Using AI for Protein Structure Prediction and Binding in Biochemistry

Eli Whiteman-Owens, Sofia Cummings, Yadira Guerrero, Madi Hoofard, Dr. Frank E. Cappuccio, Dr. Jenny A. Cappuccio

[1] CHEM 434L - Biochemistry I Lab - Department of Chemistry & Biochemistry, Humboldt State University, Arcata, CA, USA

Summary

Advances in artificial intelligence (AI) have enabled highly accurate prediction of protein tertiary structures from primary amino acid sequences. These programs have significant implications for fields such as drug discovery and biofuel engineering, as they allow for reliable modeling of protein–ligand interactions and facilitate in silico docking simulations. By improving our ability to predict how small molecules interact with protein targets, AI driven structure prediction tools accelerate the identification and optimization of functional biomolecular systems.

Introduction

- In order for proteins to perform their necessary functions they must achieve a tertiary three-dimensional structure

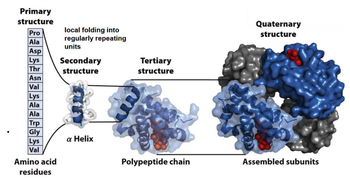


Figure 1. 4 leves of protein structure.
1° amino acid sequence
2° common folding structures
3° complete fold of one polypeptide chain
4° complet field of multiple polypeptides in a complex.
Principles of Biochemistry 2021

- The X-ray crystallography technique is responsible for solving approximately 85% of the structures we know today, but use of artificial intelligence (AI) is now informing and surpassing protein and nucleic acid structure prediction.
- AlphaFold3 (AF3 - Google DeepMind) revolutionized structure prediction for protein-ligand complexes, surpassing conventional docking tools in accuracy.
- Compared to AlphaFold2, AF3 extends past protein predictions via molecular recognition processes by generative diffusion models.
- Standardized assessment based on experimentally determined structures (CASP15)
- Training for AF2 came from PDB which is used to prove confidence metrics (pLDDT, pTM, and PAE) and enables critical interpretation by user
- AF2 utilizes neural network multiple sequence alignment (MSA) in order to assist in developing evolutionary relationships between different proteins
- AF3 requires tokens to process job requests. These tokens are used to assist in confidence metrics as opposed to actual residues. Tokens also facilitate the determination of protein structure size
- AF3 is limited in its ability to determine protein dynamics, physics of folding or binding, covalent bonds for general ligands, and hallucinatory structures. These limitations are displayed via poor confidence metric scores.
- Protenix is a comprehensive reproduction of AF3, targeted for predicting the structures and complex interactions involving proteins, ligands, and nucleic acids.
- These processes can be highly beneficial in early drug discovery in the context of structure-based drug design.
- Use of generative AI possibly creates a substantially cost-efficient, accessable, and mechanically safer way to visualize proteins in an academic setting in comparison to physical techniques like x-ray crystallography, cryo-EM, & NMR spectroscopy.

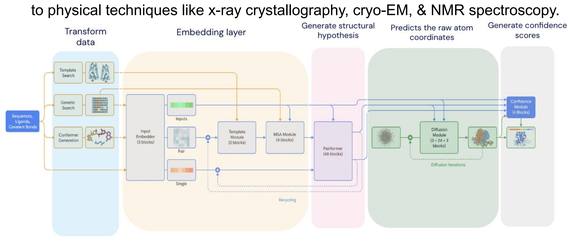
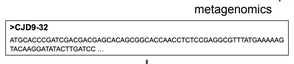


Figure 2. Inputs and outputs of the Alphafold 3 AI
Biomolecular inputs (amino acid sequences, ligands, and covalent interactions) are processed through MSA and template based feature generation and encoded into sequence and pairwise representations. These are refined by the Pairformer network to produce a structural hypothesis, which is iteratively refined by a generative diffusion model to predict atomic coordinates. Final structures are evaluated using confidence metrics, including predicted Local Distance Difference Test (pLDDT) and Predicted Aligned Error (PAE), providing estimates of local and global prediction reliability.

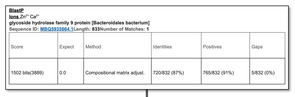
Methods

- Nucleic acid sequence obtained from metagenomics



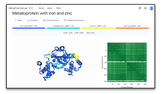
- Using EXPASY nucleic acid sequence was translated to primary structure and theoretical pI/M.W. were calculated

- Amino acid sequence inserted to BLASTP to find most likely ions based on nearest neighbor

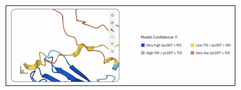


- Translated primary structure is the imported into AlphaFold server

Predicted structure generated with ions of nearest neighbor



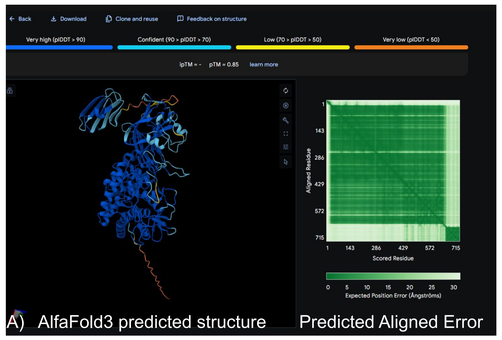
- pLDDT measures confidence in the local structure, estimating how well the prediction would agree with an experimental structure.



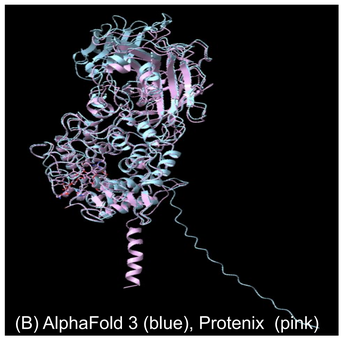
- PAE is a measure of how confident AlphaFold2 is in the relative position of two residues within the predicted structure.

Results

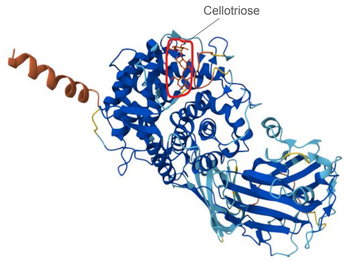
- **Sequence CJD 8-11** is predicted to be a glycosyl hydrolase family 8 enzyme with a 90% identity to *Fibrobacter succinogenes* based on BLASTp sequence homology analysis.
- Classified as an endocellulase (EC 3.2.1.4)
- Endocellulase cleave internal β(1-4) glycosidic bonds in cellulose creating new chain ends.
- This data can be utilized in the lab to recombinantly express the enzyme in *E. coli* using a plasmid expression vector
- Enzyme activity can then be assessed and potentially be useful in the production of cellulosic bioethanol.



A) AlfaFold3 predicted structure | Predicted Aligned Error



(B) AlphaFold 3 (blue), Protenix (pink)

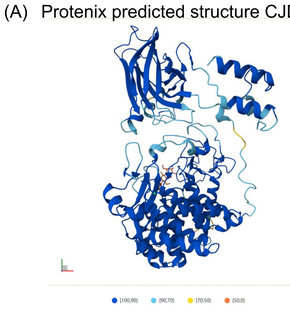


(C) Protenix predicted structure bound to cellotriose substrate

***Figure 3.* Predicted structure and comparison of sequence CJD 8-11**
(A) AlfaFold3 predicted structure, showing matched pLDDT scores above color coded to secondary structures and PAE values on right. (B) Matched overlay of AlfaFold3 (blue) and Protenix (pink) structures in ChimeraX. (C) Prontenix predicted structure with docked cellotriose (outlined in red).

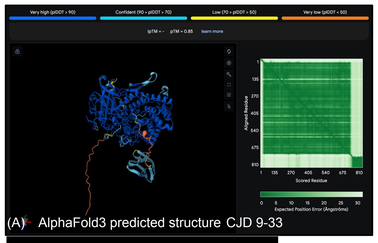
Results

(A) Protenix predicted structure CJD 9-32

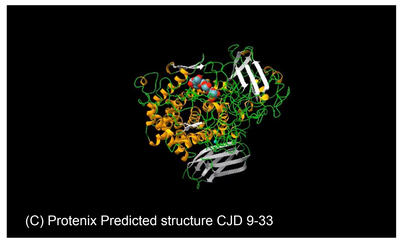


(B) CJD 9-32 cellulose & docked cellotriose

***Figure 4.* Protenix predicted structure and visualization of sequence CJD 9-32 with docked small ligand**
(A) Protenix predicted structure with pLDDT scores below and PAE values on right. (B) ChimeraX visualization of Protenix structure with Ca^{2+} (green) and Zn^{2+} (yellow) as well as docked cellotriose (cyan).



(A) AlphaFold3 predicted structure CJD 9-33



(C) Protenix Predicted structure CJD 9-33

(B) ChimeraX visualization of Alfafold3 predicted structure CJD 9-33

***Figure 5.* Predicted structure of sequence CJD 9-33. Endoglucanase. Class: Hydrolase Ions: Ca2+ and Mg2+**
(A) AlpaFold3 predicted structure showing color coded pLDDT (above) and PAE (right) confidence metrics. (B) ChimeraX visualization of the AlphaFold3 predicted structure. (C) ChimeraX visualization of Protenix predicted structure with Ca^{2+} (cyan) and Mg^{2+} (olive) bound to cellotriose (blue and red)

Conclusions

- Key differences in predictive performance indicate that AF3 demonstrates highly accurate side chain modeling and strong ligand docking
- Protenix shows higher confidence in backbone structures and protein–protein interactions, though with some low confidence structural artifacts.
- Overall, AI-driven tools are accelerating drug discovery by enabling rapid prediction of protein structures and interactions, supporting applications such as antibody development.
- **Limitations in modeling flexible ligands and allosteric systems highlight the importance of use with experimental methods such as X-ray crystallography, cryo-EM, and NMR.**

Acknowledgments

Cellulase sequences were provided by M. Hess & M. Escobar through a JGI JOINT GENOME INSTITUTE A DOE OFFICE OF SCIENCE USER FACILITY tech and the Joint Genome Institute.

The AlphaFold 3 server can be found at Google DeepMind EMBL-EBI PDBe Protein Data Bank in Europe
https://alphafoldserver.com

The Protenix server can be found at
https://protenix-server.com/login
By PXDesign: Fast, Modular, and Accurate De Novo Design of Protein Binders

References

(1) Hess et al. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science* 2011, *331* (6016), 463–467. https://doi.org/10.1126/science.1200387.
(2) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Tools for Structure Building and Analysis. *Protein Science: A Publication of the Protein Society* 2023, *32* (11), e4792. https://doi.org/10.1002/pro.4792.
(3) Morehead, Alex, et al. "Artificial Intelligence Methods for Protein Structure and Interaction Prediction: Recent Advances and Challenges." *Current Opinion in Structural Biology*, vol. 98, June 2026, p. 103247. www.sciencedirect.com/science/article/pii/S0959440X26000291, https://doi.org/10.1016/j.sbi.2026.103247. Accessed 26 Mar. 2026.
(4) Abramson, J., et al. "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3." *Nature*, vol. 630, no. 630, 8 May 2024, pp. 493–500. www.nature.com/articles/s41586-024-07487-w, https://doi.org/10.1038/s41586-024-07487-w.
(5) Graille, Marc, et al. "Best Practices of Using AI-Based Models in Crystallography and Their Impact in Structural Biology." *Journal of Chemical Information and Modeling*, vol. 63, no. 12, 12 June 2023, pp. 3637–3646, https://doi.org/10.1021/acs.jcim.3c00381.
(6) Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J. and Bodenstein, S.W., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pp.1-3.
(7) wwPDB consortium. (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 47: D520-D528 doi: https://doi.org/10.1093/nar/gky949