The Journal of Open Source Software

Af-analysis: a Python package for Alphafold analysis

Alaa Reguei [1] **and Samuel Murail** [1,2¶]

1 Université Paris Cité, Inserm, CNRS, BFA, F-75013 Paris, France **2** Ressource Parisienne en Bioinformatique Structurale (RPBS), F-75013 Paris, France ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

The publication of AlphaFold 2 (Jumper et al., 2021) has significantly advanced the field of protein structure prediction. The prediction of protein structures has long been a central challenge in the field of structural bioinformatics, with the ultimate goal of elucidating the relationship between protein structure and function (Baker & Sali, 2001; Pearce & Zhang, 2021). Accurate prediction of protein structure is essential for a number of applications, including drug discovery, protein engineering, and the study of protein-protein interactions. AlphaFold, which employs a deep learning-based approach, has demonstrated unprecedented accuracy in protein structure prediction, outperforming other contemporary methods. In this paper, we present `af-analysis`, a Python package that provides tools for the analysis of AlphaFold results. The `af-analysis` library has been designed to facilitate the analysis of many different protein structures predicted by AlphaFold and its derivatives. It provides functions for comparing predicted structures with experimental structures, visualising predicted structures, and calculating structural quality metrics.

Statement of need

With the release of AlphaFold 2 (Jumper et al., 2021) in 2021, the scientific community has achieved an unprecedented level of accuracy in predicting protein structures. Derivatives of AlphaFold 2, namely ColabFold (Mirdita et al., 2022), AlphaFold Multimer (Evans et al., 2022), AlphaFold 3 (Abramson et al., 2024) and its re-implementations such as Boltz-1 (Wohlwend et al., 2024) and Chai-1 (Discovery et al., 2024) have been developed to predict the structure of protein complexes, setting a new standard for protein-protein and protein-peptide docking.

Analysis of AlphaFold results is a crucial step in the process of utilising these predictions for scientific research. The AlphaFold software provides several excellent quality metrics that offer valuable information about the accuracy of the predicted structures. Among these scores, the predicted local distance difference test (pLDDT) is a per-residue measure of local confidence, as the predicted aligned error (PAE) provides confidence over the relative position of two residues within the predicted structure. To analyze these results, the AlphaBridge webserver (Álvarez-Salmoral et al., 2024) and the PICKLUSTER plugin (Genz et al., 2023) for the UCSF ChimeraX visualisation software were developed to characterize the different interfaces within protein complexes, and extract their respective scores.

Although these tools are very practical, Bjorn Wallner has shown that calculating 5 or 25 basic AlphaFold models may not be enough, it is sometimes necessary to generate thousands of models to obtain a few high quality models, leading to the AlphaFold derivative, AFsample (Wallner, 2023). Massive sampling altogether with multiple software usage (AFsample and ColabFold), weights and parameters has been integrated into the MassiveFold software (Raouraoua et al., 2024) and has shown performance approaching the accuracy of AlphaFold 3.

The subsequent analysis of hundreds to thousands of models can prove to be a tedious and

meticulous process, as dealing with thousands of models and different output formats can be time consuming. Furthermore, while the quality metrics produced by AlphaFold are good, additional metrics have been developed to assess the quality of the models. These include pdockq (Bryant et al., 2022), pdockq2 (Zhu et al., 2023), and LIS score (Kim et al., 2024). All of the these metrics have to be calculated from different scripts. Another point to consider is the diversity of the models. As shown in AFsample, it is sometimes necessary to compute up to tens of thousands of models and then cluster them in order to select the best ones. The `af-analysis` library has been developed to facilitate the analysis of sets of model structures and associated metrics. The library is based on the `pandas` library and is able to import AlphaFold 2 and 3, ColabFold, Boltz-1 and Chai-1 prediction directory as a `pandas` DataFrame. The library provides a number of functions to add further metrics to the DataFrame, compare models with experimental structures, visualise models, cluster models and select the best models.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Álvarez-Salmoral, D., Borza, R., Xie, R., Joosten, R. P., Hekkelman, M. L., & Perrakis, A. (2024). AlphaBridge: Tools for the analysis of predicted macromolecular complexes. *bioRxiv*. https://doi.org/10.1101/2024.10.23.619601

Baker, D., & Sali, A. (2001). Protein Structure Prediction and Structural Genomics. *Science*, *294*(5540), 93–96. https://doi.org/10.1126/science.1065659

Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), 1265. https://doi.org/10.1038/s41467-022-28865-w

Discovery, C., Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhnikov, A., & Wu, K. (2024). Chai-1: Decoding the molecular interactions of life. *bioRxiv*. https://doi.org/10.1101/2024.10.10.615955

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2022). Protein complex prediction with AlphaFold-multimer. *bioRxiv*. https://doi.org/10.1101/2021.10.04.463034

Genz, L. R., Mulvaney, T., Nair, S., & Topf, M. (2023). PICKLUSTER: A protein-interface clustering and analysis plug-in for UCSF ChimeraX. *Bioinformatics*, *39*(11), btad629. https://doi.org/10.1093/bioinformatics/btad629

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kim, A.-R., Hu, Y., Comjean, A., Rodiger, J., Mohr, S. E., & Perrimon, N. (2024). Enhanced protein-protein interaction discovery via AlphaFold-multimer. *bioRxiv*. https://doi.org/10.1101/2024.02.19.580970

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682.

https://doi.org/10.1038/s41592-022-01488-1

Pearce, R., & Zhang, Y. (2021). Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, *297*(1), 100870. https://doi.org/10.1016/j.jbc.2021.100870

Raouraoua, N., Mirabello, C., Véry, T., Blanchet, C., Wallner, B., Lensink, M. F., & Brysbaert, G. (2024). MassiveFold: Unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. *Nature Computational Science*, *4*(11), 824–828. https://doi.org/10.1038/s43588-024-00714-4

Wallner, B. (2023). AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*, *39*(9), btad573. https://doi.org/10.1093/bioinformatics/btad573

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., & Barzilay, R. (2024). Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*. https://doi.org/10.1101/2024.11.19.624167

Zhu, W., Shenoy, A., Kundrotas, P., & Elofsson, A. (2023). Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, *39*(7), btad424. https://doi.org/10.1093/bioinformatics/btad424