frontiers | Frontiers in Marine Science

TYPE Data Report
PUBLISHED 08 October 2024
DOI 10.3389/fmars.2024.1487904

OPEN ACCESS

EDITED BY
Sandie M Degnan,
The University of Queensland, Australia

REVIEWED BY
Bernard Degnan,
The University of Queensland, Australia
Octavio R. Salazar,
King Abdullah University of Science and Technology, Saudi Arabia

*CORRESPONDENCE
Zuhong Lu
zhlu@seu.edu.cn

RECEIVED 28 August 2024
ACCEPTED 23 September 2024
PUBLISHED 08 October 2024

CITATION
Zhu Y and Lu Z (2024) Proteome structuring of crown-of-thorns starfish.
Front. Mar. Sci. 11:1487904.
doi: 10.3389/fmars.2024.1487904

Proteome structuring of crown-of-thorns starfish

Yunchi Zhu and Zuhong Lu*

State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China

KEYWORDS

proteome structuring, crown-of-thorns starfish, *Acanthaster planci*, ColabFold, structural bioinformatics

1 Introduction

The crown-of-thorns starfish (COTS, *Acanthaster planci*) is a highly fecund predator of reef-building corals throughout the Indo-Pacific region (Hall et al., 2017). COTS population outbreaks cause significant damage to coral reefs, the living environment for more than 30% of marine animals and plants (Zhu et al., 2022), leading to a loss of coral cover and biodiversity. Scientists have sequenced the COTS genome (Hall et al., 2017), which provides a wealth of information on the genetic basis of COTS biology. By identifying specific genes and proteins involved in these behaviors, scientists are able to gain a deeper understanding of their reproductive strategies and the factors contributing to outbreaks, so as to develop targeted biocontrol methods such as peptide mimetics to disrupt COTS aggregation. However, the function annotation of COTS proteome turns out to be incomplete, with over 20% of proteins being annotated as "uncharacterized". Traditional sequence-based annotation methods may be insufficient for fully resolving genomes, particularly for non-model organisms. It is commonly recognized that "sequence determines structure, and structure determines function." If proteome structuring, by which sequences can be transformed into accurate structures in a high-throughput way, is as feasible as genome and transcriptome sequencing, it is believed that such approach could not only substantially aid researchers in complementing and correcting protein annotations, but also pave a new dimension for protein data mining (Tunyasuvunakool et al., 2021).

This vision is going to be realized with the help of the booming artificial intelligence (AI) technology. AI-based protein structure prediction systems represented by RoseTTAFold (Baek et al., 2021) and AlphaFold2 (Jumper et al., 2021) have brought the dawn of the high-throughput era of structural proteomics. With accuracy not inferior to traditional methods such as x-ray crystallography and cryo-electron microscopy (Baek et al., 2021), they have overwhelming advantages in cost, efficiency and ease of operation. As of July 2022, the AlphaFold Protein Structure Database (AFDB) boasts open access to a staggering collection of over 200 million protein structures (Varadi et al., 2024), marking a 1000-fold increase compared to the 50-year accumulation of the PDB. Furthermore, owing to collaborative efforts within the open-source community, many optimized versions of AlphaFold2, which are collectively referred to as AlphaFold-like systems have been developed, with ColabFold (Mirdita et al., 2022) standing out as a notable example. It significantly reduces the resource demands for protein folding, empowering more researchers to engage in personalized

structure predictions and expanding the overall data scale. Up to now, AlphaFold-like systems present to be the most robust tools for high-throughput proteome structuring (Callaway, 2023), facilitating a diverse array of research endeavors. Notable examples include ColabFold proteome CP-8382 from Southeast University (Zhu et al., 2022), structural proteome of *Sphagnum divinum* from Oak Ridge National Laboratory (Davidson et al., 2023), AlphaFold proteome of *Mnemiopsis leidyi* from National Institutes of Health (Moreland et al., 2024), etc. These explorations are progressively expanding the protein structure universe and enabling new insights into protein function and biology.

Here we present the proteome structuring of COTS. Deploying ColabFold in the Big Data Computing Center at Southeast University, we predicted 31,743 protein structures. The resulting dataset covers 60.4% of residues with a confident prediction and 35.5% with very high confidence. We also performed a preliminary structural bioinformatics analysis using several post-AlphaFold methods, including fast structure clustering, ligand transplanting and structure-based Gene Ontology (GO) annotation.

2 Materials and methods

2.1 Proteome structuring

The NCBI RefSeq of *Acanthaster planci* (GCF_001949145.1) was used as sequence source. Protein sequences were downloaded and filtered, discarding those exceeding 2,550 aa to accommodate the upper limit of GPU memory. Multiple sequence alignments (MSA) generation were conducted locally (*colabfold_search*), then MSAs in A3M format were uploaded to ColabFold 1.5.2 on the NVIDIA Tesla V100 cluster at the Big Data Computing Center of Southeast University. The parameters of ColabFold were set to *–amber, –num-recycle 3, –use-gpu-relax*, *–zip, –num-relax 1*. During the structure prediction process, the MineProt (Zhu et al., 2023) toolkit (*colabfold/import.sh –name-mode 1 –zip –relax*) was periodically executed to process predicted proteins. This included selection of best structure models with highest predicted local distance difference test (pLDDT) scores, generation of CIF files, and storage of model scores in JSON format.

2.2 Structure alignment and clustering

Foldseek (van Kempen et al., 2024) was employed for high-throughput structure alignment clustering. Predicted structures were aligned to the AlphaFold Clusters (Barrio-Hernandez et al., 2023) using *easy-search -e 0.01 -s 7.5*, and were clustered using *easy-cluster -c 0.9 -e 0.01 –min-seq-id 0.5*. Uncharacterized proteins clustered with annotated COTS proteins were selected, then their similarities to the annotated proteins were calculated by US-align (Zhang et al., 2022).

2.3 Structure-based function annotation

The representative protein structures identified by Foldseek clustering got function annotation. AlphaFill (Hekkelman et al., 2023) 2.0.0 with PDB-REDO databank (van Beusekom et al., 2018) was used to enrich them with ligands and cofactors, and DeepFRI (Gligorijević et al., 2021) was used for GO annotation.

Uncharacterized proteins with average cluster pLDDT > 70 were selected for GO enrichment analysis by the aid of clusterProfiler (Wu et al., 2021), where GO annotations with DeepFRI score lower than 0.5 were pre-filtered. The parameters were *pvalueCutoff = 0.05, pAdjustMethod = 'fdr', qvalueCutoff = 0.2*.

3 Results

The resulting dataset contains 31,743 protein structures, of which 6,338 were tagged with "uncharacterized" in RefSeq non-redundant proteins (NR) database (O'Leary et al., 2016) previously. The model confidence distribution is demonstrated in Figure 1A. 60.4% of residues have a pLDDT larger than 70, and 35.5% have a very high pLDDT over 90. A total of 20,419 protein structures attain an average pLDDT greater than 70, the commonly recognized benchmark of confident model (Tunyasuvunakool et al., 2021).

Structure alignment was made between COTS structural proteome and AlphaFold Clusters, the Foldseek clustered AFDB. It should be mentioned that Foldseek hold a comparable position in structural bioinformatics to that of BLAST in sequential bioinformatics, for they both enable the feasibility of large-scale searches in their fields. With its help, scientists have successfully clustered all of the structures in the AFDB into 2.3 million clusters, rendering localized operations of AFDB practicable. As illustrated in Figure 1B, the overall protein structural similarity between COTS and known species within the AFDB is found to be moderate to low, with more than half of alignments exhibiting TM-scores (*qtmscore* in Supplementary Table S1) below 0.5 (Zhang and Skolnick, 2004). The majority of sequence identities (*fident* in Supplementary Table S1) between these alignments are even lower, aligning with the notion that protein structure is more conserved than sequence. The phenomenon of low similarity may be attributed to AFDB's limited inclusion of *Acanthaster planci* as well as its closely related species, for the number of *Acanthaster* protein structures does not surpass 100 in the database. Therefore, the predicted COTS structural proteome from this work can currently serve as a starfish-specific extension of AFDB in view of its generally confident model scores.

Foldseek clustering of the COTS dataset was performed with reference to the construction process of AlphaFold Clusters, and 16,896 structural clusters were generated (Supplementary Table S2). 192 uncharacterized proteins were found to have high structural similarity with annotated proteins, as listed in Supplementary Table S3. The sequence identity between the majority of them and their annotated counterparts seems to be not low, suggesting that the "uncharacterized" labels for most of them may result from previous annotation omissions. At the time of the COTS genome release, sequencing data from its closely related species were not abundant in the NCBI databases, thus NCBI's automated annotation pipeline based on non-COTS sequences was inevitably to miss several proteins. Structure clustering could potentially contribute to capturing such omissions and refining annotations. Furthermore, there are

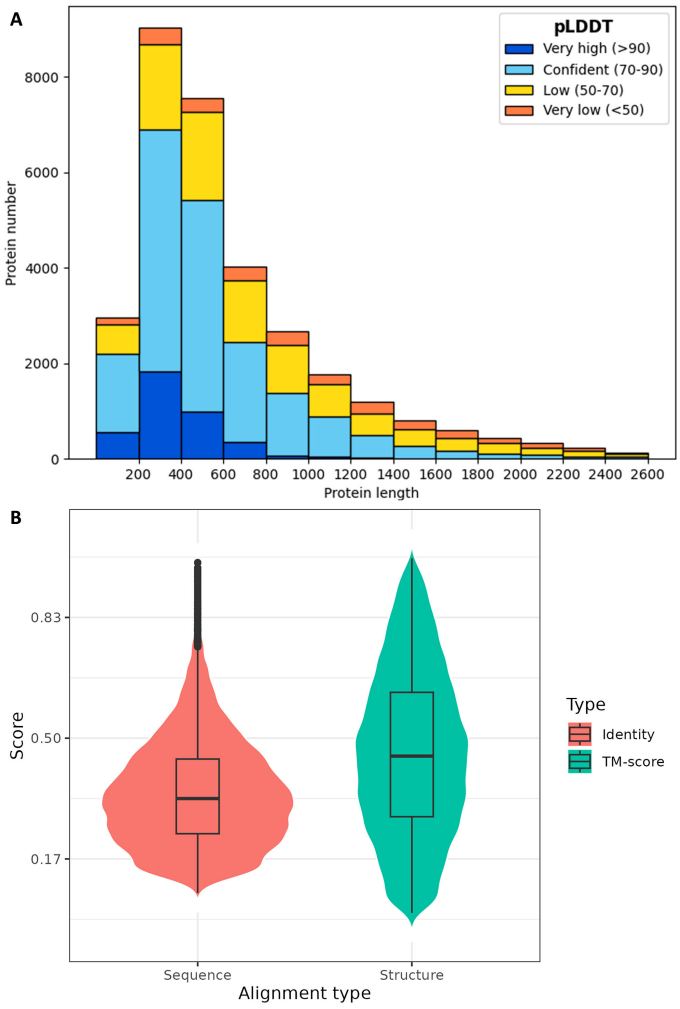

FIGURE 1
Statistics of COTS proteome structuring. **(A)** Distribution of model confidence against protein length. Horizontal axis is protein length and vertical axis is protein number. Model confidence calculated by pLDDT is color-coded. Very high: pLDDT > 90; confident: 90 > pLDDT > 70; low: 70 > pLDDT > 50; very low: pLDDT < 50. **(B)** Violin plot of structure alignment between COTS structural proteome and AlphaFold Clusters. Horizontal axis is alignment type, for Foldseek performs structure alignment while also calculating sequence alignment. Vertical axis is alignment score, with identity corresponding to sequence alignment and TM-score to structure alignment.

totally 3,836 clusters that consist solely of uncharacterized proteins, which can be designated as uncharacterized clusters (Supplementary Table S4). Only 1,045 of them are non-singleton clusters (Barrio-Hernandez et al., 2023), and 1,138 of them have average pLDDT > 70.

Representative protein of each cluster got structure-based function annotation, including ligand transplanting by AlphaFill (Supplementary Table S5) and GO annotation by DeepFRI (Supplementary Table S6). Several annotation results are visualized in Figure 2.

The AlphaFill algorithm identifies experimentally determined protein structures similar to input structure models through sequence alignment, followed by structural comparison to ascertain the positions of ligands and cofactors. Subsequently, these entities are transplanted into structure models, thereby enriching their information content. 10,171 proteins got the "filled" models, and Figure 2A shows the statistics of top-20 compounds with largest number of transplants. Except for cholesterol hemisuccinate (Y01), 19 of them are also present within the top-50 ranked compounds of the entire AlphaFill databank (alphafill.eu). Given that numerous experimentally determined protein structures derive from drug experiments, AlphaFill models may confer additional benefits in supporting the development of marine drugs against COTS outbreaks, particularly in the discovery of small molecule targets.

In the pre-AlphaFold era, DeepFRI is one of the few neural network models for GO annotation that accepts protein structure input. This feature has distinguished it in this era of high-throughput structural proteomics. 13,311 proteins got GO annotation with DeepFRI scores above 0.5, the benchmark of significant prediction (Gligorijević et al., 2021). 701 representative proteins of uncharacterized clusters with average pLDDT > 70 were selected for GO enrichment analysis, results of which are demonstrated in Figure 2B; Supplementary Table S7. Three

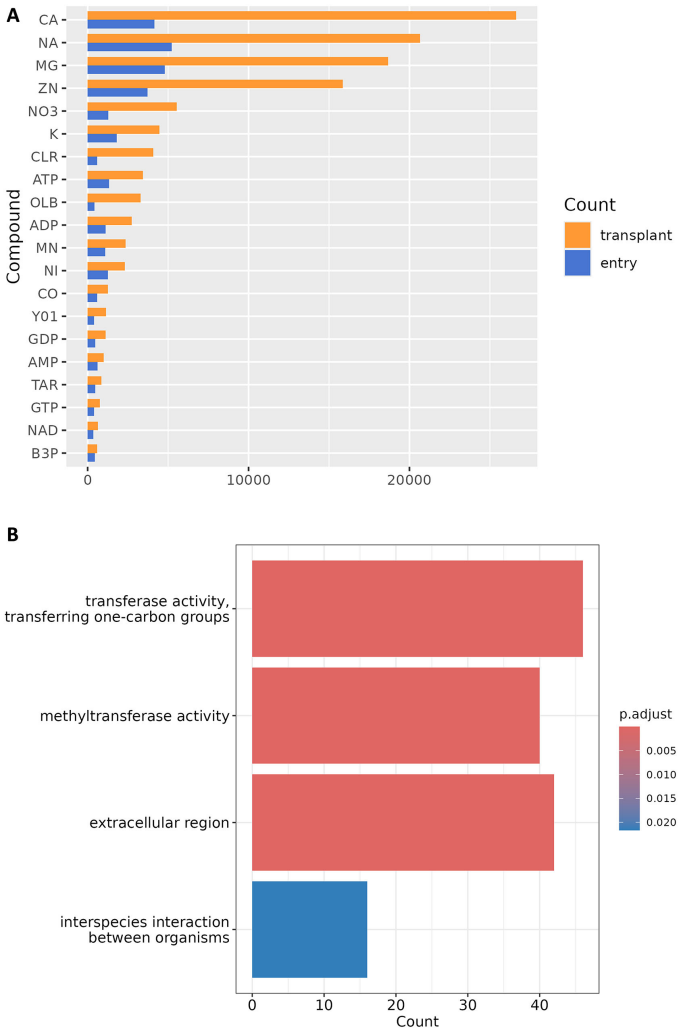

FIGURE 2
Structure-based function annotation of COTS structural proteome. **(A)** Statistics of the top-20 compounds with largest number of transplants. The horizontal axis is the number of entries and transplants, and the vertical axis is PDBe ligand code. CA, Calcium($^{2+}$) ion; NA, Sodium($^{1+}$) ion; MG, Magnesium($^{2+}$) ion; ZN, Zinc($^{2+}$) ion; NO3, Nitrate ion; K, Potassium($^{1+}$) ion; CLR, Cholesterol; ATP, Adenosine triphosphate; OLB, 1-Oleoyl-sn-glycerol; ADP, Adenosine diphosphate; MN, Manganese($^{2+}$) ion; NI, Nickel($^{2+}$) ion; CO, Cobalt($^{2+}$) ion; Y01, Cholesterol hemisuccinate; GDP, Guanosine diphosphate; AMP, Adenosine monophosphate; TAR, d-Tartaric acid; GTP, Guanosine triphosphate; NAD, Nicotinamide adenine dinucleotide; B3P, Bis-Tris propane. **(B)** Bar plot of GO enrichment analysis of 701 representative uncharacterized proteins.

groups of proteins were significantly enriched: one group might be involved in interactions between organisms (GO:0044419), another group is localized to the extracellular region (GO:0005576), and a third group probably have methyltransferase-like activity (GO:0016741 and GO:0008168). These three groups appear to have little overlap, with only XP_022108099.1, XP_022096454.1, and XP_022097466.1 (Supplementary Figure S1) being present in both GO:0044419 and GO:0005576 groups. It should be noted that these three proteins possess signal peptides detectable using SignalP (Teufel et al., 2022) but no obvious transmembrane domains, which is the feature of secreted proteins (Morin et al., 2023). Pheromone-like signals are acknowledged as pivotal in the biology of COTS, enabling the regulation of reproductive aggregations, synchronized spawning events (Jönsson et al., 2022) (Morin et al., 2024), foraging behaviors, and escape responses from predators (Hall et al., 2017). Hence, it might be important for subsequent experimental research as well as meta-analysis to ascertain the presence of these proteins within the COTS exoproteome and to elucidate their potential involvement in conspecific or interspecies communication.

4 Discussion

We succeeded to generate a predicted structural proteome of COTS with acceptable confidence. The application of post-AlphaFold structure-centric methodology not only provides evidence for our dataset's capability of complementing AFDB, but also enhances existing COTS protein annotation. It is expected for the COTS structural proteome to deepen our understanding of COTS biology and facilitating biocontrol method development, not to mention that these protein structures can be directly used as raw inputs in various computational biology tasks, obviating the need

for researchers to engage in time-consuming AlphaFold2 deployment and *de novo* structure modelling.

There is no denying that this work has several limitations. First, the report lacks a thorough interpretation of the ligand transplanting results. The software ecosystem of cheminformatics is less developed than that of bioinformatics, with even the most fundamental ID conversion tools lacking support for PDB ligand codes utilized by AlphaFill. Chemical information mining from these models will continue to pose a challenge unless there is an improvement in the productivity of cheminformatics programmers. Second, the proteome structuring failed to cover proteins too large for our GPU devices to process, a prevalent issue encountered in almost all proteome structuring efforts. It is proposed that reducing computational costs be recognized as another significant direction for the development of AlphaFold-like systems, following the improvement of accuracy and throughput. Last but not least, our COTS structural proteome is entirely a "dry lab" product, necessitating further utilization and assessment by experienced marine biologists. In summary, joint efforts should be made to apply our dataset to the control of COTS and the protection of coral reefs.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://figshare.com/, https://doi.org/10.6084/m9.figshare.26706793.v1, https://figshare.com/, https://doi.org/10.6084/m9.figshare.26779318.v1.

Author contributions

YZ: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. ZL: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Key Research and Development Project (6307030004).

Acknowledgments

We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2024.1487904/full#supplementary-material

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi: 10.1126/science.abj8754

Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L. M., Wein, T., et al. (2023). Clustering predicted structures at the scale of the known protein universe. *Nature* 622, 637–645. doi: 10.1038/s41586-023-06510-w

Callaway, E. (2023). After AlphaFold: protein-folding contest seeks next big breakthrough. *Nature* 613, 13–14. doi: 10.1038/d41586-022-04438-1

Davidson, R. B., Coletti, M., Gao, M., Piatkowski, B., Sreedasyam, A., Quadir, F., et al. (2023). Predicted structural proteome of Sphagnum divinum and proteome-scale annotation. *Bioinformatics* 39, btad511. doi: 10.1093/bioinformatics/btad511

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* 12, 3168. doi: 10.1038/s41467-021-23303-9

Hall, M. R., Kocot, K. M., Baughman, K. W., Fernandez-Valverde, S. L., Gauthier, M. E. A., Hatleberg, W. L., et al. (2017). The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature* 544, 231–234. doi: 10.1038/nature22033

Hekkelman, M. L., de Vries, I., Joosten, R. P., and Perrakis, A. (2023). AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* 20, 205–213. doi: 10.1038/s41592-022-01685-y

Jönsson, M., Morin, M., Wang, C. K., Craik, D. J., Degnan, S. M., and Degnan, B. M. (2022). Sex-specific expression of pheromones and other signals in gravid starfish. *BMC Biol.* 20, 288. doi: 10.1186/s12915-022-01491-0

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi: 10.1038/s41592-022-01488-1

Moreland, R. T., Zhang, S., Barreira, S. N., Ryan, J. F., and Baxevanis, A. D. (2024). An AI-generated proteome-scale dataset of predicted protein structures for the ctenophore Mnemiopsis leidyi. *Proteomics* 24, e2300397. doi: 10.1002/pmic.202300397

Morin, M., Jönsson, M., Wang, C. K., Craik, D. J., Degnan, S. M., and Degnan, B. M. (2023). Captivity induces a sweeping and sustained genomic response in a starfish. *Mol. Ecol.* 32, 3541–3556. doi: 10.1111/mec.16947

Morin, M., Jönsson, M., Wang, C. K., Craik, D. J., Degnan, S. M., and Degnan, B. M. (2024). Seasonal tissue-specific gene expression in wild crown-of-thorns starfish reveals reproductive and stress-related transcriptional systems. *PloS Biol.* 22, e3002620. doi: 10.1371/journal.pbio.3002620

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., et al. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* 40, 1023–1025. doi: 10.1038/s41587-021-01156-3

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596. doi: 10.1038/s41586-021-03828-1

van Beusekom, B., Touw, W. G., Tatineni, M., Somani, S., Rajagopal, G., Luo, J., et al. (2018). Homology-based hydrogen bond information improves crystallographic structures in the PDB. *Protein Sci.* 27, 798–808. doi: 10.1002/pro.3353

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., et al. (2024). Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42, 243–246. doi: 10.1038/s41587-023-01773-0

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., et al. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* 52, D368–D375. doi: 10.1093/nar/gkad1011

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2, 100141. doi: 10.1016/j.xinn.2021.100141

Zhang, C., Shine, M., Pyle, A. M., and Zhang, Y. (2022). US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* 19, 1109–1115. doi: 10.1038/s41592-022-01585-1

Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality [published correction appears in Proteins. *Proteins* 57, 702–710. doi: 10.1002/prot.20264

Zhu, Y., Liao, X., Han, T., Chen, J. Y., He, C., and Lu, Z. (2022). Utilizing an artificial intelligence system to build the digital structural proteome of reef-building corals. *Gigascience* 11, giac117. doi: 10.1093/gigascience/giac117

Zhu, Y., Tong, C., Zhao, Z., and Lu, Z. (2023). MineProt: a stand-alone server for structural proteome curation. *Database (Oxford)* 2023. doi: 10.1093/database/baad059