IMPROVING ALPHAFOLD2 PERFORMANCE WITH A GLOBAL METAGENOMIC & BIOLOGICAL DATA SUPPLY CHAIN

Geraldene Munsamy[1], Tanggis Bohnuud[1], and Philipp Lorenz[1,2]

[1]Basecamp Research
[2]To whom correspondence may be addressed: phil@basecamp-research.com

March 6, 2024

ABSTRACT

Scaling laws suggest that more than a trillion species inhabit our planet but only a miniscule and unrepresentative fraction (less than 0.00001%) have been studied or sequenced to date. Deep learning models, including those applied to tasks in the life sciences, depend on the quality and size of training or reference datasets. Given the large knowledge gap we experience when it comes to life on earth, we present a data-centric approach to improving deep learning models in Biology: We built partnerships with nature parks and biodiversity stakeholders across 5 continents covering 50% of global biomes, establishing a global metagenomics and biological data supply chain. With higher protein sequence diversity captured in this dataset compared to existing public data, we apply this data advantage to the protein folding problem by MSA supplementation during inference of AlphaFold2. Our model, BaseFold, exceeds traditional AlphaFold2 performance across targets from the CASP15 and CAMEO, 60% of which show improved pLDDT scores and RMSD values being reduced by up to 80%. On top of this, the improved quality of the predicted structures can yield better docking results. By sharing benefits with the stakeholders this data originates from, we present a way of simultaneously improving deep learning models for biology and incentivising protection of our planet's biodiversity.

1 Introduction

In the last several years we have experienced the rise of a plethora of deep learning models applied to a wide range of biological tasks [1], [2]. Of particular prominence is the protein folding problem, given its impact on structural biology and drug discovery, for which AlphaFold2, RoseTTAFold, and ESMFold, for example, are providing promising and often highly accurate predictions [3], [4], [5]. A lot of research and effort has been put into optimising the architecture of these models to improve performance [6]. However, given that these models depend on the protein sequence and structure datasets available for training, we deployed a data-centric approach towards improving deep learning models in biology, exemplified on the protein folding problem for the purpose of this study.

Previous studies have shown that with improved data quality and quantity, the error loss of transformers while training would no longer follow a power law, but rather an exponential relationship [7]. When looking at the public sequence databases available for deep learning in biology, such as UniProt, NCBI, or MGnify, scaling laws suggest that these datasets represent less than 0.000001% of life on earth [8], [9], [10]. A significant portion of sequences deposited in these databases originate from human, mammals, and model organisms that are cultivated in narrow laboratory conditions [11], [12]. Furthermore, for environmentally collected sequence data, these resources lack consistent geolocation and environmental metadata. The latter point not only precludes us from inferring how comprehensively life on earth is represented in these databases, but also raises questions relating to the governance of biological sequence data.

In the case of environmentally collected biological samples for genomic (or other -omic) purposes, ethical compliance to Access and Benefit Sharing (ABS) agreements is contingent on both explicit prior informed consent (PIC) and mutually agreed terms (MAT) that speak to the potential for commercialization, and consistent traceability from a

sequence to its sampling origin. Regulatory frameworks ensuring ethical ABS upon commercialization of biological resources have been driven on an international level by the United Nations Convention on Biological Diversity (CBD) and are documented in the 2011 Nagoya Protocol [13]. The inclusion of digital sequence information as part of ABS frameworks is an area of significant development in this context [14]. Historically, there are many instances where materials analyzed for research purposes have contributed to commercial assets of unexpectedly high value without due reconsideration of the original agreements under which the samples were accessed and what fair benefit sharing should look like, leading to controversies around biopiracy and impediments to the development of assets that could have been transformative for industry and human health [15], [16], [17], [18].

Here we describe a global metagenomics and biological data supply chain that simultaneously addresses both the issue of equitable benefit sharing of digital sequence information and the large knowledge gap we experience regarding genomic sequence diversity of life on earth with the aim to improve biological deep learning models. The genome and protein sequences as well as consistently collected metadata derived from metagenomic sampling expeditions are captured in a knowledge graph that counts 6 billion relationships at the time of writing. In the context of knowing that models like AlphaFold2 perform less well on orphan proteins for which deep multiple sequence alignments (MSAs) cannot be generated [19], we show that the performance of AlphaFold2 can be improved when MSAs are supplemented with diverse sequences from our knowledge graph. Assessing confidence and accuracy of the predicted structure, we observe the root mean squared deviation (RMSD) compared to ground-truth crystal structures being reduced by up to 80% . We display improved structure predictions for a wide range of CASP15 and CAMEO competition targets [20], [21], and demonstrate that docking performance can be improved as a result, too.

2 A global metagenomic and biological data supply chain addresses the knowledge gap of biological sequence diversity

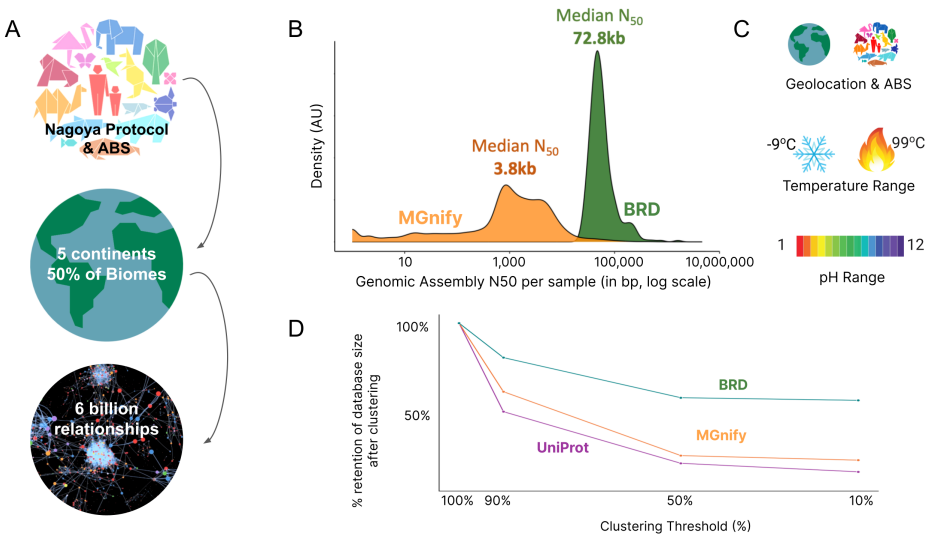


Figure 1: Strategy for accessing and organising data derived from a global metagenomic and bilogical data supply chain. A. Biological and metagenomic sequence collection strategy covering ABS agreements & Nagoya compliance; global expeditions covering 5 continents; and organisation of this data into a knowledge graph (data resource hereforth referred to as BRD). B. Metagenomic assembly length distribution as measured by the N50 value for MGnify and BRD. C. Examples of metadata and features captured in BRD that other resources lack or do not consistently display. D. Protein sequence diversity of MGnify, UniProt, and BRD, as shown by clustering the sequence content.

In order to curate genomic and biological data that are more representative of the true diversity of life on earth, we entered Access Benefit Sharing (ABS) agreements following prior informed consent (PIC) of relevant landowners and stakeholders across 23 nations on 5 continents before conducting environmental metagenomic sampling alongside geological, geographic, and chemical metadata collection (Figure 1 A). The sampling sites cover 50% of global biomes according to the WWF Ecoregion defintion [22]. Methods pertaining to the sampling, sequencing, and bioinformatic assembly and annotation following these expeditions are described in Supplementary Section A1.

We organised all genome and protein sequences alongside chemical and environmental metadata into a knowledge graph counting 6 billion relationships at the time of writing. With the downstream application of MSA generation for

AlphaFold2 predictions in mind, we wanted to ensure that the sequences for such application are derived from high-quality and long (meta)genomic assemblies. The reasoning for this is that a large portion of the sequence database content that AlphaFold2 currently derives MSAs from is MGnify, and a significant portion of the metagenomic assemblies found in MGnify are fragmented and not long enough to cover entire open-reading frames (ORF) for larger proteins [10]. The length distributions of metagenomic assemblies in MGnify and our database (from hereon referred to as BRD, Basecamp Research Data) are displayed in Figure 1B. With consistent metadata collection we were able to sample a wide range of geological and chemical environments, spanning a temperature range of -9 to 99°C (15.8 to 210°F), and a pH range of 1 to 12, as shown in Figure 1C. We then assess the diversity of the protein sequences deposited in BRD compared to MGnify and UniProt by comparing how the size of the databases collapses when clustered at 90%, 50%, and 10% (Figure 1D).

3 Improving AlphaFold2 through MSA augmentation

To leverage the sequence diversity captured in BRD for MSA supplementation during inference without sacrificing too much speed for sequence search, we clustered both MGnify and BRD at a 50% identity threshold using MMseqs2 Linclust [23]. The resulting combined metagenomic sequence dataset contained approximately 1 billion sequences. To assess whether the addition of sequences through MSA supplementation would improve AlphaFold2, we performed structural analysis on sequences from the CASP15 and CAMEO targets.

3.1 CASP15 targets

CASP (Critical Assessment of Structure Prediction) is a biennial global experiment designed to advance the state of the art in modeling the three-dimensional structure of a protein from its amino acid sequence. Organized by the scientific community, it invites participants to present their modeling predictions for a selection of proteins whose experimental structures are yet to be deposited in the Protein Data Bank (PDB) [24].

A

ID	PDB	Protein Name	Taxonomy
T1158	8SXA	MRP4 - E1202Q	*Bos taurus*
T1145	7UWW	Starch-bdgprotein	*Ruminococcus bromii L2-63*
T1134s1	7UBZ	TleTli	*Enterobacter cloacae*
T1106s2	7QIH	YscY-YscX	*Yersinia enterocolitica*
T1133	8DYS	Q9BY44	*Homo sapiens*
T1124	7UX8	MfnG	*Streptomyces drozdowiczii*
T1137s8	8FEF	MceG complex	*Mycobacterium smegmatis*
T1119	7SQ4	7sq4	synthetic construct
T1106s1	7QIH	YscY-YscX	*Yersinia enterocolitica*
T1137s2	8FEF	MceG complex	*Mycobacterium smegmatis*
T1147	8E5M	KEK_05752	*Mycobacterium thermoresistibile*
T1113	7UYX	gp2	*Bacteriophage PA1C*
T1176	8SMQ	idp97509	*Clostridioides difficile 630*

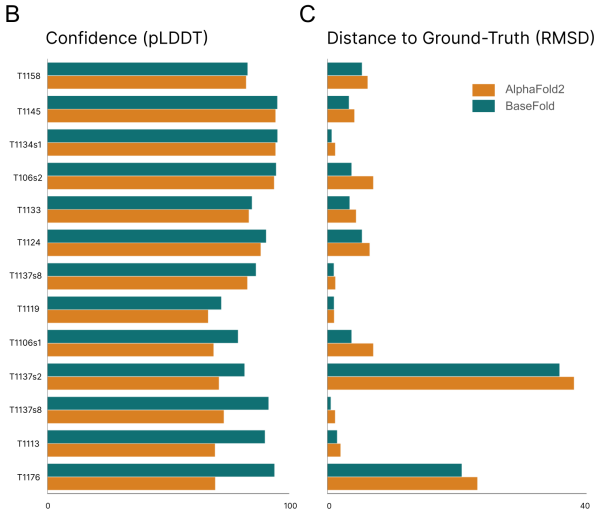


Figure 2: MSA supplementation improves AlphaFold2 performance across CASP15 targets. A. Table of targets where MSA supplementation improves both pLDDT score (shown in B) and RMSD scores (shown in C).

We predicted the structures of 49 CASP15 regular targets, among which single monomeric protein crystal structures were available in the PDB, establishing a benchmark for our comparative study. The structural predictions were evaluated using the predicted Local Distance Difference Test (pLDDT) scores, providing a per-residue confidence metric ranging from 10 to 100 [25]. Among the 49 targets analyzed, 61.22% demonstrated an improvement in pLDDT scores, with increases ranging from 0.08 to 24. The scores of these targets are provided in Supplementary Information Table 1. For the subset of targets where we did not observe an increase in pLDDT the average percentage difference was 3.1%.

Subsequently, we calculated the Root Mean Square Deviation (RMSD), which quantifies the mean distance between corresponding atoms of superimposed protein structures [26]. RMSD is a critical metric in CASP competitions for gauging the congruence of predicted protein structures with their experimentally determined counterparts. An RMSD value ranging from 0 to 3 Ångstroms denotes a high level of structural similarity, particularly in the protein backbone,

indicative of a more accurate prediction. For all targets that had an increased pLDDT score the RMSD score was computed using the SwissModel server [27] which revealed an RMSD score reduction ranging from 0.02 to 3.33. We show an overview of targets from CASP15 where MSA supplementation both improves the pLDDT and reduces the RMSD score in Figure 2. We visualized two specific examples with structural superimposition and corresponding MSA visualization as phylogenetic trees in Figure 3A and 3B.

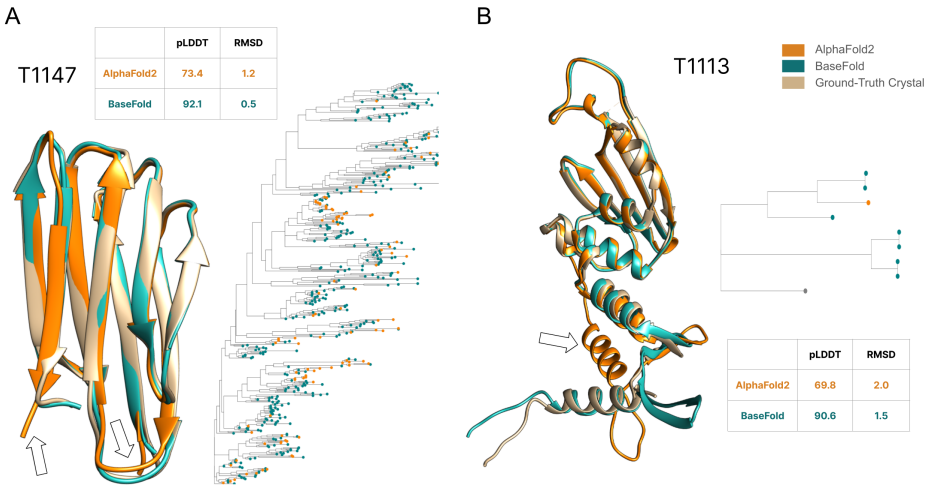


Figure 3: Structural superimposition and corresponding phylogenetic trees derived from MSAs for CASP15 targets T1147 (A) and T1131 (B). Significant discrepancies between the AlphaFold2 prediction and the crystal structure are indicated with a white arrow.

3.2 CAMEO targets

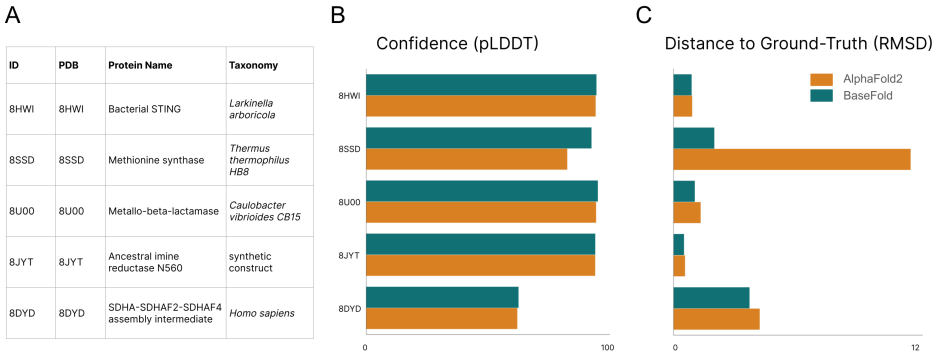


Figure 4: MSA supplementation improves AlphaFold2 performance across a range of CAMEO targets. A. Table of targets where MSA supplementation improves both pLDDT score (shown in B) and RMSD scores (shown in C).

Continuous Automated Model Evaluation (CAMEO) is an online platform that offers automated assessments of 3D protein prediction models, providing weekly updates based on sequences awaiting deposition in the PDB[28]. Expanding to address the structural bioinformatics community's evolving needs, CAMEO features a variety of assessment categories, including prediction coverage, local accuracy, and completeness, while maintaining a focus on evaluating quality estimates for protein structure predictions.

In this study, we predicted the structures of 26 CAMEO targets, predominantly comprising medium to hard difficulty levels. Notably, 57% of these targets demonstrated an increase in pLDDT scores ranging from 0.03 to 10.04. We show an overview of targets from CAMEO where MSA supplementation both improves the pLDDT and reduces the RMSD score in Figure 4 and visualized two specific examples with structural superimposition and corresponding MSA visualization as phylogenetic trees in Figure 5A and 5B.

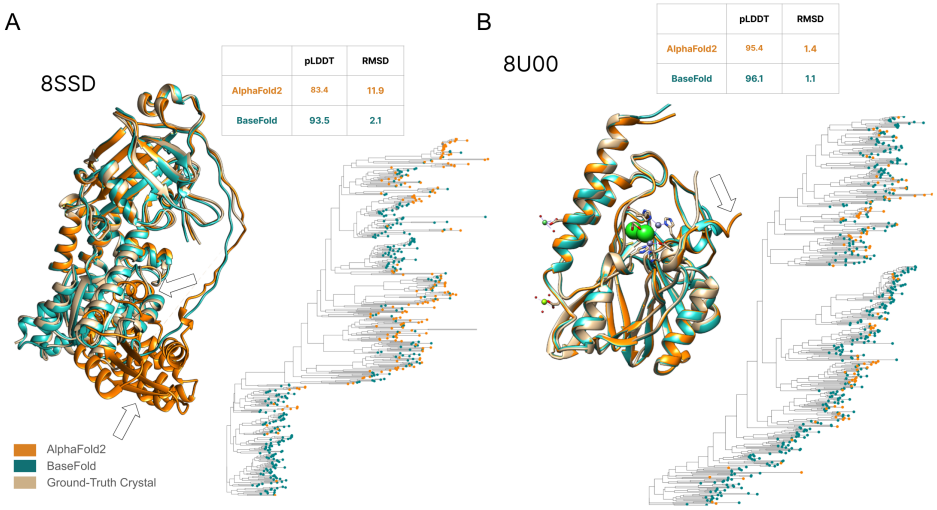


Figure 5: Structural superimposition and corresponding phylogenetic trees derived from MSAs for CAMEO targets 8SSD (A) and 8U00 (B). Significant discrepancies between the AlphaFold2 prediction and the crystal structure are indicated with a white arrow.

3.3 Improving the scale of BaseFold

Building diverse MSAs requires large compute capabilities and is time consuming. To ensure quicker iterations and greater scalability in structure predictions we refined the MSA generation step. We implemented the same strategy implemented by ColabFold [29] for the database preparation in addition to creating two environmental databases to search against which contained BRD clustered at 50% and 90% respectively. More information on the clustering of the respective BRD databases can be found in Supplementary Section A2. We ran these versions of BaseFold using the default settings and predicted the structure of 395 medium and hard CAMEO targets between 2023-02-25 to 2024-02-27 setting the template date to 2023-01-01. We visualise this in Figure 6, where we see that optimisation for speed at lower clustering thresholds does not impact performance significantly.

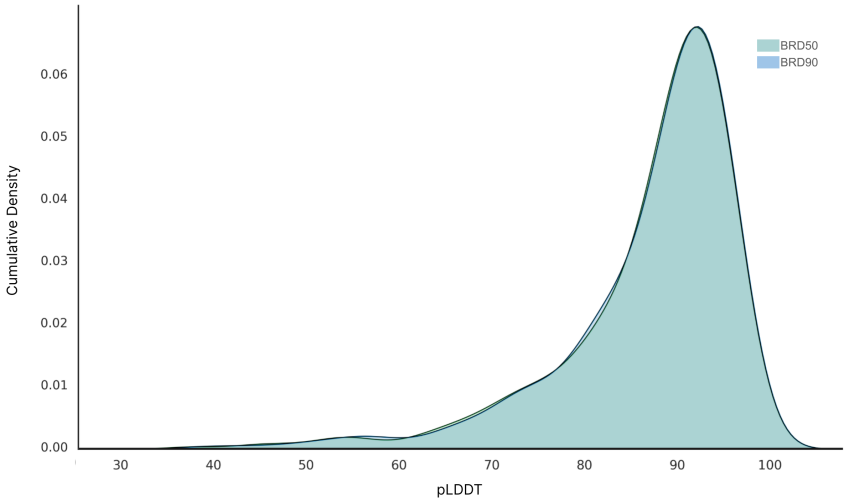


Figure 6: pLDDT distribution of CAMEO targets at 50% and 90% suggest that optimisation for speed at lower clustering thresholds does not impact performance significantly.

3.4 Molecular Docking

Across the structure prediction improvements shown above, we noticed a particularly significant improvement for CASP15 target T1124 (an L- and D-tyrosine O-methyltransferase, MFnG). Using this example, we wanted to as-

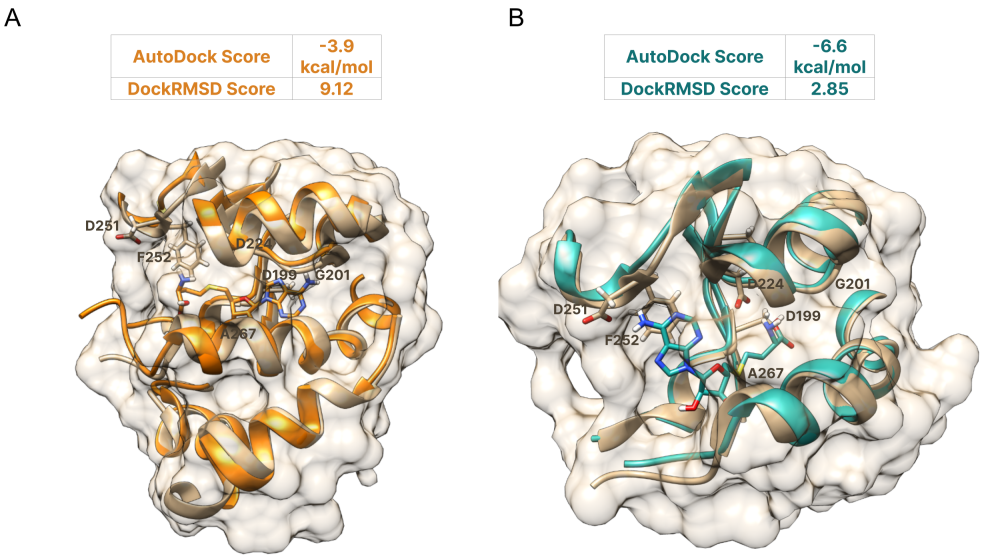


Figure 7: Docking of SAH to MFnG (an L- and D-tyrosine O-methyltransferase, T1124) is improved significantly by MSA-supplemented structure prediction Predicted structure, superimposed to the ground-truth crystal with docked substrates are shown for traditional AlphaFold2 (A) and the MSA-supplemented version (B). Active site residues are highlighted and docking scores indicated in respective tables.

sess the impact of the structural prediction methods on substrate conformation and binding affinity. We performed molecular docking on the MSA-augmented and AlphaFold2-derived structure of MFnG to its substrate S-adenosyl-L-homocysteine (SAH). Docking was performed using AutoDock Vina [30], more information of the docking protocol can be found in Supplementary Section A7. The MSA-augmented structure, when bound to SAH, achieved the lowest binding score of -6.6 kcal/mol, in comparison to the AF2 structure bound to SAH, which yielded a docking score of -3.9 kcal/mol. Further analysis of the docked complexes was assessed using DockRMSD [31]. This tool measures the RMSD between two poses of the same ligand molecule docked onto the same protein structure, without presuming a known atomic order between the two files. The DockRMSD score for the MSA-augmented structure bound to SAH was 2.85 Ångstroms, while for the AF2 structure bound to SAH, it was 9.12 Ångstroms.

4 Discussion

Relative to the magnitude of diversity of life on earth – whether taxonomically or with respect to genomic or protein sequence space – everything that has been captured in public data to date still only represents a tiny fraction. By building a data supply chain in partnership with biodiversity stakeholders we aim to leverage this data to continuously improve deep learning models in biology. Specifically for protein folding, we have demonstrated that by supplementing MSAs with diverse sequences from this supply chain, we can improve AlphaFold2 predictions for a range of targets from the CASP15 and CAMEO competitions. Depending on the protein family and breakdown of the sequence composition of the corresponding MSAs generated during inference, our supplementation approach can improve AlphaFold2 predictions substantially, with some RMSD values (deviation from the ground-truth crystal) decreasing by over 80%. We show that improvements as significant as this can also improve substrate/ligand docking performance. We envision this will positively impact enzyme engineering and drug discovery efforts.

Regarding further work, we foresee additional analysis on the sequence composition of reference databases and what the ideal breakdown of sequence space should look like, in a way that balances both inference speed and accuracy of the predicted structure. Moreover, with further data collection in alignment with the United Nations' ABS principles, we aim to continue unifying the goal of biodiversity conservation efforts with the goal of improving deep learning models in the life sciences in a data-centric manner.

5 Acknowledgements

We are deeply grateful to Noelia Ferruz, Kevin Yang, Ahir Pushpanath, and Phoebe Oldach for helpful feedback and fruitful discussions throughout the writing of this manuscript. We thank Alexandros Papadopolous in particular for

engineering support. We also want to thank Glen-Oliver Gowers, Oliver Vince, Sybil Wong, Leif Christoffersen, Bupe Mwambingu, Nadine Greenhalgh, Emma Bolton, Marlon Clarke, Ineke Knot, Neem Patel, William Chow, Carla Greco, Saif Ur-Rehman, Gus Minto-Cowcher, Keith Kam, Richard De Napoli, Gavin Ayres, Lily Goodyer Sait, and Marcus Leung.

6 References

[1] Sapoval, N., Aghazadeh, A., Nute, M.G. et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun 13, 1728 (2022). https://doi.org/10.1038/s41467-022-29268-7

[2] Khakzad H, Igashov I, Schneuing A, Goverde C, Bronstein M, Correia B. A new age in protein design empowered by deep learning. Cell Syst. 2023 Nov 15;14(11):925-939. doi: 10.1016/j.cels.2023.10.006. PMID: 37972559.

[3] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[4] Minkyung Baek et al.,Accurate prediction of protein structures and interactions using a three-track neural network.Science373,871-876(2021).DOI:10.1126/science.abj8754

[5] Zeming Lin et al.,Evolutionary-scale prediction of atomic-level protein structure with a language model.Science379,1123-1130(2023).DOI:10.1126/science.ade2574

[6] Peng Z, Wang W, Han R, Zhang F, Yang J. Protein structure prediction in the deep learning era. Curr Opin Struct Biol. 2022 Dec;77:102495. doi: 10.1016/j.sbi.2022.102495. Epub 2022 Nov 10. PMID: 36371845.

[7] Sorscher B, Geirhos R, et al. Beyond neural scaling laws: beating power law scaling via data pruning. arXiv (2023). doi: https://doi.org/10.48550/arXiv.2206.14486

[8] UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.

[9] Sayers EW, Bolton EE et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022 Jan 7;50(D1):D20-D26. doi: 10.1093/nar/gkab1112. PMID: 34850941; PMCID: PMC8728269.

[10] Richardson L, Allen B et al. MGnify: the microbiome sequence data analysis resource in 2023, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, D753–D759, https://doi.org/10.1093/nar/gkac1080

[11] Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 2016 May 24;113(21):5970-5. doi: 10.1073/pnas.1521291113. Epub 2016 May 2. PMID: 27140646; PMCID: PMC4889364.

[12] Fishman FJ, Lennon JT. Macroevolutionary constraints on global microbial diversity. Ecol Evol. 2023 Aug 8;13(8):e10403. doi: 10.1002/ece3.10403. PMID: 37560179; PMCID: PMC10408003.

[13] United Nations Convention on Biological Diversity. (2011) Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity, 2011.

[14] United Nations Convention on Biological Diversity. (2022) Decision adopted by the conference of the parties to the convention on biological diversity. Agenda Item 11: Digital sequence information on genetic resources, 2022.

[15] Wadman, M (2023) What does the historic settlement won by Henrietta Lacks's family mean for others? Science. doi: 10.1126/science.adk1834

[16] Mojica, F.J.M., D ez-Villase or, C.S., Garc a-Mart nez, J.S., and Soria, E. (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. J Mol Evol 60, 174–182.

[17] National Park Service, Yellowstone (2019). Bioprospecting. https://www.nps.gov/yell/learn/nature/bioprospecting.html

[18] Stokstad E. (2019) Major U.K. genetics lab accused of misusing African DNA. Science. doi: 10.1126/science.aba0343

[19] Michaud JM, Madani A, Fraser JS. A language model beats AlphaFold2 on orphans. Nat Biotechnol. 2022 Nov;40(11):1576-1577. doi: 10.1038/s41587-022-01466-0. PMID: 36192635; PMCID: PMC9669189.

[20] Alexander LT, Durairaj J et al. Protein target highlights in CASP15: Analysis of models by structure providers. Proteins. 2023 Dec;91(12):1571-1599. doi: 10.1002/prot.26545. Epub 2023 Jul 26. PMID: 37493353; PMCID: PMC10792529.

[21] Leemann M, Sagasta A, Eberhardt J, Schwede T, Robin X, Durairaj J. Automated benchmarking of combined protein structure and ligand conformation prediction. Proteins. 2023 Dec;91(12):1912-1924. doi: 10.1002/prot.26605. Epub 2023 Oct 26. PMID: 37885318.

[22] Olson D. M., Dinerstein E., et al. 2001. Terrestrial ecoregions of the world: a new map of life on Earth. Bioscience 51(11):933-938.

[23] Steinegger, M., Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028 (2017). https://doi.org/10.1038/nbt.3988

[24] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics. 2019;87(12):1011-1020. doi:https://doi.org/10.1002/prot.25823

[25] Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29(21):2722-2728. doi:https://doi.org/10.1093/bioinformatics/btt473 [26] Kufareva I, Abagyan R. Methods of protein structure comparison. Methods in Molecular Biology. 2012;857:231-257. doi:10.1007/978-1-61779-588-6-10

[27] Schwede T. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research. 2003;31(13):3381-3385. doi:https://doi.org/10.1093/nar/gkg520

[28] Haas J, Barbato A, Behringer D, et al. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins: Structure, Function, and Bioinformatics. 2017;86:387-398. doi:https://doi.org/10.1002/prot.25431

[29] Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nature Methods. 2022;19(June 2022):1-4. doi:https://doi.org/10.1038/s41592-022-01488-1

[30] Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2009;31(2):NA-NA. doi:https://doi.org/10.1002/jcc.21334

[31] Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. Journal of Cheminformatics. 2019;11(1). doi:https://doi.org/10.1186/s13321-019-0362-7

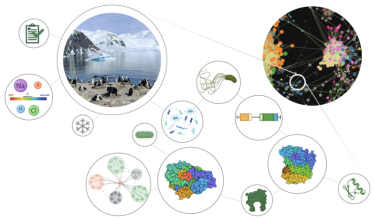
Supplementary Information

March 6, 2024

A Methods

A.1 Global Sampling & Knowledge Graph Construction

Environmental samples subjected to metagenomic sequencing were collected after receiving landowner's permission and entering access-benefit-sharing agreements with the relevant local or national authority, following Nagoya protocol guidelines. All samples were sequenced with both long-read and short-read sequencing methods applied after extraction. Alongside sample collection, we captured consistent metadata collection that include chemical, physical, weather, and geological measurements.



Sample	Genomic Assembly	ORF/Protein	Other
Geolocation	CRISPR Arrays	PFAMs, InterPro	Non-coding RNAs
Biomes	Biosynthetic Gene Clusters	EC Numbers	
Chemical elements & compound measurements	Taxonomy*	Taxonomy*	
Geological & physical measurements	Gene neighborhoods	KEGG & COG	
	Mobile Genetic Elements, Phage Insertions		

Figure 1: Visual representation of the data model for the Knowledge graph described in this study shown on the left. On the right we show a selection of information, measurements, and annotations associated with the entities in the graph. Taxonomies (*) are annotated both on the genomic assembly as well as the open reading frame (ORF) level.

We applied a custom assembly and annotation pipeline to the sequencing data which performs standard QC of sequencing reads and joint assembly of short and long reads to optimise for both low error rate and high assembly length. Open reading frames and non-coding RNAs are annotated on the genomic assemblies, along with CRISPR arrays, biosynthetic gene clusters, gene neighborhoods, mobile genetic elements, and phage integration events. The open-reading frames were translated into amino acid sequences which were subjected to comprehensive *in silico* annotations, including PFAM [1], KEGG [2], COG [3],

InterPro [4], and EC Numbers [6]. Functional annotations were performed with custom Hidden-Markov-Model-based and Deep Learning based models. Multiple custom taxonomic annotation methods were performed both on the ORF and the genomic assembly level. For an overview of these annotations and how they relate to the data model of the knowledge graph, see Figure 1.

For knowledge graph construction, we ingested all measurements, annotations, and information into a neo4j graph database.

A.2 Clustering

To assess the redundancy of protein sequence data across the Uniprot [1], [2], Mgnify [3], and Basecamp databases, we apply a hierarchical clustering strategy utilizing MMSeqs2 [4]. We initially cluster the Mgnify and Basecamp databases to a sequence identity threshold of 90%. Subsequently, we further refine the clustering of these databases down to 50% and 10% sequence identity thresholds. This stepwise reduction approach was chosen for its computational efficiency and reduced time consumption.

For the Uniprot database, we leverage the existing clustered datasets Uniref100, Uniref90, and Uniref50. These datasets provide a basis for our analysis, from which we identify the number of clusters at each identity threshold. Further, we utilize the Uniref50 clustered dataset to further cluster sequences down to a 10% identity threshold. This was achieved by adhering to the same clustering protocol used for the other datasets, which involves clustering sequences based on a 10% sequence identity and an 80% overlap with the longest sequence in the cluster.

A.3 Database Reduction For Efficient Search

To facilitate efficient searching within the environmental databases BRD v.2023.10 and Mgnify v.2022.05, we employed a database reduction strategy. This involved using the Linclust algorithm in Mmseqs2 to cluster both databases with a minimum sequence identity of 50% (–min-seq-id 0.5). This approach effectively reduced the combined database size to 239GB. After selecting cluster representatives, this process resulted in a total of 1.01 billion sequences used for the MSA supplemented flavour of AlphaFold2.

A.4 Template Search

AlphaFold2 [5] employs HHsearch [6] to scan a clustered version of the PDB (PDB70) for identifying the top 20 ranked templates. To maintain consistency with the original AlphaFold submissions for CASP15 targets, we configured the template search cutoff date to January 1, 2022. This setup was crucial to avoid any influence from newly deposited targets that might affect the predictions when using BRD.

A.5 Running BaseFold for Structural Comparison

To evaluate the full impact of the BRD on AlphaFold2 performance, we utilized AlphaFold's default settings changing only the environmental database and template search date when computing the BaseFold structures. This approach aimed to directly compare the structural predictions under standard conditions. The CASP15 AlphaFold2 structures were downloaded from the AlphaFold2 Github repository and the AlphaFold2 structures for the CAMEO targets were downloaded directly from the CAMEO website. All BaseFold model inference was run on 8 NVIDIA A100 Tensor Core GPUs with 80GB of memory.

ID	AlphaFold2	BaseFold	Difference
T1176	69.94	94.61	24.67
T1113	69.83	90.64	20.81
T1147	73.50	92.15	18.64
T1114s1	62.72	78.22	15.50
T1137s7	71.48	82.15	10.68
T1134s2	77.90	88.43	10.53
T1106s1	69.24	79.41	10.17
T1115	75.16	85.02	9.85
T1119	66.97	72.45	5.48
T1170	90.57	94.24	3.67
T1137s8	83.33	86.92	3.59
T1124	88.89	91.13	2.24
T1137s9	84.69	86.89	2.19
T1139	89.52	91.40	1.88
T1155	79.62	81.43	1.81
T1150	92.90	94.57	1.67
T1114s2	86.56	88.11	1.55
T1109	94.48	95.94	1.46
T1153	86.36	87.82	1.46
T1133	83.97	85.23	1.27
T1120	92.24	93.45	1.21
T1157s2	89.17	90.29	1.12
T1110	94.87	95.95	1.08
T1106s2	94.48	95.30	0.82
T1134s1	95.08	95.88	0.81
T1145	95.07	95.83	0.76
T1158	82.79	83.43	0.64
T1127	93.01	93.28	0.27
T1157s1	84.40	84.49	0.09

Table 1: CASP15 targets that display an increase in the pLDDT scores for BaseFold predictions compared to AlphaFold2

Target	Mgnify	BRD
T1158	163	337
T1145	426	74
T1134s1	307	193
T1106s2	163	121
T1133	101	399
T1124	149	351
T1137s8	108	397
T1106s1	19	12
T1137s2	44	456
T1147	104	396
T1113	1	7
T1176	12	3
8HWI	291	209
8SSD	174	326
8U00	105	395
8JYT	80	420
8DYD	404	96

Table 2: Sequence contributions by MGnify and BRD to the MSAs of the following CASP15 and CAMEO targets

A.6 Molecular Docking

In preparation for docking, compound SAH was supplemented with Gasteiger charges followed by the addition of non polar hydrogen atoms [7]. Docking was performed using the default setting of AutoDock Vina [8] with a random seed of 42 and exhaustiveness set to 32. The box was defined using USCF Chimera [9] based on the crystal structure MFnG (PDB:7UX8). The defined dimensions of the box were 9.64 $\times$ 14.86 $\times$ 7.69 with a grid spacing of 1 Å, centered at coordinates x = 21.83, y = 38.56, z = 14.30 to maximize the precision of the substrate positioning within the active site. In the docking process, both the protein and ligands are treated as rigid entities. Results with a positional root-mean-square deviation (RMSD) below 1.0 Å were grouped, with each cluster represented by the most favorable binding free energy. The pose with the lowest binding affinity was then selected and aligned with the receptor structure for further analysis.

B References

[1] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, *et al.* (2021) Pfam: The protein families database in 2021, Nucleic Acids Research, 49(D1), D412–D419, https://doi.org/10.1093/nar/gkaa913

[2] Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. and Ishiguro-Watanabe, M. (2023) KEGG for taxonomy-based analysis of pathways and

genomes. Nucleic Acids Res. 51, D587-D592.

[3] Michael Y Galperin, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, Eugene V Koonin, COG database update: focus on microbial diversity, model organisms, and widespread pathogens, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D274–D281, https://doi.org/10.1093/nar/gkaa1018

[4] Blum M, Chang HY, Chuguransky S, /it et al. 2021 The InterPro protein families and domains database: 20 years on. Nucleic Acids Research. Jan;49(D1):D344-D354. DOI: 10.1093/nar/gkaa977. PMID: 33156333; PMCID: PMC7778928.

[5] Webb, E. C. (1992) Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and molecular biology on the nomenclature and classification of enzymes. *Academic Press.*

[6] Apweiler R. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research. 2004;32(90001):115D119. doi:https://doi.org/10.1093/nar/gkh131

[7] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2014;31(6):926-932. doi:https://doi.org/10.1093/bioinformatics/btu739

[8] 20.Mitchell AL, Almeida A, Beracochea M, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research. 2019;48(D1). doi:https://doi.org/10.1093/nar/gkz1035

[9] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017;35(11):1026-1028. doi:https://doi.org/10.1038/nbt.3988

[10] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-589.

[11] Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20(1). doi:https://doi.org/10.1186/s12859-019-3019-7

[12] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of Cheminformatics. 2011;3(1). doi:https://doi.org/10.1186/1758-2946-3-33

[13] Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2009;31(2):NA-NA. doi:https://doi.org/10.1002/jcc.21334

[14] Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera–A visualization system for exploratory research and analysis. Journal of Computational Chemistry. 2004;25(13):1605-1612. doi:https://doi.org/10.1002/jcc.20084