



Enhanced Detection of Homology Using Artificial Intelligence in Euglenids

Mark C. Field 

Abstract

Identification of similarity between protein sequences is an important component for the assignment of function. With ever-growing databases of genome sequence, this becomes an increasing challenge, and especially in the detection of relationships between distantly related sequences, which is frequently an issue with euglenids. The introduction of artificial intelligence tools to the prediction of protein structure has been, without exaggeration, revolutionary. In particular, AlphaFold3 (AF3), the latest iteration of the AI predictor from DeepMind, a Google subsidiary, offers a potent combination of speed, accuracy, and ease-of-use, all free of charge. Here I will describe a basic workflow for the detection of low similarity between proteins, that is otherwise cryptic, using AF3, discuss how to interpret the predictions, and highlight examples of bizarre predictions or hallucinations.

Key words AI, Protein structure, Sequence evolution, AlphaFold, Homology

1 Introduction

Assigning function to a specific protein sequence usually requires multiple lines of evidence, one of which is sequence similarity. This can be in the form of overall sequence relatedness, which implies orthology or paralogy, or of more limited scope, such as the presence of a conserved domain or fold. With an explosive growth in low-cost nucleic acid sequencing providing an ever-increasing landscape for evolutionary analysis, both the opportunities and challenges for such studies have become greater. Additional challenges arise in terms of evolutionary distance of gene sequences from canonical model organisms, upon which accurate annotation is often based. In the case of euglenids and their kinetoplastid relatives this is particularly important—it is now clear that these organisms lie very close the root of the eukaryotic tree [1] and that this provides both the promise of deep evolutionary insight, coupled with the challenge of identification of homology. Given that

structural information is considerably more conserved than simple sequence [2, 3], the availability of such data is of great value in the identification of relationships.

Understanding the principles underpinning protein structure has been a long-standing goal and has included efforts to define restrictions to amide bond configuration in the Ramachandran plot [4] and physicochemical properties as used in the Kyte–Doolittle hydrophathy score [5]. However, accurate models have required a template, as used by Phyre2, but which of course have meant that the vast majority of proteins have been refractory to structural predictions [6]. The emergence of DeepMind’s AlphaFold3 (AF3) has provided a true leap forward in accuracy, speed, and user-friendliness and brings the use of structural comparisons to augment identification of relationships between proteins truly to all [7–9].

2 Materials

2.1 *A Basic Bioinformatics Environment*

Computational platform requirements can be very modest as the heavier computation is web-based. A mid-range Mac/PC and an internet connection are obviously essential, but beyond this requirements are simple. A large screen is a significant advantage, as laptop screens are restrictive for sequence comparisons and structural visualizations. Most of the software that is recommended is either available for download (with academic licenses usually free) or can be used via a web browser. Table 1 details a small number of applications that the author finds indispensable, together with some of the major global data repositories. You will also require a free Google account for running AF3.

3 Methods

3.1 *Basic Running of AlphaFold3*

1. Log in to <https://alphafoldserver.com/welcome>. Here you will find the basic start page. Paste your sequence into the box annotated as “input” and select sequence type as “protein.”
2. At the free tier you can run 30 searches in a 24-h period, with each search limited to 5000 tokens. A token is one amino acid residue or one nucleotide, or an atom of another kind of molecule. There are therefore higher costs to adding post-translational modifications such as N-glycans, if these are particularly important to your prediction, but in the vast majority of analyses, this is likely not the case.
3. The input box is somewhat unforgiving and any numbers or non-amino acid characters will result in the job failing. You can

Table 1
Software, databases, and tools

<i>Sequence manipulation</i>
<i>JalView</i> (https://www.jalview.org/)
An excellent “Swiss Army knife” of functionality, multiplatform, continually updated, slightly dated design. JalView also supports viewing of protein structure files. Online support is also offered. JalView is cross platform.
BBEdit (https://www.barebones.com/products/bbedit/index.html)
A commercial text editor/programming environment, which is actively developed and maintained. Full functionality requires a paid license, but a free version with some restricted functionality remains excellent and is sufficient for most purposes. MacOS only.
Notepad++ (https://notepad-plus-plus.org/)
Notepad++ is free and open-source, supporting several languages. It is highly customizable and suitable for various programming needs. Windows only.
Visual studio code (https://code.visualstudio.com/)
Visual studio Code from Microsoft is a free and extensible code editor for building web, desktop, and mobile applications, and allows plain text manipulation. Cross platform, so particularly useful if collaborating with folks on multiple OSs.
https://www.reverse-complement.com/cleanup.html
An indispensable tool for removing non-residue characters from sequence files. Can also perform some simple sequence manipulations.
<i>Structure display</i>
ChimeraX (https://www.cgl.ucsf.edu/chimerax/)
An excellent “Swiss Army knife” of functionality, multiplatform, continually updated and like JalView, slightly dated design. The newer versions now include the ability to implement a-fold analysis, either by retrieval of pre-predicted structures or de novo analysis using Google playbook (which requires a Google login). Chimera can visualize protein structures in a huge range of formats and display options.
PyMOL (https://pymol.org/)
A commercial program, but which has an impressive array of features for processing images for presentation and examination.
<i>Software resources</i>
European bioinformatics institute (https://www.ebi.ac.uk/)
Sequence analysis, recalculated protein structure predictions, domain annotation, and many more options.
Expasy at Swiss Institute of Bioinformatics (https://www.expasy.org/)
Offers a suite of software, together with the SwissProt database, a highly curated sequence repository.

The examples are used by the author regularly and have been arrived at by trial and error after evaluating many packages. These are all stable, curated, and unlikely to cease functionality any time soon. The list is far from exhaustive, and there are a great many other high-quality tools available, at some level which different workers prefer is one of personal choice. With one exception, all of these tools can be used free of charge

clean your sequence at <https://www.reverse-complement.com/cleanup.html> if needed.

4. You may add more than one sequence to the input field, if you believe that you are working with a heterooligomer, but the total sequence entered must not exceed the 5000 token limit. Furthermore, you may also add more than one copy of a given sequence, again being mindful of the token limitation, if you are aware that the prediction should be of a homo-oligomer.
5. Once you trigger your job by clicking the “Continue and preview job” button, you will be presented with a new window that will provide the opportunity to name the analysis. By default, each job is named by date and time. It is recommended that you rename the job with something more specific for later retrieval.
6. Once your job is completed, it will become available in the history panel. A vertical ellipsis at right provides several options, most relevant are “Open results” and “Download.” It is recommended that you download each analysis so that you may examine the predictions offline as well as archive the data.
7. The results page contains a lot of information. First, the top model prediction is displayed in an interactive window and the structure is color-coded according to confidence (Fig. 1). Coloring uses the pLDDT score, which is essentially a measure of confidence in the position of atoms on a 0–100 scale. Next to this is the predicted alignment error, a grid plot that illustrates where high confidence positions are situated for the individual residues plotted against each other. This is particularly useful in detecting if there are domains present and detected as these tend to have lower PAE scores, and can indicate if the protein has just interactions in a contiguous manner or if there is higher order structure produced.
8. The final parameters available on the results page are ipTM and pTM. These are the calculated scores for the overall accuracy, with a pTM above 0.5 indicating a likely accurate prediction. The ipTM value is similar but is only calculated for the modeling of subunits in a complex [10, 11].
9. Finally download the dataset and decompress it. This is essential for accessing additional models to the top prediction.

3.2 Interpretation

1. The following are recommendations and depending on the user’s expertise may well be supplanted by more detailed approaches.
2. The folder that you have downloaded contains 26 files; `fold_year_month_day_hour_min_full_data_0.json` to `fold_year_month_day_hour_min_full_data_4.json`, `fold_year_month_day_hour_min_model_0.cif` to `fold_year_month_day_hour_min_model_4.cif`, `fold_year_month_day_hour_min`

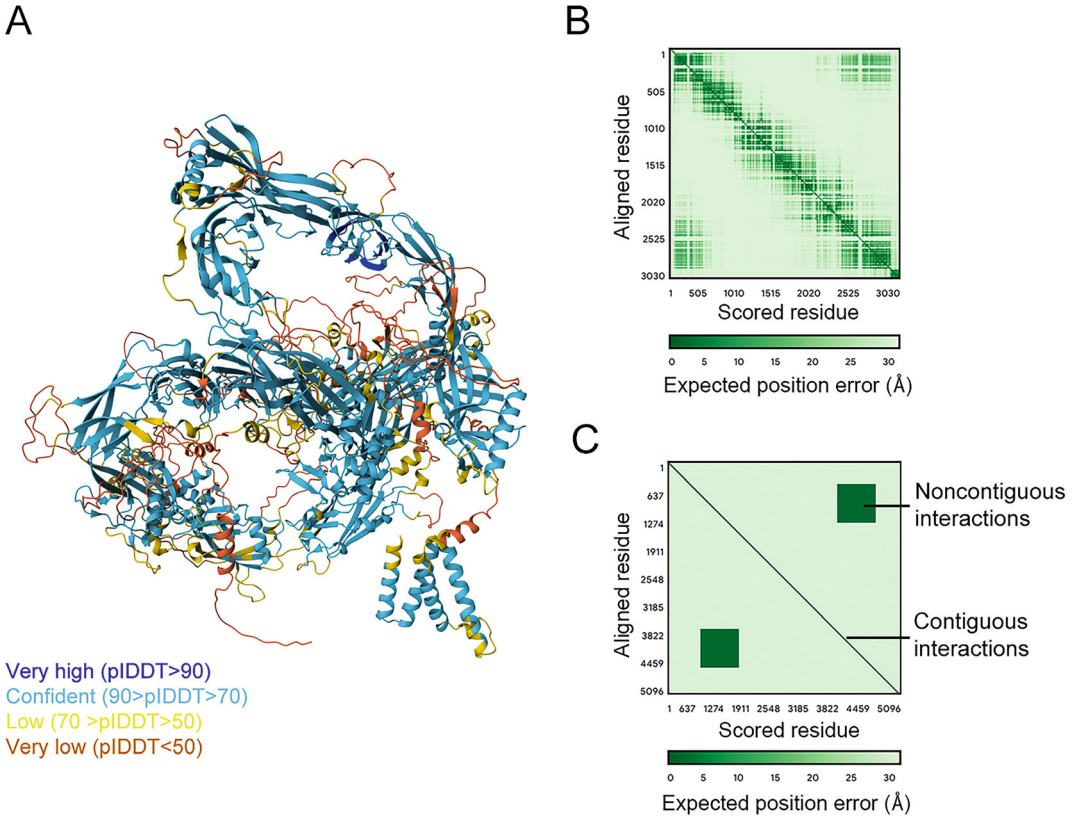


Fig. 1 Data output from Google AlphaFold 3 server. Panel (a) AF3 prediction for *Trypanosoma brucei* LAP333, revealing confident prediction scores for much of the structure. Where low confidence is displayed (pIDDT<50) is almost exclusively for linker regions and at the termini of domains (70 > pIDDT>50). Panel (b) PAE plot for TbLAP333 which is the estimate of the error in relative separation of two amino acids in the predicted structure. Higher values indicate higher predicted error which equates to lower confidence in the predicted structure. Panel (c) The PAE data can be considered also as evidence for close residue interactions, that is, within domain, which would sit upon the diagonal, or for more distant structural connections, off diagonal within the PAE plot

summary_confidences_0.json to fold_year_month_day_hour_min_summary_confidences_4.json and terms_of_use.md.

3. These correspond to five predictions of the structure. The fold_year_month_day_hour_min_full_data_0.json etcetera files contain the raw data while fold_year_month_day_hour_min_summary_confidences_0.json is a summary of the confidences of the prediction, including the pTM/ipTM scores, percent disorders and other parameters. The .json files (javascript object notation) are human-readable and can be edited with a text editor (Table 1).

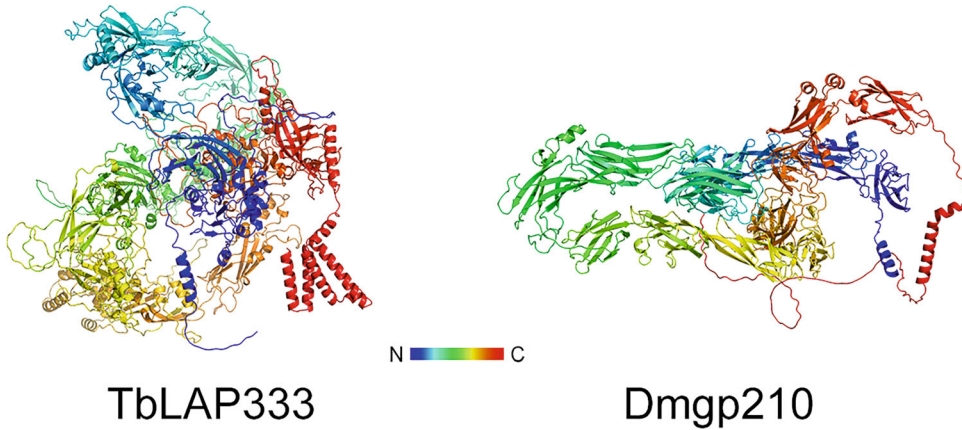


Fig. 2 Cryptic homology revealed by structure prediction. AF3 predictions for *Trypanosoma brucei* TbLAP333 (left) and *Drosophila melanogaster* Dmgrp210 (right). Gp210 is a nuclear pore complex component and crucial to assembly of the complex, acting to interface with the nuclear envelope membrane. The equivalent to gp210 in yeast is Pom152, which shares the architecture of gp210 but is quite divergent in sequence, and neither will identify each other by BLAST; the same is the case for LAP333. The trypanosome protein was identified from interactome analysis of nuclear envelope proteins [13], and localized as a nuclear envelope protein: AF3 predicts the presence of Ig domains, which suggests that LAP333, gp210 and Pom153 are homologs, despite varying considerably in molecular weight. Protein structures were rendered with PyMOL and are shown colored with rainbow shading (N-terminus blue to C-terminus red)

4. The .cif files contain the structural model data (.cif = crystallographic information file), with file 0 the most confident and file 4 the least. These files can be viewed in Chimera and/or PyMOL.
5. Standard colorization options can help to interpret the predictions. In Fig. 2 are shown two nuclear pore complex (NPC) protein predictions, from *Trypanosoma brucei* (TbLAP333) and *Drosophila melanogaster* (Dmgrp210). There is no detectable sequence similarity between these two proteins, but AF3 reveals a clear architectural relatedness in that both proteins are predominantly composed of Ig-like domains [12, 13], and together with additional evidence is sufficient to suggest a common function in anchoring the NPC to the nuclear envelope, as well as retention of architecture across eukaryotes.

3.3 Alternatives and Extending AlphaFold Access

1. Many software packages are now incorporating AF as a component, perhaps a reflection of the rapid adoption of this technology.
2. ChimeraX is a software package for the visualization of molecular structural data and is an example of software with AF connectivity. It can be downloaded from <https://www.cgl.ucsf.edu/chimerax> [14]. Download the appropriate version for your platform, unpack, and install.

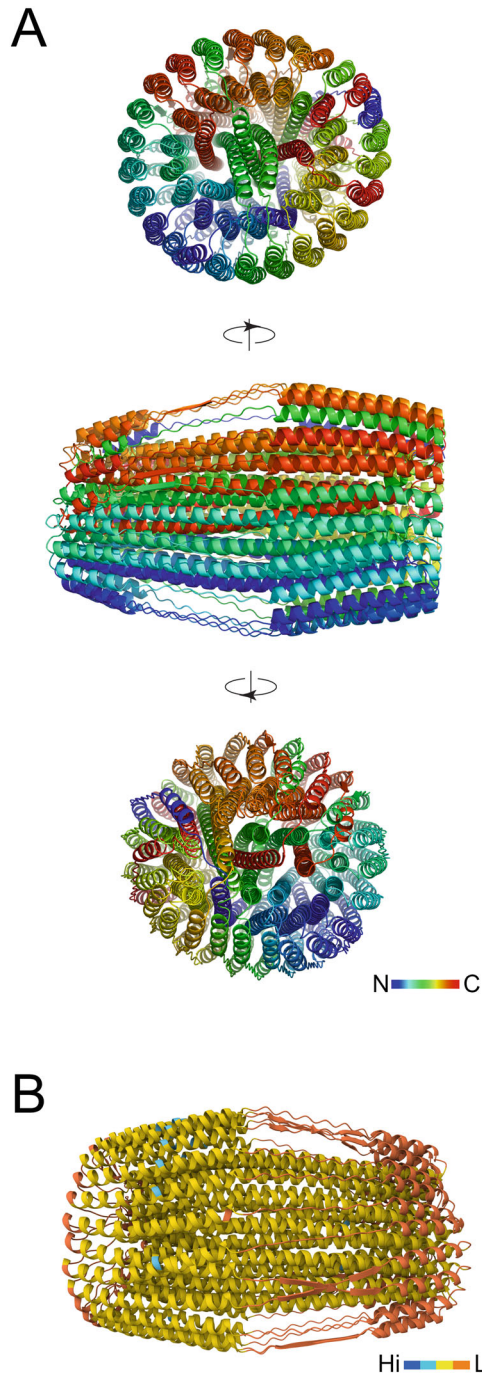


Fig. 3 Example of hallucinatory structure prediction. AF3 prediction of the homodimer of the central repeat region of *Trypanosoma brucei* NUP-1. The sequence contains 17 perfect repeats, which are coiled coil with interspersed unstructured linker regions. The predicted structure has collapsed these repeats into a nested structure, which is inconsistent with known in cellulo behavior [20]. Protein structures were rendered with PyMOL and each monomer is colored with rainbow shading (N-terminus blue to C-terminus red). In panel b, structural confidence is shown colored by pLDDT value: dark blue >90, light blue 90 > to >70, yellow >50, and orange <50. Note that essentially all of the prediction falls into the lower two confidence categories

3. The AF interface is under Tools/Predictions/AlphaFold. At the time of writing the present version of ChimeraX does not allow longer sequences to be analyzed and relies on a ColabFold Google notebook [15] that interfaces with AF2, which the author has found to be slower and less reliable than the AF3 server. However, some users may find the presence of additional tools for the visualization of structure an advantage.
4. A further alternative to calculating an AF3 prediction ab initio is to use the AlphaFold database and/or FoldSeek [16–18]. The AlphaFold database has some obvious disadvantages, such as being fully dependent on data that may not be updated frequently as well as potentially using predictions not made with the most up to date approaches; for example, at the time of writing predictions, use AF2. FoldSeek allows structural datasets to be searched easily and rapidly by innovative simplification of the structural data format used to search [17].
5. As with many predictive software, it is important to have some concept of what an expected/accurate result would be. AI software, such as ChatGPT and other LLM-based software, is known to “hallucinate,” a polite way of saying that these systems can produce bizarre results, with little or no basis in reality. Polydactyly in images of humans or felines is not uncommon. AF3 is not dissimilar as very unrealistic predictions can emerge (*see Note*).
6. An online course in the use of AF3 is offered by the European Bioinformatics Institute (EBI) and can be accessed at <https://www.ebi.ac.uk/training/online/courses/alphafold/> [19].

4 Notes

Figure 3 provides an example of such a hallucination, and is a prediction of the repeat core region of the trypanosome nuclear lamina protein NUP-1. The submission to AF3 is two copies of a 1750 residue repetitive region and which AF3 predicts to fold back on itself multiple times. Despite a clear aesthetic appeal, there is no evidence to support that such a configuration is adopted in cellulose but rather that the conformation is more likely to be open and forming fibers.

Acknowledgments

I am heavily indebted to Erin Butterfield (University of Dundee) for her careful and detailed work in structure prediction and considerable patience in educating the author in appropriate and best practice with AlphaFold.

References

1. Williamson K, Eme L, Baños H et al (2025) A robustly rooted tree of eukaryotes reveals their excavate ancestry. *Nature* 640(8060):974–981
2. Krissinel E (2007) On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* 23(6):717–723. <https://doi.org/10.1093/bioinformatics/btm006>. Epub 2007 Jan 22. PMID: 17242029
3. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5(4):823–826
4. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99
5. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157(1):105–132
6. Powell HR, Islam SA, David A et al (2025) Phyre2.2: a community resource for template-based protein structure prediction. *J Mol Biol* 23:168960
7. Abramson J, Adler J, Dunger J et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630(8016):493–500
8. Wee J, Wei GW. Benchmarking AlphaFold3's protein-protein complex accuracy and machine learning prediction reliability for binding free energy changes upon mutation. *ArXiv [Preprint]*. 2024 6:arXiv:2406.03979v1
9. Mifsud JCO, Lytras S, Oliver MR et al (2024) Mapping glycoprotein structure reveals Flaviviridae evolutionary history. *Nature* 633(8030):695–703
10. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57(4):702–710
11. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26(7):889–895
12. Upla P, Kim SJ, Sampathkumar P et al (2017) Molecular architecture of the major membrane ring component of the nuclear pore complex. *Structure* 25(3):434–445
13. Butterfield ER, Obado SO, Scutts SR et al (2024) A lineage-specific protein network at the trypanosome nuclear envelope. *Nucleus* 15(1):2310452
14. Meng EC, Goddard TD, Pettersen EF et al (2023) UCSF ChimeraX: tools for structure building and analysis. *Protein Sci* 32(11):e4792
15. Mirdita M, Schütze K, Moriwaki Y et al (2022) ColabFold: making protein folding accessible to all. *Nat Methods* 19(6):679–682
16. <https://alphafold.com/>
17. van Kempen M, Kim SS, Tumescheit C et al (2024) Fast and accurate protein structure search with FoldSeek. *Nat Biotechnol* 42(2):243–246
18. Elfmann C, Stülke J (2025) Cutting-edge tools for structural biology: bringing AlphaFold to the people. *Trends Microbiol* 33:S0966-842X(25)00110–6
19. <https://www.ebi.ac.uk/training/online/courses/alphafold/>
20. Padilla-Mejia NE, Koreny L, Holden J et al (2021) A hub-and-spoke nuclear lamina architecture in trypanosomes. *J Cell Sci* 134(12):jcs251264