https://doi.org/10.1038/s42003-024-07113-5

The Nobel Prize in Chemistry: past, present, and future of AI in biology

Luciano A. Abriata

The work by Hassabis and Jumper on protein structure prediction together with Baker's supremacy in de novo protein design set the stage for a future where AI not only deciphers biology at the atomic level but also designs new molecules for biotechnology, medicine, and beyond. I provide an overview of the recent past, the present, and the future of AI in structural biology, from how it all started with the Critical Assessment of Structure Prediction (CASP) experiments and a protein engineering lab, to how the field could further evolve with AI models that eventually "understand" biology holistically.

The 2024 Nobel Prize in Chemistry, awarded to Demis Hassabis & John Jumper from Deepmind and David Baker from the Institute for Protein Design at the University of Washington, recognizes transformative achievements in artificial intelligence-driven protein structure prediction and design. It certainly ushers in a new era for chemistry and biology, in particular acknowledging the profound impact of artificial intelligence (AI) on scientific research and on practical applications across disciplines, which was also acknowledged more broadly with the 2024 Nobel Prize in Physics.

At the core of the advancements behind the Chemistry prize is a full computational understanding of living matter at atomic level, particularly by AI models capable of predicting, analyzing and designing the 3D structures of proteins, alone or more recently forming complexes with other molecules such as nucleic acids, ions, and small ligands. Such capabilities tackle one of biology's most enduring challenges and are the reason why the irruption of AI in the field represented a revolution already since AlphaFold 2 "won" the 14th edition of CASP in 2020. After seeing at best some incremental improvements for nearly 25 years, CASP was finally giving its first really fleshy and tasty fruits.

Artificial neural networks that understand biomolecules

Deepmind had already entered CASP in its previous edition with its AlphaFold (version 1) model, which engineered the most out of the same techniques that the top academic groups were applying at the time, all mainly capitalizing on the recent breakthroughs in residue contact (and distance and orientation) prediction from multiple sequence alignments (MSAs) through coevolution calculations[1]. AlphaFold 2 was not at all a new version, but rather a full redesign and rethinking of the protein structure prediction problem, whose performance left scientists both in awe and initially frustrated after which a period of illumination came that changed the future of structural biology forever. It turns out the AlphaFold 2 paper[2] put forward several innovations that other scientists could subsequently build on. Two key innovations included an Evoformer module and the integration of attention mechanisms to model proteins as spatial graphs right as part of the AI model itself, unlike all other methods - including the first AlphaFold model - which only predicted contacts, distances and angles that were then fed into a regular proteinfolding program. In particular, the Evoformer module allowed AlphaFold 2 to process multiple sequence alignments to extract coevolutionary information in an indirect way that made the system more tolerant to problems in the MSAs. The attention mechanisms in turn allowed the system to process evolutionary relationships and physical interactions between distant residues, enabling highly accurate 3D predictions even for protein complexes. Importantly, too, the integration of the structure calculation stage as part of the neural network itself connected fluently (in mathematical terms) the input data (sequences, alignments, and 3D structures of candidate templates) with the outputs (modeled structures together with various confidence scores). This meant that the system could be run iteratively to better process the information and achieve better convergence. It was also critical from the users' points of view that AlphaFold 2 returned not just structural models but also various metrics (global Tm score, residue-wise pLDDT, and pairwise PAE scores) that are reporting the quality of its own predictions - something that CASP had always pushed for but rarely assessed[1].

The availability of such a powerful tool as AlphaFold 2 meant a dramatic acceleration of all research in structural biology, as thousands of previously unknown protein structures became accessible through computational means, especially when backed up by high-quality metrics. Rapidly, Deepmind paired with the European Bioinformatics Institute[3] to produce millions of structural models that soon became available as part of UniProt and the Protein Data Bank themselves. Far from competing against experimental structure determination methods, AlphaFold 2 became their perfect ally, boosting the efficiency of scientists and software processing experimental data by orders of magnitude. Already in CASP14, when AlphaFold 2 came out, its models helped to solve the phase problem on X-ray diffraction data available for some of the targets[4]; Cryo-EM structures can now be solved much faster when at least parts of the volumetric maps can be filled with AlphaFold 2 models to then optimize conformations as the experimental densities are fit[5]; and NMR structure determination was driven to almost full automation by tools like NMRtist especially when assisted with reliable AlphaFold 2 models[6].

Beyond AlphaFold 2's direct applications, the number of new concepts, methods and algorithms presented by the AlphaFold 2 paper inspired many academic and private groups to either recycle, build on, or adapt the new knowledge and tools into their own methods and software. That is how a burst of new tools for computational structural biology came about that facilitated all kinds of studies on biomolecular structures, from predicting interacting surfaces[7,8] or stabilizing mutations[9] given a structure to filling them with ligands[10], modeling the 3D structures of RNA (although notably, non-AI methods seemed to perform best in CASP's only assessment for RNA folding[11]), predicting structures of proteins complexed with non-protein molecules (pioneered by Baker with RoseTTAFold-AllAtoms[12]), processing MSAs to explore protein structure and evolution[13], and designing proteins in whole new ways[14–17]—the latter developed below.

Expanding AI to all biomolecules

While the initial focus of AlphaFold and similar models was on predicting the structure of proteins, the latest advancements have expanded their scope to other biomolecules, including nucleic acids, ions, lipids, and other small molecules. This broader application marks a critical shift from studying proteins in isolation to modeling complex molecular environments, and promises a new revolution in biology as the new generation of AI models can essentially understand all the different kinds of molecules and interactions relevant to biology.

The first program capable of parsing and modeling more than protein atoms was RoseTTAFold-AllAtoms from the Baker lab[12]. Then, AlphaFold 3[18] came out in an extremely simple-to-use web server within the Google domain, but with serious limitations that the community did not welcome: no source code, only a limited number of jobs per day and only for academic not-for-profit work, and only handling a limited set of small molecules and ions despite the program's intrinsic capabilities to actually handle, in theory at least, any small molecule. Today, new programs are coming out that incorporate these "all atoms" functionalities in more permissive ways, such as Chai-1[19] from Chai Discovery, which can be executed locally or through a web interface similar to AlphaFold 3's but without limitations on the small molecule inputs, accepting any molecule provided as a Simplified Molecular Input Line Entry System (SMILES) string.

These all-atoms models not only advance ways to model life at atomic level like never before, but also stand as new ways for computers to assist drug development. While the "canonical" protocol for testing whether a ligand binds to a target protein involves knowing their 3D structures and sampling possible binding poses in silico, with little hope for any required conformational changes to take place during the docking procedure, the new AI models can simultaneously sample ligand and protein target conformations as they are "co-folded". As these all-atoms models become more efficient, we can expect a shift toward AI-driven drug discovery through the "co-folding method". This will have profound implications for the pharma, biotechnology and healthcare, likely reducing costs and experimental research time in drug development pipelines. This application is so important that various companies are working on it and CASP started dedicating a specific track to this problem since its 15th edition[20].

Understanding protein structure enables protein engineering

Prof. David Baker's pioneering efforts in de novo protein design[21], initially void of any AI methods at its core but in the last years largely relying on them especially through its RoseTTAFold[12] and MPNN methods[14,22], set the stage for a future where AI not only deciphers natural biology but also designs new molecular entities for use in biotechnology, medicine, and beyond. Baker's group at the University of Washington's Institute for Protein Design pioneered methods to create novel proteins from scratch, a feat that became significantly more powerful with the advent of AI - especially diffusion models to design protein conformations in space[23] and message-passing neural networks to produce sequences that fold into the designed structures[14,22]. Along key proof of concept and concrete applications of these AI-based tools from the Baker lab, we count with efficiently designing new enzymes[24] or stabilizing existing ones[25], crafting complex multiprotein assemblies[26], designing multi-state proteins[27], engineering binders with therapeutic applications, and building protein crystals of use in material sciences, to mention some notable examples.

Proteins designed with AI methods are already proving powerful, for example as multivalent single-chain proteins of potential use as vaccines[28], as soluble analogs of membrane proteins to facilitate their study[29], and as high-affinity binders useful for therapies or as sensors[30]. Broader applications include protein function regulation through designed binders, even engineered clinical antibodies[31], enzyme stabilization and computational evolution[32], etc.

The future of AI in structural biology and of "holistic" AI models for biology

CASP16 is now rolling, with results expected for late 2024 and promising an assessment of the state of the art of structure prediction beyond static tertiary protein structures. As CASP15 revealed, modeling of multimeric assemblies still needs some tweaks, and now that protein-only modeling is close to solved, the new frontiers await modeling ligand binding to proteins, multiple protein conformations, and nucleic acid folding, all already tackled in CASP15. Besides, CASP16 reintroduced the track assessing integrative modeling, which involves modeling typically large multicomponent complexes from sparse and varied data and after years of rather poor results[33] could take new heights as AI methods step in. All these special evaluation tracks in CASP16 point at the direction in which the field of computational structural biology will progress next, likely also carrying along that of protein design and, importantly, of small molecule discovery and drug development.

Another big piece of the AI-for-biology picture is that of multimodal foundational models for biology trained on, for the moment, massive amounts of DNA, RNA and protein sequences. Training on protein sequences "only" already proved useful to predict protein structures and detect structure-consistent evolutionary relationships, with Meta's ESM-Fold at the pinnacle[13]. Meanwhile, foundational models centered around Biology's central dogma hold promise for new applications in genomics, transcriptomics and proteomics[34].

Next, multimodal foundational models that span also molecular structure are perfectly foreseeable with current technologies. Such models could bring a whole new series of tools to interrogate and understand biology holistically, for example explaining complex changes in gene expression patterns in molecular and structural terms and then inferring what molecular effectors could restore the disrupted pathways.

Luciano A. Abriata

School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. e-mail: luciano.abriata@epfl.ch

Received: 16 October 2024; Accepted: 21 October 2024;
Published online: 29 October 2024

References

1. Abriata, L. A., Tamò, G. E. & Dal Peraro, M. A Further Leap of Improvement in Tertiary Structure Prediction in CASP13 Prompts New Routes for Future Assessments. *Proteins Struct. Funct. Bioinforma.* **87**, 1100–1112 (2019).
2. Jumper, J. et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Varadi, M. et al. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
4. Millán, C. et al. Assessing the Utility of CASP14 Models for Molecular Replacement. *Proteins Struct. Funct. Bioinforma.* **89**, 1752–1769 (2021).
5. Terwilliger, T. C. et al. Improved AlphaFold Modeling with Implicit Experimental Information. *Nat. Methods* **19**, 1376–1382 (2022).

6. Klukowski, P., Riek, R. & Güntert, P. Time-Optimized Protein NMR Assignment with an Integrative Deep Learning Approach Using AlphaFold and Chemical Shift Prediction. *Sci. Adv.* **9**, eadi9323 (2023).
7. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. PeSTo: Parameter-Free Geometric Deep Learning for Accurate Prediction of Protein Binding Interfaces. *Nat. Commun.* **14**, 2175 (2023).
8. Gainza, P. et al. De Novo Design of Protein Interactions with Learned Surface Fingerprints. *Nature* **617**, 176–184 (2023).
9. Diaz, D. J. et al. Stability Oracle: A Structure-Based Graph-Transformer Framework for Identifying Stabilizing Mutations. *Nat. Commun.* **15**, 6170 (2024).
10. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: Enriching AlphaFold Models with Ligands and Cofactors. *Nat. Methods* **20**, 205–213 (2023).
11. Das, R. et al. Assessment of Three-Dimensional RNA Structure Prediction in CASP15. *bioRxiv*, https://www.biorxiv.org/content/10.1101/2023.04.25.538330v1 (2023).
12. Generalized biomolecular modeling and design with RoseTTAFold All-Atom | Science. https://www.science.org/doi/abs/10.1126/science.adl2528. Accessed 2024-10-13.
13. Lin, Z. et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **379**, 1123–1130 (2023).
14. Dauparas, J. et al. Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN. *Science* **378**, 49–56 (2022).
15. Krapp, L. F. et al. Context-Aware Geometric Deep Learning for Protein Sequence Design. *Nat. Commun.* **15**, 6273 (2024).
16. Ingraham, J. B. et al. Illuminating Protein Space with a Programmable Generative Model. *Nature* **623**, 1070–1078 (2023).
17. Pacesa, M., et al. BindCraft: One-Shot Design of Functional Protein Binders. bioRxiv, https://doi.org/10.1101/2024.09.30.615802 (2024).
18. Abramson, J. et al. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
19. Discovery (Chai), C., et al. Chai-1: Decoding the Molecular Interactions of Life. *bioRxiv*, https://doi.org/10.1101/2024.10.10.615955 (2024).
20. Robin, X. et al. Assessment of Protein-Ligand Complexes in CASP15. *Proteins* **91**, 1811–1821 (2023).
21. Kuhlman, B. et al. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302**, 1364–1368 (2003).
22. Dauparas, J., et al. Atomic Context-Conditioned Protein Sequence Design Using LigandMPNN. *bioRxiv*, https://doi.org/10.1101/2023.12.22.573103 (2024).
23. Watson, J. L. et al. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
24. Yeh, A. H.-W. et al. De Novo Design of Luciferases Using Deep Learning. *Nature* **614**, 774–780 (2023).
25. Sumida, K. H. et al. Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
26. Wicky, B. I. M. et al. Hallucinating Symmetric Protein Assemblies. *Science* **378**, 56–61 (2022).
27. Lisanza, S. L., et al. Multistate and Functional Protein Design Using RoseTTAFold Sequence Space Diffusion. *Nat. Biotechnol*. 1–11, https://doi.org/10.1038/s41587-024-02395-w (2024).
28. Castro, K. M., et al. Accurate Single Domain Scaffolding of Three Non-Overlapping Protein Epitopes Using Deep Learning. *bioRxiv*, https://doi.org/10.1101/2024.05.07.592871 (2024).
29. Goverde, C. A. et al. Computational Design of Soluble and Functional Membrane Protein Analogues. *Nature* **631**, 449–458 (2024).
30. Vázquez Torres, S. et al. De Novo Design of High-Affinity Binders of Bioactive Helical Peptides. *Nature* **626**, 435–442 (2024).
31. Shanker, V. R., Bruun, T. U. J., Hie, B. L. & Kim, P. S. Unsupervised Evolution of Protein and Antibody Complexes with a Structure-Informed Language Model. *Science* **385**, 46–53 (2024).
32. Yang, J., Li, F.-Z. & Arnold, F. H. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. *ACS Cent. Sci.* **10**, 226–241 (2024).
33. Tamò, G. E., Abriata, L. A., Fonti, G. & Dal Peraro, M. Assessment of Data-Assisted Prediction by Inclusion of Crosslinking/Mass-Spectrometry and Small Angle X-Ray Scattering Data in the 12th Critical Assessment of Protein Structure Prediction Experiment. *Proteins Struct. Funct. Bioinforma.* **86**, 215–227 (2018).
34. Nguyen, E., et al. Sequence Modeling and Design from Molecular to Genome Scale with Evo. *bioRxiv*, https://doi.org/10.1101/2024.02.27.582234 (2024).

Competing interests

L.A.A. is an Editorial Board Member for *Communications Biology*, but was not involved in the editorial review of, nor the decision to publish this article. The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Luciano A. Abriata.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.