PLOS ONE

RESEARCH ARTICLE

A resource for improved predictions of *Trypanosoma* and *Leishmania* protein three-dimensional structure

Richard John Wheeler *

Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, United Kingdom

* richard.wheeler@ndm.ox.ac.uk

Citation: Wheeler RJ (2021) A resource for improved predictions of *Trypanosoma* and *Leishmania* protein three-dimensional structure. PLoS ONE 16(11): e0259871. https://doi.org/10.1371/journal.pone.0259871

Editor: Vyacheslav Yurchenko, University of Ostrava, CZECH REPUBLIC

Received: September 9, 2021

Accepted: October 27, 2021

Published: November 11, 2021

Data Availability Statement: The protein sequence database is available from Zenodo (record 5563074, doi:10.5281/zenodo.5563074).

Funding: RJW 211075/Z/18/Z Wellcome Trust https://wellcome.org/ The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

AlphaFold2 and RoseTTAfold represent a transformative advance for predicting protein structure. They are able to make very high-quality predictions given a high-quality alignment of the protein sequence with related proteins. These predictions are now readily available *via* the AlphaFold database of predicted structures and AlphaFold or RoseTTAfold Colaboratory notebooks for custom predictions. However, predictions for some species tend to be lower confidence than model organisms. Problematic species include *Trypanosoma cruzi* and *Leishmania infantum*: important unicellular eukaryotic human parasites in an early-branching eukaryotic lineage. The cause appears to be due to poor sampling of this branch of life (Discoba) in the protein sequences databases used for the AlphaFold database and ColabFold. Here, by comprehensively gathering openly available protein sequence data for Discoba species, significant improvements to AlphaFold2 protein structure prediction over the AlphaFold database and ColabFold are demonstrated. This is made available as an easy-to-use tool for the parasitology community in the form of Colaboratory notebooks for generating multiple sequence alignments and AlphaFold2 predictions of protein structure for *Trypanosoma*, *Leishmania* and related species.

Introduction

Machine learning approaches to protein structure prediction have crossed a critical success threshold. While predicting the three-dimensional structure of a protein from sequence alone is still unsolved problem, a multiple sequence alignment (MSA) of the target protein sequence with related proteins provides key additional information. Cutting edge approaches using such MSAs now have the potential to reach very high accuracy. MSAs are the input for AlphaFold2 [1] and RoseTTAfold [2], with AlphaFold2 reaching the highest accuracy prediction at the most recent Critical Assessment of protein Structure Prediction (CASP) competition (CASP14 [3])–an accuracy comparable to experimental protein structure determination. AlphaFold2-predicted structures for the near-whole proteome of 21 species [4] has been made publicly available, and tools like ColabFold [5] and the official AlphaFold Colaboratory notebook [6] make custom predictions easily accessible.

Trypanosomatids pose a challenge because of their large evolutionary distance from common model eukaryotes [7, 8]. This order includes important unicellular human, animal and plant parasites, including the human infective *Trypanosoma cruzi*, *Trypanosoma brucei* and many human-infective *Leishmania* species. *T. cruzi* and *Leishmania infantum* are the most deadly of these species and were included in the initial 21 AlphaFold whole proteome predictions [1, 4]. Trypanosomatids are members of an early-branching eukaryote linage (Discoba) which also includes the less common, but still deadly, pathogen *Naegleria fowleri*. Other speciose Discoba lineages are Euglena and Diplonema, unicellular aquatic organisms and important and abundant auto- and heterotrophic plankton respectively. An initial inspection of the AlphaFold database (alphafold.ebi.ac.uk) suggested protein structure prediction accuracy for *T. cruzi* and *L. infantum* is often low–particularly for kinetoplastid specific proteins–based on self-reported prediction quality scores. Many of these proteins are vital, like the unconventional kinetochore proteins [9].

Discoba diversity is less well sampled by genomes and transcriptomes than lineages like plants or metazoa, making construction of deep MSAs more difficult. This is important as MSAs encode additional structural information beyond the primary protein sequence alone: They capture evidence for co-evolution of different regions of the primary sequence which may correspond to proximity or interaction in the three-dimensional structure. AlphaFold2 and RoseTTAFold prediction of protein structure is greatly improved by high MSA quality and depth, with high MSA coverage critical for high confidence predictions [1, 2]. While new approaches [10] are trying to move beyond multiple sequence alignments, MSAs will remain a powerful source of information.

Currently, the input databases for the AlphaFold database and the ColabFold notebook are of UniRef [11] plus environmental sample sequence databases (BFD, Uniclust and MGNify [12–14]). However, these databases have relatively poor coverage of Discoba. It appears that a significant quantity of genomic and transcriptomic data available in the community genome resource TriTrypDB [15, 16], the NCBI genome [17], transcriptome shotgun assembly (TSA) [18] and sequencing read archive (SRA) [19] databases are not incorporated. It seemed likely that an improved database is a simple opportunity to improve protein MSAs for protein structure predictions for *Trypanosoma*, *Leishmania* and other Discoba species.

Here, protein sequence data was gathered into a comprehensive Discoba database and the ColabFold MMSeqs2-based pipeline [5, 20] was modified to also include the result of a HMMER search of Discoba. Using a test set of 30 *L. infantum* proteins, MSA coverage was always improved, leading to increased AlphaFold2 prediction accuracy in 2/3 of cases. Improvements were greatest for kinetoplastids-specific proteins, with dramatic improvements often possible. The necessary tools to make similar protein structure predictions have been made openly available: The Discoba protein sequence database (for custom searches and MSA generation), Colaboratory notebooks for generating MSAs by HMMER or MMSeqs2 (for use in AlphaFold2 or RoseTTAFold implementations), and a standalone Colaboratory notebook for AlphaFold2 structure predictions based on ColabFold incorporating a search of the Discoba database. These tools are available at github.com/zephyris/discoba_alphafold.

Methods

Discoba sequence data

Predicted protein sequences were gathered from 243 Discoba transcriptomes or genomes (S1 Table): 160 transcriptomes and 83 genomes. 152 from cultured populations (almost all axenic) and 91 from single cell samples. 238 were assembled giving a good number of predicted

protein sequences (>500). The full set of sequences have been deposited as a Zenodo dataset (version 1.0.0) [21].

Predicted proteins from genome sequencing were gathered from two sources: TriTrypDB: All 53 trypanosomatid species available in TriTrypDB [15,16] release 53, using the provided predicted protein sequences on TriTrypDB where available. For the 17 without predicted protein sequences the translation of all predicted open reading frames (ORFs) over 100 amino acids were used, as kinetoplastids typically have compact genomes with short intergenic sequences and extremely low occurrence of introns [22]. NCBI Genomes: 32 genomes for Discoba species. For the 14 with predicted protein sequences on NCBI the existing prediction was used. For the 18 without predictions, the translation of all ORFs over 100 amino acids were used.

Sequencing read archive (SRA): 17 whole genome sequencing (WG-seq) datasets from axenic cultures of Discoba species and 32 single cell WG-seq datasets. For each, assembly was carried out using Velvet [23, 24] (see Genome assembly) and all predicted ORFs over 100 amino acids were used.

Predicted proteins from transcriptome sequencing were gathered from three sources: Transcriptome shotgun assembly (TSA) database: 11 transcriptomes for Discoba species, using protein sequence predicted by TransDecoder [25]. Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP): 2 transcriptomes for Discoba species, using the provided protein sequences which were predicted using TransDecoder. NCBI SRA: 19 mRNA-seq datasets from axenic cultures of Discoba species, 2 mRNA-seq datasets from mixed cultures including a Discoba species and 59 single cell mRNA-seq datasets. For each, transcriptome assembly was carried out using Trinity [26–28] followed by protein sequence prediction with TransDecoder [25] (see Transcriptome assembly).

Transcriptome assembly

Transcriptome assembly from RNA-seq data used a standardised pipeline, with the same approach used for axenic culture, mixed culture and single cell transcriptomic data. Reads were first error corrected using Rcorrector v1.0.4 [29, 30] (using Jellyfish v2.3.0 [31, 32]) and corrected reads tidied using TranscriptomeAssemblyTools [33]. Any remaining adaptor sequences were trimmed using TrimGalore v0.6.0 [34] (using Cutadapt v2.8 [35, 36]) then an assembly was generated using Trinity v2.12.0 [26–28]. Many of these species use polycistronic transcription with a single spliced leader sequence trans-spliced onto the start of all mRNAs. As such common sequences may affect assembly, a two-step approach was used. First, a trial assembly using 1,000,000 reads (or all reads if fewer were available) was generated and the common spliced leader sequence identified using a custom script. Cutadapt was then used to trim reads to remove the spliced leader, then a final assembly was generated using 40,000,000 reads (or all reads if fewer were available). Very similar transcript sequences were removed using cd-hit-est v4.8.1 (part of CD-HIT [37, 38]) then remaining sequences translated to predicted proteins using TransDecoder v5.5.0 [25] LongOrfs.

Genome assembly

Genome assembly from WG-seq data also used a standardised pipeline. For single cell genomic data, reads were first error corrected using Rcorrector v1.0.4 [29, 30] (using Jellyfish v2.3.0 [31, 32]) and TranscriptomeAssemblyTools [33]. For all assemblies, any remaining adaptor sequences were trimmed using TrimGalore v0.6.0 [34] (using Cutadapt v2.8 [35, 36]) then an assembly was generated using Velvet v1.2.10 [23, 24] using all available reads. As expected coverage and insert size are not necessarily known, a refinement step was used. Reads were aligned

to the assembly using bwa mem v0.7.17 [39, 40] and insert size and mean coverage determined using samtools v1.10 [41, 42], then a final assembly was generated using Velvet including these parameters and a minimum coverage threshold of 0.25 the mean trial assembly coverage. All open reading frames ≥300 bp (all three frames, both strands) were identified using a custom script.

Orthology

Protein orthogroups were identified using OrthoFinder v2.5.4 [43–45] (using diamond v2.0.5.143 [46, 47] and FastME 2.1.4 [48]). Reciprocal best protein sequence search hits were carried out using diamond v2.0.5.143 [46, 47] with no e-value cut-off. OrthoFinder and reciprocal best sequence search hit analysis were carried out on a diverse set of 77 UniProt reference eukaryote proteomes [49]: UP000001450, UP000002729, UP000007800, UP000012073, UP000054560, UP000000437, UP000001542, UP000005203, UP000008144, UP000013827, UP000059680, UP000000539, UP000001548, UP000005226, UP000008153, UP000014760, UP000179807, UP000000559, UP000001593, UP000005640, UP000008493, UP000018208, UP000186817, UP000000560, UP000001926, UP000006548, UP000008524, UP000023152, UP000218209, UP000000561, UP000001940, UP000006671, UP000008743, UP000027080, UP000247409, UP000000589, UP000001950, UP000006727, UP000008827, UP000030693, UP000265515, UP000000600, UP000002195, UP000006729, UP000009022, UP000030746, UP000265618, UP000000759, UP000002296, UP000006906, UP000009138, UP000036983, UP000316726, UP000000803, UP000002311, UP000007110, UP000009168, UP000037460, UP000323011, UP000000819, UP000002485, UP000007241, UP000009170, UP000051952, UP000324585, UP000001357, UP000002494, UP000007305, UP000009377, UP000054408, UP000444721, UP000001449, UP000002640, UP000007799, UP000011083, UP000054558. This includes 6 kinetoplastid species (*Bodo saltans*, *Leishmania infantum*, *Leishmania mexicana*, *Perkinsela sp*., *Trypanosoma brucei brucei* and *Trypanosoma cruzi*), which were used as the basis for identifying kinetoplastid specific proteins.

Intrinsically disordered domains

Intrinsically disordered domains were predicted using IUPred2A [50] using a score threshold of 0.5 for classification of a residue as disordered.

AlphaFold2 predictions

Existing AlphaFold predictions of protein structures for *Leishmania infantum* (UP000008153) *Trypanosoma cruzi* (UP000002296) and *Mus musculus* (UP000000589) were taken from alphafold.ebi.ac.uk [1], last updated using AlphaFold v2.0 2021-07-01. Per residue and global predicted local distance difference test score (pLDDT) was taken from the mmCIF file, predicted average error (pAE) from the error json file.

ColabFold predictions were made using an unmodified version of ColabFold [5, 20], with the default MSA pipeline, a MMseqs2 [51] search of UniRef [11] and environmental sample sequence databases [12–14]. Predictions were done using AlphaFold2 parameters from 2021-07-14, not using Amber [52] relaxation and not using PDB [53] templates. Due to GPU memory availability in Google Colaboratory, predictions were restricted to proteins with ≤800 amino acids.

AlphaFold2 predictions incorporating the new Discoba protein sequence database were carried out using a modified version of ColabFold [5, 20]. The MMseqs2 [51] search of UniRef [11] and environmental sample sequence databases [12–14], was supplemented with a HMMER (part of HH-suite) [54] search of the Discoba protein database described here. A

MMSeqs2 search of the Discoba protein database was also trialled but use of HMMER for Discoba searches typically gave slightly higher pLDDT, presumably as AlphaFold v2.0 was trained using HMMER MSAs. Predictions were again done using AlphaFold2 parameters from 2021-07-14, not using Amber relaxation and not using PDB templates.

The test set of *L*. *infantum* proteins were selected, using a random number generator, from the proteins meeting the criteria for each group, see Results for selection criteria. Randomly selected conserved genes: A4HU53, A4I2E1, A4HUD2, A4IA46, A4I444, A4IC57, A4IAB2, A4I7M6, A4I0C5, A4HTD2. Randomly selected not conserved genes: A4I944, A4HYM2, A4I1S6, A4I5D0, A4IAG0, A4I787, A4I9X8, A4IB72, A4I5C1, A4HS18. Randomly selected not conserved 'promising' (see Results) genes: E9AGZ8, A4HW74, A4HZS9, A4I0P7, A4I2Z9, A4IDS7, A4HRK9, A4IBK2, A4I4D7, E9AGB8.

Results

AlphaFold2 self-reports confidence in predictions through two measures: predicted local distance difference test score (pLDDT) [55], a per residue 0 to 100 score with high values showing higher confidence, and predicted average error (pAE), a per residue pair distance score with low values showing lower error. Many AlphaFold2-predicted protein structures for *Leishmania infantum* and *Trypanosoma cruzi* (from the AlphaFold database alphafold.ebi.ac.uk [1, 4]) had high pLDDT and low pAE. This is an impressive achievement, however using the Alphafold database pLDDT proteome-wide performance of AlphaFold2 can be evaluated quantitatively. For comparison mouse (*Mus musculus*) was selected as, unlike human proteins, predictions were carried out without special treatment. Overall, *L*. *infantum* and *T*. *cruzi* protein structure predictions are skewed to lower pLDDT (Fig 1A).

Lower pLDDT could be explained by more disordered protein domains, as predictions for these regions correlate with low pLDDT [1]. *L*. *infantum* and *T*. *cruzi* proteins do not have a markedly different predicted degree of disorder to *M*. *musculus* (Fig 1B) although, as expected [1], pLDDT had a negative correlation with disorder score in all three species (Fig 1C). Alternatively, it may be a limitation due to the depth of the input protein MSAs. pLDDT correlated with number of orthologs detected using OrthoFinder [43–45] on a set of 77 reference proteins of diverse eukaryotes (Fig 1D), indicating that MSA depth is a likely explanation.

Unlike *M*. *musculus*, the distribution of number of orthologs for *L*. *infantum* and *T*. *cruzi* was strongly bimodal with many having fewer than 10. These proteins had, on average, markedly lower pLDDTs (Fig 1D). Analysis using more stringent measures of protein specificity to the kinetoplastids showed a similar pattern: Kinetoplastids-specific proteins were identified as those with only reciprocal best sequence search hits among the kinetoplastids (Fig 1E) or those with only orthogroup members among the kinetoplastids (Fig 1F), and both sets had low pLDDT (Fig 1E, 1F). Overall, this confirms that MSA quality is likely the limiting factor for many *T*. *cruzi* and *L*. *infantum* protein structure predictions.

As much protein sequence data as possible was therefore gathered for Discoba species, drawing upon both TriTrypDB [15, 16] (well known to the *Trypanosoma* and *Leishmania* community), and lesser known, unpublished or very recent data available *via* nucleotide sequencing databases, gathered using the NCBI taxonomy browser (see Methods, S1 Table). Many of these proteomes are in UniParc, but seemingly not used to build UniRef100 which is one of the key databases used by the AlphaFold database and ColabFold. Unlike many applications, precise knowledge about sample or species identity, high sample purity and high transcriptome/genome coverage are not critical–therefore the gather was as inclusive as possible. This database ultimately included 238 predicted proteomes, representing 1.45 billion amino acids across 4.3 million protein sequences.

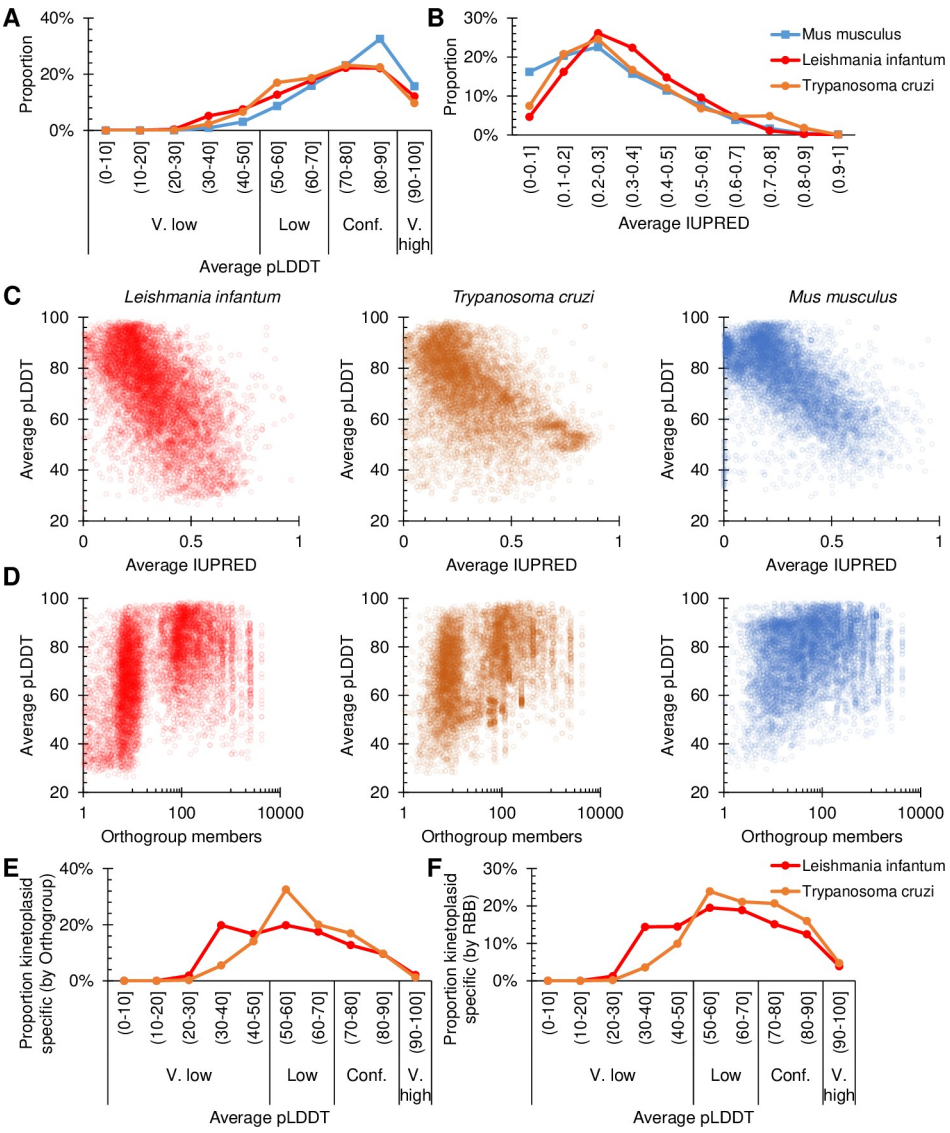

Fig 1. Proteome-wide quality of protein structure predictions of kinetoplastid proteins in comparison to mouse proteins in the AlphaFold database. A) Distribution of per-protein average pLDDT for all *L. infantum* (7924), *T. cruzi* (19024) and, for comparison, *M. musculus* (21588) proteins, from the AlphaFold database [1, 4]. Scores for very low, low, confident and very high confidence categories are the same as used on the website. **B)** Distribution of per-protein average IUPred score for the same three species. **C)** Correlation of per-protein average pLDDT with IUPred score for the same three species. **D)** Correlation of per-protein average pLDDT with number of orthologs (total number of orthogroup members determined from a diverse set of eukaryotes, see Methods). A random number between 0 and 1 was added to each ortholog count to better represent point density at low ortholog numbers. **E,F)** Distribution of per-protein average pLDDT for all *L. infantum* and *T. cruzi* proteins lacking an ortholog outside of the kinetoplastids, as determined by either **E)** orthogroup members only in kinetoplastid species (1509 and 7181 proteins respectively) or **F)** reciprocal best protein sequence search hits only in kinetoplastids species (2361 and 11723 proteins respectively).

https://doi.org/10.1371/journal.pone.0259871.g001

To benchmark any improvements over the AlphaFold database predictions a set of 30 *L. infantum* proteins were selected: 10 random proteins which have orthologs in many diverse eukaryotes, 10 random proteins which appear unique to the kinetoplastid lineage (no orthogroup members outside the kinetoplastids) and 10 random proteins which appeared 'promising' and likely to have globular domains but with a low pLDDT in the AlphaFold

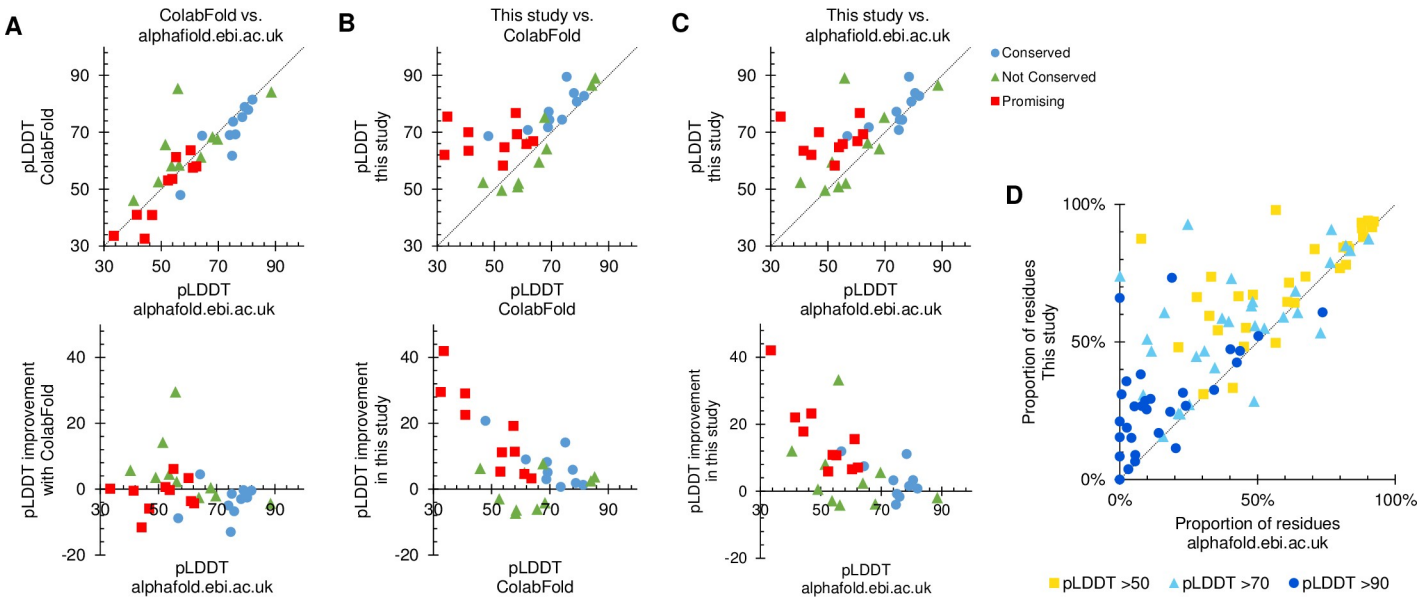

Fig 2. Improved AlphaFold2 predictions using ColabFold and a wider set of Discoba sequences for MSAs. A-C) Comparison of per-protein average pLDDT for 30 test proteins, 10 random widely conserved proteins, 10 random kinetoplastids-specific proteins and 10 'promising' kinetoplastid-specific proteins which appeared likely to improve with additional MSA sequences. **A)** Unmodified ColabFold in comparison to the AlphaFold database plotted as: Top, raw pLDDTs. Points to the top left of the diagonal represent improved (higher pLDDT) predictions. Bottom, change in pLDDT. Points above the horizontal axis represent improvement. **B)** This study (ColabFold with HMMER search of additional Discoba sequences) in comparison to unmodified ColabFold. **C)** This study in comparison to the AlphaFold database. **D)** The same comparison as C) but plotting the proportion of residues over different threshold pLDDT values instead of mean pLDDT.

https://doi.org/10.1371/journal.pone.0259871.g002

database. The latter were selected based on size (avoiding small proteins, ≳300 amino acids), lack of low complexity or repetitive regions (≲30% unstructured and manually avoiding repeats), orthologs in few species (≲10), without numerous paralogs, and low average pLDDT (≲60).

To carry out AlphaFold2 protein structure predictions ColabFold was selected as a fast but high accuracy and accessible AlphaFold2 implementation [5, 20]. As expected, unmodified ColabFold gave per-protein mean pLDDTs comparable to, but on average slightly lower than, the AlphaFold database for the test proteins (Fig 2A). Lower pLDDTs may be through ColabFold's use of MMSeqs2 rather than HMMER, on which AlphaFold2 was originally trained, for MSAs. ColabFold was then modified to generate a HMMER-generated MSA from the Discoba database and append this to the default MSA, before carrying out the AlphaFold2 prediction. This extended MSA improved mean pLDDT for a large majority of protein structure predictions, whether compared to the AlphaFold database or unmodified ColabFold, with less confident predictions seeing the largest improvement (Fig 2B and 2C). pLDDT increase occurred at all confidence levels within a protein. Using the confidence thresholds in the AlphaFold database, the proportion of residues over the threshold for a low confidence (>50), confident (>70, high confidence in backbone structure) and high confidence (>90, likely correct side-chain rotamers) prediction almost all increased for a large majority of proteins (Fig 2D).

Improvement was most marked among the test proteins not conserved outside of the kinetoplastids, especially the ones selected as 'promising' (Fig 2A). Inspection of these predictions showed a range of improvements, best interpreted from plots of pAE which show a pairwise measure of predicted error in residue-residue spacing. Improvements included overall large decreases in pAE (Fig 3A), the first high confidence prediction of any folds (Fig 3B) and the prediction of a single high confidence domain (one contiguous block of low pAE) rather than two subdomains (Fig 3C).

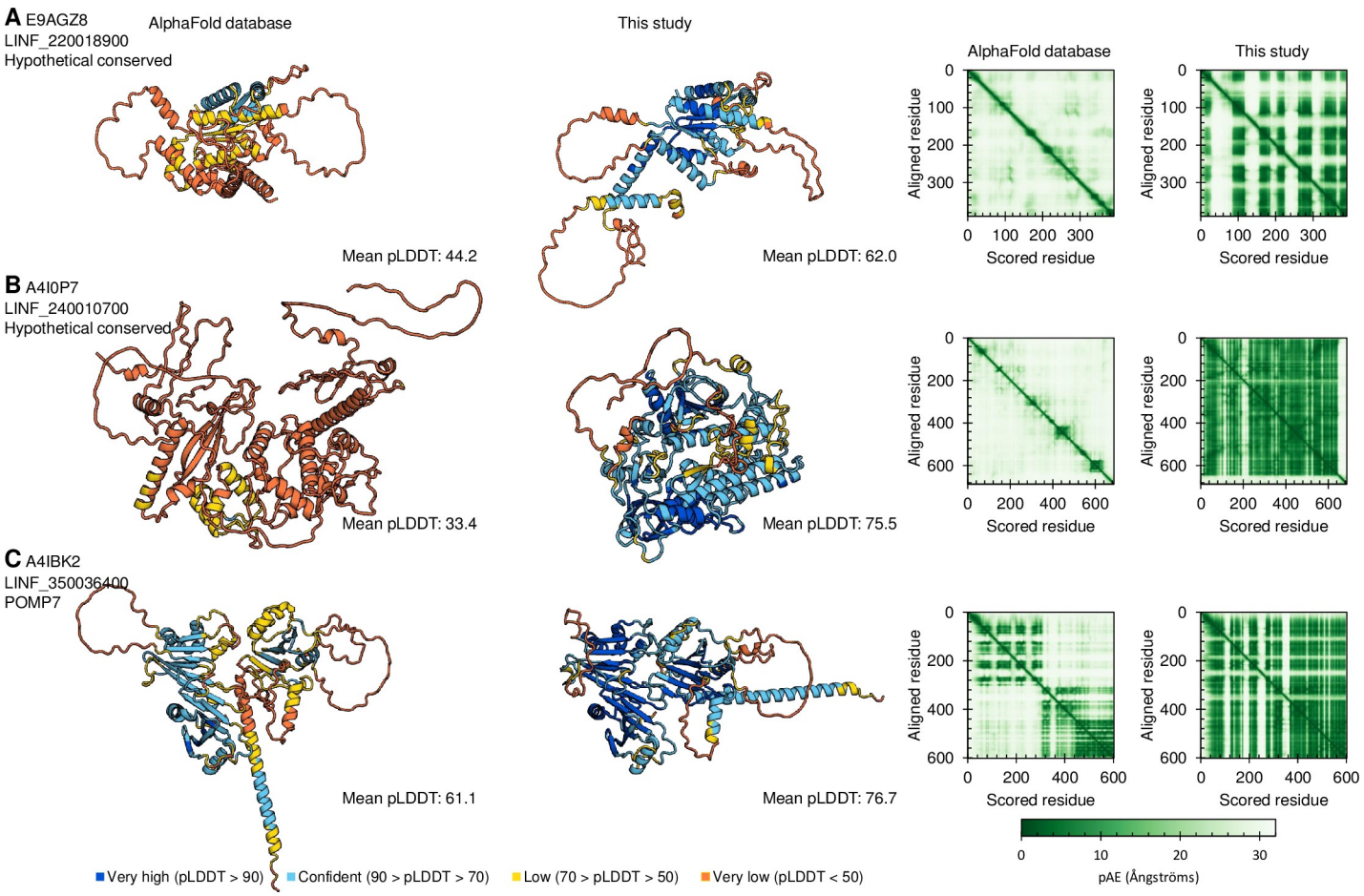

Fig 3. Example *L. infantum* proteins showing significant improvement in structure prediction over the AlphaFold database. Predicted protein structures for three example *L. infantum* proteins showing, from left to right, the AlphaFold database structure, the predicted structure using ColabFold supplemented with a HMMER search of the Discoba database described in this study, the pairwise pAE for the AlphaFold database structure and the pAE for the structure predicted in this study.

https://doi.org/10.1371/journal.pone.0259871.g003

Discussion

This work shows that significant improvement in the pLDDT and pAE of AlphaFold2 structure predictions is possible for *Trypanosoma* and *Leishmania* proteins, relative to the publicly available AlphaFold database at alphafold.ebi.ac.uk [1, 4] and the open tool ColabFold [5, 20] (Fig 2). This was simply by designing a protein sequence database for MSA generation more appropriate for the Discoba branch of eukaryotic life. Easy-to-use tools for MSA generation and AlphaFold2 structure prediction exploiting this Discoba protein sequence database have been made available at github.com/zephyris/discoba_alphafold. Even in these early-branching eukaryotes, the huge advance AlphaFold (and RoseTTAfold) represent can therefore, to a great extent, translate protein structure determination into a genome and transcriptome sequencing problem. Although, experimental protein structure determination will continue to be vital to confirm predictions, explore dynamics and complexes, etc.

Structure prediction improvement was most marked for proteins specific to the kinetoplastids. A large proportion of trypanosomatid parasites' genomes falls into this group–several hundred to thousands depending on definition (Fig 1E and 1F). Many of these proteins lack any domains detectable by primary sequence (sometimes called the 'dark proteome') making a structure prediction a first insight into potential function. However, improvement in protein

structure prediction at all levels are valuable. It may allow a high-confidence prediction of vital kinetoplastid proteins with orthologs in many parasite species, allowing analysis of high specificity small molecule docking. This is of potential importance for drug development.

Improvement is certainly not guaranteed for any individual protein: Proteins well conserved across diverse eukaryotes will already have deep MSAs giving high confidence structural prediction (*cf*. Fig 1D). Proteins which have intrinsically disordered domains (eg. many RNA binding proteins) or only gain structure as part of a multisubunit structure (eg. many ribosome proteins) are unlikely to see significant improvement (*cf*. Fig 1C). Also, proteins which are extremely fast-evolving, or recent innovation found only in a few species, are less likely to benefit. Properties of kinetoplastids chromosome organisation may enable future developments. The order of genes on chromosomes is well conserved [56], sometimes providing additional information which allows identification of orthologs which are difficult or impossible to detect based on primary sequence alone (eg. Basalin [57]) which may allow even deeper MSA generation.

Overall, this work highlights both the importance of sequencing diverse organisms, for example animal pathogens related to human pathogens and non-pathogenic relatives, and ensuring that this data is made available through nucleotide sequencing, genome and proteome databases. It also emphasises that protein sequence data need to be carefully gathered before embarking on important or large-scale structural predictions: Even carefully selected representative databases often retain biases towards model organisms. Similar protein structure prediction improvements are likely possible for other branches of eukaryotic life.

Supporting information

S1 Table. Genome and transcriptome data used to generate the Discoba protein database.
(DOCX)

Acknowledgments

I would like to particularly thank the numerous research groups who have made their sequencing data available through publicly available databases.

Author Contributions

Data curation: Richard John Wheeler.

Formal analysis: Richard John Wheeler.

Funding acquisition: Richard John Wheeler.

Investigation: Richard John Wheeler.

Methodology: Richard John Wheeler.

Project administration: Richard John Wheeler.

Resources: Richard John Wheeler.

Software: Richard John Wheeler.

Validation: Richard John Wheeler.

Visualization: Richard John Wheeler.

Writing – original draft: Richard John Wheeler.

Writing – review & editing: Richard John Wheeler.

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 1–11. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
2. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021. https://doi.org/10.1126/science.abj8754 PMID: 34282049
3. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, and Bioinformatics. [cited 17 Aug 2021]. https://doi.org/10.1002/prot.26171 PMID: 34218458
4. AlphaFold Protein Structure Database. [cited 17 Aug 2021]. Available: https://alphafold.ebi.ac.uk/
5. Ovchinnikov S, Mirdita M, Steinegger M. ColabFold—Making protein folding accessible to all. 2021 [cited 11 Aug 2021]. https://doi.org/10.5281/zenodo.5123297
6. AlphaFold. DeepMind; 2021. Available: https://github.com/deepmind/alphafold
7. Butenko A, Opperdoes FR, Flegontova O, Horák A, Hampl V, Keeling P, et al. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. BMC Biology. 2020; 18: 23. https://doi.org/10.1186/s12915-020-0754-1 PMID: 32122335
8. Maslov DA, Opperdoes FR, Kostygov AY, Hashimi H, Lukeš J, Yurchenko V. Recent advances in trypanosomatid research: genome organization, expression, metabolism, taxonomy and evolution. Parasitology. 2019; 146: 1–27. https://doi.org/10.1017/S0031182018000951 PMID: 29898792
9. Akiyoshi B, Gull K. Discovery of Unconventional Kinetochores in Kinetoplastids. Cell. 2014; 216: 379–391. https://doi.org/10.1016/j.cell.2014.01.049 PMID: 24582333
10. Chowdhury R, Bouatta N, Biswas S, Rochereau C, Church GM, Sorger PK, et al. Single-sequence protein structure prediction using language models from deep learning. 2021 Aug p. 2021.08.02.454840. https://doi.org/10.1101/2021.08.02.454840
11. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research. 2017; 45: D170–D176. https://doi.org/10.1093/nar/gkw1081 PMID: 27899574
12. Levy Karin E, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020; 8: 48. https://doi.org/10.1186/s40168-020-00808-x PMID: 32245390
13. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research. 2020; 48: D570–D578. https://doi.org/10.1093/nar/gkz1035 PMID: 31696235
14. Delmont TO, Gaia M, Hinsinger DD, Fremont P, Guerra AF, Eren AM, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. 2020 Oct p. 2020.10.15.341214. https://doi.org/10.1101/2020.10.15.341214
15. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Research. 2009; 38: D457–D462. https://doi.org/10.1093/nar/gkp851 PMID: 19843604
16. Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, et al. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. 2017; 45: D581–D591. https://doi.org/10.1093/nar/gkw1105 PMID: 27903906
17. NCBI Genome. [cited 31 Aug 2021]. Available: https://www.ncbi.nlm.nih.gov/genome/
18. Transcriptome Shotgun Assembly Sequence Database. [cited 31 Aug 2021]. Available: https://www.ncbi.nlm.nih.gov/genbank/tsa/
19. NCBI Sequencing Read Archive. [cited 31 Aug 2021]. Available: https://www.ncbi.nlm.nih.gov/sra
20. Mirdita M, Ovchinnikov S, Steinegger M. ColabFold—Making protein folding accessible to all. 2021 Aug p. 2021.08.15.456425. https://doi.org/10.1101/2021.08.15.456425
21. Wheeler RJ. Discoba protein sequences for protein structure predictions. Zenodo; 2021. https://doi.org/10.5281/zenodo.5563074
22. Daniels J-P, Gull K, Wickstead B. Cell biology of the trypanosome genome. Microbiol Mol Biol Rev. 2010; 74: 552–569. https://doi.org/10.1128/MMBR.00024-10 PMID: 21119017
23. Velvet. 2021. Available: https://github.com/dzerbino/velvet
24. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821–829. https://doi.org/10.1101/gr.074492.107 PMID: 18349386
25. TransDecoder. 2020. Available: https://github.com/TransDecoder/TransDecoder

26. Trinity RNA-Seq. 2021. Available: https://github.com/trinityrnaseq/trinityrnaseq
27. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 2013; 8. https://doi.org/10.1038/nprot.2013.084 PMID: 23845962
28. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011; 29: 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440
29. Rcorrector. 2021. Available: https://github.com/mourisl/Rcorrector
30. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience. 2015; 4: 48. https://doi.org/10.1186/s13742-015-0089-y PMID: 26500767
31. Jellyfish. 2021. Available: https://github.com/gmarcais/Jellyfish
32. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27: 764–770. https://doi.org/10.1093/bioinformatics/btr011 PMID: 21217122
33. harvardinformatics/TranscriptomeAssemblyTools. Harvard Informatics; 2020. Available: https://github.com/harvardinformatics/TranscriptomeAssemblyTools
34. Krueger F. Trim Galore. 2021. Available: https://github.com/FelixKrueger/TrimGalore
35. Cutadapt. 2021. Available: https://github.com/marcelm/cutadapt
36. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17: 10–12. https://doi.org/10.14806/ej.17.1.200
37. cd-hit. 2021. Available: https://github.com/weizhongli/cdhit
38. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28: 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610
39. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013 [cited 11 Aug 2021]. Available: http://arxiv.org/abs/1303.3997
40. bwa. 2021. Available: https://github.com/lh3/bwa
41. samtools. 2021. Available: https://github.com/samtools/samtools
42. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021; 10. https://doi.org/10.1093/gigascience/giab008 PMID: 33590861
43. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology. 2015; 16: 157. https://doi.org/10.1186/s13059-015-0721-2 PMID: 26243257
44. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019; 20: 238. https://doi.org/10.1186/s13059-019-1832-y PMID: 31727128
45. OrthoFinder. 2021. Available: https://github.com/davidemms/OrthoFinder
46. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021; 18: 366–368. https://doi.org/10.1038/s41592-021-01101-x PMID: 33828273
47. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015; 12: 59–60. https://doi.org/10.1038/nmeth.3176 PMID: 25402007
48. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution. 2015; 32: 2798–2800. https://doi.org/10.1093/molbev/msv150 PMID: 26130081
49. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2021; 49: D480–D489. https://doi.org/10.1093/nar/gkaa1100 PMID: 33237286
50. Erdős G, Dosztányi Z. Analyzing Protein Disorder with IUPred2A. Current Protocols in Bioinformatics. 2020; 70: e99. https://doi.org/10.1002/cpbi.99 PMID: 32237272
51. Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics. 2019; 35: 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057 PMID: 30615063
52. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology. 2017; 13: e1005659. https://doi.org/10.1371/journal.pcbi.1005659 PMID: 28746339
53. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol. 2003; 10: 980–980. https://doi.org/10.1038/nsb1203-980 PMID: 14634627

54. 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: 473. https://doi.org/10.1186/s12859-019-3019-7 PMID: 31521110

55. 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: 2722–2728. https://doi.org/10.1093/bioinformatics/btt473 PMID: 23986568

56. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al. Gene synteny and evolution of genome architecture in trypanosomatids. Molecular and Biochemical Parasitology. 2004; 134: 183–191. https://doi.org/10.1016/j.molbiopara.2003.11.012 PMID: 15003838

57. Dean S, Moreira-Leite F, Gull K. Basalin is an evolutionarily unconstrained protein revealed via a conserved role in flagellum basal plate function. Bastin P, Akhmanova A, editors. eLife. 2019; 8: e42282. https://doi.org/10.7554/eLife.42282 PMID: 30810527