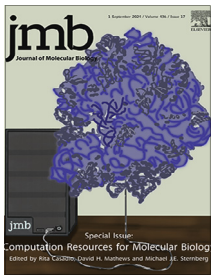

Alpha&ESMhFolds: A Web Server for Comparing AlphaFold2 and ESMFold Models of the Human Reference Proteome

Matteo Manfredi [1,†], Castrense Savojardo [1,*,†], Georgii Iardukhin [1,†], Davide Salomoni [2], Alessandro Costantini [2], Pier Luigi Martelli [1,*] and Rita Casadio [1]

*1 - **Biocomputing Group,** Dept. of Pharmacy and Biotechnology, University of Bologna, Italy*
*2 - **INFN-CNAF,** Bologna, Italy*

***Correspondence to Castrense Savojardo and Pier Luigi Martelli:** castrense.savojardo2@unibo.it (C. Savojardo), pierluigi.martelli@unibo.it (P.L. Martelli)*
https://doi.org/10.1016/j.jmb.2024.168593
Edited by Michael Sternberg

Abstract

We develop a novel database Alpha&ESMhFolds which allows the direct comparison of AlphaFold2 and ESMFold predicted models for 42,942 proteins of the Reference Human Proteome, and when available, their comparison with 2,900 directly associated PDB structures with at least a structure to sequence coverage of 70%. Statistics indicate that good quality models tend to overlap with a TM-score >0.6 as long as some PDB structural information is available. As expected, a direct model superimposition to the PDB structure highlights that AlphaFold2 models are slightly superior to ESMFold ones. However, some 55% of the database is endowed with models overlapping with TM-score <0.6. This highlights the different outputs of the two methods. The database is freely available for usage at https://alpha-esmhfolds.biocomp.unibo.it/.

Introduction

Recently published results from the Critical Assessment of methods for Structure Prediction (CASP 15, https://predictioncenter.org/casp15/index.cgi) confirm the relevance of Artificial Intelligence (AI)-based modelling on the accuracy of protein structure prediction.[1] On some one hundred assembly targets the impressive performance of the submitting community was due to methods differently based on DeepMind's AlphaFold2 (https://alphafold.ebi.ac.uk/),[2] which in CASP14 paved the way to highly accurate large-scale structure predictions.[3,4] AlphaFold2 contains deep neural networks (transformers) trained to produce protein structures from amino acid sequences, multiple sequence alignments, and homologous proteins. As an alternative, recent methods take advantage of protein Language Models (pLMs),[5] and sequence embeddings to develop end-to-end models of protein structures, such as ESMFold (https://esmatlas.com/).[5] CASP 15 indicates that when proteins are better annotated at the level of the protein family, AlphaFold2 models are superior to those generated by purely embedding systems, as expected given its development and the large amount of pre-existing information it takes as input.

In this paper we develop a novel database, Alpha&ESMhFolds, where we store two models per human sequence. Models are downloaded from https://alphafold.ebi.ac.uk/, the model database developed by Google DeepMind and EMBL-EBI,[6] and paired with models generated in house with ESMFold. Our database gives the unique opportunity for a given human protein to directly compare the two models, and when available, to compare them with the associated PDB structure (https://www.rcsb.org/).[7,8] Our database,

which presently contains 85,884 models for 42,942 proteins of the human Reference Proteome, is a unique resource for investigating when both methods give similar or dissimilar good/bad quality models. The information is particularly interesting in the absence of an experimental reference structure, allowing a direct evaluation and comparison of the high-quality predicted regions for each protein. The models can support research in different areas, such as active site conservations, functional annotation for specific biological processes, mapping of disease related variations and new developments for improving protein structure prediction, considering the different approaches of the prediction methods.

The database is freely available at https://alpha-esmhfolds.biocomp.unibo.it/.

Materials and Methods

Protein database

Human proteins adopted in this study were extracted from the human Reference Proteome (UP000005640), available at UniProt,[9] release 2023_03, as of January 2023 (80,581 proteins, https://ftp.uniprot.org/pub/). From the initial set, we excluded fragments, short peptides (shorter than 50 residues) and sequences for which AlphaFold2[2] models were not available in AlphaFoldDB.[6,10] We ended up with 42,942 protein sequences, which we also modelled in house with ESMFold.[5]

When necessary for data analysis, we performed a clustering procedure to reduce internal redundancy. To this aim, we ran MMseqs2[11] with a sequence identity threshold set to 50% over an alignment coverage of 70%. The number of non-redundant protein sequences turns out to be 23,939 (Table S1). In the set, 13,489 proteins are reviewed and present in UniProt/SwissProt.

In order to retrieve the best available structure from the Protein Data Bank (PDB, https://www.rcsb.org/ we extracted structural data associated to our dataset by adopting the PDBe entry-based SIFTS REST API (available at https://www.ebi.ac.uk/pdbe/api/sifts.html),[12] which recovers structures mapping to a UniProt accession sorted by coverage of the protein and resolution (Table S1). After considering only structures with residues in the protein sequence for which atomic coordinates are available (by excluding potential artifacts in the PDB (e.g. tags), unmodeled regions (i.e., gaps) and sorting signals (N-terminal signal and/or transit peptides reported with experimental evidence in the UniProt entry)), and constraining a minimum PDB coverage of at least 70%, we were able to retain 2,900 PDB chains, of which 2,404 are associated to non-redundant proteins in Swiss-Prot. This data set was then adopted as structural ground truth to assess predictive performance of both AlphaFold2 and ESMFold (Table S1). The minimum structure to sequence coverage is constrained in order to avoid the association of a sequence to small structural fragment/s.

For sake of extending the set of structural data associated to human proteins, we took advantage of the basic principles and rules of template search in building by comparison[13]. We aligned with MMSeqs2 the remaining 40,042 human proteins of our set not endowed with a directly associated PDB structure (Table S1) to the PDB database, requiring a minimum sequence identity of 50% (an arbitrary threshold for functional conservation) over an alignment coverage of at least 70%. After this search, we retrieved PDB templates for other 5,736 additional human sequences.

Characteristics of the human data set modelled with Alphafold2 and ESMFold

Human proteins modelled with AlphaFold2 and ESMFold span a wide range of lengths, from the shortest ones, accounting for 51 residues (A0A286YFK9, Small integral membrane protein 38) to the longest ones including 1,836 residues (P35499, Sodium channel protein type 4 subunit alpha). Evidently, the set comprises both monodomain and polydomain proteins. Figure S1 shows the protein length distribution of reviewed (SwissProt) and unreviewed (TrEMBL) entries in the dataset (Table S1, Figure S1).

For 2,409 non-redundant proteins, a PDB chain covering at least 70% of the sequence is available (Table S1). The vast majority of included structures were obtained with X-ray diffraction after crystallization (1,643 chains, about 68% of the dataset), followed by Electron Microscopy (EM, 689 chains, about 29%) and Nuclear Magnetic Resonance (NMR, 77 chains, about 3%).

Structural alignments and scoring metrics

We adopted Foldseek[14] to produce pairwise structural alignments between pairs of models predicted with AlphaFold2 and ESMFold, and between the computed models and the corresponding PDB chain, when available. The Foldseek program was run using the alignment type set to the TM-align algorithm (option --alignment-type 1) and disabling prefiltering of results (option --prefilter-mode 2). The remaining program parameters were left to default values.

Standard metrics were always adopted to score the structural similarity of two superimposed structures, including the template modelling score (TM-score)[15] and the Root Mean Square Deviation (RMSD). When comparing models, we also adopted the Global Distance Test (GDT) score,[16,17] which is well suited for comparing two predicted models of the same protein sequence. RMSD values were directly retrieved from Foldseek output and are computed considering Cα atoms of the structural aligned regions. TM-score and GDT

score consider backbone Cα atoms and are computed using the ClusCo tool.[17]

Computational infrastructures

The computational infrastructure employed in this work to generate EMSFold models has been made accessible through the Istituto Nazionale di Fisica Nucleare (INFN, https://home.infn.it/it/). Specifically, the INFN's primary data processing and computing research centre, CNAF (https://www.cnaf.infn.it/en/), supplied the necessary cloud resources via the INFN private cloud, known as INFN Cloud (https://www.cloud.infn.it/architecture/). The virtual machine (VM) used in this study operates on the ×86_64 architecture and runs the Ubuntu 22.04 operating system. It is equipped with 8 CPU cores, each being an Intel Xeon Processor (Cascadelake) with 1 thread per core and 1 core per socket. It boasts 32 GB of CPU RAM and is equipped with a single NVIDIA V100 Tensor Core GPU boasting 32 GB of RAM. Its storage infrastructure includes a hard disk drive (HDD) with a total capacity of 97 GB and a 3.5 TB additional storage space.

ESMFold runtime is strongly correlated with sequence length, ranging from a few seconds for very short sequences to about 60 s for sequences with 600 residues and up to 800 s for longer sequences (about 1800 residues). Structure alignments and superpositions are performed in a few seconds by Foldseek for each pair of structures considered (either model-vs-model or model-vs-PDB).

The model database is now hosted at the computational infrastructure of the Bologna Biocomputing Group (https://www.biocomp.unibo.it/).

Web server implementation

The web application has been implemented using the Python Django application server (https://www.djangoproject.com, version 4.0.4). The user interface adopts the Bootstrap frontend toolkit (https://getbootstrap.com, version 5.3.2). Results of the queries in the search page are displayed with DataTables (https://datatables.net, version 2.0.2). For molecular visualizations, we adopt the PDBe Mol* viewer JavaScript plugin (https://github.com/molstar/pdbe-molstar, version 3.1.3). The backend database has been implemented using the PostgreSQL DBMS (version 13) and it stores all data and files displayed on the web server. All queries to the database are implemented adopting the Python Psycopg library (version 2.9.9).

Results

Introducing Alpha&ESMhFolds, a database of AlphaFold2 and ESMFold paired models for human proteins

Following AlphaFold2,[2] the predicted local distance difference test (pLDDT) reliably predicts the Cα local distance difference test (lDDT-Cα) accuracy of the corresponding prediction. The pLDDT score computed by AlphaFold2 is also computed by ESMFold (https://esmatlas.com/),[5] and indicates a per-residue confidence score between 0 and 100 (see also https://alphafold.ebi.ac.uk/faq#faq-12).

In Table 1, we organise paired models as a function of the relative fraction of high-quality residues (with pLDDT value >70, High or Confident), considering four possible intervals (up to 25%, 50%, 75%,100% of the protein sequence, respectively). Alpha&ESMhFolds contains 85,884 models of 42,942 human proteins computed with AlphaFold2 and ESMFold. Their superimposition is evaluated with the TM-score (which ranges from 0 to 1), 0.6 being the threshold among good and bad superimposition.[15,18]

In Table 1, for each interval of high confidence residues in both models, we show the number of proteins, and among round brackets the number of PDB available structure/s in the subset, the median value of TM-scores and the number of model pairs (proteins) with TM-score > 0.6. Considering that the total number of modelled proteins is 42,942, 65.5% of the models (cells along the main diagonal) have similar quality; 33% of the models have TM-score > 0.6 and 26% of the models are high quality (pLDDT > 70 for at least 75% of the residues)[6,10] and superimposable. Off diagonal, we find models which are predicted with a larger fraction of high-quality residues by ESMFold than AlphaFold2 or vice versa. Interestingly enough, 17 proteins have ESMFold models which superimpose with a TM-score > 0.6 with low quality AlphaFold2 models, while the opposite happens for 47 proteins. The number of PDB structures in the whole set is 2,900, 2,792 of which (96.3%) are included in the four labelled cells with an empty dot (°) of Table 1. Figure S2 shows the TM-score distribution of paired models as a function of bins of protein length. The median TM-score and data set median value are also included. It appears that the highest model superimposition (TM-score median value >0.6) includes proteins with length ranging from about 150 to 650 residues.

Benchmarking AlphaFold2/ESMFold models on the corresponding PDB structures

In this section, we consider models of AlphaFold2 and ESMFold of 2,404 non redundant proteins

Table 1 Overview of model quality comparison in Alpha&ESMhFolds.

		ESMFold			
AlphaFold2	#of high quality residues in the protein (%)	[0%, 25%)	[25%, 50%)	[50%, 75%)	[75%, 100%]
	[0%, 25%)	5717 (1, 0.21, 31)*	465 (0, 0.32, 15)*	164 (0, 0.41, 23)*	62 (1, 0.52, 17)*
	[25%, 50%)	3132 (4, 0.28, 49)*	3391 (2, 0.36, 101)*	1203 (4, 0.45, 143)*	157 (4, 0.62, 87)*
	[50%, 75%)	1302 (6, 0.35, 89)*	1946 (27, 0.47, 390)*	°6477 (111, 0.58, 2996)*	°2841 (136, 0.69, 1940)*
	[75%, 100%]	421 (23, 0.36, 47)*	514 (36, 0.55, 211)*	°2597 (310, 0.72, 1901)*	°12553 (2235, 0.86, 11182)*

Squared and round brackets indicate closed and open endpoints, respectively.
Per residue confidence is set to local pLDDT > 70.[2] #Intervals of percentage high-quality residues in the protein. Proteins with a direct PDB structure are 2,900 (Table S1).
* Values among brackets are: (1) number of proteins with PDB, (2) median model-vs-model of TM-score values in the set, (3) number of proteins for which model-vs-model TM-score > 0.6.
° Cells with the highest number of PDBs (92.3%). The total number of models is twofold the number of human proteins (42,942).

endowed with PDB structure covering at least 70% of the sequence, and present in SwissProt (Table S1). For sake of scoring the methods on the same protein, we performed a side-by-side comparison of AlphaFold2 and ESMFold models and obtained the scatter plot of Figure 1. Here, we group data by considering the TM-score difference of both models with respect to the corresponding

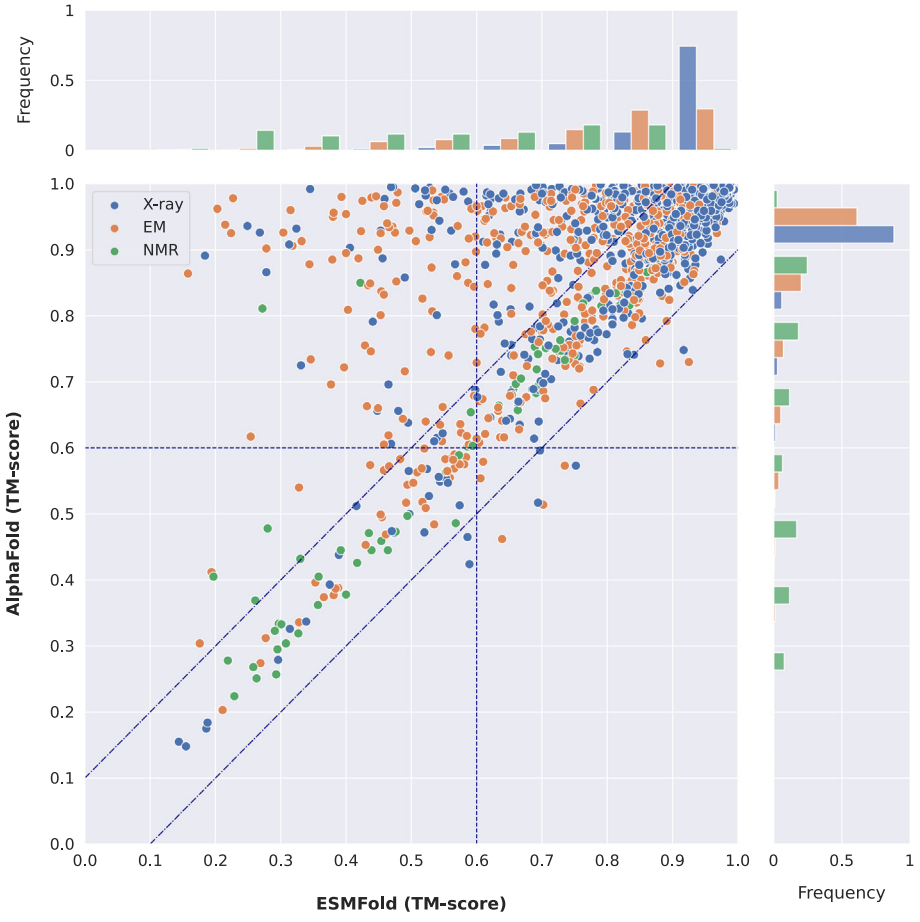

Figure 1. Scatter plot of ESMFold against AlphaFold2 TM-scores of the models with respect to the corresponding PDB chains (2,404). Colour codes are according to the methods adopted for structure determination (Materials and Methods, 2.2). Most of the proteins (1,973 out of 2,404 proteins, 81%) are included in the highlighted region inside the two parallel dashed lines, with a TM-score difference <0.1 to the corresponding PDB chains.

PDB structure and colour code according to the different methods adopted for structural determination (Green: NMR; Orange: EM; Blue: Xray after crystallization; see Materials and Methods, 2.2). Each dot represents a protein for which AlphaFold2 and ESMFold TM-scores are computed against the corresponding PDB chain structure. For most of the proteins (1,973 out of 2,404 proteins (81%), included in the highlighted region), the performance of the two methods is highly similar, with an absolute value of the difference between TM-scores towards the experimental structure equal or lower than 0.1. Considering the TM-score threshold of 0.6, 133 AlphaFold2 models are better than ESMFold ones, and 8 ESMFold models are better than AlphaFold2 ones, confirming that AlphaFold2 is overall performing better than ESMFold.

The relevance of structural information on model building

Finally, we consider the whole set of proteins (42,942), 2,900 of which have direct PDB chains. The remaining set of proteins (42,042) is aligned towards PDB by constraining 50% sequence identity over 70% of alignment coverage (see Materials and Methods, 2.1). The remaining human sequences can be clustered according to three possible cases: (i) proteins with human templates; (ii) proteins with templates external to our human database; (iii) proteins without templates. In Figure 2, we plot the distribution of TM-score values of models with a directly associated PDB structure (in green) and of the three different groups of proteins as described above (blue, orange, and red, respectively). It appears that, when structural information is available, the model-vs-model TM-scores are higher (TM-score > 0.6). When structural information is lacking (red bars), TM-score values shift towards lower figures, peaking around 0.25–0.30. Results indicate that about 68% of the models (for 30,574 proteins) diverge (with a TM-score below 0.6) when structural information is absent. Overall, 23,720 proteins (of which only 12% retain structural information) have models with a TM-score < 0.6.

Web Server Description

In Alpha&ESMhFolds, AlphaFold2 and ESMFold models of human proteins with details on statistics and links to relevant materials are available through a web application, accessible at https://alpha-esmhfolds.biocomp.unibo.it/.

From the home page of the web server, it is possible to query the model database either by UniProt accession or using a valid protein sequence in FASTA format. In the former case, the accession is searched against the database and, if present, all data for the entry are shown. In the latter case, the query sequence is aligned against all the sequences present in our database using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi), by considering an E-value threshold of 10^{-3},

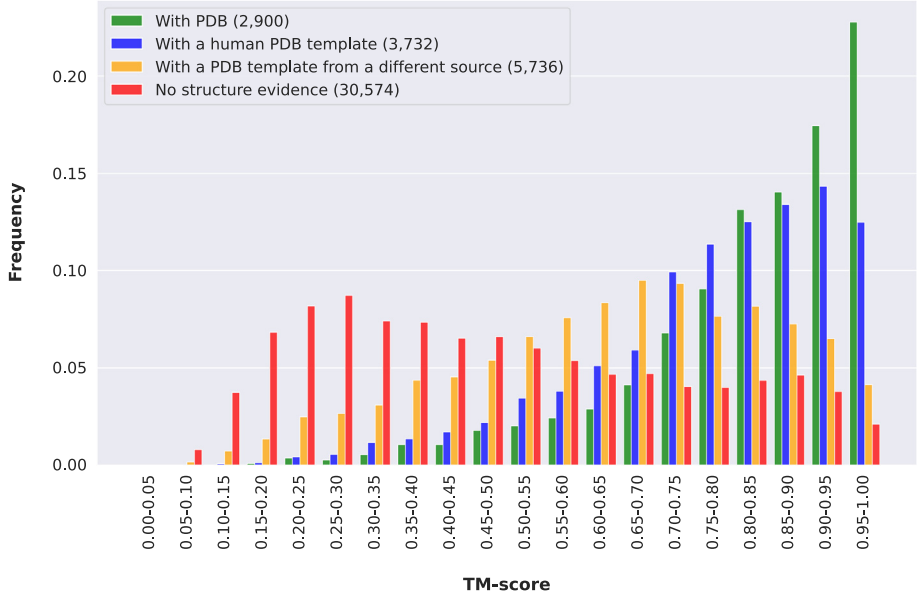

Figure 2. The distribution of TM-scores between AlphaFold2 and ESMFold models of each protein is plotted as a function of the TM score value. The green distribution considers proteins endowed with a direct PDB (2,900). The blue distribution includes proteins (3,732) with a human PDB template; the orange one, proteins with PDB templates from a different source (5,736). The red distribution considers proteins (30,574) for which no template is available (according to our PDB coverage criteria, see Materials and Methods).

and a list of possible entries is returned, each one with a link to access the corresponding file in UniProt/PDB. If a perfect match is found, the corresponding entry will be the only one displayed. Considering UniProt as a reference database, by constraining a global sequence identity of the query to the targets of at least 50% with an alignment coverage of at least 70%, a list containing the 10 best hits is displayed. When no significant hit fulfilling the constraints is found, a list is generated with at most 10 targets which share an identity lower than 50% with the query and a warning message is displayed. When no target is found, the page shows an error message prompting the user to return to the home page.

Alternatively, from the search page, it is possible to query the database by adopting different criteria. Entries can be filtered based on the gene name and TM-scores between the ESMFold and AlphaFold2 models. Additionally, results can be restricted to only proteins endowed with a PDB structure, in which case a specific PDB accession can be searched, or thresholds for the TM-scores between the PDB and each model can be selected. The results of the query are displayed in a tabular format, providing links to access the page of each specific entry as well as the possibility to download the full list.

For all entries in the model database, the result page shows at the top a table that contains general information on the selected protein, including links to the corresponding UniProt page and, if available, to the 3D structure in PDB. If no PDB structure is available, but the sequence is highly similar (sequence identity $\geq$80% and coverage $\geq$70%) to an entry endowed with an experimental structure, a cross-link to the putative template is shown. The complete list of information can be downloaded in JSON format (at the web server buttons are highlighted).

For each modelled human protein, a tab displays the comparison between the AlphaFold2 and ESMFold computed structures. The tab includes: (i) The sequence alignment obtained from the superimposition of the two structural models (Supplementary Figure 3A). Residues are colour coded according to the model confidence (pLDDT); a green bar highlights residues which correctly match at the same positions; (ii) The graphical structure superimposition of the two models (Supplementary Figure 3B). Two different colours are adopted to distinguish ESMFold models (green) and AlphaFold2 models (purple). The graphical viewer is our implementation of PDBe Mol* and can be similarly interacted with (see original documentation at https://molstar.org/viewer-docs; some operations are not available in our viewer). Additionally, residues shown in the sequence alignment can be clicked to zoom in on the corresponding position; (iii) Model quality statistics and Alignment statistics (Supplementary Figure 3C), which include the number of residues with pLDDT greater than given thresholds (>50, 70 and 90, respectively) and different scoring values (TM-score, RMSD, GDT), to represent the level of agreement between the two predicted models. The individual models (PDB format), the superimposed models (PDB format) and the sequence alignments (fasta-like format including the two gapped sequences) are available for download (https://alpha-esmhfolds.biocomp.unibo.it/, see the Help page).

When an entry is endowed with a PDB chain whose coverage to sequence is greater than 70%, two similar tabs show the comparison between the experimental structure and each predicted model (Supplementary Figure 3D). In the graphical viewer, PDB chains are coloured in white.

We supply a help page, including four examples derived from Table1. Specifically, we show two proteins (P07902 and Q96P20) with high-quality models (derived from proteins clustered in the bottom-right cell of Table 1 with more than 75% of residues with pLDDT > 70). Both proteins are endowed with high-coverage 3D structures and Alpha&ESMhFolds provides the model-to-model and the models-to-structure superimpositions. In the case of P07902, the two models superimpose well (TM-score = 0.9) and both agree with the experimental structure (TM-score = 0.92 and TM-score = 0.97 for ESMFold and AlphaFold2, respectively). Conversely, the models of Q96P20 poorly superimpose (TM-score = 0.58); comparison with the experimental structure shows that ESMFold is better performing than AlphaFold2 (TM-score = 0.73 and TM-score = 0.57, respectively). Furthermore, we show other two proteins (Q9HD87 and Q9NVL8) with low-quality models (derived from the top-left cell in Table 1 with less than 25% of residues with pLDDT > 70). No structure is available for the two proteins and Alpha&ESMhFolds provides the model-to-model superimpositions. In the case of Q9HD87, the two models superimpose well (TM-score = 0.72, among the 31 proteins within brackets, out of the total 5,717); in the case of Q9NVL8, the models largely diverge (TM-score = 0.2).

Regular updates are foreseen every six months following major releases of UniProt.

Conclusions and Perspectives

Alpha&ESMhFolds is a novel database which handle, for a given human protein sequence, both AlphaFold2 and ESMFold models. This allows a direct comparison of the two models, and their superimposition to the PDB structure when available.

Recently, at CASP 15, it has been demonstrated on a small data set, including some 100 targets from different organisms, that AlphaFold2 and methods based on AlphaFold2 are outperforming methods based on pLMs and embedding procedures like ESMFold. Here we focus on a large fraction of the human reference proteome (42,942 proteins), to realize on a much larger scale which of the two methods is more reliable. As previously discussed,[2,19] AlphaFold2 is trained with transformers on a precomputed information including for a protein sequence, multiple sequence alignments, correlated mutations in the family and contact maps of family templates. On the other hand, ESMFold[5,19] takes as input embedded sequences relying on a pLM which carries along the information derived from billions of sequences, and some selected structures from the PDB and from AlphaFold2 models (included to reach a suited structural level of information).[5] In the prediction phase AlphaFold2 searches for templates in the protein family of a given sequence, whereas ESMFold takes as input the sequence embedding. It is therefore not surprising that AlphaFold2 computes better models than ESMFold when structural information is present for the family and deteriorates when little information is available.[4] Model statistics in our database support the expected conclusion, considering the data shown in Figure 1 and derived by comparing both models with 2,404 human PDB structures. It is noticeable that 81% of the models are superimposable with their associated PDB chains with a TM-score difference <0.1. This indicates that, as anticipated[5,20], embeddings carry along evolutionary and structural information.

What is really interesting is that models computed with both methods overlap to a good extent (TM score > 0.6) for 45% of the protein set. This is so particularly when structural information is somehow available (Figure 2). However, 55% of the human protein set models diverge (TM score < 0.6) at decreasing structural content information. Which is the most reliable model in this region? The question is open, as much as the question of how protein flexibility can affect the whole scenario[21]. More structural data are necessary to solve these issues. For the time being, each superimposition per se allows a direct view of modelling and is open for further investigation.

Funding

The work was supported by the European Union – NextGenerationEU through the Italian Ministry of University and Research under the projects "Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics" (ElixirxNextGenIT)" (Investment PNRRM4C2-I3.1, Project IR_0000010, CUP B53C22001800006) and "HEAL ITALIA" (Investment PNRR-M4C2-I1.3, Project PE_00000019, CUP J33C22002920006).

CRediT authorship contribution statement

Matteo Manfredi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Formal analysis, Data curation. **Castrense Savojardo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Formal analysis, Data curation, Conceptualization. **Georgii Iardukhin:** Software, Methodology, Investigation, Data curation. **Davide Salomoni:** Resources, Methodology, Conceptualization. **Alessandro Costantini:** Resources, Methodology. **Pier Luigi Martelli:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Rita Casadio:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization.

DATA AVAILABILITY

No data was used for the research described in the article.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2024.168593.

Received 25 March 2024;
Accepted 30 April 2024;
Available online 6 May 2024

Keywords:
protein structure prediction;
model superimposition;
embedding;
protein language model;
database

† Equally contributed.

References

1. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J., (2023). Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins: Struct.*

Funct. Bioinf. **91**, 1539–1549. https://doi.org/10.1002/prot.26617.
2. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
3. Ozden, B., Kryshtafovych, A., Karaca, E., (2023). The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from CASP15. *Proteins: Struct. Funct. Bioinf.* **91**, 1636–1657. https://doi.org/10.1002/prot.26598.
4. Simpkin, A.J., Mesdaghi, S., Sánchez Rodríguez, F., Elliott, L., Murphy, D.L., Kryshtafovych, A., Keegan, R.M., Rigden, D.J., (2023). Tertiary structure assessment at CASP15. *Proteins Struct. Funct. Bioinf.* **91**, 1616–1635. https://doi.org/10.1002/prot.26593.
5. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., et al., (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130. https://doi.org/10.1126/science.ade2574.
6. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., et al., (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444. https://doi.org/10.1093/nar/gkab1061.
7. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242. https://doi.org/10.1093/nar/28.1.235.
8. wwPDB Consortium, (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47** (D1), D520–D528. https://doi.org/10.1093/nar/gky949.
9. The UniProt Consortium, (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531. https://doi.org/10.1093/nar/gkac1052.
10. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., et al., (2021). Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596. https://doi.org/10.1038/s41586-021-03828-1.
11. Steinegger, M., Söding, J., (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnol.* **35**, 1026–1028. https://doi.org/10.1038/nbt.3988.
12. Armstrong, D.R., Berrisford, J.M., Conroy, M.J., Gutmanas, A., Anyango, S., Choudhary, P., Clark, A.R., Dana, J.M., et al., (2019). PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.*, gkz990. https://doi.org/10.1093/nar/gkz990.
13. Lesk, A.M., (2019). Introduction to bioinformatics. Oxford University Press, Oxford, United Kingdom.
14. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., Steinegger, M., (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnol.*, 1–4. https://doi.org/10.1038/s41587-023-01773-0.
15. Zhang, Y., Skolnick, J., (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Struct. Funct. Bioinf.* **57**, 702–710. https://doi.org/10.1002/prot.20264.
16. Zemla, A., (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374. https://doi.org/10.1093/nar/gkg571.
17. Jamroz, M., Kolinski, A., (2013). ClusCo: clustering and comparison of protein models. *BMC Bioinf.* **14**, 62. https://doi.org/10.1186/1471-2105-14-62.
18. Zhang, Y., Skolnick, J., (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302–2309. https://doi.org/10.1093/nar/gki524.
19. Kandathil, S.M., Lau, A.M., Jones, D.T., (2023). Machine learning methods for predicting protein structure from single sequences. *Curr. Opin. Struct. Biol.* **81**, 102627. https://doi.org/10.1016/j.sbi.2023.102627.
20. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., et al., (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **10**, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.
21. Tesei, G., Trolle, A.I., Jonsson, N., Betz, J., Knudsen, F.E., Pesce, F., Johansson, K.E., Lindorff-Larsen, K., (2024). Conformational ensembles of the human intrinsically disordered proteome. *Nature 626*, 897–904. https://doi.org/10.1038/s41586-023-07004-5.