Petascale Homology Search for Structure Prediction

Sewon Lee[1,#], Gyuri Kim[1,#], Eli Levy Karin[2], Milot Mirdita[1], Sukhwan Park[3], Rayan Chikhi[4], Artem Babaian[5,6], Andriy Kryshtafovych[7], Martin Steinegger[1,3,8,9] ✉

[1] School of Biological Sciences, Seoul National University, Seoul 08826, South Korea
[2] ELKMO, Copenhagen 2720, Denmark
[3] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, South Korea
[4] Institut Pasteur, Université Paris Cité, G5 Sequence Bioinformatics, 75015 Paris, France
[5] Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada
[6] Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada
[7] Genome Center, University of California, Davis, California 95616, USA
[8] Artificial Intelligence Institute, Seoul National University, Seoul 08826, South Korea
[9] Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, South Korea

[#] These authors contributed equally
Correspondence to: martin.steinegger@snu.ac.kr

Abstract

The recent CASP15 competition highlighted the critical role of multiple sequence alignments (MSAs) in protein structure prediction, as demonstrated by the success of the top AlphaFold2-based prediction methods. To push the boundaries of MSA utilization, we conducted a petabase-scale search of the Sequence Read Archive (SRA), resulting in gigabytes of aligned homologs for CASP15 targets. These were merged with default MSAs produced by ColabFold-search and provided to ColabFold-predict. By using SRA data, we achieved highly accurate predictions (GDT_TS > 70) for 66% of the non-easy targets, whereas using ColabFold-search default MSAs scored highly in only 52%. Next, we tested the effect of deep homology search and ColabFold's advanced features, such as more recycles, on prediction accuracy. While SRA homologs were most significant for improving ColabFold's CASP15 ranking from 11th to 3rd place, other strategies contributed too. We analyze these in the context of existing strategies to improve prediction.

Introduction

Determining the 3D structure of proteins is of great importance to many research fields, encompassing cancer drug discovery (Ren et al. 2023; Borkakoti and Thornton 2023), pesticide development, and crop improvement (Koesoema 2022). It also plays a crucial role in the design of sensors and enzymes (Pereira et al. 2021), as well as numerous other applications, as reviewed by Pearce and Zhang (2021).

Traditionally, protein structures have been solved using laborious techniques, such as X-ray crystallography, resulting in just under 200,000 structures in over 50 years of communal effort (Berman et al. 2000; Subramaniam and Kleywegt 2022). Resolved structures are routinely deposited in the Protein Data Bank (wwPDB consortium 2019). The demanding experimental process has motivated the development of computational tools as a less burdensome alternative for structure prediction. Since 1994, the Critical Assessment of protein Structure Prediction (CASP) has aimed to identify state-of-the-art computational methods by competition (Moult et al. 1995). The organizers of CASP provide the participants with protein sequences, whose structures were experimentally solved but not yet deposited in the PDB. The solved structures are unknown to the organizers, assessors as well as to the participants. Also, the group identities are kept anonymous from the assessors, therefore the competition is considered double-blinded, ensuring fairness.

Originally, computational prediction methods could be divided into two main groups: template-based modeling (TBM) and free modeling (FM). However, in the past decade, the lines between the groups have been blurred (Bertoline et al. 2023). TBM is a broad category in which known structures are used as templates to predict the structure of query proteins, based on the sequence similarity between them. In its simplest form, a similarity between a single query and a match from the PDB serves as the base for projecting the match's structure onto the query. The inaugural software MODELLER (Sali and Blundell 1993) and other tools that followed have made use of this principle - see Pearce and Zhang (2021) for a review.

Given the limited size of the PDB and its bias towards model organisms (Orlando et al. 2016), detecting remote sequence homology is crucial. To that end, increasingly sensitive search methods have been developed. The first step forward

was taken by algorithms like BLAST (Altschul et al. 1990), which directly compare the query sequence to the reference database. PSI-BLAST (Altschul et al. 1997) improved upon this by computing a multiple sequence alignment (MSA) of the query and its best BLAST hits and calculating a position-specific scoring matrix from that. This generalization of the query is used for a sensitive search of the reference. This approach was further refined by using probabilistic hidden Markov models (HMMs) (Krogh et al. 1994) in tools like HMMer (Eddy 2011). Another significant advancement came with HHsearch (Söding 2005), which expressed both query and reference as HMMs, markedly improving search sensitivity. This underpinned the success of HHpred (Hildebrand et al. 2009) in the CASP9 challenge (Moult et al. 2011). A further development, HHblits (Remmert et al. 2011; Steinegger et al. 2019a) accelerated the HMM-HMM comparison allowing to query databases with millions of HMMs like the Uniclust30 (Mirdita et al. 2017), a clustered version of the Uniprot (UniProt Consortium 2023) to generate diverse query MSAs.

Due to its unprecedented sensitivity, HHpred has transformed CASP in two ways. First, many methods competing in CASP have incorporated HHblits/HHsearch or other tools to identify distant structural homologs (Bertoline et al. 2023). Second, CASP has started using it for classifying target domains in subsequent competitions (Kinch et al. 2011). Specifically, domains in targets for which HHpred could identify a homolog in the PDB, are considered by CASP as “TBM target domains”, while the others - as “FM target domains”.

Recent advances in deep learning have been harnessed by various methods for protein structure prediction (Torrisi et al. 2020). Undoubtedly, the most revolutionary of these is AlphaFold2 (Jumper et al. 2021), which won the CASP14 challenge by a significant margin (Kryshtafovych et al. 2021), reaching experimental accuracy for over two-thirds of the targets. Despite its success, AlphaFold2's prediction accuracy is not without its limitations. Most notably, it relies on its input MSA (Mirdita et al. 2022) diversity, experiencing a significant drop in prediction accuracy when the median number of diverse sequences in the MSA is 30 or less (Jumper et al. 2021). This finding is in agreement with previous studies on the importance of distant homologs to structure prediction (Kuhlman and Bradley 2019; Ashkenazy et al. 2009). However, the ability to construct a deep MSA depends not only on the sensitivity of search algorithms, such as HHblits, but also on the potential pool of sequences, i.e., the reference database.

Metagenomics allows for sequencing uncultivable organisms directly from the environment, significantly expanding the repertoire of protein sequences deposited in scientific databases. In recent years, metagenomic sequences have shown great potential in increasing the fraction of proteins, whose structure can be modeled accurately (Söding 2017; Ovchinnikov et al. 2017; Yang et al. 2021; Wang et al. 2019). Of note, the largest metagenomic database used in these studies is the IMG/M, which contains 27 tera base pairs (Chen et al. 2023b).

It is therefore not surprising that the top scoring servers in the most recent CASP15 challenge were based on AlphaFold2 and included metagenomic sequences in their constructed MSAs (Table 1). An example for such a server is ColabFold (Mirdita et al. 2022). ColabFold takes as input a query protein sequence(s), whose structure is to be predicted. Its first step, denoted here as CF-search, implements a procedure for collecting homologs of the query using MMseqs2 (Steinegger and Söding 2017). CF-search starts by querying the input against the UniRef30 database (Mirdita et al. 2017) and computing profiles from the hits. Next, CF-search queries these profiles against one of two metagenomic databases, which were constructed as part of the ColabFold release: BFD/MGnify and ColabFoldDB (the default reference database). As detailed in Table 2, BFD/MGnify contains 513 million non-redundant proteins from the union of the BFD (Jumper et al. 2021) and MGnify (Richardson et al. 2023) databases. ColabFoldDB expanded the BFD/MGnify with various environmental proteins, resulting in ~740 million proteins. Following the search, an MSA is computed from the detected homologs and finally, in a step denoted here as CF-predict, the MSA is provided as input to the AlphaFold2 models.

In this study we examined different strategies to improve protein structure prediction along three axes. The first two focused on adding homologs to MSAs used for protein structure prediction and the third - on utilizing advanced features of CF-predict. The first and main axis is the breadth of the search, where we studied the impact of a much more systematic inclusion of metagenomic sequences on prediction accuracy. Over 37 peta base pairs are publicly available through the Sequence Read Archive (SRA), the world's largest metagenomic database (Katz et al. 2022). Recently, Serratus, a tool for a high-throughput search of the SRA, was introduced (Edgar et al. 2022). Here, we constructed MSAs based on Serratus-mined homologs of the CASP15 targets and merged them with the default MSAs produced by CF-search. In the second axis of this study, we further

enhanced the merged MSAs by searching for distant homologs of their sequences using HHblits against the BFD. The third axis concerns tuning the advanced parameters of ColabFold to control the use of templates and multimer models and the number of recycles. We then provided MSAs produced by each of these strategies as input to CF-predict and compared the resulting prediction accuracy to that measured with default CF-search MSAs as well as to the leading CASP15 servers. For fair comparison, we ensured all databases used in this study excluded any sequences deposited after the start of CASP15 competition (May 2022).

Our results show that adding SRA-mined homologs improves prediction accuracy for 61% of the examined targets. Tuning advanced features of CF-predict, especially adding more recycles, also contributed to better prediction. By combining the different strategies, ColabFold's CASP15 ranking among servers on non-easy template targets increased from 11th to 3rd place, indicating the vast potential of large-scale sequence exploration for better structure prediction.

Results and Discussion

1. Homolog search and MSA construction

The entry point to this study was a list of 126 targets provided by CASP15. We excluded from this list all targets, which were RNA, heteromers, canceled by CASP or indicated as auxiliary structure for ligand prediction, leaving 77 targets. Each of these targets had one or more domains, which are divided into categories by CASP as follows: FM, FM/TBM, TBM-hard and TBM-easy. We used CF-search to query the targets against ColabFoldDB (Fig. 1 ①), resulting in 77 MSAs, denoted here as *cfdb* MSAs.

1.1. Thousand times broader search

Our next goal was to expand the search beyond ColabFoldDB and explore the SRA. With its 37 peta base pairs of publicly available data, the SRA is orders of magnitude bigger than any previously-used metagenomic resource, including ColabFoldDB (Table 2). We queried the 77 CASP15 targets using Serratus against over half of the publicly available SRA, comprising 22 peta base pairs, organized in over five million SRA runs. Reads aligned to the CASP15 protein sequence queries, were examined in the context of their run and assembled using rnaviralSpades (Meleshko et al. 2021), mounting to over a hundred gigabytes of assembled data.

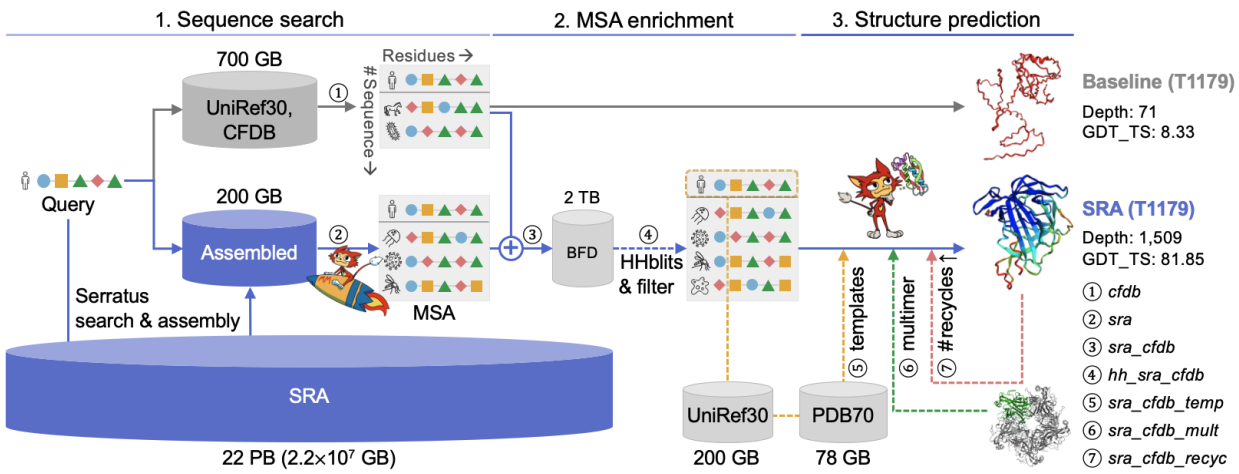

Figure 1. MSA enrichment using SRA and other strategies to improve protein structure prediction. Workflow of the different strategies examined in this study ①~⑦. All strategies construct an MSA (but differ in the homology DBs they utilize) and provide it to CF-predict (but differ in the way they tune its parameters). The size of each homology DB is denoted close to it. The baseline MSA (*cfdb* MSA, ①) is constructed by CF-search. The SRA-detected homologs are aligned to create ② using MMseqs2. The *sra_cfdb* MSA (③) is constructed by combining ① and ②. The *hh_sra_cfdb* MSA (④) is constructed by querying ③ against UniRef30 and BFD using HHblits. Strategies ⑤, ⑥, ⑦ refer to the following CF-predict options: use of templates, multimer (homo-oligomer) modeling, and 12 recycles (instead of the default 3). Before being provided to CF-predict, each MSA is filtered based on the sequence identity between its members and the query.

Each CASP15 target was then queried against a reference protein database created from its Serratus-produced assembled proteins using MMseqs2 (Steinegger and Söding 2017). The identified homologs were then aligned by MMseqs2 to create an MSA (Fig. 1 ②), denoted as *sra* MSA. As a first indication of the tremendous capacity of the SRA, we found that more than half of the targets, which were composed solely of TBM-easy domains, could be matched with at least 1,000,000 homologs, some even exceeding 100,000,000 (Fig. 2A). Processing MSAs with millions of sequences poses a heavy computational burden. Thus, we opted to exclude from this study targets, which only contained TBM-easy domains, focusing on the remaining 46 targets, which had 62 non-TBM-easy domains: 39 FM, 8 FM/TBM, and 15 TBM-hard. On average, the number of homologs detected per domain doubled from 106,586 in the CFDB to 274,231 when including the SRA results (Fig. 2A).

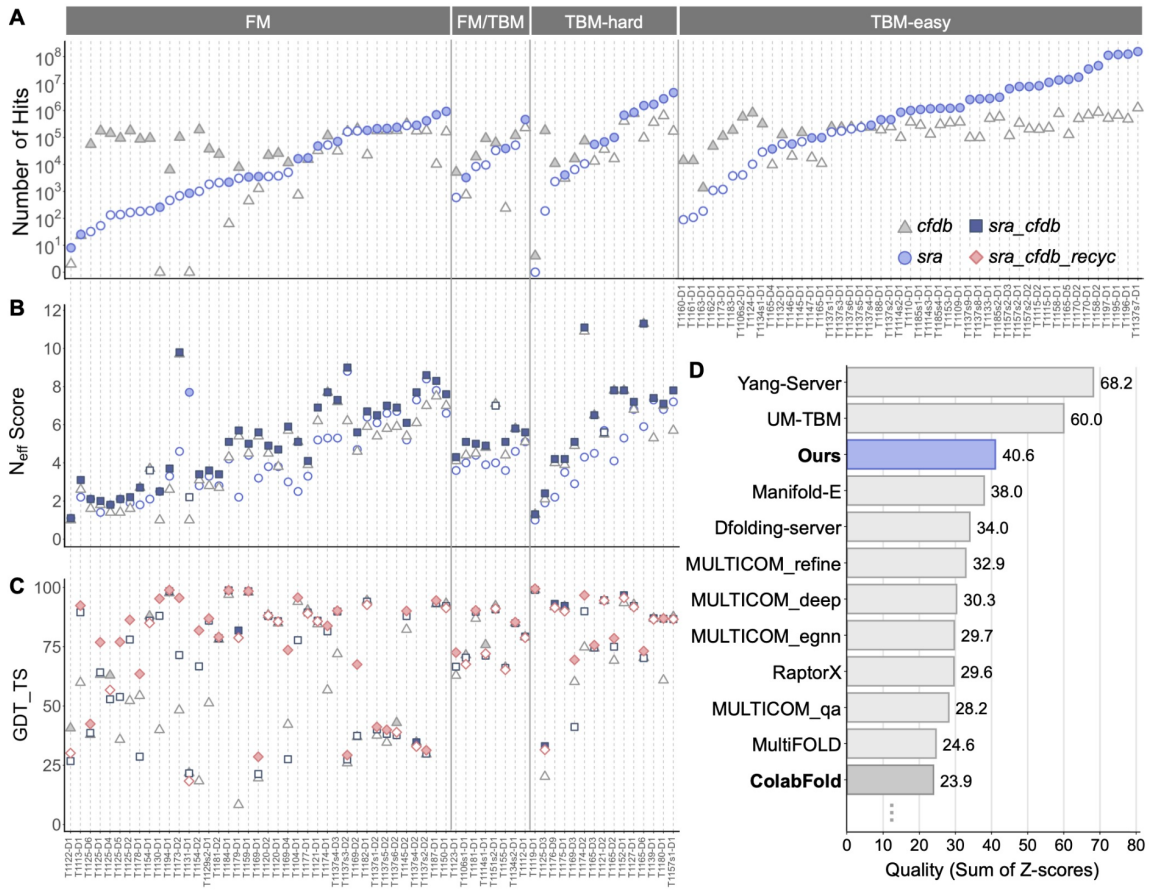

Figure 2. Effect of ColabFold parameters on structure prediction accuracy.
(A) Comparison of homology search of 109 domains of 77 CASP15 targets. Each mark denotes the number of hits found for each target domain using either CF-search against CFDB (triangle) or MMseqs2 against SRA-mined and assembled proteins (circle) before the MSA filtering step. **(B)** N_{eff} scores of the different MSAs computed for each domain. The MSA with the most homologs and the highest N_{eff} is indicated with a filled mark in panels A and B, respectively. **(C)** Structure prediction of 62 target domains in the categories: FM, FM/TBM, and TBM-hard was evaluated based on GDT_TS scores of three prediction strategies: *cfdb* MSA, *sra_cfdb* MSA and *sra_cfdb_recyc*. The best-scoring strategy for each target domain is indicated with filled marks. **(D)** Prediction performance comparison between server groups in CASP15. The x-axis refers to the Sum Z (> 0.0) in Table 3. The score of this study is from the Model1 in Table 3. Here, ColabFold refers to the performance of the server group submitted in CASP15.

1.2. Diving deeper

Serratus' ability to scan the SRA in feasible time comes at the expense of its sensitivity to detect remote homologs. Specifically, it is limited in its ability to detect sequences with less than 50% identity to the query (Edgar et al. 2022). This prompted us to search deeply for remote homologs. To that end, we merged for each target its *cfdb* and *sra* MSAs (Fig. 1 ③), resulting in an *sra_cfdb* MSA. We then ran HHblits with each *sra_cfdb* MSA as input against UniRef30 and BFD (Fig. 1 ④), setting parameters to include all sequences in the output without any filtering and ensure maximum sensitivity (for a full parameter description, see Supp. information of Jumper et al. 2021). The resulting MSAs, denoted as *hh_sra_cfdb* MSAs, contained the input sequences as well as the homologs detected by HHblits.

To filter each input MSA based on the sequence identity (seqid) between its members and the query, we utilized a newly introduced filter module called filtera3m in MMseqs2, which implements ColabFold's MSA filtering strategy. We removed unlikely homologs (seqid < 0.2), non-informative homologs (seqid > 0.95) and redundant sequences, keeping the most diverse and informative set of sequences in the MSA.

The filtered *cfdb* MSAs had on average 2,395 sequences, *sra_cfdb* MSAs - 5,731 and *hh_sra_cfdb* MSAs - 8,133. We used HHmake (Steinegger et al. 2019a) to compute the number of effective sequences (N_{eff}), where higher values indicate less similarity between the sequences and more diverse MSAs. Here, a more moderate increase was observed with the average N_{eff} score, rising from 4.87 for *cfdb* MSA to 5.43 for *sra_cfdb* MSA and to 7.26 for *hh_sra_cfdb* MSA (Fig. 2B).

2. The effect of homologs on structure prediction

In order to measure the effect of the various homolog collection strategies on structure prediction, we provided CF-predict with different input MSAs. These included the *sra_cfdb* and *hh_sra_cfdb* MSAs as well as their controls, which did not include Serratus-detected homologs from the SRA: *cfdb* and *hh_cfdb* MSAs. The *hh_cfdb* MSAs were produced in a similar manner to *hh_sra_cfdb* MSAs, using *cfdb* MSAs, rather than *sra_cfdb* MSAs, as the input to HHblits.

For each strategy, five models were produced and the best one was selected according to its computed predicted local distance difference test (pLDDT) score

(Jumper et al. 2021). For the selected models, we measured the domain level accuracy, using the GDT_TS score (Zemla 2003). Compared to using *cfdb* MSAs, *sra_cfdb* MSAs significantly improved GDT_TS scores (Table 3), increasing scores for 38 out of 62 domains (Fig. 2C). On the other hand, using *hh_sra_cfdb* MSAs did not lead to a significant improvement over *sra_cfdb* MSAs (Table 3) and is therefore omitted from the figure.

We further examined the relative performance of each strategy compared to other CASP15 servers using Z-scores, as follows. For each strategy, we deducted from its GDT_TS scores the mean servers' GDT_TS score and divided it by the servers' standard deviation. The sum of non-negative Z-scores and the average GDT_TS of all evaluated target domains were used as representative scores for each strategy (Table 3). As expected, the addition of *sra* to *cfdb* MSAs substantially improved the performance, increasing the sum of Z-scores from 17.09 to 30.30 (Table 3). On average, no significant improvement was observed when adding HHblits-detected homologs (Table 3). However, specific domains gained a substantial improvement by including these homologs, indicating a variable effect for each target. A notable example with the highest improvement is target T1178-D1, where running HHblits increased GDT_TS from 28.61 for *sra_cfdb* to 84.17 for *hh_sra_cfdb* MSAs.

3. Tuning parameters

In addition to homolog collection strategies, we examined the impact of three advanced features of CF-predict: using templates, using multimer models, and increasing the number of recycles. The strategies corresponding to these features were built on the MSAs constructed in previous steps and are denoted: *sra_cfdb_temp*, *sra_cfdb_recyc* and *sra_cfdb_mult* and their control: *sra_cfdb*. Other strategies, taken by the leading CASP15 servers are detailed in Table 4.

3.1. Leveraging templates

Two out of the five AlphaFold2 models require structural features as input. Setting the “templates” parameter (Fig. 1 ⑤) changes the default behavior of CF-predict from using mock templates to querying the PDB70 (Steinegger et al. 2019a) using the UniRef30-based constructed profile as input and returning hits, which are later aligned by HHsearch.

Adding templates did not improve the accuracy compared to using default parameters (Table 3). However, it should be noted that our examined targets are FM, FM/TBM and TBM-hard, which are classified as difficult targets to find templates for. Applying templates to TBM-easy targets might have different effects on prediction performance.

3.2. Multimer modeling

We also tested the impact of setting on the "multimer" option (Fig. 1 ⑥). This changes the default behavior of CF-predict from treating the input query sequence as a monomer to considering it to be a part of a complex. This has the potential of stabilizing the structure, thereby improving prediction accuracy. We applied multimer modeling only for the homo-oligomer targets, based on the stoichiometry provided by CASP and used the multimer model weights from AlphaFold-multimer version 2 (Evans et al. 2022).

Using multimer modeling had a diverse range of effects. Overall, it did not make a significant improvement over *sra_cfdb* MSAs (Table 3). However, it drastically increased or decreased prediction performance for certain targets. Two notable examples stand out: T1178-D1 achieved the highest improvement, with its GDT_TS score soaring from 28.61 to 61.02 after incorporating multimer modeling. Conversely, the GDT_TS score of target T1174-D1 experienced the largest decrease, dropping from 81.48 to 59.84.

3.3. Adding more recycles

Through the "recycle" parameter (Fig. 1 ⑦), CF-predict allows setting the number of iterations in which a prediction will be re-fed to the AlphaFold2 models. By default, this value is set to 3, but additional recycles have the potential to improve prediction accuracy (Mirdita et al. 2022). Thus, when exploring this option, we set it to 12. The recently released version 3 of AlphaFold-multimer (Evans et al. 2022) uses up to 20 recycle iterations, with early stopping if a model has already converged.

Increasing recycles significantly improved the prediction accuracy compared to the control *sra_cfdb* with default parameters (Table 3). As depicted in Figure 2C, *sra_cfdb_recyc* MSA scored higher than *cfdb* MSAs and *sra_cfdb* MSAs in 34 domains and achieved high-accuracy (GDT_TS > 70) in 72% of the 62 domains.

4. Strategy selection and comparison with CASP15 servers

Finally, among the seven examined strategies, we selected the one with the highest pLDDT for each target domain, denoted here as Model1. We then compared the performance of Model1 with the leading server groups in CASP15, including the original ColabFold server, which is similar to this study's *cfdb* MSAs (Fig. 2D). Model1 resulted in an average GDT_TS of 75.56 and sum of Z-scores of 40.56, increasing 9.76 and 23.47 units from the baseline *cfdb* MSAs, respectively. Notably, 46 out of 62 domains (74%) of Model1 achieved high-accuracy scores (GDT_TS > 70), compared to 52% of the *cfdb* MSAs. When comparing with other server groups based on the sum of Z-scores, Model1 would have ranked 3rd, outperforming the ColabFold original server, which ranked 11th in CASP15 among server-only groups on non-easy targets.

In order to examine the validity of using pLDDT as selection criterion, we compared for each domain the GDT_TS score of Model1 and the highest GDT_TS score (Model_best), across all strategies. In case of perfect agreement between pLLDT and GDT_TS, these values should be equal. However, we observed a disagreement for 37 out of 62 target domains, resulting in a notable increase in Model_best with the average GDT_TS reaching 78.13 and the sum of Z-scores reaching 52.31. This disparity between pLDDT and GDT_TS highlights the challenge in selecting the best model. For instance, choosing the strategy with the highest pLDDT for target T1104-D1 yielded a GDT_TS score of 77.56 (*hh_sra_cfdb* MSA), while the actual best GDT_TS score was 95.73 (*sra_cfdb_recyc* MSA).

To address this discrepancy, there were attempts to use alternative model selection (ranking) methods in CASP15, instead of relying solely on pLDDT (Table 4). For instance, the MULTICOM servers (Liu et al. 2023) utilized APOLLO (Wang et al. 2011), DeepRank (Renaud et al. 2021), and EnQA (Chen et al. 2023a) for ranking models, and MultiFOLD (McGuffin et al. 2023) employed ModFOLDdockR (Edmunds et al. 2023) for both scoring and ranking purposes. Further developments of ranking methods are needed to improve the accuracy and reliability of model selection for structure prediction.

Concluding remarks

In this study, we have shown the importance of a comprehensive inclusion of metagenomic sequences from the SRA for improving protein structure prediction.

Our results highlight the large variation in the number of homologs found for different targets. For example, over 100 million environmental homologs were found for T1137s7, T1195, T1196, and T1197 - the same order of magnitude as the entire UniProt database. However, there are some targets for which few matches were detected, possibly due to the limited sensitivity of the mining procedure.

Serratus, the tool used for mining the SRA has impressive capabilities, but also significant constraints. It is limited to detecting homologs, which have about 50% sequence identity to the query, missing the full potential of homologs from the twilight zone (Rost 1999). Fast and more sensitive search methods are thus required to further improve our ability to exploit the SRA. Additionally, using Serratus in a similar manner to this study is likely to cost thousands of dollars (Edgar et al. 2022) and this cost could become limiting with the expected continued exponential growth of the SRA.

We further investigated the impact of advanced CF-predict features on structure prediction performance. While adding more recycles led to improvement, using multimer models and templates did not contribute significantly. Nonetheless, each feature may have varying effects on different targets, as demonstrated by some notable examples. Thus, it is highly recommendable to experiment with different combinations of these features to optimize performance.

In conclusion, while limitations persist, advancements in metagenomic data mining tools, coupled with a blend of automated and human-guided predictions, promise exciting prospects for the future. Further, the results of this research underline the necessity for diversification in methodologies used in protein structure prediction. Finally, the disparity in MSA coverage among different targets stresses the importance of individual target evaluation and tailored approaches. As the field continues to evolve, we anticipate these findings to contribute to the ongoing quest for accurate protein structure prediction.

Data availability

The data that support the findings of this study are available at https://doi.org/10.5281/zenodo.8126538.

Acknowledgments

MS acknowledges the support by the National Research Foundation of Korea, grants [2020M3-A9G7-103933, 2021-R1C1-C102065, 2021-M3A9-I4021220], Samsung DS research fund and the Creative-Pioneering Researchers Program through Seoul National University. MM acknowledges support by the National Research Foundation of Korea (grant RS-2023-00250470). Computing resources were provided by the University of British Columbia Community Health and Wellbeing Cloud Innovation Centre, powered by AWS. RC was supported by ANR grants [ANR-19-CE45-0008, ANR-22-CE45-0007, PIA/ANR16-CONV-0005, ANR-19-P3IA-0001], and H2020 Marie Skłodowska-Curie grants [No 956229 and 872539]. AK acknowledges the support of the US National Institute of General Medical Sciences (NIGMS/NIH), grant R01GM100482.

Table 1. Use of homology algorithms and databases among leading CASP15 servers[a].

Name	Rank[b]	Homology search alg.	Ref. sequence DBs[c]
Yang-Server	1	HHblits, MMseqs2, manual?	MGnify, UniRef30
UM-TBM	2	DeepMSA, LOMETS	BFD, IMG/M, Metaclust, MetaSource, Uniclust30, UniRef90, Tara?
Manifold-E	3	HHblits, hmmsearch, Jackhmmer	BFD, Uniprot, UniRef30, UniRef90
DFolding	4	CRFalign, Jackhmmer, HHblits, HHpred, Kalign	BFD, Uniclust30, UniRef90, MGnify
MULTICOM	5-7,9	DeepMSA (modified), Foldseek, HHblits, Jackhmmer, MMseqs2	BFD, ColabFoldDB, IMG/M, MGnify, Uniclust30, UniRef90, Uniprot
RaptorX	8	HHblits, Jackhmmer	BFD, SMAG+MetaEuk+TOPAZ+MGV+GPD+IMG/M (in-house HHblits DB), MGnify, Uniclust30, UniRef90
MultiFOLD	10	MMseqs2, UniRef30	ColabFoldDB
ColabFold	11	MMseqs2	ColabFoldDB, UniRef30

[a] The information about the servers was mined from the CASP15 abstract book.
[b] The rank refers to the "server only" performance on protein targets, excluding the TBM-easy category.
[c] A detailed overview of the reference sequence DBs is provided in Table 2.

Table 2. Size and composition of reference databases used by leading CASP15 servers.

DB name	Reference[a]	Source; processing	ca. # seqs	env[b]
Type: Protein				
UniProt/Swiss-Prot	UniProt Consortium 2023	Experiments; manual annotation + redundancy reduction	$< 10^6$	no
RefSeq	UniProt Consortium 2023	NCBI; annotation + redundancy reduction	$> 100 \times 10^6$	no
UniProt/TrEMBL	UniProt Consortium 2023	EMBL-Bank/GenBank/DDBJ; annotation	$> 100 \times 10^6$	no
UniParc	UniProt Consortium 2023	UniProt + RefSeq + other sources	$> 500 \times 10^6$	no
UniRef100	UniProt Consortium 2023	UniProt + selected UniParc; redundancy reduction	$> 100 \times 10^6$	no
UniRef90	UniProt Consortium 2023	UniProt; clustering at 90%	$> 100 \times 10^6$	no
UniRef30	Mirdita et al. 2017	UniProt; clustering at 30%	$> 10 \times 10^6$	no
Uniclust30	Mirdita et al. 2017	UniProt; clustering at 30%	$> 10 \times 10^6$	no
Type: Predicted Protein				
GPD	Camarillo-Guerrero et al. 2021	Human gut bacteriophages; prediction + clustering	$< 10^6$	yes
MGV	Nayfach et al. 2021	Human gut viruses, mostly DNA-viruses; prediction + clustering	$< 10^6$	yes
TOPAZ	Alexander et al. 2022	TARA oceans expedition: enriched for Eukaryotes; prediction	$> 5 \times 10^6$	yes
SMAG	Delmont et al. 2022	TARA oceans expedition: sunlit ocean; prediction + annotation	$\sim 10 \times 10^6$	yes
MetaEukDB	Levy Karin et al. 2020	TARA oceans expedition: eukaryotic metagenomics; prediction	$> 5 \times 10^6$	yes
SRC	Steinegger et al. 2019b	Metagenomic soil samples; protein assembly	$> 500 \times 10^6$	yes
MERC	Steinegger et al. 2019b	TARA oceans expedition: metatranscriptomic data; protein assembly	$> 100 \times 10^6$	yes
MetaClust	Steinegger and Söding 2018	JGI's metagenomic and metatranscriptomic data; prediction + clustering	$> 500 \times 10^6$	yes
MGnify	Richardson et al. 2023	ENA + microbiome studies; EMBL-EBI pipeline	$> 1000 \times 10^6$	yes
BFD	Jumper et al. 2021	Various sources: TrEMBL + Swissprot + SRC + MERC + Metaclust; clustering	$> 500 \times 10^6$	yes
BFD/MGnify	Mirdita et al. 2022	Various sources: BFD + MGnify; clustering	$> 500 \times 10^6$	yes
ColabFoldDB	Mirdita et al. 2022	Various sources: BFD/MGnify + MetaEukDB + TOPAZ + MGV + GPD + SMAG; clustering	$> 500 \times 10^6$ (ca. 0.5 T base pairs)	yes
IMG/M	Chen et al. 2023b	Various sources: uncult. genomes, isolates, metagenomes, metatranscriptomes; JGI's pipeline	$> 50{,}000 \times 10^6$ (27 T base pairs)	yes

[a] References are to the latest version of each DB. The servers detailed in Table 1 may have used older versions of these resources. Some resources also rely on others for their construction, meaning an older version was used.

[b] Mostly environmental proteins.

Table 3. Effects of different strategies on protein structure prediction performance

CFDB	SRA	HHblits	templates	multimer	12 recycles	GDT_TS[b]	Sum Z (> 0.0)[c]
✓						65.80 ± 25.58	17.09
✓	✓					70.97 ± 24.40[d]	30.30
✓		✓				67.36 ± 25.15	15.82
✓	✓	✓				71.44 ± 23.30[e]	26.32
✓	✓		✓			70.94 ± 23.99	31.60
✓	✓			✓		70.98 ± 24.12	30.97
✓	✓				✓	75.54 ± 22.17[f]	40.59
Model1[a]						75.56 ± 22.12	40.56

[a] Model1 is the set of predicted-best strategies among all examined strategies, chosen based on pLDDT for each target domain.
[b] GDT_TS scores are represented as mean ± standard deviation.
[c] Z-scores are calculated from each target domain's GDT_TS, based on the server groups' mean and standard deviation. Sum Z (> 0.0) refers to the sum of positive Z-scores. All strategies were compared to their controls using the Wilcoxon signed-rank test. No significant improvement ($p > 0.05$) was found except for:
[d] *sra_cfdb* over *cfdb* MSAs ($p = 0.0116$)
[e] *hh_sra_cfdb* over *hh_cfdb* MSAs ($p = 0.0126$)
[f] *sra_cfdb_recyc* over *sra_cfdb* MSAs ($p = 0.0003$)

Table 4. Use of prediction algorithms and strategies among leading CASP15 servers[a].

Name	Monomer Alg.	Monomer strategies	Strategy category	Multimer Alg.	Multimer strategies	Comment
Yang-Server	trRosettaX2, AF2	AF2 when trRosettaX2 is not satisfactory	Modeler selection	AF-Multimer	Template search with HHsearch	
		AF2's Evoformer in trRosettaX2	Modeler selection		Disable MSA pairing	
		Generate using HHblits & MMseqs2	MSA			
UM-TBM	AF2, I-TASSER	Generate using DeepMSA	MSA	NA	NA	Multistep pipeline
		Detect multi-domain templates using LOMETS	Templates			
		I-TASSER REMC simulation	Model refinement			
		MD simulation	Model refinement			
Manifold-E	AF2 (reimplemented)	Generate with HHblits & JackHMMER	MSA	AF-Multimer (reimplemented)	Not specified separately from mono.	Reimplement in PyTorch
		Train models with different configs: # seqs, # templates, etc.	AF2 configuration			
		Modify predict params: MSA sampling, templates, # recycles	AF2 configuration			
		OpenMM & AMBER99 force field	Model relaxation			
DFolding	AF2 (modified)	Modify AF2's torsion & FAPE loss functions	AF2 modification	AF2 Complex (modified)	Use non-paired MSA	Multistep pipeline DeepFold
		Introduce loss functions for sidechain and secondary structure	AF2 modification		Generate subcomplexes by domains for large targets and combine using Modeller	
		Train models using uni-fold and protein chains from PDB40	AF2 configuration			
		Search MSA features using HHpred, kalign and HHblits	MSA			
		Replace template features with CRFalign	Templates			
		MD simulation	Model refinement			
MULTICOM	AF2	Generate using HHblits, JackHMMER, MMseqs2 and DeepMSA	MSA	AF-Multimer	Rank multimer models using MMalign	Different strategies tested in different MULTICOM servers
		Augment AF2 templates with one found by searching an in-house template DB	Templates			
		Augment AF2's MSAs with IMG/M homologs if depth < 200	MSA			
		Rank AF2's models using APOLLO, DeepRank and EnQA	Model ranking			
		Refine using a method based on FoldSeek	Model refinement			

RaptorX	AF2 (modified)	Generation like AF2, augmentation with homologs from in-house metagenomic DB if shallow	MSA	Not specified	Not specified	
		Template DBs: PDB70, PDB100, and DistillPDB (predicted decoy structures by AF2)	Templates			
		TMalign and DeepAlign to find and align templates based on highest pLDDT (iterative)	Templates			
		Modify AF2: use a linear layer to integrate scalar, point, and pair attention values in IPA model	AF2 modification			
		Modify AF2 feature module and train 4 combinations: MSA, MSA + template, MSA + MSAtransformer + template, MSA + MSAtransformer + template + AF2	AF2 configuration			
		Train single sequence model with included ESM-1b language model embeddings	Custom model			
		Top-5 highest pLDDT and replace one model with TM-based model clustering center model	Model ranking			
MultiFOLD	AF2 (modified ColabFold)	Use templates if seq length < 1000	Templates	AF-Multimer		Local-ColabFold (ColabFold, AF-Multimer)
		Use ModFOLDdockR to score and rank models	Model ranking			
		Use AMBER if seq length < 1000	Model relaxation			
		Use 12 recycles if seq length < 1000	AF2 configuration			
ColabFold	AF2	Use MMseqs2 to search and align	MSA	AF-Multimer-v2		
		Use 12 recycles, templates, ensembles	AF2 configuration			
		OpenMM/Amber	Model relaxation			

[a] The information about the servers was extracted from the CASP15 abstract book.

References

Alexander H, Hu SK, Krinos AI, Pachiadaki M, Tully BJ, Neely CJ, Reiter T. 2022. Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. *bioRxiv* 2021.07.25.453713. https://www.biorxiv.org/content/10.1101/2021.07.25.453713v2 (Accessed July 2, 2023).

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Ashkenazy H, Unger R, Kliger Y. 2009. Optimal data collection for correlated mutation analysis. *Proteins* **74**: 545–555.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242.

Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. 2023. Before and after AlphaFold2: An overview of protein structure prediction. *Front Bioinform* **3**: 1120370.

Borkakoti N, Thornton JM. 2023. AlphaFold2 protein structure prediction: Implications for drug discovery. *Curr Opin Struct Biol* **78**: 102526.

Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. *Cell* **184**: 1098–1109.e9.

Chen C, Chen X, Morehead A, Wu T, Cheng J. 2023a. 3D-equivariant graph neural networks for protein model quality assessment. *Bioinformatics* **39**. http://dx.doi.org/10.1093/bioinformatics/btad030.

Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter SJ, Webb C, Wu D, et al. 2023b. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res* **51**: D723–D732.

Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, Eren AM, Kourlaiev A, d'Agata L, Clayssen Q, et al. 2022. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom* **2**: 100123.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.

Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, et al. 2022. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**: 142–147.

Edmunds NS, Alharbi SMA, Genc AG, Adiyaman R, McGuffin LJ. 2023. Estimation of model accuracy in CASP15 using the ModFOLDdock server. *Proteins*. http://dx.doi.org/10.1002/prot.26532.

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, et al. 2022. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034. https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2 (Accessed June 26, 2023).

Hildebrand A, Remmert M, Biegert A, Söding J. 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins* **77 Suppl 9**: 128–132.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589.

Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. 2022. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res* **50**: D387–D390.

Kinch LN, Shi S, Cheng H, Cong Q, Pei J, Mariani V, Schwede T, Grishin NV. 2011. CASP9 target classification. *Proteins* **79 Suppl 10**: 21–36.

Koesoema AA. 2022. Protein structure determination as a powerful tool for the sustainable development of agriculture field (and its potential relevance in Indonesia). *IOP Conf Ser Earth Environ Sci* **978**: 012021.

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**: 1501–1531.

Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. 2021. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**: 1607–1617.

Kuhlman B, Bradley P. 2019. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* **20**: 681–697.

Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**: 48.

Liu J, Guo Z, Wu T, Roy RS, Chen C, Cheng J. 2023. Improving AlphaFold2-based Protein Tertiary Structure Prediction with MULTICOM in CASP15. *bioRxiv* 2023.05.01.538929. https://www.biorxiv.org/content/10.1101/2023.05.01.538929v1 (Accessed July 10, 2023).

McGuffin LJ, Edmunds NS, Genc AG, Alharbi SMA, Salehe BR, Adiyaman R. 2023. Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. *Nucleic Acids Res* **51**: W274–W280.

Meleshko D, Hajirasouliha I, Korobeynikov A. 2021. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. *Bioinformatics* **38**: 1–8.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods* **19**: 679–682.

Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* **45**: D170–D176.

Moult J, Fidelis K, Kryshtafovych A, Tramontano A. 2011. Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins* **79 Suppl 10**: 1–5.

Moult J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**: ii–v.

Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, et al. 2021. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* **6**: 960–970.

Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* **355**: 294–298.

Pearce R, Zhang Y. 2021. Toward the solution of the protein structure prediction problem. *J Biol Chem* **297**: 100870.

Pereira JM, Vieira M, Santos SM. 2021. Step-by-step design of proteins for small molecule interaction: A review on recent milestones. *Protein Sci* **30**: 1502–1520.

Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**: 173–175.

Renaud N, Geng C, Georgievska S, Ambrosetti F, Ridder L, Marzella DF, Réau MF, Bonvin AMJJ, Xue LC. 2021. DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. *Nat Commun* **12**: 7068.

Ren F, Ding X, Zheng M, Korzinkin M, Cai X, Zhu W, Mantsyzov A, Aliper A, Aladinskiy V, Cao Z, et al. 2023. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chem Sci* **14**: 1443–1452.

Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell LJ, et al. 2023. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* **51**: D753–D759.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* **12**: 85–94.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**: 779–815.

Söding J. 2017. Big-data approaches to protein structure prediction. *Science* **355**: 248–249.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951–960.

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019a. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**: 473.

Steinegger M, Mirdita M, Söding J. 2019b. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* **16**: 603–606.

Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. *Nat Commun* **9**: 2542.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**: 1026–1028.

Subramaniam S, Kleywegt GJ. 2022. A paradigm shift in structural biology. *Nat Methods* **19**: 20–23.

Torrisi M, Pollastri G, Le Q. 2020. Deep learning methods in protein structure prediction. *Comput Struct Biotechnol J* **18**: 1301–1310.

UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**: D523–D531.

Wang Y, Shi Q, Yang P, Zhang C, Mortuza SM, Xue Z, Ning K, Zhang Y. 2019. Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biol* **20**: 229.

Wang Z, Eickholt J, Cheng J. 2011. APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics* **27**: 1715–1716.

wwPDB consortium. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**: D520–D528.

Yang P, Zheng W, Ning K, Zhang Y. 2021. Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. *Proc Natl Acad Sci U S A* **118**. http://dx.doi.org/10.1073/pnas.2110828118.

Zemla A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**: 3370–3374.