International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 1, Issue 2 December 2024 ISSN: 3048-975X
https://ijrasht.com/

AI-Driven Protein Structure Prediction: A Review of Deep Learning Methods for Protein Folding and Drug Design

Hardik, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822427.cse@pietgroup.co.in

Sahil, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822412.cse@pietgroup.co.in

Simran, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822840.cse@pietgroup.co.in

Yash, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822349.cse@pietgroup.co.in

Neha Sharma, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India neha.cse-aiml@piet.co.in

Abstract— Predicting a protein's three-dimensional structure from its amino acid sequence, long regarded as one of the central unsolved problems in computational biology, has been transformed over the past several years by deep learning systems capable of achieving accuracy approaching that of experimental structure determination for many protein families. This paper reviews the methods underlying this transformation, tracing the field's progression from early template-based and co-evolution-based approaches through end-to-end deep learning architectures that jointly reason over sequence, evolutionary, and geometric information. We describe the architectural innovations that enabled this progress, including attention-based processing of evolutionary sequence alignments and geometry-aware structure modules, and examine extensions to protein complex prediction, protein language models, and generative de novo protein design. We review applications in drug discovery, protein function annotation, protein engineering, disease mechanism elucidation, and vaccine and antibody design, summarize the benchmarks used to evaluate structure prediction methods, and present a case study illustrating how a predicted structure is used within a drug discovery pipeline. We discuss remaining challenges, including prediction accuracy for intrinsically disordered regions, multi-state and dynamic conformational prediction, and experimental validation bottlenecks, and outline future directions connecting structure prediction with generative protein design and integrative structural biology.

Index Terms—protein structure prediction, protein folding, deep learning, AlphaFold, drug discovery, protein design, structural bioinformatics

I. INTRODUCTION

A protein's biological function is determined largely by its three-dimensional structure, which in turn is determined by the sequence of amino acids composing the protein, a relationship first established through classical experiments demonstrating that a denatured protein can spontaneously refold into its native structure without any additional cellular machinery beyond the information encoded in its own amino acid sequence. Predicting this three-dimensional structure computationally from sequence alone, commonly referred to as the protein folding problem, has been recognized as a central challenge in computational biology for more than half a century, motivated both by fundamental scientific interest in the physical process of protein folding and by the substantial practical value that accurate, rapid structure prediction would offer for drug discovery, protein engineering, and basic biological research. For decades, progress on this problem was comparatively incremental, with the most reliable prediction methods relying on homology modeling, which builds a structural model for a target protein based on the experimentally known structure of an evolutionarily related protein, an approach fundamentally limited to targets with a sufficiently similar protein already characterized experimentally. Progress accelerated substantially beginning in the late 2010s and especially in the early 2020s, when deep learning systems combining attention-based neural network architectures with evolutionary and geometric information achieved prediction accuracy for many protein families that approaches the accuracy of experimental structure determination methods, a development widely regarded within structural biology as a landmark achievement for the application of AI to a fundamental scientific problem.

This paper reviews the methods, applications, and open challenges of AI-driven protein structure prediction. Section II reviews the background of the protein folding problem and the historical progression of computational approaches to it. Section III presents a taxonomy of structure prediction methodologies. Section IV describes core algorithmic methods in technical detail. Section V surveys applications across drug discovery and biological research. Section VI summarizes benchmarks, centered on the long-running Critical Assessment of Protein Structure Prediction competition. Section VII offers a comparative discussion. Section VIII presents a case study illustrating structure

prediction within a drug discovery workflow. Section IX discusses implementation tooling. Section X considers ethical implications. Section XI discusses open challenges, Section XII outlines future directions, and Section XIII concludes.

Beyond the specific scientific achievement of accurate structure prediction, this development is also frequently cited as a demonstration case for the broader potential of AI to accelerate scientific discovery, discussed in related surveys within this series, and the methodological lessons from this success, including the importance of incorporating domain-specific biological and physical knowledge alongside general deep learning architectures, have informed subsequent AI-for-science research well beyond structural biology specifically.

II. BACKGROUND AND MOTIVATION

A. The Protein Folding Problem

A protein is synthesized as a linear chain of amino acid residues that, under physiological conditions, spontaneously folds into a specific, generally stable three-dimensional structure within a timescale of milliseconds to seconds, despite the astronomically large number of theoretically possible three-dimensional conformations available to even a modestly sized amino acid chain, an apparent paradox regarding how folding occurs so rapidly that has itself been an important subject of biophysical research independent of the computational prediction problem addressed in this paper. The resulting native structure typically corresponds to, though is not always precisely identical to, a global or near-global minimum of the protein's conformational free energy landscape, shaped by the specific sequence of amino acid side chains and their physicochemical interactions with each other and with the surrounding solvent.

Computational structure prediction from sequence alone requires either directly modeling this underlying physical folding process, an approach that remains computationally prohibitive for all but the smallest proteins using current simulation methods given the vast timescale and conformational search space involved, or instead learning statistical patterns relating sequence to structure from the large and growing body of experimentally determined protein structures, an approach that underlies essentially all practically successful structure prediction methods, including the deep learning methods that are the primary focus of this paper.

B. Why Deep Learning Succeeded Where Earlier Methods Struggled

Deep learning approaches to structure prediction succeeded in part because they could effectively exploit two complementary sources of information that earlier methods struggled to combine fully: evolutionary co-variation patterns across large sets of related protein sequences, which encode indirect structural information because residues that are close together in the folded three-dimensional structure tend to co-evolve to maintain compatible physicochemical interactions even as the overall sequence diverges across evolutionary time, and directly learned geometric reasoning capable of translating these statistical sequence patterns into a physically consistent three-dimensional atomic structure.

Earlier co-evolution-based methods could extract useful residue-contact information from evolutionary sequence data using comparatively simple statistical techniques, but converting this contact information into an accurate, complete three-dimensional structure still generally required a separate, often computationally expensive structure assembly or refinement procedure. The key architectural innovation of modern deep learning structure prediction systems, described in detail in Section IV, was to unify evolutionary sequence analysis and three-dimensional geometric structure generation within a single, jointly trained neural network architecture, allowing information to flow bidirectionally between sequence-level statistical patterns and three-dimensional geometric constraints throughout the prediction process, rather than processing these two sources of information in a more rigid, sequential pipeline.

C. Historical Context

The Critical Assessment of Protein Structure Prediction (CASP), a biennial blind prediction competition established in the mid-1990s in which participating research groups predict the structures of proteins whose experimental structures have been determined but not yet publicly released, has served for decades as the primary independent benchmark tracking progress in the field, providing an unusually rigorous, community-wide evaluation standard rarely matched in other areas of computational biology. Progress through the 2000s and 2010s was comparatively steady but incremental, dominated by template-based and co-evolution-based methods whose accuracy, while useful for many

practical applications, still fell well short of experimental structure determination accuracy for proteins lacking closely related experimentally characterized homologs.

The introduction of deep learning systems combining attention-based sequence processing with specialized geometric structure modules produced a marked, widely discussed discontinuity in CASP performance in the most recent competitions, with prediction accuracy for many target proteins reaching a level competitive with experimental structure determination methods such as X-ray crystallography and cryo-electron microscopy, a result that was met with considerable attention from both the structural biology research community and the broader scientific community as an unusually clear demonstration of deep learning's potential to make qualitative rather than merely incremental progress on a long-standing scientific problem.

The rapid, widely publicized release of a comprehensive database of predicted structures spanning essentially the entire human proteome, along with predicted structures for numerous other organisms of scientific interest, further accelerated adoption of these methods across the broader biological research community, allowing researchers studying a protein of interest to consult a predicted structure immediately rather than waiting, potentially indefinitely, for experimental structure determination, a shift in practical research workflow that has been compared in significance, within structural biology circles, to the transformative effect of large-scale genome sequencing on molecular biology research more broadly in preceding decades.

III. A TAXONOMY OF PROTEIN STRUCTURE PREDICTION APPROACHES

Protein structure prediction methods can be organized according to the primary source of information they exploit to constrain the predicted structure. Template-based methods, also called homology modeling, construct a predicted structure based on the experimentally known structure of an evolutionarily related protein identified through sequence similarity search, providing high accuracy when a sufficiently close structural template is available but offering little guidance for proteins without close experimentally characterized homologs. Co-evolution-based methods instead extract structural constraints, particularly residue-residue contact information, directly from statistical patterns of correlated sequence variation across a large multiple sequence alignment of evolutionarily related sequences, without requiring an experimentally solved template structure.

End-to-end deep learning methods, the primary focus of this paper, integrate evolutionary sequence information, geometric structural reasoning, and, in many current architectures, physical and chemical constraints within a single neural network architecture trained end-to-end to predict atomic coordinates directly from sequence and evolutionary information, achieving the strongest current prediction accuracy across the broadest range of target proteins. A further, increasingly prominent category, protein language model based methods, forgo explicit multiple sequence alignment construction in favor of representations learned by large protein language models pretrained on massive protein sequence databases, trading some prediction accuracy in certain regimes for substantially reduced computational cost and latency, particularly valuable for large-scale structure prediction applied to millions of sequences. Table I summarizes this taxonomy.

TABLE I. TAXONOMY OF PROTEIN STRUCTURE PREDICTION APPROACHES

Approach	Primary Information Source	Representative Methods	Key Limitation
Template-based / homology modeling	Known homologous structures	SWISS-MODEL, MODELLER	Requires close structural template
Co-evolution-based	Multiple sequence alignment statistics	Direct coupling analysis, contact prediction methods	Requires deep, diverse sequence alignment
End-to-end deep learning	Sequence + evolutionary + geometric	AlphaFold2, RoseTTAFold	Computationally intensive MSA search
Protein language model based	Learned sequence representations	ESMFold, single-sequence predictors	Somewhat lower accuracy for some targets

IV. CORE METHODS

A. Template-Based and Homology Modeling

Template-based modeling identifies one or more experimentally characterized proteins structurally similar to the target protein, typically via sequence similarity search, and constructs a structural model for the target by aligning its sequence to the template structure and adjusting the template's backbone and side-chain conformations to accommodate sequence differences between target and template. This approach can achieve high accuracy when a close structural template exists, but its accuracy degrades substantially, and eventually fails outright, as sequence similarity to the best available template decreases, motivating the co-evolution-based and deep learning methods discussed in subsequent subsections for targets lacking sufficiently close experimentally characterized homologs.

B. Co-evolution and Multiple Sequence Alignment Based Methods

Co-evolution-based structure prediction relies on the observation that pairs of amino acid residues that are in close physical contact within a protein's folded structure tend to exhibit correlated patterns of amino acid substitution across evolutionarily related sequences, since a substitution at one contacting residue that disrupts a favorable physicochemical interaction is more likely to persist through evolutionary selection if compensated by a corresponding substitution at its contacting partner residue. Statistical techniques such as direct coupling analysis attempt to disentangle these direct co-evolutionary couplings from the confounding effect of indirect, transitively propagated correlations across a multiple sequence alignment, producing a predicted residue-residue contact map that can subsequently be used to constrain three-dimensional structure assembly.

The accuracy of co-evolution-based methods depends critically on the depth and evolutionary diversity of the available multiple sequence alignment for a target protein family, since statistically reliable co-evolutionary signal requires a sufficiently large and diverse set of related sequences; protein families with few known evolutionary relatives, sometimes termed evolutionarily shallow families, have historically posed a persistent challenge for co-evolution-based methods and remain a comparatively more difficult case for modern deep learning methods as well, though to a lesser degree than for purely statistical co-evolution approaches.

C. End-to-End Deep Learning Architectures

Modern end-to-end deep learning structure prediction architectures process a target protein's multiple sequence alignment together with, in many systems, structural templates of related proteins where available, using an attention-based neural network architecture that iteratively refines two complementary internal representations: a representation of pairwise relationships between all residues in the target protein, capturing predicted geometric relationships such as inter-residue distance and orientation, and a representation of the evolutionary sequence alignment itself, allowing information to flow between evolutionary statistical patterns and geometric structural hypotheses across multiple rounds of processing.

A specialized structure module subsequently translates these refined internal representations directly into three-dimensional atomic coordinates, using a geometry-aware neural network component explicitly designed to respect the physical constraints of protein backbone geometry, such as consistent bond lengths and angles, and trained using a loss function that directly penalizes deviation between predicted and experimentally determined atomic coordinates. Critically, these architectures also produce a calibrated per-residue confidence estimate alongside each predicted structure, allowing users to assess which regions of a predicted structure are likely to be accurate and which regions, often corresponding to flexible loops or regions lacking strong evolutionary or structural constraint, should be treated with greater caution.

D. Protein Language Models for Structure Prediction

An alternative and complementary line of research uses large protein language models, pretrained using masked sequence modeling objectives directly on massive databases of protein sequences without any explicit multiple sequence alignment construction, as the primary source of evolutionary and structural information for downstream structure prediction. Because these models encode evolutionary and structural patterns implicitly within their learned parameters from pretraining, structure prediction using a protein language model can proceed directly from a single target sequence without the computationally expensive multiple sequence alignment search required by co-evolution-based and many end-to-end deep learning methods, substantially reducing prediction latency and enabling large-scale structure prediction across millions of sequences, at some cost to prediction accuracy, particularly for evolutionarily shallow protein families where the implicit evolutionary knowledge encoded during pretraining may be less complete.

E. Protein Complex and Interaction Prediction

Extensions of single-chain structure prediction architectures to protein complexes, in which two or more distinct protein chains, or multiple copies of the same chain, jointly fold into a stable multi-chain assembly, predict not only the structure of each individual chain but the geometric arrangement of chains relative to one another, a task complicated by the fact that co-evolutionary signal must now be identified across chain boundaries, requiring correctly paired multiple sequence alignments between interacting sequences from different species, itself a nontrivial computational sub-problem addressed through specialized sequence-pairing heuristics.
Accurate protein complex structure prediction has substantial practical value for understanding protein-protein interaction networks and for structure-based drug design targeting protein-protein interaction interfaces, though prediction accuracy for complexes generally remains somewhat lower than for single-chain structures, particularly for transient or weak protein-protein interactions that may not follow the same strong evolutionary co-variation patterns as stable, obligate protein complexes that have been under sustained evolutionary selection to maintain their interaction interface.

F. Generative Protein Design

Building on the success of structure prediction, generative protein design methods address the inverse problem: generating a novel amino acid sequence, or directly generating a novel three-dimensional backbone structure together with a compatible sequence, expected to fold into a desired target structure or to perform a desired biochemical function not necessarily present in any naturally occurring protein. Diffusion-based generative models adapted from image generation research have been applied directly to three-dimensional protein backbone generation, iteratively refining a noisy initial backbone structure toward a physically plausible, designable target structure, which is then paired with a sequence design model that identifies an amino acid sequence predicted to fold stably into the generated backbone, extending the discovery-oriented generative design paradigm discussed in related AI-for-science research specifically to protein engineering applications.

V. APPLICATIONS

A. Drug Discovery and Target Identification

Accurate structure prediction for drug target proteins, particularly those lacking any experimentally determined structure, has substantially expanded the set of proteins amenable to structure-based drug design methods such as virtual screening and structure-guided medicinal chemistry, which rely on a reasonably accurate three-dimensional model of the target protein's binding site to computationally evaluate candidate drug molecules prior to committing to expensive experimental synthesis and testing.
Predicted structures are also used to support target identification and validation earlier in the drug discovery pipeline, for example by predicting whether a protein implicated by genetic or genomic studies in a particular disease has a druggable binding pocket amenable to small-molecule intervention, informing prioritization decisions regarding which of many candidate disease-associated proteins merit further experimental investment as a potential drug target.

B. Protein Function Annotation

A substantial fraction of proteins identified through genomic sequencing lack experimentally characterized function, and predicted structure provides valuable indirect evidence for inferring likely biochemical function, since proteins with similar three-dimensional structure, even in the absence of strong sequence similarity, frequently share related biochemical functions, allowing large-scale structural prediction across entire genomes to support systematic function annotation efforts that would be infeasible using experimental structure determination alone given the scale of modern sequencing data.
Large-scale structural annotation has also enabled systematic structural comparison across entire predicted proteomes, supporting discovery of previously unrecognized structural relationships between proteins that share little or no detectable sequence similarity but nonetheless adopt closely related three-dimensional folds, occasionally revealing unexpected evolutionary or functional relationships between proteins previously studied in entirely separate biological contexts.

C. Protein Engineering and De Novo Design

Generative protein design methods, described in Section IV-F, are increasingly used to engineer proteins with novel or enhanced functions not present in naturally occurring proteins, including enzymes with improved catalytic efficiency or novel substrate specificity, and binding proteins engineered to bind a specified target molecule with high affinity and specificity, applications with relevance to industrial biotechnology, agriculture, and therapeutic protein development.

D. Understanding Disease Mechanisms

Predicted structures support mechanistic understanding of how specific disease-associated genetic mutations affect protein function, for example by revealing whether a mutation is likely to destabilize a protein's overall fold, disrupt a specific functionally important binding interface, or have comparatively little structural consequence, informing both basic understanding of disease mechanism and, in some cases, prioritization of therapeutic strategies targeted to the specific structural consequence of a given disease-associated mutation.
This structural perspective on disease mutations is particularly valuable for rare genetic diseases, where the number of affected individuals may be too small to support extensive functional experimental characterization of every observed mutation, and where a structurally informed hypothesis regarding likely mutational consequence can help prioritize which variants merit more detailed experimental follow-up among the often large number of genetic variants identified through clinical sequencing of an affected individual.

E. Vaccine and Antibody Design

Structure prediction and protein complex modeling support vaccine antigen design by helping researchers identify structurally stable, immunologically relevant regions of a pathogen protein suitable for use as a vaccine antigen, and support antibody engineering by modeling the structural interaction between a candidate antibody and its target antigen, accelerating early-stage design and evaluation of candidate antibody therapeutics prior to more resource-intensive experimental characterization.

VI. BENCHMARKS AND EVALUATION

The Critical Assessment of Protein Structure Prediction (CASP), introduced in Section II-C, remains the primary independent benchmark for evaluating protein structure prediction methods, using a blind prediction format in which participating groups predict the structures of target proteins whose experimental structures have been determined but withheld from public release until after prediction submissions close, providing an unusually rigorous safeguard against the risk of models having been trained, even inadvertently, on the specific evaluation targets. Prediction accuracy in CASP is typically quantified using the Global Distance Test score, which measures the fraction of predicted residue positions falling within specified distance thresholds of their true experimentally determined positions after optimal structural superposition.
Beyond CASP, the Continuous Automated Model Evaluation platform provides ongoing, ungated benchmark evaluation of structure prediction servers against newly deposited experimental structures as they become publicly available, complementing the biennial CASP competition with more frequent evaluation. For protein complex prediction specifically, dedicated benchmark sets of experimentally determined protein-protein and protein-ligand complex structures are used to evaluate complex prediction accuracy, typically using interface-focused accuracy metrics that specifically assess the correctness of the predicted inter-chain binding interface rather than only overall structural accuracy.

TABLE II. REPRESENTATIVE BENCHMARKS FOR PROTEIN STRUCTURE PREDICTION

Benchmark	Format	Primary Metric
CASP	Biennial blind prediction competition	Global Distance Test (GDT) score
CAMEO	Continuous, ungated server evaluation	GDT and related structural accuracy scores

Benchmark	Format	Primary Metric
Protein complex benchmark sets	Curated complex structure datasets	Interface accuracy, DockQ score
CASP refinement/design categories	Specialized sub-challenges	Category-specific accuracy metrics

VII. COMPARATIVE ANALYSIS AND DISCUSSION

Template-based modeling remains highly reliable and computationally inexpensive when a close experimentally characterized structural template is available, but its applicability is fundamentally limited to targets with such a template, an increasingly narrow constraint as end-to-end deep learning methods have expanded reliable prediction to a far broader range of target proteins including those lacking any closely related experimental structure. Co-evolution-based methods provided an important conceptual and methodological bridge toward modern deep learning approaches but have been largely superseded in practical use by end-to-end deep learning systems that achieve substantially higher accuracy by jointly modeling evolutionary and geometric information rather than processing them in separate sequential stages.

Between end-to-end deep learning methods and protein language model based methods, the central practical trade-off concerns computational cost versus accuracy: end-to-end methods relying on explicit multiple sequence alignment search generally achieve somewhat higher accuracy, particularly for evolutionarily shallow protein families, but require substantially more computation per prediction due to the cost of multiple sequence alignment construction, whereas protein language model based methods sacrifice some accuracy in exchange for dramatically faster, single-sequence prediction well suited to very large-scale structure prediction applications across millions of sequences, such as genome-wide structural annotation efforts.

TABLE III. COMPARATIVE SUMMARY OF STRUCTURE PREDICTION APPROACH FAMILIES

Approach	Typical Accuracy	Computational Cost	Best Suited For
Template-based modeling	High if close template exists	Very low	Targets with known close homolog
Co-evolution-based	Moderate	Moderate (MSA-dependent)	Historical/legacy pipelines
End-to-end deep learning	Highest overall	High (MSA search + inference)	Maximum accuracy, novel folds
Protein language model based	Slightly lower on average	Low (single-sequence)	Large-scale, rapid screening

VIII. ILLUSTRATIVE CASE STUDY

Consider a representative drug discovery scenario in which researchers have identified, through genomic association studies, a protein strongly implicated in a particular disease but lacking any experimentally determined structure. Using an end-to-end deep learning structure prediction method, researchers generate a predicted structure for the target protein, along with the per-residue confidence estimate produced alongside the prediction, and observe high confidence across the protein's core catalytic domain but comparatively lower confidence in a peripheral loop region, consistent with that region's known biological flexibility.

Focusing analysis on the high-confidence catalytic domain, researchers identify a candidate binding pocket suitable for small-molecule intervention and use the predicted structure to conduct a virtual screening campaign, computationally evaluating a large library of candidate small molecules for predicted binding compatibility with the identified pocket, prioritizing a manageable shortlist of top-ranked candidates for experimental testing rather than testing the full candidate library directly, substantially reducing the experimental cost and time required to identify a promising initial lead compound relative to experimental screening conducted without the benefit of a structural model.

Following identification of a promising lead compound through this virtual screening process, researchers may pursue experimental structure determination, for example using X-ray crystallography or cryo-electron microscopy, of the protein in complex with the identified lead compound, both to confirm the predicted binding mode and to guide subsequent structure-based optimization of the compound's binding affinity and selectivity, illustrating the broader pattern, discussed in related AI-for-science research, in which computational prediction accelerates and focuses, but does not entirely replace, the experimental validation that remains central to translating a computational hypothesis into an actual therapeutic candidate.

IX. IMPLEMENTATION CONSIDERATIONS AND TOOLING

A. Software Availability and Computational Requirements

Several leading end-to-end deep learning structure prediction systems have been released as open-source software together with pretrained model weights, substantially broadening practical access to high-accuracy structure prediction capability for the broader research community relative to a scenario in which such capability remained proprietary to the original developing organization. Running these methods still typically requires access to specialized hardware accelerators for practical inference speed, together with access to large sequence and structural databases required for multiple sequence alignment and template search, motivating the development of web-based prediction servers and cloud-hosted implementations that lower the practical barrier to entry for researchers lacking dedicated local computational infrastructure.

B. Integration into Structural Biology Workflows

Predicted structures are increasingly integrated directly into experimental structural biology workflows, for example serving as an initial structural model to accelerate the process of fitting an experimentally determined electron density map, such as from cryo-electron microscopy, into an interpretable atomic structure, a task that traditionally relied more heavily on manual model building or template-based approaches, illustrating a complementary rather than purely substitutive relationship between computational prediction and experimental structure determination methods within modern structural biology practice.

X. ETHICAL AND SOCIETAL CONSIDERATIONS

The dual-use potential of protein structure prediction and generative protein design tools, discussed more broadly in related AI-for-science research, applies with particular salience to this field, since the same computational capability that supports beneficial drug and vaccine design could, in principle, be misused to support the design of harmful biological agents, motivating deliberate consideration of dual-use risk in decisions regarding open publication and release of increasingly capable structure prediction and generative protein design tools and models.

A further consideration concerns equitable access to the benefits of this technology, since the pharmaceutical and biotechnology applications that stand to benefit most directly from accelerated structure-based drug discovery have historically prioritized diseases prevalent in wealthier markets, and ensuring that improved structure prediction and drug design capability translates into meaningful progress on neglected diseases disproportionately affecting lower-income populations will likely require deliberate policy and funding interventions beyond the underlying technological capability itself, which alone does not guarantee equitable application toward the diseases of greatest global public health burden.

XI. CHALLENGES AND OPEN PROBLEMS

Despite substantial overall progress, several categories of structure prediction remain more challenging than the well-studied, relatively rigid single-domain proteins on which current methods achieve their strongest performance. Intrinsically disordered proteins and regions, which do not adopt a single stable three-dimensional structure under physiological conditions but instead sample a broad ensemble of conformations, are poorly represented by current methods that are fundamentally designed to predict a single static structure, motivating distinct methodological approaches specifically suited to representing conformational ensembles rather than a single fixed conformation.

A second challenge concerns proteins that adopt multiple distinct functional conformational states, such as proteins that undergo significant conformational change upon ligand binding or post-translational modification, since current

methods trained predominantly on single static experimental structures do not naturally represent this conformational multiplicity, and predicting which of potentially several relevant functional states corresponds to a given biological context remains an active area of methodological development. A third challenge concerns evolutionarily shallow protein families lacking sufficient related sequences to support strong evolutionary signal, where prediction accuracy, while improved relative to purely co-evolution-based methods, remains lower than for well-studied protein families with abundant evolutionary sequence data.

A fourth challenge concerns experimental validation capacity: as computational structure prediction and generative design methods generate an increasingly large volume of candidate structures and designed proteins, experimental validation capacity, including protein expression, purification, and structure determination or functional characterization, has increasingly become a comparative bottleneck relative to computational hypothesis generation, echoing similar experimental validation bottlenecks discussed in broader AI-for-science research and motivating continued investment in high-throughput experimental characterization methods capable of validating computational predictions at a correspondingly increased scale.

XII. FUTURE RESEARCH DIRECTIONS

One prominent direction extends structure prediction methods toward modeling conformational ensembles and dynamic protein behavior rather than a single static structure, potentially by adapting generative modeling techniques to sample multiple plausible conformational states for a given target protein rather than producing a single deterministic prediction, addressing the multi-state prediction challenge discussed in Section XI.

A second direction focuses on deeper integration between structure prediction and generative protein design, moving beyond the current largely sequential pipeline of first predicting structure and separately designing sequences toward unified models capable of jointly reasoning over desired function, structure, and sequence within a single generative framework, potentially accelerating iterative design cycles in which a desired function directly informs generated structure and sequence in a single coherent computational step rather than through separate, loosely coupled prediction and design stages.

A further direction concerns improving prediction accuracy and reliability for the challenging categories discussed in Section XI, including intrinsically disordered regions and evolutionarily shallow protein families, potentially through specialized architectures or training strategies specifically targeted at these historically underperforming categories rather than relying solely on continued general-purpose scaling of existing architectures. Finally, continued development of high-throughput experimental validation methods, potentially themselves increasingly automated and AI-assisted, echoing the self-driving laboratory concept discussed in related AI-for-science research, is likely to remain an important complement to continued computational methodological progress, ensuring that experimental validation capacity does not become an increasingly binding constraint on translating computational structure prediction and design advances into validated scientific and therapeutic outcomes.

XIII. CONCLUSION

AI-driven protein structure prediction represents one of the clearest and most widely recognized successes of deep learning applied to a long-standing fundamental scientific problem, achieving prediction accuracy for many protein families that approaches experimental structure determination methods and substantially expanding the practical reach of structure-based approaches across drug discovery, protein function annotation, protein engineering, and disease mechanism research. This survey reviewed the field's progression from template-based and co-evolution-based methods through end-to-end deep learning architectures and protein language model based approaches, examined applications spanning drug discovery, function annotation, protein engineering, disease research, and vaccine design, and presented a case study illustrating structure prediction's role within a drug discovery workflow. While significant challenges remain, particularly for intrinsically disordered proteins, multi-state conformational prediction, and the experimental validation bottleneck that increasingly constrains translation of computational predictions into validated outcomes, the continued integration of structure prediction with generative protein design and high-throughput experimental methods suggests that this field will remain a central and rapidly advancing pillar of AI-driven structural biology and drug discovery in the years ahead.

REFERENCES

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly

accurate protein structure prediction with AlphaFold," Nature, vol. 596, 2021.
[2] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," Science, vol. 373, 2021.
[3] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, et al., "Protein complex prediction with AlphaFold-Multimer," bioRxiv preprint, 2021.
[4] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science, vol. 379, 2023.
[5] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proceedings of the National Academy of Sciences, vol. 118, 2021.
[6] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) — Round XIV," Proteins, vol. 89, 2021.
[7] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (CASP) — Round XV," Proteins, vol. 91, 2023.
[8] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, et al., "SWISS-MODEL: Homology modelling of protein structures and complexes," Nucleic Acids Research, vol. 46, 2018.
[9] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," Journal of Molecular Biology, vol. 234, 1993.
[10] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," Proceedings of the National Academy of Sciences, vol. 108, 2011.
[11] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3D structure computed from evolutionary sequence variation," PLOS ONE, vol. 6, 2011.
[12] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," PLOS Computational Biology, vol. 13, 2017.
[13] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, "Improved protein structure prediction using predicted interresidue orientations," Proceedings of the National Academy of Sciences, vol. 117, 2020.
[14] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A. W. R. Nelson, A. Bridgland, et al., "Improved protein structure prediction using potentials from deep learning," Nature, vol. 577, 2020.
[15] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, et al., "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space," Nucleic Acids Research, vol. 50, 2022.
[16] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Zidek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, et al., "Highly accurate protein structure prediction for the human proteome," Nature, vol. 596, 2021.
[17] J. Watson, D. Juergens, N. Bennett, B. Trippe, J. Yim, H. Eisenach, W. Ahern, A. Borst, R. Ragotte, L. Milles, et al., "De novo design of protein structure and function with RFdiffusion," Nature, vol. 620, 2023.
[18] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, et al., "Robust deep learning-based protein sequence design using ProteinMPNN," Science, vol. 378, 2022.
[19] B. L. Hie, S. Candido, Z. Lin, O. Kabeli, R. Rao, N. Smetanin, T. Sercu, and A. Rives, "A high-level programming language for generative protein design," bioRxiv preprint, 2022.
[20] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, and J. Peng, "High-resolution de novo structure prediction from primary sequence," bioRxiv preprint, 2022.
[21] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, and D. Baker, "Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models," bioRxiv preprint, 2022.
[22] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. B. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, "Protein structure determination using metagenome sequence data," Science, vol. 355, 2017.
[23] D. Baker and A. Sali, "Protein structure prediction and structural genomics," Science, vol. 294, 2001.
[24] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Research, vol. 28, 2000.
[25] M. Baek and D. Baker, "Deep learning and protein structure modeling," Nature Methods, vol. 19, 2022.
[26] A. Pakhrin, B. Shrestha, B. Adhikari, and D. B. Kc, "Deep learning-based advances in protein structure prediction," International Journal of Molecular Sciences, vol. 22, 2021.
[27] C. Robertson, S. Yeh, and R. Das, "Deep learning-guided design of dynamic proteins," Nature Reviews Molecular Cell Biology, vol. 24, 2023.
[28] T. Bepler and B. Berger, "Learning the protein language: Evolution, structure, and function," Cell Systems, vol. 12, 2021.
[29] M. AlQuraishi, "End-to-end differentiable learning of protein structure," Cell Systems, vol. 8, 2019.
[30] S. Alford, A. Leaver-Fay, J. Jeliazkov, M. O'Meara, F. DiMaio, H. Park, M. Shapovalov, P. Renfrew, V. Mulligan, K. Kappel, et al., "The Rosetta all-atom energy function for macromolecular modeling and design," Journal of Chemical Theory and Computation, vol. 13, 2017.
[31] D. Akdel, D. E. V. Pires, E. Porta Pardo, J. Janes, A. O. Zalevsky, B. Meszaros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, et al., "A structural biology community assessment of AlphaFold2 applications," Nature Structural & Molecular Biology, vol. 29, 2022.
[32] M. Baek, I. Anishchenko, I. R. Humphreys, Q. Cong, D. Baker, and F. DiMaio, "Efficient and accurate prediction of protein structure using RoseTTAFold2," bioRxiv preprint, 2023.
[33] P. Bryant, G. Pozzati, and A. Elofsson, "Improved prediction of protein-protein interactions using AlphaFold2," Nature Communications, vol. 13, 2022.
[34] R. Evans et al., "Protein complex prediction with AlphaFold-Multimer," bioRxiv preprint, 2021.
[35] M. Kryshtafovych, K. Fidelis, and J. Moult, "Progress from CASP1 to CASP14," Proteins, vol. 89, 2021.
[36] J. Haas, A. Barbato, D. Behringer, G. Studer, S. Roth, M. Bertoni, K. Mostaguir, R. Gumienny, and T. Schwede, "Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12," Proteins, vol. 86, 2018.
[37] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests," Bioinformatics, vol. 29, 2013.
[38] S. Basu and B. Wallner, "DockQ: A quality measure for protein-protein docking models," PLOS ONE, vol. 11, 2016.
[39] A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. Chen, V. Vacic, Z. Obradovic, and V. N. Uversky, "The unfoldomics decade:

An update on intrinsically disordered proteins," BMC Genomics, vol. 9, 2008.
[40] J. A. Ruff and R. Pappu, "AlphaFold and implications for intrinsically disordered proteins," Journal of Molecular Biology, vol. 433, 2021.
[41] C. B. Anfinsen, "Principles that govern the folding of protein chains," Science, vol. 181, 1973.
[42] H. S. Chan, S. Zhang, S. Takada, and K. A. Dill, "Cooperativity, local-nonlocal coupling, and nonnative interactions: Principles of protein folding from coarse-grained models," Annual Review of Physical Chemistry, vol. 62, 2011.