Received: 19 October 2023 | Revised: 11 January 2024 | Accepted: 12 January 2024

DOI: 10.1002/pmic.202300397

DATASET BRIEF

Proteomics
Proteomics and Systems Biology

An AI-generated proteome-scale dataset of predicted protein structures for the ctenophore *Mnemiopsis leidyi*

R. Travis Moreland[1] | **Suiyuan Zhang**[1] | **Sofia N. Barreira**[1] | **Joseph F. Ryan**[2] | **Andreas D. Baxevanis**[1]

[1]Center for Genomics and Data Science Research, Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA

[2]Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, Florida, USA

Correspondence
Andreas D. Baxevanis, National Human Genome Research Institute, National Institutes of Health, Building 12A, Room 4009, Bethesda, MD 20892 USA.
Email: andy@mail.nih.gov

Funding information
National Human Genome Research Institute, National Institutes of Health, Grant/Award Number: ZIA HG000140; Allen Distinguished Investigator Award, Paul G. Allen Frontiers Group, Paul G. Allen Family Foundation

Abstract

This Dataset Brief describes the computational prediction of protein structures for the ctenophore *Mnemiopsis leidyi*. Here, we report the proteome-scale generation of 15,333 protein structure predictions using AlphaFold, as well as an updated implementation of publicly available search, manipulation, and visualization tools for these protein structure predictions through the *Mnemiopsis* Genome Project Portal (https://research.nhgri.nih.gov/mnemiopsis). The utility of these predictions is demonstrated by highlighting comparisons to experimentally determined structures for the light-sensitive protein mnemiopsin 1 and the ionotropic glutamate receptor (iGluR). The application of these novel protein structure prediction methods will serve to further position non-bilaterian species such as *Mnemiopsis* as powerful model systems for the study of early animal evolution and human health.

KEYWORDS
evolutionary biology, protein structure prediction, structural annotation, visualization

The question of how proteins fold based on their amino acid sequences has been a long-standing challenge in the field of structural biology and, by extension, to the study of key biological questions in fields such as developmental and evolutionary biology. Historically, tertiary protein structures have been experimentally determined using labor-intensive techniques such as x-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (CryoEM) [1]. Further, many large proteins are simply too large to be analyzed by standard NMR approaches or they may be too disordered, making them poor candidates for study via traditional x-ray crystallographic methods [2, 3]. While breakthroughs in genomic technologies have led to the generation of chromosome-length sequencing data for an ever-increasing number of biologically informative species, the gap between sequence-based and structure-based data continues to grow, hampering the ability of investigators to look beyond the traditional set of model organisms to answer key questions in human biology and human health [4–5].

Fortunately, recent developments in artificial intelligence (AI) are advancing the application of neural network and machine learning methodologies to the analysis of whole-genome sequencing data, particularly through the use of predictive methods such as AlphaFold [1–3, 6–7] to generate structures for the proteins encoded within these genomes. That said, the significant amount of computational power required to generate these structural data continues to pose a significant roadblock to large-scale studies. Our access to NIH's Biowulf supercomputing resource has allowed us to overcome this barrier and apply these new methodologies to our whole-genome sequencing data of the ctenophore *Mnemiopsis leidyi* [8]. Given that ctenophore species represent the earliest-branching extant animals, they are evolutionarily informative and provide important insights into the evolution of animal multicellularity [9–12]. Our structural predictions, which consumed over 58,000 GPU hours, have produced a valuable dataset that has the potential to advance evolutionary, developmental, and

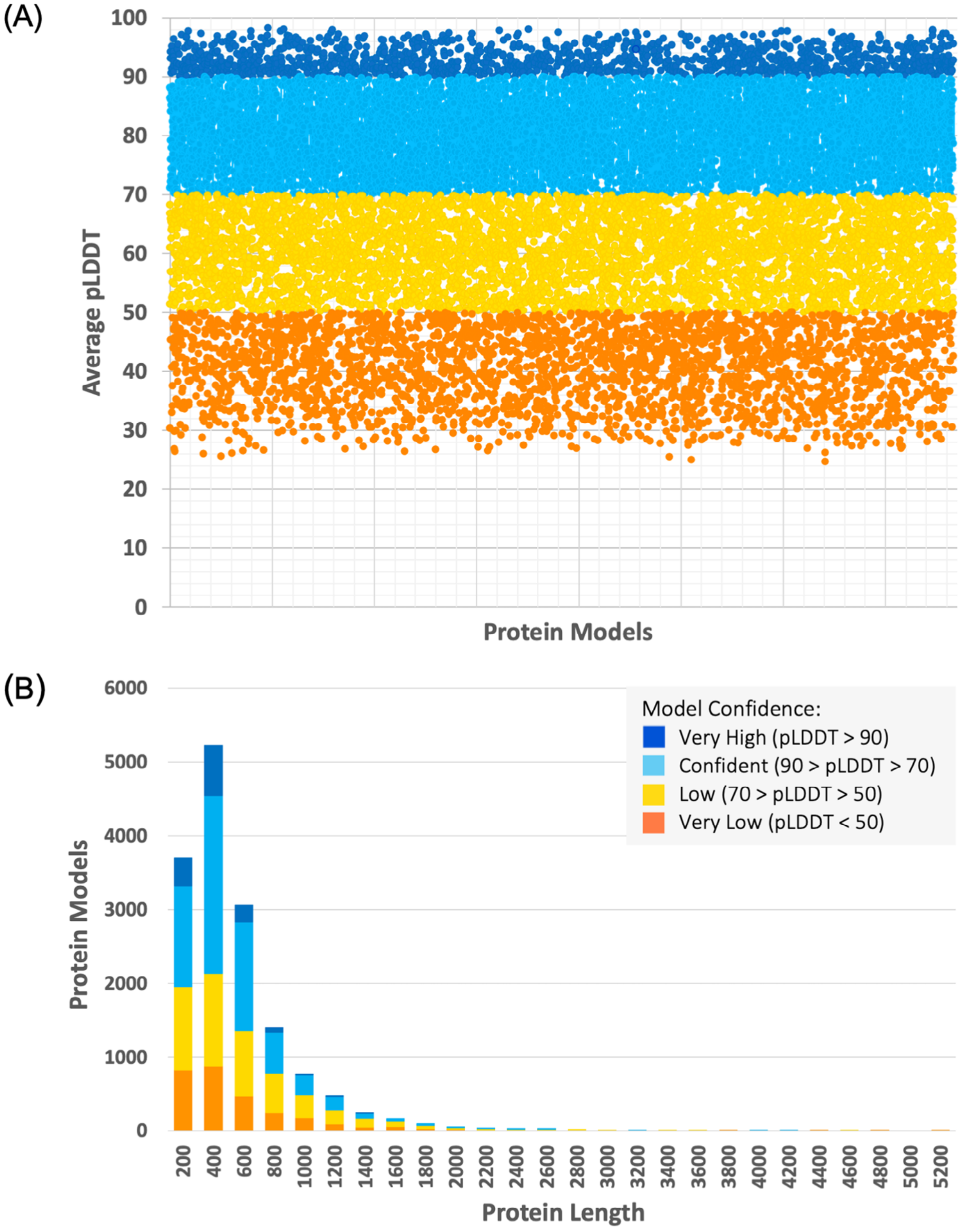
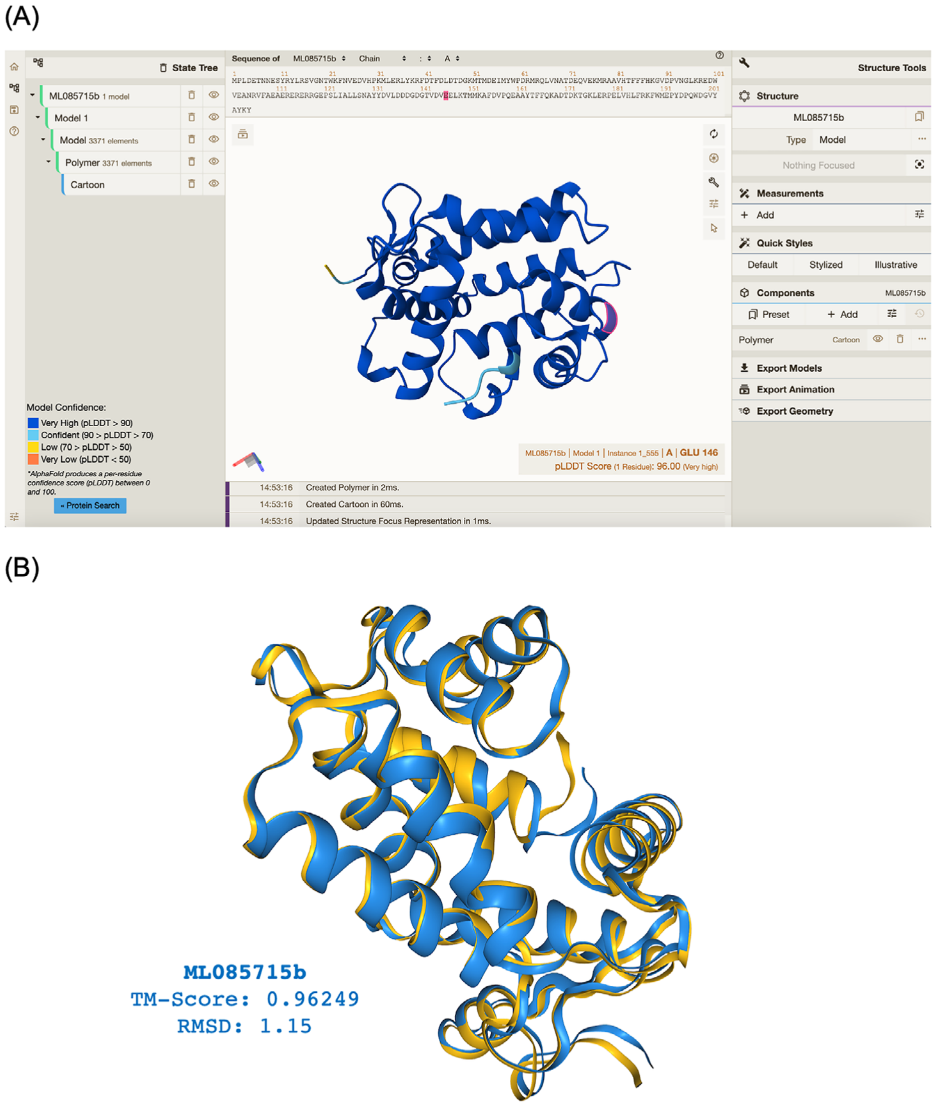


FIGURE 1 Predicted local distance difference test (pLDDT) scores for AlphaFold predictions of the *Mnemiopsis* proteome. (A) Scatterplot of average pLDDT scores for each of the 15,333 predicted protein models. Each point in the graphic represents the score for a single structure prediction. Approximately half of all predictions were above the 70% confidence threshold: 8.9% *Very High*, with pLDDT > 90 (dark blue); 41.7% *Confident*, with pLDDT between 70 and 90 (light blue). The remaining protein models fell below the 70% confidence threshold (30.1% *Low*, with pLDDT between 50 and 70; 19.3% *Very Low*, with pLDDT < 50). (B) Distribution of average model confidence as a function of protein length. Protein lengths are binned in incremental ranges of 200 amino acids. Within each group, the number of models falling into each confidence level are color-coded by pLDDT score. Higher confidence levels decrease as protein sequence length increases.

comparative genomic studies, as well as further positioning these non-traditional animal models as tractable models for studying important classes of human disease [4].

The full set of 16,548 *Mnemiopsis* protein models were downloaded from the *Mnemiopsis* Genome Project (MGP) Portal (https://research.nhgri.nih.gov/mnemiopsis) [13], filtering out 1212 possible gene joins from the initial gene prediction set. The remaining 15,336 protein sequences were then analyzed using AlphaFold version 2.3.1, resulting in the successful prediction of 15,333 protein structures covering 99.98% of the *Mnemiopsis* proteome. Each individual analysis produces a predicted local distance difference test (pLDDT) score that indicates, on a scale of 0 to 100, how well a predicted structure would agree with an experimentally determined structure [7], with higher scores indicating that the protein backbone is correctly predicted and that side chains are oriented properly. Here, the predicted structure confidence score distribution ranged from 24.7 to 98.2 (Figure 1A). Approximately half of all structure predictions were above the 70% confidence threshold (8.9% *Very High*, with a pLDDT > 90; 41.7% *Confident*, with a pLDDT



State Tree

ML085715b 1 model
Model 1
Model 3371 elements
Polymer 3371 elements
Cartoon

Model Confidence:
🟦 Very High (pLDDT > 90)
🟦 Confident (90 > pLDDT > 70)
🟨 Low (70 > pLDDT > 50)
🟧 Very Low (pLDDT < 50)
*AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100.

« Protein Search

Sequence of ML085715b Chain : A

```
1         11        21        31        41        51        61        71        81        91        101
MPLDETNNESYRYLRSVGNTWKFNVEDVHPKMLERLYKRFDTFDLDTDGKMTMDEIMYWPDRMRQLVNATDEQVEKMRAAVHTFFFHKGVDPVNGLKREDW
          111       121       131       141       151       161       171       181       191       201
VEANRVFAEAERERERRGEPSLIALLSNAYYDVLDDDGDGTVDVEELKTMMKAFDVPQEAAYTFFQKADTDKTGKLERPELVHLFRKFWMEPYDPQWDGVY
AYKY
```

ML085715b | Model 1 | Instance 1_555 | A | GLU 146
pLDDT Score (1 Residue): 96.00 (Very high)

```
14:53:16   Created Polymer in 2ms.
14:53:16   Created Cartoon in 60ms.
14:53:16   Updated Structure Focus Representation in 1ms.
```

Structure Tools

Structure
ML085715b
Type Model
Nothing Focused

Measurements
+ Add

Quick Styles
Default Stylized Illustrative

Components ML085715b
Preset + Add
Polymer Cartoon

Export Models
Export Animation
Export Geometry

FIGURE 2 Visualization and accuracy assessment of the bioluminescent protein mnemiopsin 1. (A) Searches for specific predicted protein structures can be performed by following the *View Protein Structures* link in the sidebar of the MGP Portal. Once a specific *Mnemiopsis* protein is selected, a new window depicting the predicted protein structure will appear. Users can zoom, rotate, stylize, and download the protein models. Shown here is the structure of ML085715b, the bioluminescent protein mnemiopsin 1. Models are color-coded based on per-residue model confidence scores (pLDDTs) determined by AlphaFold. In this case, most residues are indicated in dark blue to reflect a *Very High* pLDDT (96.06 on a scale of 100). (B) Foldseek-generated superimposition of the solved structure of the calcium-activated photoprotein mnemiopsis 1 (PDB:5VP3, yellow) with the AlphaFold-predicted structure of mnemiopsin photoprotein 1 ML085715b (blue).

between 70 and 90). The remaining protein models fell below the 70% confidence threshold (30.1% *Low*, with a pLDDT between 50 and 70; 19.3% *Very Low*, with a pLDDT < 50). To determine whether proteins having low pLDDT scores were ctenophore-specific, BLASTP searches were performed against UniProt using the sequences of the 152 protein structures having a pLDDT $\leq$ 30 as the query, yielding only 12 statistically significant hits. BLASTP searches of the remaining 140 proteins against the *Beroe ovata* protein database (http://ryanlab.whitney.ufl.edu/bovadb/) produced 133 statistically significant hits, suggesting that these proteins are indeed ctenophore-specific (Table S1).

In general, and as one would expect, as the length of a protein increases, so too does the computational time required to predict its structure. This was evident during our study, as the three proteins for which we were not able to successfully generate a protein structure were all greater than 5700 amino acids in length. Further, the relative percentages of pLDDT classified as *Very High* or *Confident* trends lower as the length of the protein increases (Figure 1B). Unsurprisingly, we also noted difficulty in predicting structures for both intrinsically disordered proteins (IDPs) and structured proteins having intrinsically disordered regions (IDPRs). On this point, Ruff et al. [14] previously reported that roughly 30% of the residues across predicted human protein structures tend to have pLDDT scores below 50, consistent with disorder estimates for the entire human proteome [15]. Mammalian proteomes generally contain 35%–45% IDPRs, making it

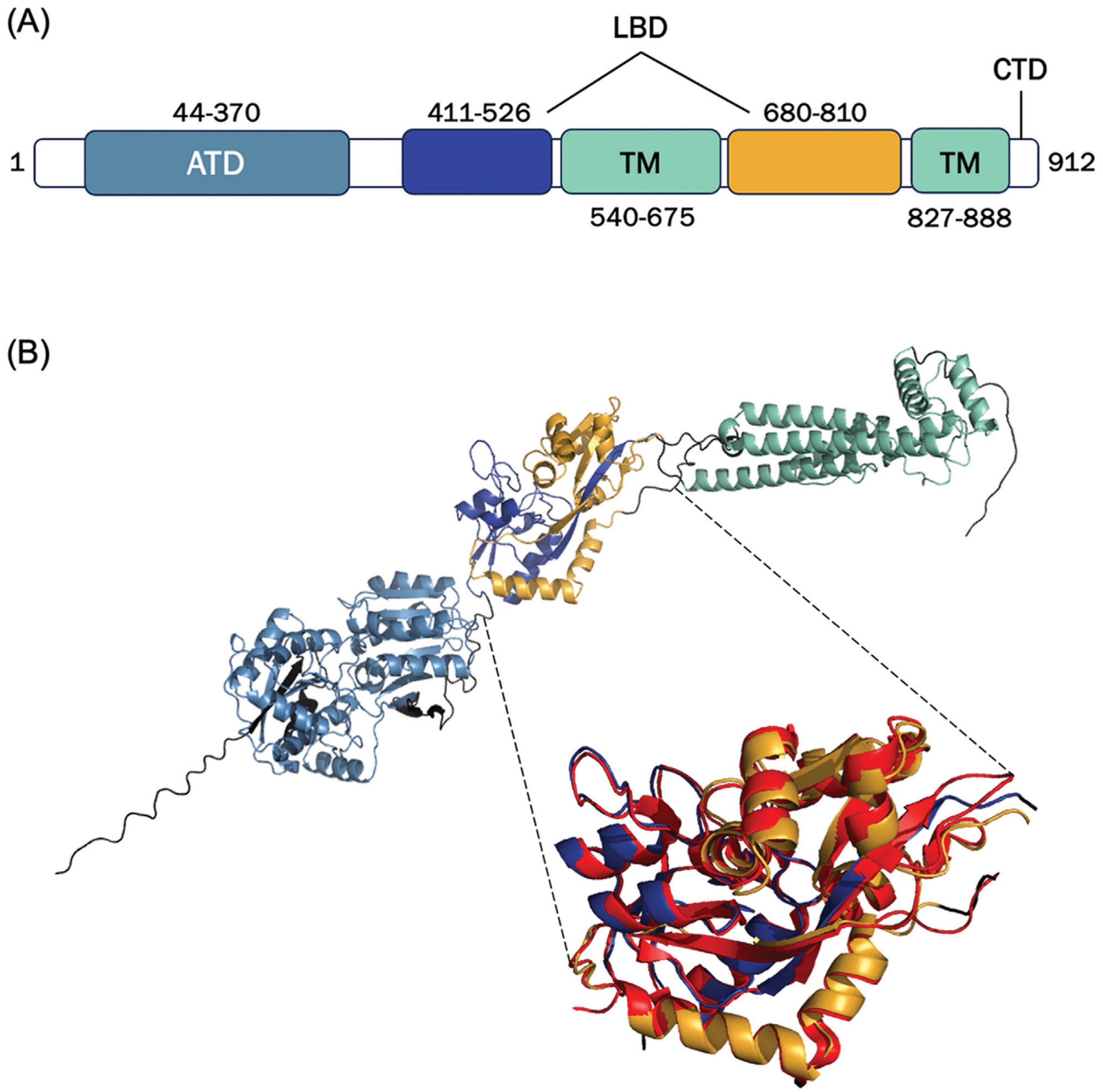


FIGURE 3 Comparison of the experimentally determined iGluR protein structure with a sequence-based AlphaFold prediction. (A) Domain architecture of the *Mnemiopsis* iGluR protein. ATD, amino-terminal domain; CTD, cytoplasmic C-terminal domain; LBD, ligand binding domain; TM, transmembrane regions. (B) The AlphaFold-predicted structure of the full-length ML032222a iGluR protein. The inset shows the superimposition of the experimentally determined structure of 4YKJ onto the prediction (RMSD = 0.924). Colors shown in the structures correspond to the colors used in the domain architecture schematic in Panel A. Structures were aligned and colored using PyMOL [28].

notoriously difficult to successfully predict high-confidence protein structures [14–16]. Accordingly, the proteome-scale structural predictions reported here should be used alongside additional functional data when considering future experimental design.

The protein structure predictions generated in the course of this study have been integrated into the publicly available *Mnemiopsis* Genome Project Portal and can be accessed through the *View Protein Structures* link in the left sidebar. Predicted protein structures can be visualized through a local implementation of Mol* Viewer [17] by searching for a *Mnemiopsis* protein identifier (e.g., ML085715b). Once the viewer shown in Figure 2A launches, users can zoom in and out, rotate, stylize, and download the structural models. Each predicted protein structure model is color-coded based on its pLDDT score. Clicking directly on a structure highlights the selected region in pink, with information on a selected amino acid (including its pLDDT score) shown in the box below the structure.

We assessed the accuracy of our AlphaFold-predicted structures using Foldseek which, rather than using a traditional (and computationally costly) superimposition approach, instead relies upon a structural alphabet based on tertiary interactions to deduce structural similarity and, by extension, assign putative protein function [18, 19]. The first of these is the calcium-activated photoprotein mnemiopsin 1, for which a crystal structure at 2.5 Å resolution was reported by Molakarimi et al. (PDB:5VP3) [20], the top sequence-based hit of this protein to the sequences in our *Mnemiopsis* database (ML085715b, mnemiopsin photoprotein 1). Our AlphaFold-generated PDB file for this *Mnemiopsis* photoprotein was then submitted to the Foldseek web server (https://search.foldseek.com) as a query against the PDB100 structural database using the program's 3Di/AA local alignment mode. Not surprisingly, as AlphaFold employs PDB data as a training set, the top hit to PDB100 was to the 5VP3 structure referenced above (100% identity, 5.94×10^{-27} *E*-value). The Foldseek-generated superimposition of the *Mnemiopsis* photoprotein with the experimentally determined photoprotein structure is illustrated in Figure 2B, showing a highly similar structural alignment with limited dissociation at the protein ends. Foldseek uses a template modeling (TM) score to indicate the degree of topological similarity between the two structures, a 0–1 scale where 1 represents a perfect match between the two structures being compared [21]. Here, the two compared structures had an almost perfect TM score (0.96). Coupled with the very low root-mean-square deviation (RMSD) of 1.15 between these two structures reported by Foldseek, as well as the *Very High* AlphaFold pLDDT average confidence score (96.06), AlphaFold has accurately predicted the protein structure of mnemiopsin 1 over its entire length.

The ionotropic glutamate receptors (iGluRs) are fast-acting ligand-gated ion channel receptors that have shown a remarkable evolutionary expansion. In humans, these proteins mediate fast excitatory synaptic transmission in the central nervous system and are located in both neuronal and non-neuronal cells [22]. The overall iGluR domain architecture consists of an amino-terminal domain, a ligand-binding domain (LBD) that is bisected by a transmembrane domain containing a pore-loop ion channel, a second transmembrane domain, and a carboxy-terminal domain (Figure 3A). The initial annotation of the *Mnemiopsis* genome identified 16 candidate iGluR genes [8]. Subsequently, Alberstein et al. [23] crystallized the LBD for one such gene product, ML032222a, then experimentally determined its structure by x-ray diffraction [23]. This structure was used to establish the molecular mechanism for the selective binding of glycine (rather than glutamate) via a ctenophore-specific interdomain salt bridge, as well as identify structural similarities to N-methyl-D-aspartate (NMDA) receptors that play an important role in excitatory neurotransmission, defects in which have been implicated in numerous neurodegenerative and cognitive disorders [24–26]. Similar to the approach used in CASP assessments [27], the AlphaFold-generated PDB file based on the sequence of ML032222a was used as the Foldseek query to search the PDB100 structural database using the 3Di/AA local alignment mode. As expected, the top hits were to the corresponding LBD domain within the solved x-ray structures of iGluR in *Mnemiopsis*: 4YKJ, 5CMB, 4YKK, and 5CMC, all with *E*-values below 1×10^{-38}. Using AlphaFold, we were able to generate a full-length predicted structure for ML032222a with a *Confident* pLDDT score of 84.23 (Figure 3B). PyMOL (28) was then used to superimpose the structure of the highest-scoring experimentally determined structural variant (4YKJ) onto the AlphaFold structure prediction for ML032222a, as illustrated in the inset of Figure 3B; a detailed superimposition is provided in Figure S1. A significant degree of structural similarity can be seen between the experimentally determined LBD and the same region as predicted by AlphaFold (RMSD = 0.924). This comparison of a partially solved protein structure to a full-length structure prediction provides an illustrative example of how computationally predicted structures can build upon our current knowledge of protein structure, especially in cases where the labor-intensive nature of solving a protein structure experimentally may be prohibitive.

Taken together, the whole-genome sequencing data available for this emerging model organism, along with the structural dataset generated in the course of this study, provide a powerful example of how experimental design can move from sequence to structure—and, by extension, to *function*—through the application of accessible computational techniques such as AlphaFold. This study serves as a model for extending the utility of whole-genome sequence data being generated for an ever-increasing number of organisms at whole-chromosome scale, data that can form the foundation for future experimental studies across the spectrum of biomedical science.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the National Institutes of Health, National Human Genome Research Institute (ZIA HG000140 to A.D.B.) and by an Allen Distinguished Investigator Award, a Paul G. Allen Frontiers Group-advised grant of the Paul G. Allen Family Foundation to J.F.R. This work also utilized the computational resources of the NIH Biowulf high-performance computing cluster (https://hpc.nih.gov). We thank Wolfgang Resch of the High-Performance Computing Services Group of the NIH Center for Information Technology for his assistance and guidance in optimizing the AlphaFold runs on Biowulf. We also thank Alberto Rivera for his assistance in creating Figure S1.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The predicted protein structures generated in the course of this study are publicly available through the *Mnemiopsis* Genome Project Portal, located at https://research.nhgri.nih.gov/mnemiopsis.

ORCID

R. Travis Moreland https://orcid.org/0009-0007-1793-7413
Joseph F. Ryan https://orcid.org/0000-0001-5478-0522
Andreas D. Baxevanis https://orcid.org/0000-0002-5370-0014

REFERENCES

1. Bertoline, L. M. F., Lima, A. N., Krieger, J. E., & Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, *3*, 1120370.

2. Obermayer, A., & Uversky, V. N. (2021). Solving Protein Structure with AI: Viva AlphaFold and Co! *Current Protein & Peptide Science*, *22*(12), 823–826.
3. Puthenveetil, R., & Vinogradova, O. (2019). Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments. *Journal of Biological Chemistry*, *294*(44), 15914–15931.
4. Maxwell, E. K., Schnitzler, C. E., Havlak, P., Putnam, N. H., Nguyen, A. D., Moreland, R. T., & Baxevanis, A. D. (2014). Evolutionary profiling reveals the heterogeneous origins of classes of human disease genes: Implications for modeling disease genetics in animals. *BMC Evolutionary Biology*, *14*, 212.
5. Russell, J. J., Theriot, J. A., Sood, P., Marshall, W. F., Landweber, L. F., Fritz-Laylin, L., Polka, J. K., Oliferenko, S., Gerbich, T., Gladfelter, A., Umen, J., Bezanilla, M., Lancaster, M. A., He, S., Gibson, M. C., Goldstein, B., Tanaka, E. M., Hu, C. K., & Brunet, A. (2017). Non-model model organisms. *BMC Biology*, *15*(1), 55.
6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.
7. Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., Kondrashov, F. A., & Ivankov, D. N. (2023). Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS One*, *18*(3), e0282689.
8. Ryan, J. F., Pang, K., Schnitzler, C. E., Nguyen, A. D., Moreland, R. T., Simmons, D. K., Koch, B. J., Francis, W. R., Havlak, P., Comparative Sequencing Program, N., Smith, S. A., Putnam, N. H., Haddock, S. H., Dunn, C. W., Wolfsberg, T. G., Mullikin, J. C., Martindale, M. Q., & Baxevanis, A. D. (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science (New York, N.Y.)*, *342*(6164), 1242592.
9. Reitzel, A. M., Ryan, J. F., & Tarrant, A. M. (2012). Establishing a model organism: A report from the first annual Nematostella meeting. *BioEssays*, *34*, 158–161.
10. Pang, K., & Martindale, M. Q. (2008). Comb jellies (ctenophora): A model for Basal metazoan evolution and development. *CSH Protocols*, *2008*, pdbemo106.
11. Pang, K., & Martindale, M. Q (2008). Ctenophores. *Current Biology*, *18*, R1119–R1120.
12. Plickert, G., Frank, U., & Muller, W. A (2012). Hydractinia, a pioneering model for stem cell biology and reprogramming somatic cells to pluripotency. *International Journal of Developmental Biology*, *56*, 519–534.
13. Moreland, R. T., Nguyen, A. D., Ryan, J. F., Schnitzler, C. E., Koch, B. J., Siewert, K., Wolfsberg, T. G., & Baxevanis, A. D. (2014). A customized Web portal for the genome of the ctenophore *Mnemiopsis leidyi*. *BMC Genomics [Electronic Resource]*, *15*, 316.
14. Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, *433*(20), 167208.
15. Deiana, A., Forcelloni, S., Porrello, A., & Giansanti, A. (2019). Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One*, *14*(8), e0217889.
16. Bondos, S. E., Dunker, A. K., & Uversky, V. N. (2021). On the roles of intrinsically disordered proteins and regions in cell communication and signaling. *Cell Communication and Signaling:CCS*, *19*(1), 88.
17. Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koča, J., & Rose, A. S. (2021). Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, *49*(W1), W431–W437.
18. Van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01773-0
19. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L. M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., & Steinegger, M. (2023). Clustering predicted structures at the scale of the known protein universe. *Nature*, *622*(7983), 637–645.
20. Molakarimi, M., Gorman, M. A., Mohseni, A., Pashandi, Z., Taghdir, M., Naderi-Manesh, H., Sajedi, R. H., & Parker, M. W. (2019). Reaction mechanism of the bioluminescent protein mnemiopsin1 revealed by X-ray crystallography and QM/MM simulations. *The Journal of Biological Chemistry*, *294*(1), 20–27.
21. Zhang, Y. Z., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, *57*(4), 702–710.
22. Traynelis, S. F., Wollmuth, L. P., McBain, C. J., Menniti, F. S., Vance, K. M., Ogden, K. K., Hansen, K. B., Yuan, H., Myers, S. J., & Dingledine, R. (2010). Glutamate receptor ion channels: Structure, regulation, and function. *Pharmacological Reviews*, *62*(3), 405–496.
23. Alberstein, R., Grey, R., Zimmet, A., Simmons, D. K., & Mayer, M. L. (2015). Glycine activated ion channel subunits encoded by ctenophore glutamate receptor genes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), E6048–E6057.
24. Burnashev, N., & Szepetowski, P. (2015). NMDA receptor subunit mutations in neurodevelopmental disorders. *Current Opinion in Pharmacology*, *20*, 73–82.
25. Chan, S. Y., Matthews, E., & Burnet, P. W. (2017). ON or OFF?: Modulating the N-Methyl-D-Aspartate receptor in major depression. *Frontiers in Molecular Neuroscience*, *9*, 169.
26. Amin, J. B., Moody, G. R., & Wollmuth, L. P. (2021). From bedside-to-bench: What disease-associated variants are teaching us about the NMDA receptor. *The Journal of Physiology*, *599*(2), 397–416.
27. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins*, *89*(12), 1607–1617.
28. The PyMOL Molecular Graphics System, Version 2.6.0 Schrödinger, LLC.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202300397 in the Supporting Information section at the end of the article.

How to cite this article: Moreland, R. T., Zhang, S., Barreira, S. N., Ryan, J. F., & Baxevanis, A. D. (2024). An AI-generated proteome-scale dataset of predicted protein structures for the ctenophore *Mnemiopsis leidyi*. *Proteomics*, e2300397. https://doi.org/10.1002/pmic.202300397