challenging to leverage evolutionary data effectively in these specific areas. Regardless of whether the tools are directly inspired by the architecture of AF or push toward alternative methods, it is evident that the following AF unresolved issues are now central to the ongoing efforts in structural biology.

### Protein ligands and cofactors

The precise prediction of protein–ligand complexes plays a crucial role in overcoming various challenges in targeted therapies. As anticipated, after CASP14, several AI-based methods for the universal modeling of proteins in combination with small molecules have been developed [87,92]. While the performance of such methods is still not outperforming classical, physico-chemical-based approaches, it is expected that the parallel growth of both protein structure prediction and protein–ligand docking will lead to mutual benefits in the accurate prediction of protein–ligand complexes.

The inability to predict proteins in complex with ligands and cofactors, which is one of the main pinpointed limitations of AF, has been partially addressed by AlphaFill [32], which builds on AFDB. With this protocol, AF models can be enriched by tapping into the extensive resources of the PDB-REDO [93] and CoFactor databases [94]. The protocol, validated for correlation in terms of root mean square deviation (RMSD) with the experimental structure both globally and locally, involves transplanting the most common ligands and cofactors from a sequence homologous to the AF model. This is achieved through a local structural superimposition of the model and the homologous experimental structure.

### Multidomain modeling

Homology modeling tools have emerged as an effective complement in scenarios where AF may not provide complete solutions. Structure prediction of GPCRs make up a paradigmatic case for the quality assessment of the predictions for a specific protein family, for which the structure determination has been critical so far. A comparative study involving AF, RoseTTAFold and MODELLER [95], confirmed the expected higher performance of MODELLER, whenever a high-quality structural homolog is used as a template. Since this higher performance is evident when assessing interdomain positioning, a combined approach, utilizing both template-based and template-free methods, can yield effective results. Indeed, on this concept, recent pipelines have been developed to leverage the strengths of both

approaches. One such example is AlphaMod [96], an automated pipeline that fuses AF with MODELLER, a well-established template-based modeling software. In a similar fashion, MoDAFold [97], combines AF with MD simulations to predict the structure of missense proteins with higher accuracy.

To address the problem of *ab-initio* multidomain protein modeling, DeepAssembly, a new computational protocol for assembling multidomain proteins and complexes, was recently developed [77]. DeepAssembly uses a deep learning network to predict interdomain interactions, and then employs a population-based evolutionary algorithm to assemble domains into complete structures. This approach outperforms AF in predicting interdomain distances in multidomain proteins and improves accuracy for low-confidence structures in the AFDB.

### Protein–protein complexes

The “Assembly” category in CASP competitions has shown a notable upward trajectory since its introduction in CASP12 [98]. Even though it saw limited participation, it presented a significant opportunity for progress and quickly captured the interest of the scientific community, witnessing a surge in engagement in the subsequent years [99], and has now become one of the most hyped categories. Notably, Deepmind group did not take part in this competition, as their AF multimer version was not competitive enough [100]. However, building on the limits of AF, the quality of predictions within this category had improved substantially in CASP15 [101], with the first-ranked DMFold-Multimer [78], heavily influenced by the AF structural module. AlphaFold-predicted pairwise subunit interactions can also be exploited for assembly prediction, as shown in new advancements that focused on MSA sampling, for example, AFsample and MULTICOM [74,76,102].

PROTAC modeling can be considered another pertinent example of the importance of protein complex prediction, even for nonphysiological interactions, in which AF fails to obtain accurate predictions [103]. Historically centered around a limited set of E3 ligases, the field is witnessing a shift with the discovery of new E3 ligase structures, opening avenues for their utilization and the rational design of ligands [104]. Yet, a notable gap remains in predicting the proper orientation of the E3–ligase complex concerning the target protein, which is essential for establishing a solid foundation for the design of ligands and linkers. Thus far, addressing such a challenge has involved utilizing a combination of tools for protein structure prediction, such as RosettaFold, along with protein docking

techniques [105]. However the shift toward universal modeling is expected to provide a method to consider the orientation of the E3 Ligase–Target protein complex and the design of an appropriate bivalent ligand, at once.

### Protein–nucleic acids complexes

Understanding protein–nucleic acid interactions is particularly vital for decoding complex biological processes such as gene expression and genome repair, but accurate prediction presents significant computational hurdles due to the complexity of biomolecular interactions [106]. This knowledge is also pivotal for precision applications in genome editing, such as the engineering of Cas proteins, which are integral to technologies such as CRISPR–Cas editing [107].

On the wave of universal modeling, the concepts and techniques underlying AF and RoseTTAFold have been extended to the prediction of the structures of nucleic acids and protein–nucleic acid complexes, leading to the development of AlphaFold3 (AF3) [75] and RosettaFoldNA [86], followed by RoseTTAFold All-Atom [85]. Based on the RosettaFold three-track neural networks for molecule representation, RoseTTAFold All-Atom enhances this framework by incorporating atomic-level details and chemical representations across various dimensions into a diffusion model. As input representation, the first track of RoseTTAFold All-Atom encodes the sequence information of proteins and nucleic acids, including amino acid types and nucleotide bases. For nonpolymer atoms, it encodes their chemical element type. The second track represents pairwise information between atoms, including chemical bonds and distances. The last track includes the 3D coordinates of atoms or residues, along with information about chirality. The network of RoseTTAFold All-Atom employs attention mechanisms to weigh the importance of different input features, allowing for dynamic and context-dependent learning, and iteratively refines the predicted structure by updating the 3D coordinates based on the information from all three tracks. This integration significantly boosts the resolution and accuracy of the predicted molecular structures.

The AF3 architecture builds upon its predecessor by incorporating diffusion models as a generative module specifically designed for 3D structure generation [75]. Input data need additional preprocessing given the different molecular types the model has to handle. To do so, the raw inputs, the MSA, and the ligand conformers are converted into three different embeddings, namely the “Input,” “Pair,” and “Single” representations. The “Input” representation includes basic

atomic and residue information such as type, position, and charge. The “Single” representation groups atoms by their amino acids or nucleotides, adding contextual information. The “Pair” representation captures spatial relationships in protein and DNA/RNA sequences, enriched with template and co-evolutionary data. The “Pairformer,” which receives the enriched “Pair” representation, refines single and pair representations through recycling steps to produce a structural hypothesis, which then conditions the diffusion module for generating 3D coordinates. Despite the significance of AF’s universal modeling generalization, the release of AF3 was notably different. The absence of source code and server constraints make it difficult to understand how AF3 generalizes and prevent its verification and replication.

Of note, such advancements boosted the research toward the application of deep learning methods also for predicting the secondary structures of DNA and RNA alone [108–111].

### Protein dynamics

Exploring the dynamics of proteins has consistently presented a complex challenge, not just within the realm of computational predictions but also in experimental approaches. This challenge is linked to the prediction of protein complexes, as the interaction with other molecules often induces significant conformational changes in proteins [112]. The issue has gained renewed attention with the advent of AF, which, despite its advancements, tends to favor predictions biased toward more commonly represented conformations in its training data [113,114]. In scenarios where conformational changes are subtle, implementing postprediction processing techniques and enhancing the accuracy of preliminary models (i.e., “refinement”) emerges as a viable strategy to mitigate this limitation. CASP10 assessment of the refinement category [115] has highlighted the effectiveness of refinement methods, especially those employing molecular dynamics, in producing conformations that in their highest accuracy find also suited application in MR.

In the era following CASP-15, the endeavor to encompass the vast diversity of protein structural conformations can be expanded to most difficult and generalizable tasks. The approaches proposed leverage on tuning the MSA, that is, masking some positions, to guide the AF algorithm toward various conformational states [116–120]. The adoption of the flow-matching method has significantly enhanced the accuracy of predicting protein conformational ensembles. This approach involves training



generative models to closely replicate the distribution of protein conformations observed in experimental data or simulations [121,122]. Another recent study showed that machine learning models can be trained on simulation data to directly create realistic protein structures without the need for extensive sampling, which significantly reduces computational cost [123]. The authors demonstrated this with a model called idpGAN, trained on coarse-grained simulations of intrinsically disordered peptides. The model can predict new structures for sequences it has not seen before, proving its ability to generalize beyond the training data.

This evolution in strategy underscores the ongoing effort to refine computational tools for protein structure prediction, aiming for a more comprehensive and accurate representation of protein behavior.

### Orphan proteins

Methods such as AF, which rely on co-evolution information extracted from Multiple Sequence Alignments (MSA), face challenges when predicting the structure of orphan proteins that lack accurate MSAs. This limitation has spurred the development of MSA-free approaches that offer computationally efficient and highly accurate alternatives for predicting orphan protein structures. Notable examples of these advancements include ESM-Fold [31] and OmegaFold [83], which leverage language models to bypass the limitations of MSA dependence. These models have shown promising results in predicting orphan protein structures, providing a valuable tool for understanding these previously enigmatic proteins.

A significant breakthrough in this field has been achieved with the large-scale application of ESMFold, whose training data retrieval was inspired by AF. This effort culminated in the creation of their Metagenomic Atlas, a comprehensive repository of over 700 million predicted structures, including numerous orphan proteins. This atlas represents a major step forward in our understanding of the vast and diverse world of orphan proteins, offering valuable insights into their structures and potential functions.

Furthermore, the success of ESMFold and OmegaFold in predicting orphan protein structures has paved the way for further research and development in this area. Ongoing efforts are focused on refining these models, exploring novel MSA-free approaches, and expanding the Metagenomic Atlas to include an even wider range of orphan proteins. Notable examples of these approaches are represented by HelixFold-Single [81] and RGN [89]. The ultimate goal is to develop robust and reliable tools that can accurately predict

the structures of all orphan proteins, unlocking their structural/functional peculiarities and contributing to our understanding of the complex biological processes they are involved in [124].

### Paths for the new era in structural biology

The advent of AF has ushered in a transformative era in structural biology, shifting the focus from merely predicting existing protein structures to the exploration and design of novel biomolecules. Indeed, while the interest in protein design and early successes can be dated back to several decades ago [125], with AI's unprecedented accuracy in predicting protein structures, researchers are now entering the era that the early work envisaged—where new proteins beyond the confines of known structures can be designed for practical applications and uses. This newfound capability opens exciting avenues, for example, engineering proteins with therapeutic potential, and crafting antibodies with enhanced specificity. The fusion of computational prediction and experimental validation is poised to revolutionize drug discovery, protein engineering, and synthetic biology, ultimately leading to the development of innovative therapeutics and biomaterials [126–128].

### *De novo* protein design

With 20 naturally occurring amino acids, a protein consisting of 100 amino acids could theoretically manifest in  $20^{100}$  different sequence variations. Given the diversity of protein sizes, the theoretical number of possible proteins far exceeds the number of proteins identified by nature [129]. *De novo* protein design leverages computational algorithms and biophysical principles to engineer novel proteins with tailored functions that, in billions of years of tinkering, Nature has never produced. Computational tools, often grounded in physics-based energy functions and machine learning models, enable the exploration of vast sequence spaces and the prediction of protein structures. Recently, several protein design tools have been introduced by researchers (Table 3), which showcase substantial progress in the field. For a detailed review of *de novo* protein design, see Ref. [145]. Here, we will focus on two very recent and state-of-the-art advancements in all-atoms approaches, that is, RFDiffusion All-Atom [85,144] and ESM3 [135].

Recently, diffusion models have started to be used in protein and peptide design [146,147]. These models,

**Table 3.** Deep-learning-based tools for the *de novo* protein and peptide design. Summary of the deep-learning-based tools developed for protein/peptide design. The “Method” column highlights the key distinguishing feature or unique aspect of each model. The aim is categorized as follows: AD, antibody design; PD, protein design; PepD, peptide design.

Tool	Method	Release year	Aim	Ref.
ABlooper	Equivariant graph neural networks	2022	AD	[130]
Chroma	ChromaBackbone	2023	PD	[131]
DeepAb	Deep residual network	2022	AD	[132]
DeepH3	Deep residual network	2020	AD	[133]
EigenFold	Harmonic diffusion	2023	PD; PepD	[134]
ESM3	Generative language model	2024	PD; PepD	[135]
EvoDiff	Diffusion model	2023	PD	[136]
FoldingDiff	Transformer	2022	PD; PepD	[137]
FrameDiPT	SE(3) graph-based diffusion model	2024	PD, PepD	[138]
GENIE	IPA, Evoformer	2022	PD	[139]
GRU-based VAE	Variational autoencoders	2024	PepD	[140]
HelixDiff	Diffusion model	2024	PepD	[141]
HelixGAN	Generative adversarial network	2023	PepD	[142]
IgFold	AntiBERTy language model	2023	AD	[132]
MaSIF-Seed	Geometric deep-learning	2023	PepD	[143]
RFdiffusion	RosettaFold2	2023	PD; PepD	[144]

originally popularized in the field of generative art for their ability to create detailed and high-fidelity images [148,149], offer several advantages that make them suitable for protein modeling. Diffusion models work by iteratively refining an initial random structure into a coherent final one through a process that simulates the gradual “denoising” of a protein’s atomic coordinates [150]. This iterative refinement allows for capturing the nuanced details of protein structures that are crucial for understanding their functions and interactions. While it is generally true that most neural networks are adept at capturing the global properties of protein structures [151], diffusion models offer an edge in generating high-diversity folds, which in turn can be conditioned through a wide variety of inputs or design objectives [136]. In this field, the enhanced RFdiffusion All-Atom model incorporates diverse biological building blocks such as DNA, RNA, ions, and small molecules, expanding the scope of protein design possibilities. When coupled to other tools such as ProteinMPNN [152], LigandMPNN [153], and AF (see, e.g., a design pipeline of heme-binding proteins, available at: [https://github.com/ikalvet/heme\\_binder\\_diffusion](https://github.com/ikalvet/heme_binder_diffusion)), it opens avenues for designing proteins with unprecedented sequences, structures, and functions. ESM3 [135] is another cutting-edge AI model that can understand and design protein sequences, structures, and functions, using a frontier multimodal generative language model. The latter has been trained on a massive dataset of entries and can be prompted with any combination of sequence, structure, or function information to generate new proteins. Notably, it can

generate proteins with characteristics not seen in nature, demonstrating its creativity in problem-solving. As an example of its capabilities, ESM3 generated a new green fluorescent protein (esmGFP), which is significantly different from any known natural protein. This level of novelty is comparable to the amount of change that occurs in natural proteins over hundreds of millions of years of evolution. This demonstrates the potential of ESM3 as a powerful tool in protein engineering.

These techniques are revolutionizing protein engineering by enabling the rapid design of novel proteins with desired properties and, most importantly, have been experimentally validated, which in the end serves as the ultimate benchmark for the efficacy of predictive tools.

### Antibody design

Structure prediction of antibodies could be considered as a specialized area of protein structure prediction. In developing therapeutic antibodies, vaccines, and treatments for autoimmune disorders, the structural prediction of antibodies has historically depended on homology modeling, due to their highly evolutionarily conserved Y-shaped scaffold [102]. However, predicting antibody structures, especially the highly variable complementary determining regions (CDR)-H3 loop [154], remains challenging. Several methods have been developed to address this, utilizing both *ab initio* protocols and machine learning techniques. *Ab initio* protocols such as OptCDR [155], RosettaAntibody [156], and AbDesign [157] tackle this problem by

redesigning CDRs to enhance antibody stability and affinity by optimizing conformational and free energy changes in specific residues. RosettaAntibody, for example, can perform *de novo* antibody design or affinity maturation of existing antibodies by classifying the antibody into regions, including the framework, canonical loops, and HCDR3 loop.

For more complex tasks, innovative antibody design protocols have emerged, including DeepH3, DeepAb, IgFold, and ABlooper, showcasing the power of machine learning techniques in antibody structural predictions [130,132,133,158]. DeepH3, a deep residual neural network, identifies near-native CDR-H3 loops and improves the average RMSD of prediction compared to the standard Rosetta energy function. ABlooper employs Equivariant Graph Neural Networks to predict CDR structures, producing accurate antibody models efficiently.

Furthermore, pretrained language models have proven effective in inferring full atomic-level protein structures. DeepAb leverages an antibody pretrained language model with recurrent neural network to reconstruct the entire antibody variable region, generating more precise structures compared with alternatives. IgFold, inspired by DeepAb, utilizes a pretrained language model trained on natural antibody sequences and graph networks to directly predict backbone atom coordinates, offering high speed, accuracy, and nanobody modeling capabilities.

While the remarkable advancements in domain prediction contribute to highly accurate models of immunoglobulin domains or target epitopes, the precise orientation and interaction between the CDR loops and the target epitope are areas that require further refinement [159–161]. These aspects of antibody–antigen interaction are more likely to benefit from the latest breakthroughs in predicting multimeric complexes, highlighting a crucial direction for future advancements in antibody design [130,162,163]. Sculptor is a new algorithm that addresses this challenge using deep generative design to create antibodies that bind to specific epitopes [164]. It does this by jointly searching for the best positions, interactions, and shapes of the protein scaffold. It then designs a protein backbone that complements the target.

In summary, while accurately predicting antibody structures, especially CDR-H3, remains a challenge, significant progress has been made through the development of various computational methods and the integration of machine learning techniques and pretrained language models. These advancements hold promise for accelerating antibody design and

engineering efforts, ultimately contributing to the development of more effective therapeutic antibodies.

## Peptide design

The development of therapeutic peptides hinges on the ability to design peptidic binders that target specific proteins of interest. Traditionally, peptide design has relied on a combination of rational design, simulation, and screening techniques [165]. Similarly, early AI-based approaches to peptide design, which were adapted from protein design methods, employed various techniques such as inverse design (e.g., ProteinMPNN [152]), peptide-specific methods (e.g., PepMLM [166]), and generative models (e.g., MaSIF-Seed [143]). These approaches all aimed to design new peptide binders starting from the target protein. By leveraging different architectures, it is now possible to focus on the *de-novo* design of peptide sequences. These approaches are particularly valuable for their ability to capture the distribution of amino acids that confer a set of functionalities and activities, such as antimicrobial, anticancer, immunogenic properties, or signal peptide functions. Variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models have emerged as viable options. As an example, a recent multistep sequence generation algorithm was proposed [140]. The deep learning-based generative model Gated Recurrent Unit based variational autoencoder (GRU-based VAE) and the Metropolis Hasting (MH) sampling algorithm efficiently generate new peptide sequences. The binding affinity of generated peptides is then evaluated using physics-based methods, such as molecular dynamics (MD) simulations. Several GANs have been trained for peptide design, tailored to specific use cases such as immunogenic, antimicrobial [167–169], and antiviral peptides [170]. Other examples of GANs, such as HelixGAN [142], have been specifically trained to focus on the design of helical peptides, and similarly, a diffusion model called HelixDiff [141] has been developed with the same objective. In a recent work [137], a diffusion-based model (FoldingDiff) generating high-quality backbone peptides (up to 128 residues) via a procedure inspired by the natural folding process, is presented. FoldingDiff uses a sequence of dihedral angles capturing the relative orientation of the constituent backbone atoms and generates stable folded peptides by denoising from an unfolded structure. Moreover, the development of combined and fine-tuned approaches is becoming increasingly common. Latent space diffusion models, in particular, are gaining traction, and AMP-Diffusion [171] exemplifies

their application by harnessing the power of latent representations and the flexibility of diffusion processes to enhance the generation of antimicrobial peptides.

### Accessible software and hardware in protein structure prediction

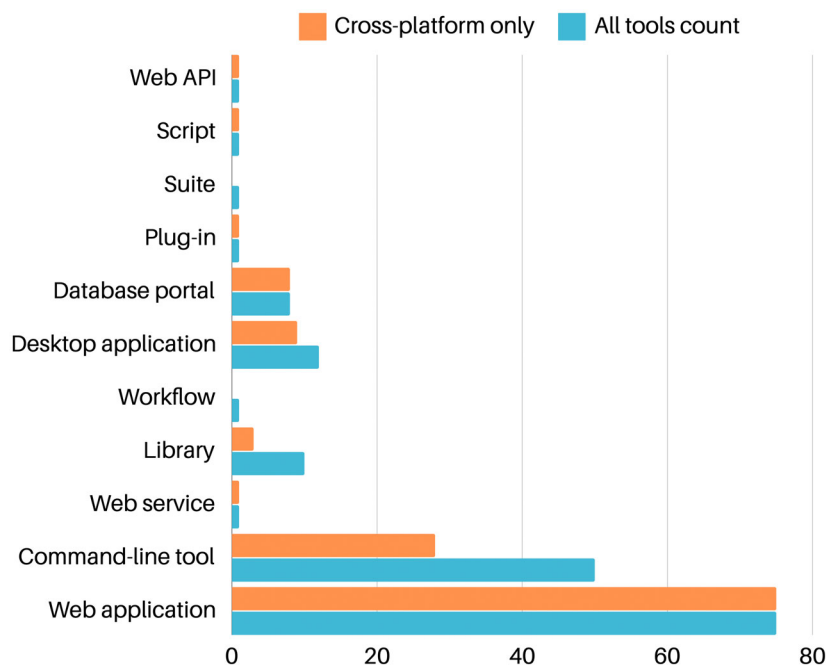
With the widespread adoption and success of AI and computer science techniques in biology, it has become crucial to enable access to such prediction tools and technologies for a broad audience with limited familiarity with bioinformatics and software engineering. Ensuring easy access of AF and related tools to researchers from diverse backgrounds can permit new and diverse complex biological questions to be asked, and a fruitful *commubium* of AI and human expertise to be reached. Lowering such barriers and identifying the pivotal key factors essential for ensuring the success of a software release for scientists, clinicians, and students alike revolves around accessibility, user-friendliness, and comprehensibility [172,173]. For example, factors such as Graphical User Interfaces (GUIs) significantly reduce the barrier to software use, especially when coupled with User Experience/User Interface (UX/UI) studies and tutorials covering all aspects of user interaction [174,175].

Analyzing the features of the other tools developed in the realm of protein structure prediction [176], the distribution of tool types reveals a strong inclination toward Web Application (Fig. 2). These are noted for their accessibility, yet they come with drawbacks such

as server-side dependence and limited control. Command-Line Tools also feature prominently, showcasing their utility for batch processing. It emerges that Desktop Applications, which would offer unmatched control and independence from server constraints, are very limited as they face significant hurdles in cross-platform compatibility. Widely used molecular graphics viewers constitute an exception. Indeed, the integration of AF has been promptly pursued in tools such as ChimeraX [40].

An example of a free interface for AF is seen with ColabFold [66] that, other than featuring an easy-to-use Colab notebook, implemented a faster MSA step and a way to customize the MSA. Delving into structure prediction interfaces, we can find some widely used web servers, that is, Phyre2 [177] and SwissModel [178], along with some Desktop Applications, like PyMod [179].

Now that AI models have become predominant in this field, the issue of accessibility is no longer solely related to the concept of GUIs and similar interfaces. The “black box” factor also comes into play, referring to the inability to explain what occurs during the process. Therefore, there is a growing need to develop Explainable AI (XAI) solutions that provide transparent insights into the decision-making processes of these models, fostering trust, accountability, and understanding among users. Even if some attempts are ongoing [180], transparent AI solutions still need some time to become predominant in this domain.



**Fig. 2.** Overview of tools categorized under the “Structure Analysis” EDAM (Ontology of bioscientific Data Analysis and Management) “topic” tag in the Bio.Tools database; ([81], accessed on February, 2024). For each tool type, the plot displays the number of occurrences, along with the count of tools that are cross-platform.

While there is little room for improvement in increasing the modularity of environments that implement algorithms akin to AF, or similar “black box” models, there is potential in harnessing software development that facilitates a unified application of structural bioinformatics protocols, fostering a user-driven methodology. However, there is a trend toward developing “blind” software—applications that take inputs and produce outputs without requiring users to understand the underlying processes. While this approach promotes efficiency and caters to users of different expertise levels, it risks discouraging deep understanding of the methodological principles. In structural bioinformatics, this may lead to insufficient appreciation of physicochemical principles such as protein folding and interactions, essential for accurate interpretation of results.

When examining the process of predicting 3D structures and its various stages, it becomes clear that an insufficient understanding and interaction between the user and the algorithm often results in suboptimal predictions [181]. A key aspect of this process is the selection of a structural template. This decision should be predominantly influenced by the user’s expertise and judgment, rather than relying solely on sequence identity percentage metrics. Factors other than sequence similarity are functional characteristics, unique structural motifs or conformations, solvent composition, pH values, and the interaction with additional binding entities.

Another fundamental yet often overlooked and non-optimized step is the quality of the MSA prior to the modeling step. A well-built MSA can effectively link sequence information with the structural elements of the proteins, increasing the accuracy of the final model. As a remark of the importance of such a step, still in the post-AF era, there are witnesses of effort in developing tools for facilitating the manipulation of multiple sequence alignments [182]. Manipulation here refers to refining the alignment, correcting gaps, and mis-alignments, to ensure that it accurately reflects the evolutionary and functional relationships among the sequences. This consideration becomes particularly pertinent when addressing processes such as the prediction of multidomain and/or multimeric structures. As highlighted earlier, a significant challenge persists in realizing *ab-initio* predictions of protein complexes. In addressing this challenge, a fusion of *ab-initio* and homology modeling protocols, approached with a user-driven perspective, can be a potential alternative by leveraging the possibility to integrate a variety of information sources.

In order to translate these objectives into reality, the figure of the Research Software Engineer (RSE) surges

as central, a hybrid professional embodying the confluence of software engineering and scientific inquiry [183].

As software may become more accessible, computational resources must do so too. Protein structure prediction and design are recognized as Nondeterministic Polynomial-hard problems [184], necessitating exponential computational efforts with traditional techniques. High-Performance Computing (HPC) significantly influences this field, as many prediction algorithms benefit from parallelization across HPC’s multiple processors [185]. Similarly, the rapid advancement in Graphical Processing Units (GPUs) enables efficient execution of these complex tasks, especially with deep learning algorithms [186]. Consequently, adapting existing tools for both parallel and GPU computing has become widespread.

## Conclusions and future perspectives

As extensively discussed in previous papers [187–193], the arrival of AF has fundamentally transformed our approach to structural biology. In this review, we have focused on the aspects most significantly influenced by the release of AF, examining limits and new opportunities, changes from the methodological perspectives, some state-of-the-art applications of particular interest, and software development viewpoints. Until 2021, efforts have primarily been directed toward accurately predicting naturally occurring proteins, the imminent solution to this issue now shifts focus toward a variety of distinctly different domains: protein design, synthetic biology, AI-driven drugs and antibodies design, integrative structural biology.

Each of these areas has had to adapt to the sudden availability of advanced structural information, now made readily accessible to a broad audience. This accessibility has not only democratized the field but also spurred a wave of innovation, necessitating a re-evaluation of existing practices and the development of completely new methodologies to fully leverage the potential of this groundbreaking tool. A promising future direction is to integrate AI with quantum computing frameworks [185]. The development of hybrid quantum-classical solvers, such as QPacker in Rosetta software [186], exemplifies this innovative approach, reshaping our understanding of complex energy landscapes in protein structures.

## Acknowledgements

This work was supported by European Research Council (ERC) Advanced Grant HOLO-GT [101019289]; AIRC

(Associazione Italiana Ricerca sul Cancro) MFA2017 [20447]; “Progetti Ateneo” Sapienza University of Rome [RM1221815D52AB32]; Italy Ministry of University and Research PRIN [CUP: 2022N3JXLA].

## Conflict of interest

The authors declare no conflict of interest.

## Author contributions

SR and MP collected, analyzed and interpreted the data, and wrote the paper. FM collected the data. AP wrote the paper, supervised, and conceived the project.

## References

- 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.
- Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM and Lupas AN (2021) High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687–1699.
- Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, 439–444.
- Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J *et al.* (2024) AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* **52**, 368–375.
- Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L *et al.* (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, 296–303.
- Porta-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**, e1009818.
- David A, Islam S, Tankhilevich E and Sternberg MJE (2022) The AlphaFold database of protein structures: a biologist’s guide. *J Mol Biol* **434**, 167336.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
- Sippl MJ, Lackner P, Domingues FS and Koppensteiner WA (1999) An attempt to analyse progress in fold recognition from CASP1 to CASP3. *Proteins Suppl* **3**, 226–230.
- Huang YJ, Zhang N, Bersch B, Fidelis K, Inouye M, Ishida Y, Kryshtafovych A, Kobayashi N, Kuroda Y, Liu G *et al.* (2021) Assessment of prediction methods for protein structures determined by NMR in CASP14: impact of AlphaFold2. *Proteins* **89**, 1959–1976.
- Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**, 285–289.
- Johnson MS, Srinivasan N, Sowdhamini R and Blundell TL (1994) Knowledge-based protein modelling. *Crit Rev Biochem Mol Biol* **29**, 1–68.
- Rost B and Sander C (1996) Bridging the protein sequence–structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* **25**, 113–136.
- Sali A and Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815.
- Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE and Baker D (2001) Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins Suppl* **5**, 119–126.
- Rohl CA, Strauss CE, Misura KM and Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66–93.
- Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40.
- Zhang Y and Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA* **101**, 7594–7599.
- Göbel U, Sander C, Schneider R and Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317.
- Giraud BG, Heumann JM and Lapedes AS (1999) Superadditive correlation. *Phys Rev* **59**, 4983–4991.
- Afonnikov DA, Kondrakhin YV, Titov II and Kolchanov NA (1997) Detecting direct correlation between positions in multiple alignment of amino-acid sequences. In *Computer Science and Biology. Genome Informatics: Function, Structure, Phylogeny* (Frishman D and Mewes HW, eds), pp. 87–98. Proceedings of the German Conference on Bioinformatics.
- Weigt M, White RA, Szurmant H, Hoch JA and Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA* **106**, 67–72.
- Burger L and van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* **6**, e1000633.
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R and Sander C (2011) Protein 3D

- structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766.
- 25 Skwark MJ, Raimondi D, Michel M and Elofsson A (2014) Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol* **10**, e1003889.
  - 26 Wang S, Sun S, Li Z, Zhang R and Xu J (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* **13**, e1005324.
  - 27 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.
  - 28 Mirabello C and Wallner B (2019) rawMSA: end-to-end deep learning using raw multiple sequence alignments. *PLoS One* **14**, e0220182.
  - 29 Kryshtafovych A, Antczak M, Szachniuk M, Zok T, Kretsch RC, Rangan R, Pham P, Das R, Robin X, Studer G *et al.* (2023) New prediction categories in CASP15. *Proteins* **91**, 1550–1557.
  - 30 Kryshtafovych A, Schwede T, Topf M, Fidelis K and Moult J (2021) Critical assessment of methods of protein structure prediction (CASP)–round XIV. *Proteins* **89**, 1607–1617.
  - 31 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.
  - 32 Hekkelman ML, de Vries I, Joosten RP and Perrakis A (2023) AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods* **20**, 205–213.
  - 33 Dobson L, Szekeres LI, Gerdán C, Langó T, Zeke A and Tusnády GE (2023) TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. *Nucleic Acids Res* **51**, D517–D522.
  - 34 Pei J and Cong Q (2023) AFTM: a database of transmembrane regions in the human proteome predicted by AlphaFold. *Database (Oxford)* **2023**, baad008.
  - 35 Agirre J, Atanasova M, Bagdonas H, Ballard CB, Baslé A, Beilstein-Edmands J, Borges RJ, Brown DG, Burgos-Mármol JJ, Berrisford JM *et al.* (2023) The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr D* **79**, 449–461.
  - 36 Simpkin AJ, Thomas JMH, Keegan RM and Rigden DJ (2022) MrParse: finding homologues in the PDB and the EBI AlphaFold database for molecular replacement and more. *Acta Crystallogr D* **78**, 553–559.
  - 37 Keegan RM and Winn MD (2008) MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr D* **64**, 119–124.
  - 38 Oeffner RD, Croll TI, Millán C, Poon BK, Schlicksup CJ, Read RJ and Terwilliger TC (2022) Putting AlphaFold models to work with phenix.process\_predicted\_model and ISOLDE. *Acta Crystallogr D* **78**, 1303–1314.
  - 39 Simpkin AJ, Caballero I, McNicholas S, Stevenson K, Jiménez E, Sánchez Rodríguez F, Fando M, Uski V, Ballard C, Chojnowski G *et al.* (2023) Predicted models and CCP4. *Acta Crystallogr D* **79**, 806–819.
  - 40 Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH and Ferrin TE (2021) UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82.
  - 41 Emsley P and Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D* **60**, 2126–2132.
  - 42 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J and Steinegger M (2024) Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246.
  - 43 Holm L (2022) Dali server: structural unification of protein families. *Nucleic Acids Res* **50**, W210–W215.
  - 44 Waterhouse AM, Procter JB, Martin DMA, Clamp M and Barton GJ (2009) Jalview version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191.
  - 45 Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, Velankar S, Burley SK, Koča J and Rose AS (2021) Mol\* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* **49**, W431–W437.
  - 46 Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A and Adams PD (2018) Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr D* **74**, 531–544.
  - 47 Poon BK, Terwilliger TC and Adams PD (2024) The Phenix-AlphaFold webservice: enabling AlphaFold predictions for use in Phenix. *Protein Sci* **33**, e4992.
  - 48 Pfab J, Phan NM and Si D (2021) DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proc Natl Acad Sci USA* **118**, e2017525118.
  - 49 Giri N and Cheng J (2023) Improving protein–ligand interaction modeling with cryo-EM data, templates, and deep learning in 2021 ligand model challenge. *Biomolecules* **13**, 132.
  - 50 He J, Lin P, Chen J, Cao H and Huang SY (2022) Model building of protein complexes from intermediate-resolution cryo-EM maps with deep learning-guided automatic assembly. *Nat Commun* **13**, 4066.
  - 51 Varadi M and Velankar S (2023) The impact of AlphaFold protein structure database on the fields of life sciences. *Proteomics* **23**, e2200128.

- 52 Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L *et al.* (2023) InterPro in 2022. *Nucleic Acids Res* **51**, 418–427.
- 53 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res* **49**, 412–419.
- 54 Bairoch A and Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48.
- 55 Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al.* (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, 447–452.
- 56 Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J *et al.* (2023) Ensembl 2023. *Nucleic Acids Res* **51**, 933–941.
- 57 UniProt Consortium (2023) UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* **51**, 523–531.
- 58 Aderinwale T, Bharadwaj V, Christoffer C, Terashi G, Zhang Z, Jahandideh R, Kagaya Y and Kihara D (2022) Real-time structure search and structure classification for AlphaFold protein models. *Commun Biol* **5**, 316.
- 59 Faezov B and Dunbrack RL Jr (2023) AlphaFold2 models of the active form of all 437 catalytically competent human protein kinase domains. *bioRxiv*. doi: [10.1101/2023.07.21.550125](https://doi.org/10.1101/2023.07.21.550125)
- 60 Piovesan D, Del Conte A, Clementel D, Monzon AM, Bevilacqua M, Aspromonte MC, Iserle JA, Orti FE, Marino-Buslje C and Tosatto SCE (2023) MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res* **51**, 438–444.
- 61 Jakubec D, Skoda P, Krivak R, Novotny M and Hoksza D (2022) PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Res* **50**, 593–597.
- 62 Saito M, Xu P, Faure G, Maguire S, Kannan S, Altae-Tran H, Vo S, Desimone A, Macrae RK and Zhang F (2023) Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620**, 660–668.
- 63 Terwilliger TC, Afonine PV, Liebschner D, Croll TI, McCoy AJ, Oeffner RD, Williams CJ, Poon BK, Richardson JS, Read RJ *et al.* (2023) Accelerating crystal structure determination with iterative AlphaFold prediction. *Acta Crystallogr D* **79**, 234–244.
- 64 McCoy AJ, Sammito MD and Read RJ (2022) Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallogr D* **78**, 1–13.
- 65 Millán C, Keegan RM, Pereira J, Sammito MD, Simpkin AJ, McCoy AJ, Lupas AN, Hartmann MD, Rigden DJ and Read RJ (2021) Assessing the utility of CASP14 models for molecular replacement. *Proteins* **89**, 1752–1769.
- 66 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.
- 67 Rantos V, Karius K and Kosinski J (2022) Integrative structural modeling of macromolecular complexes using Assemblin. *Nat Protoc* **17**, 152–176.
- 68 Alshammari M, Wriggers W, Sun J and He J (2022) Refinement of AlphaFold2 models against experimental and hybrid cryo-EM density maps. *QRB Discov* **3**, e16.
- 69 DiIorio MC and Kulczyk AW (2023) Novel artificial intelligence-based approaches for ab initio structure determination and atomic model building for cryo-electron microscopy. *Micromachines (Basel)* **14**, 1674.
- 70 Ziegler SJ, Mallinson SJB, St. John PC and Bomble YJ (2021) Advances in integrative structural biology: towards understanding protein complexes in their cellular context. *Comput Struct Biotechnol J* **19**, 214–225.
- 71 Terwilliger TC, Poon BK, Afonine PV, Schlicksup CJ, Croll TI, Millán C, Richardson JS, Read RJ and Adams PD (2022) Improved AlphaFold modeling with implicit experimental information. *Nat Methods* **19**, 1376–1382.
- 72 Fontana P, Dong Y, Pi X, Tong AB, Hecksel CW, Wang L, Fu TM, Bustamante C and Wu H (2022) Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. *Science* **376**, eabm9326.
- 73 Chen J, Fruhauf A, Fan C, Ponce J, Ueberheide B, Bhabha G and Ekiert DC (2023) Structure of an endogenous mycobacterial MCE lipid transporter. *Nature* **620**, 445–452.
- 74 Wallner B (2023) AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573.
- 75 Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500.
- 76 Shor B and Schneidman-Duhovny D (2024) CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. *Nat Methods* **21**, 477–487.
- 77 Xia Y, Zhao K, Liu D, Zhou X and Zhang G (2023) Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning. *Commun Biol* **6**, 1221.



- 78 Zheng W, Wuyun Q, Li Y, Zhang C, Freddolino PL and Zhang Y (2024) Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods* **21**, 279–289.
- 79 Weissenow K, Heinzinger M, Steinegger M and Rost B (2022) Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. *bioRxiv*. doi: [10.1101/2022.11.14.516473](https://doi.org/10.1101/2022.11.14.516473)
- 80 Lee JH, Yadollahpour P, Watkins A, Frey NC, Leaver-Fay A, Ra S, Cho K, Gligorijevic V, Regev A and Bonneau R (2023) EquiFold: protein structure prediction with a novel coarse-grained structure representation. *bioRxiv*. doi: [10.1101/2022.10.07.511322](https://doi.org/10.1101/2022.10.07.511322)
- 81 Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, Zhu K, Zhang X, Wu H, Li H *et al.* (2023) A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat Mach Intell* **5**, 1087–1096.
- 82 Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P and Elofsson A (2022) Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat Commun* **13**, 6028.
- 83 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. *bioRxiv*. doi: [10.1101/2022.07.21.500999](https://doi.org/10.1101/2022.07.21.500999)
- 84 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.
- 85 Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR *et al.* (2024) Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, ead12528.
- 86 Baek M, McHugh R, Anishchenko I, Jiang H, Baker D and DiMaio F (2024) Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat Methods* **21**, 117–121.
- 87 Bryant P, Kelkar A, Guljas A, Clementi C and Noé F (2024) Structure prediction of protein–ligand complexes from sequence information with Umol. *Nat Commun* **15**, 4536.
- 88 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I (2017) Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- 89 Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, Rochereau C, Ahdriz 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.
- 90 Das R, Kretsch RC, Simpkin AJ, Mulvaney T, Pham P, Rangan R, Bu F, Keegan RM, Topf M, Rigden DJ *et al.* (2023) Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins* **91**, 1747–1770.
- 91 Robin X, Studer G, Durairaj J, Eberhardt J, Schwede T and Walters WP (2023) Assessment of protein–ligand complexes in CASP15. *Proteins* **91**, 1811–1821.
- 92 Cai H, Shen C, Jian T, Zhang X, Chen T, Han X, Yang Z, Dang W, Hsieh CY, Kang Y *et al.* (2023) CarsiDock: a deep learning paradigm for accurate protein–ligand docking and screening based on large-scale pre-training. *Chem Sci* **15**, 1449–1471.
- 93 Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund AC, Blanchet C, Bongcam-Rudloff E, Combet C, Da Costa AL, Deleage G *et al.* (2009) PDB\_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Cryst* **42**, 376–384.
- 94 Fischer JD, Holliday GL and Thornton JM (2010) The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics* **26**, 2496–2497.
- 95 Lee C, Su BH and Tseng YJ (2022) Comparative studies of AlphaFold, RoseTTAFold and Modeller: a case study involving the use of G-protein-coupled receptors. *Brief Bioinform* **23**, bbac308.
- 96 Gil Zuluaga FH, D’Arminio N, Bardozzo F, Tagliaferri R and Marabotti A (2023) An automated pipeline integrating AlphaFold 2 and MODELLER for protein structure prediction. *Comput Struct Biotechnol J* **21**, 5620–5629.
- 97 Zheng L, Shi S, Sun X, Lu M, Liao Y, Zhu S, Zhang H, Pan Z, Fang P, Zeng Z *et al.* (2024) MoDAFold: a strategy for predicting the structure of missense mutant protein based on AlphaFold2 and molecular dynamics. *Brief Bioinform* **25**, bbae006.
- 98 Lafita A, Bliven S, Kryshtafovych A, Bertoni M, Monastyrskyy B, Duarte JM, Schwede T and Capitani G (2018) Assessment of protein assembly prediction in CASP12. *Proteins* **86**, 247–256.
- 99 Ozden B, Kryshtafovych A and Karaca E (2021) Assessment of the CASP14 assembly predictions. *Proteins* **89**, 1787–1799.
- 100 Evans R, O’Neill M, Pritzel A, Antropova N, Senior A, Green T, Židek A, Bates R, Blackwell S, Yim J *et al.* (2021) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. doi: [10.1101/2021.10.04.463034v2](https://doi.org/10.1101/2021.10.04.463034v2)
- 101 Elofsson A (2023) Progress at protein structure prediction, as seen in CASP15. *Curr Opin Struct Biol* **80**, 102594.
- 102 Liu J, Guo Z, Wu T, Roy RS, Chen C and Cheng J (2023) Improving AlphaFold2-based protein tertiary

- structure prediction with MULTICOM in CASP15. *Commun Chem* **6**, 188.
- 103 Pereira GP, Gouzien C, Souza PCT and Martin J (2024) AlphaFold-Multimer struggles in predicting PROTAC-mediated protein–protein interfaces. *bioRxiv*. doi: [10.1101/2024.03.19.585735](https://doi.org/10.1101/2024.03.19.585735)
- 104 Liu Y, Yang J, Wang T, Luo M, Chen Y, Chen C, Ronai Z, Zhou Y, Ruppin E and Han L (2023) Expanding PROTACtable genome universe of E3 ligases. *Nat Commun* **14**, 6509.
- 105 Zaidman D, Prilusky J and London N (2020) PROsettaC: Rosetta based modeling of PROTAC mediated ternary complexes. *J Chem Inf Model* **60**, 4894–4903.
- 106 Chiu TP, Rao S and Rohs R (2023) Physicochemical models of protein–DNA binding with standard and modified base pairs. *Proc Natl Acad Sci USA* **120**, e2205796120.
- 107 Zhao L, Koseki SRT, Silverstein RA, Amrani N, Peng C, Kramme C, Savic N, Pacesa M, Rodríguez TC, Stan T *et al.* (2023) PAM-flexible genome editing with an engineered chimeric Cas9. *Nat Commun* **14**, 6175.
- 108 Li J, Chiu TP and Rohs R (2024) Predicting DNA structure using a deep learning method. *Nat Commun* **15**, 1243.
- 109 Shulgina Y, Trinidad MI, Langeberg CJ, Nisonoff H, Chithrananda S, Skopintsev P, Nissley AJ, Patel J, Boger RS, Shi H *et al.* (2024) RNA language models predict mutations that improve RNA function. *bioRxiv*. doi: [10.1101/2024.04.05.588317](https://doi.org/10.1101/2024.04.05.588317)
- 110 Gong T and Bu D (2024) Language models enable zero-shot prediction of RNA secondary structure including pseudoknots. *bioRxiv*. doi: [10.1101/2024.01.27.577533](https://doi.org/10.1101/2024.01.27.577533)
- 111 Nguyen E, Poli M, Durrant MG, Thomas AW, Kang B, Sullivan J, Ng MY, Lewis A, Patel A, Lou A *et al.* (2024) Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv*. doi: [10.1101/2024.02.27.582234](https://doi.org/10.1101/2024.02.27.582234)
- 112 Jisna VA and Jayaraj PB (2021) Protein structure prediction: conventional and deep learning perspectives. *Protein J* **40**, 522–544.
- 113 Lane TJ (2023) Protein structure prediction has reached the single-structure frontier. *Nat Methods* **20**, 170–173.
- 114 Sala D, Engelberger F, Mchaourab HS and Meiler J (2023) Modeling conformational states of proteins with AlphaFold. *Curr Opin Struct Biol* **81**, 102645.
- 115 Nugent T, Cozzetto D and Jones DT (2014) Evaluation of predictions in the CASP10 model refinement category. *Proteins* **82**, 98–111.
- 116 Sala D, Hildebrand PW and Meiler J (2023) Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties. *Front Mol Biosci* **10**, 1121962.
- 117 Del Alamo D, Sala D, Mchaourab HS and Meiler J (2022) Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11**, e75751.
- 118 Heo L and Feig M (2022) Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873–1885.
- 119 Stein RA and Mchaourab HS (2022) SPEACH\_AF: sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLoS Comput Biol* **18**, e1010483.
- 120 Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, Ovchinnikov S, Colwell L and Kern D (2024) Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839.
- 121 Jing B, Berger B and Jaakkola T (2024) AlphaFold meets flow matching for generating protein ensembles. *arXiv*: 2402.04845.
- 122 Yim J, Campbell A, Mathieu E, Foong AYG, Gastegger M, Jiménez-Luna J, Lewis S, Satorras VG, Veeling BS, Noé F *et al.* (2024) Improved motif-scaffolding with SE(3) flow matching. *arXiv*: 2401.04082v1.
- 123 Janson G, Valdes-Garcia G, Heo L and Feig M (2023) Direct generation of protein conformational ensembles via machine learning. *Nat Commun* **14**, 774.
- 124 Masrati G, Landau M, Ben-Tal N, Lupas A, Kosloff M and Kosinski J (2021) Integrative structural biology in the era of accurate structure prediction. *J Mol Biol* **433**, 167127.
- 125 Regan L, Caballero D, Hinrichsen MR, Virrueta A, Williams DM and O’Hern CS (2015) Protein design: past, present, and future. *Biopolymers* **104**, 334–350.
- 126 Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, Heymach J, Le X, Yan H and Alam T (2023) AI in drug discovery and its clinical relevance. *Heliyon* **9**, e17575.
- 127 Harrer S, Menard J, Rivers M, Green DVS, Karpiak J, Jeliazkov R, Shapovalov MV, del Alamo D and Sternke MC (2024) Artificial intelligence drives the digital transformation of pharma. In *Artificial Intelligence in Clinical Practice* (Krittawanong C, ed.), pp. 345–372. Academic Press, Cambridge, MA.
- 128 Almubarak HF, Tan W, Hoffmann AD, Wei J, El-Shennawy L, Squires JR, Sun Y, Dashzeveg NK, Simonton B, Jia Y *et al.* (2024) Physics-driven structural docking and protein language models accelerate antibody screening and design for broad-spectrum antiviral therapy. *bioRxiv*. doi: [10.1101/2024.03.01.582176](https://doi.org/10.1101/2024.03.01.582176)
- 129 No authors listed (2024) Spotlight on protein structure design. *Nat Biotechnol* **42**, 157.
- 130 Ruffolo JA, Chu LS, Mahajan SP and Gray JJ (2023) Fast, accurate antibody structure prediction from deep

- learning on massive set of natural antibodies. *Nat Commun* **14**, 2389.
- 131 Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER *et al.* (2023) Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078.
- 132 Ruffolo JA, Sulam J and Gray JJ (2021) Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 100406.
- 133 Ruffolo JA, Guerra C, Mahajan SP, Sulam J and Gray JJ (2020) Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* **36**, i268–i275.
- 134 Jing B, Erives E, Pao-Huang P, Corso G, Berger B and Jaakkola T (2023) EigenFold: generative protein structure prediction with diffusion models. *arXiv*: 2304.02198.
- 135 Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M *et al.* (2024) Simulating 500 million years of evolution with a language model. *bioRxiv*. doi: [10.1101/2024.07.01.600583](https://doi.org/10.1101/2024.07.01.600583)
- 136 Alamdari S, Thakkar N, van den Berg R, Lu AX, Fusi N, Amini AP and Yang KK (2023) Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*. doi: [10.1101/2023.09.11.556673](https://doi.org/10.1101/2023.09.11.556673)
- 137 Wu KE, Yang KK, Berg R, van den Zou JY, Lu AX and Amini AP (2022) Protein structure generation via folding diffusion. *arXiv*: 2209.15611.
- 138 Zhang C, Leach A, Makkink T, Arbesú M, Kadri I, Luo D, Mizrahi L, Krichen S, Lang M, Tovchigrechko A *et al.* (2023) FrameDiPT: SE(3) diffusion model for protein structure inpainting. *bioRxiv*. doi: [10.1101/2023.11.21.568057](https://doi.org/10.1101/2023.11.21.568057)
- 139 Lin Y and AlQuraishi M (2023) Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv*: 2301.12485.
- 140 Chen S, Lin T, Basu R, Ritchey J, Wang S, Luo Y, Li X, Pei D, Kara LB and Cheng X (2024) Design of target-specific peptide inhibitors using generative deep learning and molecular dynamics simulations. *Nat Commun* **15**, 1611.
- 141 Xie X, Valiente PA, Kim J and Kim PM (2024) HelixDiff, a score-based diffusion model for generating all-atom  $\alpha$ -helical structures. *ACS Cent Sci* **10**, 1001–1011.
- 142 Xie X, Valiente PA and Kim PM (2023) HelixGAN a deep-learning methodology for conditional de novo design of  $\alpha$ -helix structures. *Bioinformatics* **39**, btad036.
- 143 Gainza P, Wehrle S, Van Hall-Beauvais A, Marchand A, Scheck A, Hartevelde Z, Buckley S, Ni D, Tan S, Sverrisson F *et al.* (2023) De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184.
- 144 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**, 1089–1100.
- 145 Kortemme T (2024) De novo protein design-from new structures to programmable functions. *Cell* **187**, 526–544.
- 146 Yim J, Stärk H, Corso G, Jing B, Barzilay R and Jaakkola TS (2024) Diffusion models in protein structure and docking. *Wiley Interdiscip Rev Comput Mol Sci* **14**, e1711.
- 147 Wang XF, Tang JY, Liang H, Sun J, Dorje S, Peng B, Ji XW, Li Z, Zhang XE and Wang DB (2024) ProT-Diff: a modularized and efficient approach to de novo generation of antimicrobial peptide sequences through integration of protein language model and diffusion model. *bioRxiv*. doi: [10.1101/2024.02.22.581480](https://doi.org/10.1101/2024.02.22.581480)
- 148 Ho J, Jain A and Abbeel P (2020) Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* **33**, 6840–6851.
- 149 Dhariwal P and Nichol A (2021) Diffusion models beat GANs on image synthesis. *Adv Neural Inf Process Syst* **34**, 8780–8794.
- 150 Guo Z, Liu J, Wang Y, Chen M, Wang D, Xu D and Cheng J (2024) Diffusion models in bioinformatics and computational biology. *Nat Rev Bioeng* **2**, 136–154.
- 151 Lin Y, Lee M, Zhang Z and AlQuraishi M (2024) Out of many, one: designing and scaffolding proteins at the scale of the structural universe with Genie 2. *arXiv*: 2405.15489.
- 152 Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N *et al.* (2022) Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56.
- 153 Dauparas J, Lee GR, Pecoraro R, An L, Anishchenko I, Glasscock C and Baker D (2023) Atomic context-conditioned protein sequence design using LigandMPNN. *bioRxiv*. doi: [10.1101/2023.12.22.573103](https://doi.org/10.1101/2023.12.22.573103)
- 154 Chen H, Fan X, Zhu S, Pei Y, Zhang X, Zhang X, Liu L, Qian F and Tian B (2024) Accurate prediction of CDR-H3 loop structures of antibodies with deep learning. *Elife* **12**, RP91512.
- 155 Pantazes RJ and Maranas CD (2010) OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng Des Sel* **23**, 849–858.
- 156 Sircar A, Kim ET and Gray JJ (2009) RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res* **37**, W474–W479.
- 157 Lipsh-Sokolik R, Listov D and Fleishman SJ (2021) The AbDesign computational pipeline for modular

- backbone assembly and design of binders and enzymes. *Protein Sci* **30**, 151–159.
- 158 Abanades B, Georges G, Bujotzek A and Deane CM (2022) ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* **38**, 1877–1880.
- 159 Yin R and Pierce BG (2024) Evaluation of AlphaFold antibody-antigen modeling with implications for improving predictive accuracy. *Protein Sci* **33**, e4865.
- 160 Polonsky K, Pupko T and Freund NT (2023) Evaluation of the ability of AlphaFold to predict the three-dimensional structures of antibodies and epitopes. *J Immunol* **211**, 1578–1588.
- 161 Xu Z, Davila A, Wilamowski J, Teraguchi S and Standley DM (2022) Improved antibody-specific epitope prediction using AlphaFold and AbAdapt. *ChemBiochem* **23**, e202200303.
- 162 Schoeder CT, Schmitz S, Adolf-Bryfogle J, Sevy AM, Finn JA, Sauer MF, Bozhanova NG, Mueller BK, Sangha AK, Bonet J *et al.* (2021) Modeling immunity with Rosetta: methods for antibody and antigen design. *Biochemistry* **60**, 825–846.
- 163 Khan A, Cowen-Rivers AI, Grosnit A, Deik DG, Robert PA, Greiff V, Smorodina E, Rawat P, Akbar R, Dreckowski K *et al.* (2023) Toward real-world automated antibody design with combinatorial Bayesian optimization. *Cell Rep Methods* **3**, 100374.
- 164 Eguchi RR, Choe CA, Parekh U, Khalek IS, Ward MD, Vithani N, Bowman GR, Jardine JG and Huang PS (2022) Deep generative design of epitope-specific binding proteins by latent conformation optimization. *bioRxiv*. doi: [10.1101/2022.12.22.521698](https://doi.org/10.1101/2022.12.22.521698)
- 165 Chang L, Mondal A and Perez A (2022) Towards rational computational peptide design. *Front Bioinform* **2**, 1046493.
- 166 Chen T, Pertsemlidis S, Watson R, Kavirayuni VS, Hsu A, Vure P, Pulugurta R, Vincoff S, Hong L, Wang T *et al.* (2023) PepMLM: target sequence-conditioned generation of peptide binders via masked language modeling. *arXiv*: 2310.03842v2.
- 167 Li G, Iyer B, Prasath VBS, Ni Y and Salomonis N (2021) DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Brief Bioinform* **22**, bbab160.
- 168 Van Oort CM, Ferrell JB, Remington JM, Wshah S and Li J (2021) AMPGAN v2: machine learning-guided design of antimicrobial peptides. *J Chem Inf Model* **61**, 2198–2207.
- 169 Zervou MA, Doutsis E, Pantazis Y and Tsakalides P (2024) De novo antimicrobial peptide design with feedback generative adversarial networks. *Int J Mol Sci* **25**, 5506.
- 170 Surana S, Arora P, Singh D, Sahasrabudde D and Valadi J (2023) PandoraGAN: generating antiviral peptides using generative adversarial network. *SN Comput Sci* **4**, 607. doi: [10.1007/s42979-023-02203-3](https://doi.org/10.1007/s42979-023-02203-3)
- 171 Chen T, Vure P, Pulugurta R and Chatterjee P (2024) AMP-diffusion: integrating latent diffusion with protein language models for antimicrobial peptide generation. *bioRxiv*. doi: [10.1101/2024.03.03.583201](https://doi.org/10.1101/2024.03.03.583201)
- 172 Bolchini D, Finkelstein A, Perrone V and Nagl S (2009) Better bioinformatics through usability analysis. *Bioinformatics* **25**, 406–412.
- 173 Goldman M and Gehlenborg N (2021) Making tools that people will use: user-centered design in computational biology research. *Pac Symp Biocomput* **26**, 346–350.
- 174 Pavelin K, Cham JA, de Matos P, Brooksbank C, Cameron G and Steinbeck C (2012) Bioinformatics meets user-centred design: a perspective. *PLoS Comput Biol* **8**, e1002554.
- 175 Javahery H, Seffah A and Radhakrishnan T (2004) Beyond power: making bioinformatics tools user-centered. *Commun ACM* **47**, 59–62.
- 176 Ison J, Ienasescu H, Chmura P, Rydza E, Ménager H, Kalaš M, Schwämmle V, Grüning B, Beard N, Lopez R *et al.* (2019) The bio.tools registry of software tools and data resources for the life sciences. *Genome Biol* **20**, 164.
- 177 Kelley LA, Mezulis S, Yates CM, Wass MN and Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845–858.
- 178 Schwede T, Kopp J, Guex N and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* **31**, 3381–3385.
- 179 Janson G and Paiardini A (2021) PyMod 3: a complete suite for structural bioinformatics in PyMOL. *Bioinformatics* **37**, 1471–1472.
- 180 Zhou Z, Hu M, Salcedo M, Gravel N, Yeung W, Venkat A, Guo D, Zhang J, Kannan N and Li S (2023) XAI meets biology: a comprehensive review of explainable AI in bioinformatics applications. *arXiv*: 2312.06082.
- 181 Fiser A (2010) Template-based protein structure modeling. *Methods Mol Biol* **673**, 73–94.
- 182 Fraenkel AS (1993) Complexity of protein folding. *Bull Math Biol* **55**, 1199–1210.
- 183 Horsfall D, Cool J, Hettrick S, Pisco AO, Hong NC and Haniffa M (2023) Research software engineering accelerates the translation of biomedical research for health. *Nat Med* **29**, 1313–1316.
- 184 Pierce NA and Winfree E (2002) Protein design is NP-hard. *Protein Eng* **15**, 779–782.
- 185 Zhong B, Su X, Wen M, Zuo S, Hong L and Lin L (2021) ParaFold: paralleling AlphaFold for large-scale predictions. *arXiv*: 2111.06340. doi: [10.48550/arXiv.2111.06340](https://doi.org/10.48550/arXiv.2111.06340)
- 186 MacCarthy EA, Zhang C, Zhang Y and Kc DB (2022) GPU-I-TASSER: a GPU accelerated I-TASSER protein structure prediction tool. *Bioinformatics* **38**, 1754–1755.

- 187 Bertoline LMF, Lima AN, Krieger JE and Teixeira SK (2023) Before and after AlphaFold2: an overview of protein structure prediction. *Front Bioinform* **3**, 1120370.
- 188 Xu T, Xu Q and Li J (2023) Toward the appropriate interpretation of AlphaFold2. *Front Artif Intell* **6**, 1149748.
- 189 Pakhrin SC, Shrestha B, Adhikari B and Kc DB (2021) Deep learning-based advances in protein structure prediction. *Int J Mol Sci* **22**, 5553.
- 190 Aithani L, Alcaide E, Bartunov S, Cooper CDO, Doré AS, Lane TJ, Maclean F, Rucktooa P, Shaw RA and Skerratt SE (2023) Advancing structural biology through breakthroughs in AI. *Curr Opin Struct Biol* **80**, 102601.
- 191 Borkakoti N and Thornton JM (2023) AlphaFold2 protein structure prediction: implications for drug discovery. *Curr Opin Struct Biol* **78**, 102526.
- 192 Khatami MH, Mendes UC, Wiebe N and Kim PM (2023) Gate-based quantum computing for protein design. *PLoS Comput Biol* **19**, e1011033.
- 193 Mulligan VK, Melo H, Merritt HI, Slocum S, Weitzner BD, Watkins AM, Renfrew PD, Pelissier C, Arora PS and Bonneau R (2020) Designing peptides on a quantum computer. *bioRxiv*. doi: [10.1101/752485](https://doi.org/10.1101/752485)