Computational and Structural Biotechnology Journal 27 (2025) 461–466

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

AlphaFold2 and ESMFold: A large-scale pairwise model comparison of human enzymes upon Pfam functional annotation

Matteo Manfredi [a,b], Gabriele Vazzana [a,b], Castrense Savojardo [a,b], Pier Luigi Martelli [a,b,*], Rita Casadio [a,c,*]

[a] Biocomputing Group, University of Bologna, Italy
[b] Dept. of Pharmacy and Biotechnology, University of Bologna, Italy
[c] the Alma Climate Institute, University of Bologna, Italy

ARTICLE INFO

Keywords:
Human enzyme structural and functional annotation
Pfam domains
Enzyme Active site
AlphaFold2
ESMFold
Human Reference Proteome

ABSTRACT

AlphaFold2 predicts protein structures from structural and functional knowledge. Alternatively, ESMFold does the same adopting protein language models. Here, we map available Pfam domains on pairs of models of the human reference proteome computed with both procedures and we compare the mapped regions relevant for functional annotation. We find that, rather irrespectively of the global superimposition of the pairwise models, Pfam-containing regions overlap with a TM-score above 0.8 and a predicted local distance difference test (pLDDT) which is higher than the rest of the modeled sequence. This indicates that both methods are similarly performing in modeled regions that overlap Pfam domains, carrying structural and functional information, with pLDDT values slightly higher for AlphaFold2. The mapping of 9834 Pfam domains also allows the location of 2578 active sites in 3382 enzymes of the human proteome, including 807 proteins for which the active site is not reported in UniProt.

1. Introduction

Automated sequencing techniques generate a huge and increasing gap between the number of protein sequences deposited in databases, and their available three-dimensional structures ([1,2] and references therein). Recently, AlphaFold2 from DeepMind has been proposed as a useful tool for filling this gap in UniProt files ([3], https://www.uniprot.org). AlphaFold2 is a deep machine learning method trained on protein multiple sequence alignments, protein contact maps, correlated mutations, and protein family templates to infer protein structures [4,5]. However, when basic structural and functional information is missing, convincing and complete models are still poorly predicted. As a recent alternative, ESMFold adopts a protein "embedded" representation, derived from protein language models generated after filtering hundreds of millions of protein sequences, to compute the protein structure ([6] and references therein). Basically, both methods rely on an enormous computational power to extract evolutionary information at the base of concepts such as protein families and superfamilies, which routinely allowed the development of very successful methods for protein structure prediction including that of building by comparison [7]. AlphaFold2-based methods have been proven superior to ESMFold in the international benchmark of CASP15 [8], where however a limited number of structures were tested. For the sake of comparing both approaches, we recently generated a database of models for the human reference proteome, in which each human protein is endowed with AlphaFold2 and ESMFold models [9]. We commented before on the relatively better performance in structure prediction of AlphaFold2 when protein family templates are known [9]. Here, we focus on human enzymes and their functional annotations. In UniProt, human enzymes derive their functional annotation from the Association-Rule-Based Annotator (ARBA) rule system (https://www.uniprot.org/help/arba), "a multiclass learning system trained on expertly annotated entries" that unifies both InterPro signatures [10] and Pfam models [11], relying on the notion that proteins in families and superfamilies conserve functional and structural domains [12].

A protein domain is a compact three-dimensional region of the folded polypeptide chain that is self-stabilizing and that can fold independently from the rest [12]. Many proteins consist of several domains, and a domain may be shared by different proteins so that they can be considered as building blocks that molecular evolution adopted to

* Corresponding authors at: Biocomputing Group, University of Bologna, Italy.
E-mail addresses: pierluigi.martelli@unibo.it (P.L. Martelli), rita.casadio@unibo.it (R. Casadio).

https://doi.org/10.1016/j.csbj.2025.01.008
Received 21 October 2024; Received in revised form 8 January 2025; Accepted 13 January 2025
Available online 14 January 2025

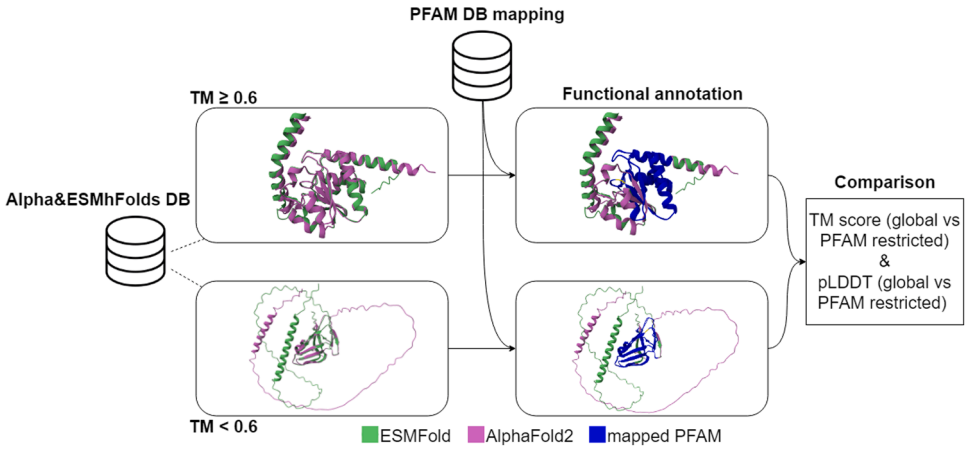

Fig. 1. Workflow of the annotation procedure adopted in this work. Alpha&ESMhFolds DB (https://alpha-esmhfolds.biocomp.unibo.it/); Pfam DB (https://pfam-docs.readthedocs.io). For TM-score and pLDDT definitions see the Material and Methods section.

generate proteins with different functions [12]. Domains can be detected with protein multiple structural and sequence alignments [13]. Traditionally, they have been modeled with hidden Markov models (HHMs) [14], and the Pfam database of protein-conserved domains is available ([11], https://pfam-docs.readthedocs.io). In the enzyme world, Pfam domains can include active sites, superfamily or family signatures, and/or structural characteristics ([10], https://www.ebi.ac.uk/interpro/). The procedure that we apply here is therefore based on mapping Pfam domains carrying along their annotation on our pairwise AlphaFold2 and ESMFold models of human enzymes out of the human reference proteome, focusing on those containing an active site and their comparison. We find that independently of the pairwise model superimposition, mapped Pfam regions are structurally well-predicted and superimposed. This indicates that both predictors in the Pfam regions are similarly effective in grasping functional and structural features.

2. Materials and methods

The dataset adopted for the analysis comprises 6956 human enzymes, all the proteins endowed with an EC number included in the Alpha&ESMhFolds database [9]. Alpha&ESMhFolds is a collection of 42,942 proteins extracted from the human Reference Proteome (UP000005640, available at UniProt [3], release 2023_03 of January 2023), which for each protein provides the respective AlphaFold2 and ESMFold models. AlphaFold2 models are downloaded from the AlphaFoldDB ([5,15] https://alphafold.com, accessed in January 2023), and their ESMFold counterpart is computed in-house [9]. A fraction of the total enzyme database (1314 proteins of the 6956) is endowed with a PDB [16] three-dimensional structure covering at least 70 % of the protein sequence. Among the remaining 5643 proteins, 3037 are included in Swiss-Prot, the manually curated part of UniProt. The remaining 2606 are listed in TrEMBL, the automatically annotated part of UniProt.

For each enzyme, we extract the set of annotations present in Pfam [11] downloading data available at the website (https://pfam-docs.readthedocs.io, version 37.0 accessed in September 2024). We collected 14,122 Pfam entries, including 9834 domains, 3391 families, 769 repeats, 93 short motifs, 19 conserved intrinsically disordered regions, and 16 coiled-coil regions, all documented also in InterPro (https://www.ebi.ac.uk/interpro/). Notably, the Pfams we extracted are the same as those annotated in the UniProt database. Additionally, we ran the PfamScan tool to annotate 2578 Pfam entries with a reported active site. In total, 5684 residues were determined to be part of an active site with an average of 2.2 residues per active site (1642 active sites with 1 residue, 1230 with 2, 461 with 3, 47 with 4, the active site of domain PF00561 in the protein Q9H4I8 which has 5, and the active site

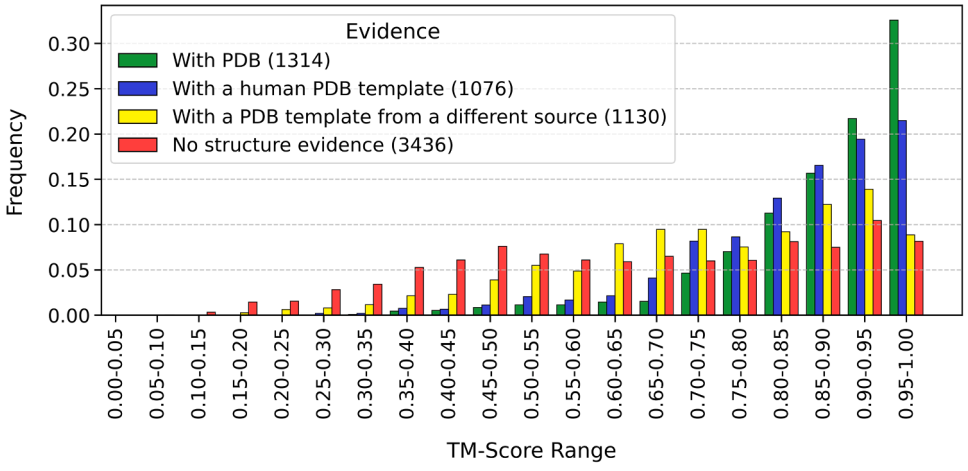

Fig. 2. Distribution of pairwise AlphaFold and ESMFold models of human enzymes in our database ([9], available at https://alpha-esmhfolds.biocomp.unibo.it/)) as a function of their superimposition as evaluated with the TM-score. Colors distinguish models with an underlying PDB structure (green), with a human PDB template (blue), with a PDB template from other organisms (yellow), and without structural evidence (red). It appears that 49.4 % of the 6959 pairwise models do not have PDB structural templates.

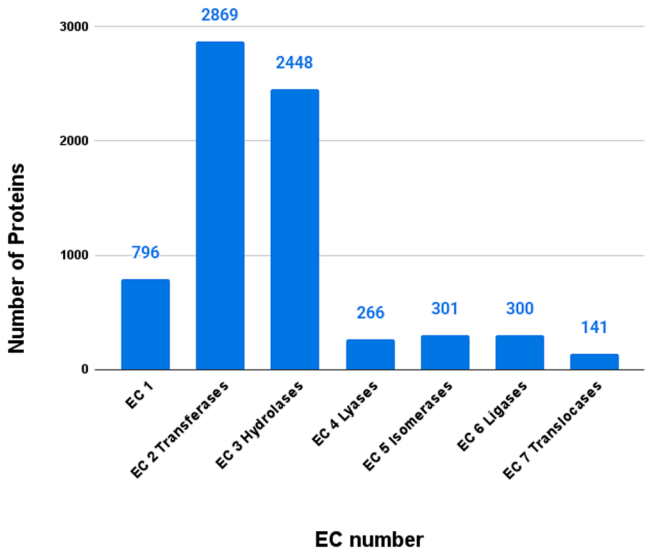

Fig. 3. Histogram showing the number of proteins associated with each EC number. 151 proteins out of the 6956 enzymes in our dataset are associated with more than one EC number. The multi-EC combinations in the dataset and the relative number of proteins is reported here: EC 2–3: 71; EC 2–4: 12; EC 1–2: 9; EC 3–4: 7; EC 4–5: 7; EC 1–3: 6; EC 1–4: 6; EC 1–5: 5; EC 2–6: 5; EC 3–7: 3; EC 3–5: 3; EC 2–5: 2; EC 4–6: 1; EC 1–3–6: 4; EC 1–2–4: 3; EC 1–2–5: 2; EC 2–3–4: 2; EC 1–4–5: 2; EC 1–2–3: 1.

of domain PF01536 in the protein P17707 which has 6).

For each protein, the Alpha&ESMhFolds database [9] provides the per-residue pLDDT of the AlphaFold2 and ESMFold models (a self-assessed measure by both methods of the reliability of the predicted model [17]), as well as the TM-score between the two models (a metric to estimate the similarity between the two 3D models [18]). Here, we focus on computing the local TM-score of the regions of the proteins that are covered by a Pfam entry, obtained by manually extracting the relevant regions from the predicted models. Evaluation is performed with the Foldseek program [13], which computes the structural superimposition of the models and their similarity. The Foldseek program was run using the options "–alignment-type 1" (alignment type set to the TM-align algorithm) and "–prefilter-mode 2" (disabling prefiltering of results). The remaining program parameters were left to default values.

For the sake of reproducibility, we provide a GitHub repository accessible at https://github.com/MatteoManfredi/pfam_models. It contains detailed instructions to run a script that computes the results reported in this manuscript starting from the list of UniProt identifiers of our enzyme dataset. The script can also be used to reproduce our analysis starting from a different list of UniProt identifiers, as long as those are included in our web server Alpha&ESMhFolds which provides the pairwise models.

3. Results and discussion

Proteins Enzyme Commission (EC) numbers are routinely assigned with the UniProt automatic annotations pipeline (https://www.uniprot.org/help/biocuration), which leverages the ARBA Rule system (https://www.uniprot.org/help/arba). In this paper, we are interested in understanding the ability of AlphaFold2 and ESMFold models to capture the functional features of human enzymes. In Fig. 2 we show the distribution of enzyme models as a function of the computed TM-score, color-coded depending on their structural evidence. It appears that 49.4 % of the pairwise models (6956) are without a PDB structural template.

Summing up, human enzyme model pairs can be grouped into three categories: i) models with an underlying PDB structure of the enzyme, ii) models without PDB in the background which superimpose (TM-score ≥ 0.6), or iii) which do not superimpose (TM-score <0.6). Then, after mapping the Pfam database on the models, we compute TM-scores at the Pfam region and we evaluate the local quality of the prediction by averaging the per-residue pLDDT score over the Pfam region (see Materials and Methods for definitions and Fig. 1).

Table 1
The human enzyme set as distributed on the 6 types of Pfam entries.

Pfam Type[a]	# Entries in Pfam database	# Unique Pfam in the HES[b]	# Enzymes with Pfam	# Pfam Occurrences	Range of Pfam lengths[c]
Domains	9147	1517	5204	9834	16–713
Domains with annotated active site	773	249	2459	2577	34–655
Families	11,536	683	2738	3391	11–1444
Repeats	859	78	324	769	14–517
Short Motifs	122	21	77	93	15–61
Intrinsically disordered regions	122	9	19	19	60–165
Coiled-coil regions	193	8	14	16	35–331

= Number of

[a] Pfam classifies its entries into 6 types (see Materials and Methods and https://pfam-docs.readthedocs.io). We additionally distinguish among the "Domains" those containing an active site, as annotated by the PfamScan tool.

[b] HES = Human Enzyme Set comprising 6956 enzymes.

[c] We report the minimum and maximum length of the Pfam types included in our dataset, (number of residues). See Figure S1 for distributions.

Fig. 3 shows the distribution of the human protein enzymes as a function of the seven first levels of EC numbers. The most populated classes are Transferases and Hydrolases. Traditionally, ARBA rules include, among other features, Pfam domains. In the automatic annotation process at UniProt, the biocuration system (https://www.uniprot.org/help/biocuration) filters protein sequences and assigns the functional annotation, inclusive of the EC number, provided that the sequence meets a given set of features. In Table 1 we distribute our enzyme database according to the different Pfam types, as described on the Pfam website (see Materials and Methods). The Domain type is particularly interesting for our analysis since it contains information on the functional annotation and, when known, also on the active sites, which are conserved in the families. In the human enzyme set, 1517 unique Pfam domains are shared by 5204 enzyme proteins for a total of 9834 occurrences. The domain lengths range from 16 up to 713 residues (length distribution is shown in Figure S1). Included in domains are those containing an active site (249) which map into 2459 human enzymes, with lengths ranging from 34 to 655 (See Figure S1 for distribution), and 99 % of these carry structural information. Along with domains, Pfam models are available for family signatures, repeats, short motifs, intrinsically disordered regions, and coiled-coil regions (https://pfam-docs.readthedocs.io/en/latest/summary.html). These Pfam models, in principle, can shed light on functionality, although in a more general way. In Table 1 we list Pfam types, sorted by decreasing number of human enzyme proteins of our data set.

According to the workflow of Fig. 1, after mapping the Pfam database on the pairwise models, we compare the TM-scores of the global models with those evaluated for the Pfam regions. Similarly, we compare the local quality of the prediction of the global models with that restricted to the Pfam regions (by computing the pLDDT score). Results are shown in Fig. 4 and Table S2. TM-scores of the pairwise models are higher when the models have an underlying PDB structure (1047 enzymes); when structural information is lacking, 2766 enzymes have global models with a pairwise TM-score ≥ 0.6, and 1391 enzymes have global models with TM-score < 0.6. Considering the Pfam domains, TM-scores remain high when evaluated on the region of the mapped domain, irrespective of the superimposition of global enzyme

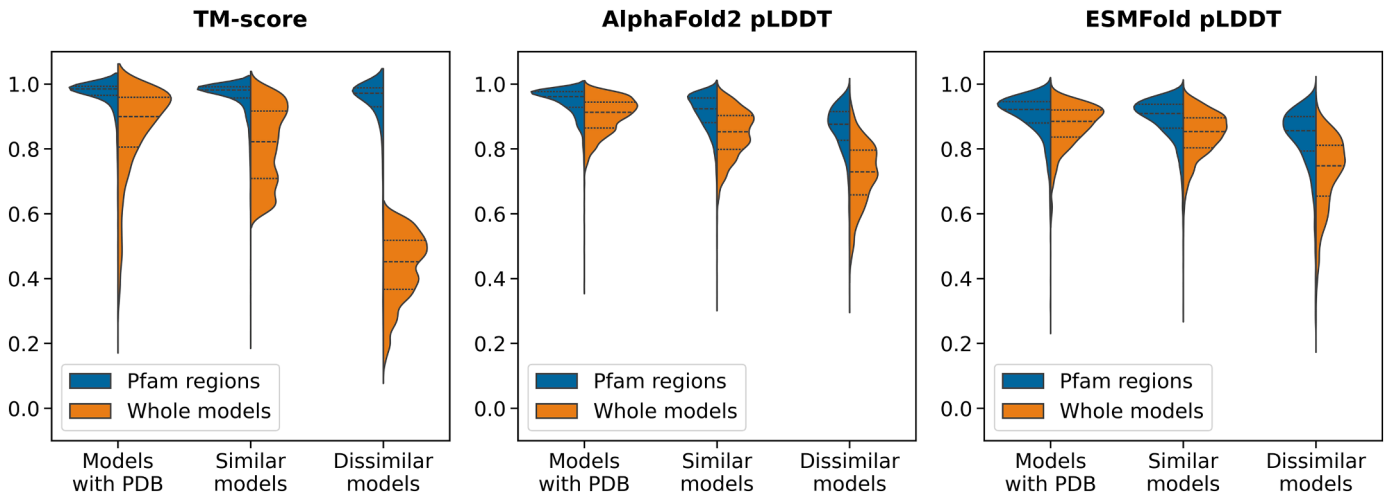

Fig. 4. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam domains. From left to right, on the y-axis we report the TM-score obtained when superimposing the two models, the mean pLDDT for AlphaFold2 models, and the mean pLDDT for ESMFold models. On the x-axis, we report observations for three subsets of enzymes, respectively those with a known PDB structure in the database, those with similar models (TM-score $\geq$ 0.6), and those with dissimilar models (TM-score $<$ 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

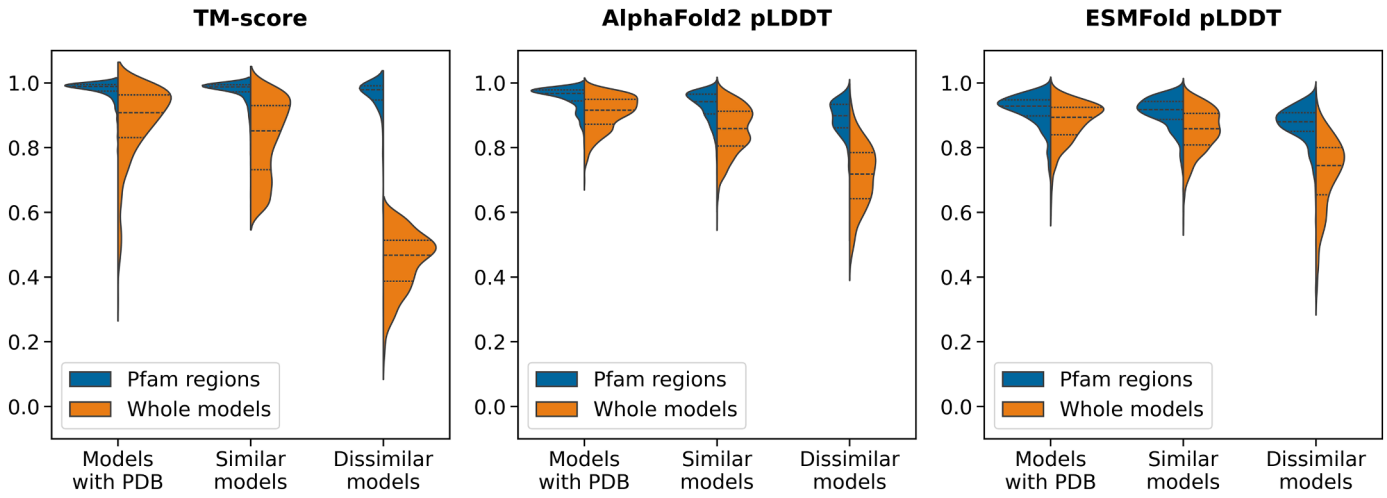

Fig. 5. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam domains with an active site. From left to right, on the y-axis we report the TM-score obtained when superimposing the two models, the mean pLDDT for AlphaFold2 models, and the mean pLDDT for ESMFold models. On the x-axis, we report observations for three subsets of enzymes, respectively those with a known PDB structure in the database, those with similar models (TM-score $\geq$ 0.6), and those with dissimilar models (TM-score $<$ 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

pairwise models (Fig. 4). The same holds for the predicted local evaluation of the quality of the models (pLDDT), which is slightly higher in AlphaFold2 than in ESMFold.

When we restrict the analysis to the enzymes with an active site, derived by mapping Pfam domains containing an active site, again we can observe (Fig. 5 and Table S3) that the superimposition of the Pfam regions (TM-score) in the pairwise models increases rather irrespectively of the superimposition and the quality of the global enzyme models. This is particularly evident when the global models diverge (TM-score $<$ 0.6, Fig. 6). Interestingly, our mapping performed with PfamScan allows the annotation of 807 proteins with 858 active sites from 117 Pfam domains. These annotations are lacking in the associated files at UniProt.

According to the Pfam definition, a family includes proteins that share a common evolutionary origin "as reflected in their related functions, sequences or structure" (https://pfam-docs.readthedocs.io/en/latest/summary.html). We therefore mapped Pfam family models into our dataset of enzymes and compared the pairwise models. Noticeably, as reported in Table 1, 509 enzymes map more than one Pfam family (for details see Table S1). We found that the mapped Pfam region overlap (higher TM-score than that of the global models) and the local quality of the self-evaluation of the predictions is higher than average (Fig. 7 and Table S4). Finally, we consider other Pfam models including repeats, short motifs (including metal binding sites), intrinsically disordered regions, and coiled-coiled regions. Even in these cases (with the exception of disordered regions), the shared Pfam regions among the pairwise models have better quality than the overall global models (see Tables S5, S6, S7, and S8). The number of enzymes that map these Pfam types is progressively decreasing (Table 1), indicating that the annotation carried along pertains to a minor fraction of our human enzyme dataset.

4. Conclusions

In this paper, we functionally annotate our dataset of pairwise models of human enzymes derived from the human reference proteome.

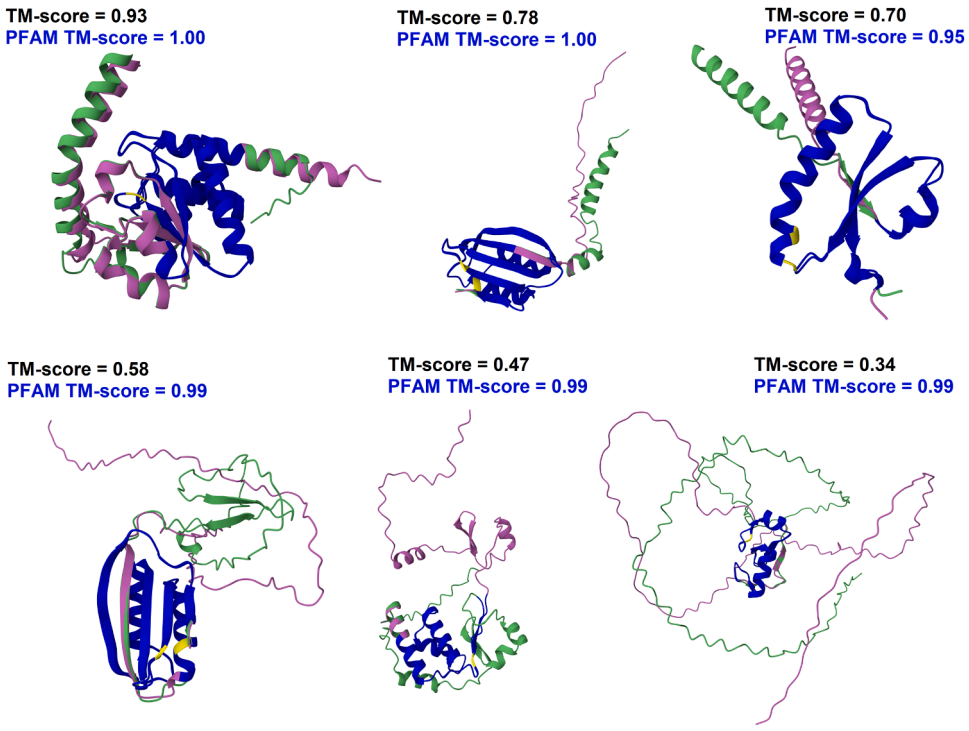

Fig. 6. Examples of enzymes at different levels of model TM-scores showing that the superimposition remains good in the region covered by a Pfam domain. In all images, the green model is obtained with ESMFold and the purple one with AlphaFold2. We highlight the regions covered by a Pfam domain in blue and the active sites in yellow. In the image, we report the TM-score of the AlphaFold2 and ESMFold models (in black) and the TM-score computed on the region covered by the domain (in blue). From left to right, top to bottom, we show: "Phosphatidylglycerophosphatase and protein-tyrosine phosphatase 1" (UniProt accession: Q8WUK0, EC 3.1.3.27), with a "Dual specificity phosphatase, catalytic" domain (Pfam accession: PF00782) covering residues 88–184 (active site in position 132); "Acylphosphatase" (UniProt accession: G3V2U7, EC 3.6.1.7), with a "Acylphosphatase" domain (Pfam accession: PF00708) covering residues 41–127 (active site in positions 54 and 72); "Protein disulfide-isomerase" (UniProt accession: H3BVI1, EC 5.3.4.1), with a "Thioredoxin" domain (Pfam accession: PF00085) covering residues 32–87 (active site in positions 53 and 56); "Acylphosphatase" (UniProt accession: U3KQL2, EC 3.6.1.7), with a "Acylphosphatase" domain (Pfam accession: PF00708) covering residues 94–170 (active site in positions 97 and 115); "Dual specificity protein phosphatase" (UniProt accession: E9PSD4, EC 3.1.3.16), with a "Dual specificity phosphatase, catalytic" domain (Pfam accession: PF00782) covering residues 81–135 (active site in position 87); "E3 ubiquitin-protein ligase RNF4" (UniProt accession: P78317, EC 2.3.2.27), with a "Ring finger" domain (Pfam accession: PF13639) covering residues 131–177 (active site in position 132).

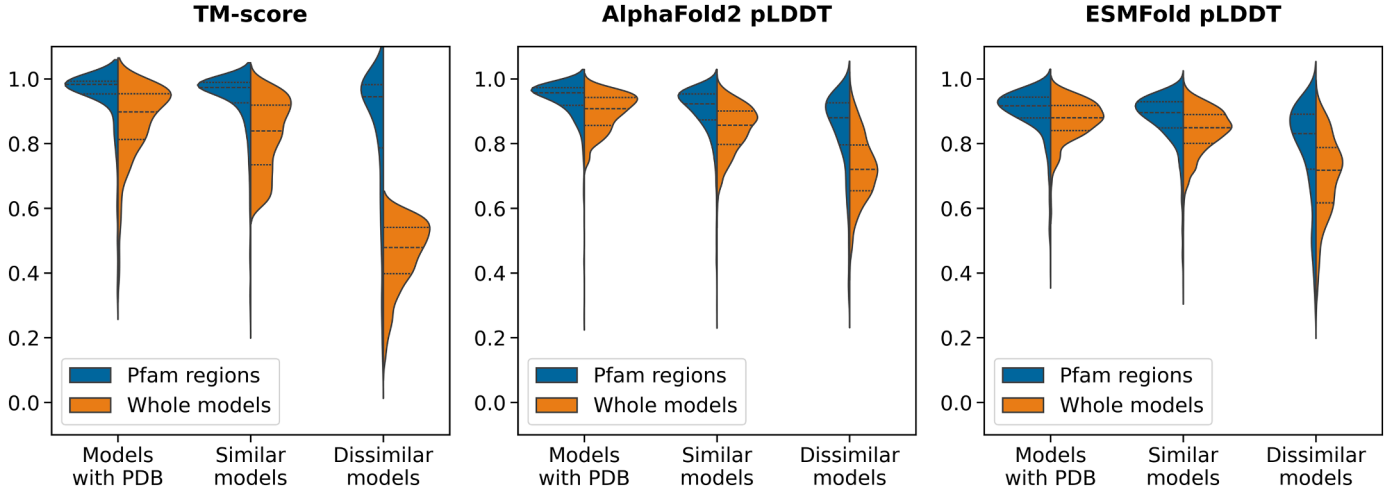

Fig. 7. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam families. From left to right, on the y-axis we report the TM-score obtained when superimposing the two models, the mean pLDDT for AlphaFold2 models, and the mean pLDDT for ESMFold models. On the x-axis, we report observations for three subsets of enzymes, respectively those with a known PDB structure in the database, those with similar models (TM-score ≥ 0.6), and those with dissimilar models (TM-score < 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

Global models are generated with AlphaFold2 and ESMFold, two different methods relying the first on deep learning of different protein features including evolution information derived from multiple sequence alignments, correlated mutations, and contact maps, the second on the protein sequence representation with embeddings derived from protein large language models [5,6,9]. For functional annotation

we took advantage of Pfam models, casting into HMMs local structural and/or functional conservation highlighted by grouping proteins into families and superfamilies (clans) [11]. What we obtain is interesting, particularly when considering pairwise global models of enzymes without a PDB reference structure (81 % of the entire enzyme dataset). Rather independently of the superimposition of the global enzyme models that we evaluate with the TM-score and group in two sets (TM-score higher or lower than 0.6 [9]), Pfam regions overlap, as demonstrated by the Pfam-restricted TM-score values. In these regions, the predicted local evaluation of the quality of the models (pLDDT) is higher than the quality of the global model. Our results suggest that both AlphaFold2 and ESMFold methods are equally good in grasping the information carried out by Pfam models.

Interestingly, our procedure allows mapping structural and functional information in enzyme domains where the active site is present, as detected by PfamScan. For 807 human enzymes (whose list is available as a supplementary file), the functional annotation of the active site is not yet present in the associated UniProt file.

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

Gabriele Vazzana: Visualization, Investigation, Formal analysis. **Castrense Savojardo:** Writing – review & editing, Software, Methodology, Funding acquisition. **Matteo Manfredi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Pier Luigi Martelli:** Writing – review & editing, Funding acquisition, Conceptualization. **Rita Casadio:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Data curation, Conceptualization.

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. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.01.008.

References

[1] Kandathil SM, Lau AM, Jones DT. Machine learning methods for predicting protein structure from single sequences. Curr Opin Struct Biol 2023;81:102627.
[2] Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Syst 2021;12:654–69. e3.
[3] Consortium UniProt. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51. D523–31.
[4] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[5] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res 2024;52:D368–75.
[6] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30.
[7] Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–6.
[8] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XV. Proteins 2023;91:1539–49.
[9] Manfredi M, Savojardo C, Iardukhin G, Salomoni D, Costantini A, Martelli PL, et al. Alpha&ESMhFolds: a web server for comparing AlphaFold2 and ESMFold models of the human reference proteome. J Mol Biol 2024;436:168593.
[10] Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res 2023;51:D418–27.
[11] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49:D412–9.
[12] Lesk A. Introduction to bioinformatics. 5th ed. London, England: Oxford University Press; 2019.
[13] van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 2024;42:243–6.
[14] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis. Cambridge, England: Cambridge University Press; 2007.
[15] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596:590–6.
[16] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[17] 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;29:2722–8.
[18] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2007;68. 1020–1020.