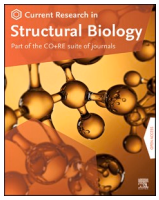
Current Research in Structural Biology 9 (2025) 100167

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology



Research Article

Evaluation of the structural models of the human reference proteome: AlphaFold2 versus ESMFold

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

[a] Biocomputing Group, Dept. of Pharmacy and Biotechnology, University of Bologna, Italy
[b] Biocomputing Group, AlmaClimate Interdepartmental Center, University of Bologna, Italy
[c] Institute of Biomembrane and Bioenergetics, Italian National Research Council (IBIOM-CNR), Italy

ARTICLE INFO

Handling Editor: Arun Prasad Pandurangan

Keywords:
Human protein structure prediction
AI computed structural models
Structural model quality assessment
Model secondary structure
Model solvent accessibility
AlphaFold2
ESMFold
QATEN

ABSTRACT

The human reference proteome is routinely modelled with predictive tools such as AlphaFold2. We recently released a database in which, for each human protein, the AlphaFold2 model is paired with its ESMFold counterpart. The two predictive methods take advantage of different procedures and it is interesting to compare them in relation to their quality, particularly when an experimental protein structure is not available. Here, we select three state-of-the-art quality assessment methods and we adopt them to compare 42,942 pairs of models. This procedure helps to find the most reliable models for human proteins, particularly for the set of proteins for which structure prediction methods give dissimilar results. We obtain that when predicted structures are similar, AlphaFold2 models consistently receive higher scores than the ESMFold counterparts. When predicted structures differ, the ESMFold model is the best choice for 49 % of the proteins according to a consensus of the three QA tools.

1. Introduction

Artificial Intelligence (AI)-based methods are ever more relevant to scientific fields of study, including structural biology. The release of AlphaFold2 (Jumper et al., 2021) allowed a huge leap in the ability to generate high-quality predicted 3D models of proteins. After their impressive results at the 14th edition of the Critical Assessment of methods for Structure Prediction (CASP14, https://predictioncenter.org/), many groups worked on either fine-tuning the capabilities of AlphaFold2 or on generating alternative methods (Kryshtafovych et al., 2021). Among the last, ESMFold takes advantage of protein Language Models (pLMs) to achieve good performances without the dependency on Multiple Sequence Alignments (MSAs), allowing it to be faster and to generate better models for protein sequences lacking known homologs (Lin et al., 2023).

Despite the huge success of these combined efforts, which was confirmed by the results of the CASP15 (Kryshtafovych et al., 2023), many proteins remain hard to predict, and different methods may perform better or worse in different settings. This raises the need to compare models when generated with different methods, to identify which can be trusted and which should be disregarded. To this aim, we recently released a novel database collecting AlphaFold2 and ESMFold models for 42,942 proteins covering the human Reference Proteome, called Alpha&ESMhFolds (Manfredi et al., 2024). The database is accessible as a web server at https://alpha-esmhfolds.biocomp.unibo.it/ and it allows end users to examine the superimposition between the two models for a protein of interest, reporting statistics useful to determine how similar they are. This is of particular relevance, since on a large scale we observed that the similarity between two models (as measured by the TM-score (Zhang and Skolnick, 2004)) strongly correlates with the reliability of either one (as evaluated by the self-assessed pLDDT (Mariani et al., 2013) that AlphaFold2 and ESMFold provide as output) (Manfredi et al., 2024). Our resource helps to distinguish proteins where the two methods provide similar results from proteins where the two methods disagree (models of 23,720 proteins, out of 42,942, have a TM-score < 0.6). In the latter case, either model should be adopted with caution and the problem of identifying which, if any, of the two models is of higher quality is open. Here, we apply Quality Assessment (QA) methods of predicted protein models in order to score which of the two models for any human protein is more reliable.

This article is part of a special issue entitled: AI in structural biology published in Current Research in Structural Biology.

* Corresponding author.
** Corresponding author. Biocomputing Group, AlmaClimate Interdepartmental Center, University of Bologna, Italy.
E-mail addresses: castrense.savojardo2@unibo.it (C. Savojardo), rita.casadio@unibo.it (R. Casadio).

https://doi.org/10.1016/j.crstbi.2025.100167
Received 26 September 2024; Received in revised form 21 March 2025; Accepted 28 April 2025
Available online 22 May 2025

QA methods released in the literature differ both in scope and approach. They are routinely trained to compute local metrics, like pLDDT or the S-score (Ray et al., 2012), which can then be pooled to produce a per-protein score. Other methods are also trained to directly predict global scores such as GDT-TS (Zemla, 2003). QA methods can be grouped into single-model and multi-model methods. The latter requires a pool of different models for the same protein to be analysed together and they are trained to extract from them a general consensus. The methods achieve high correlations between their predicted scores and the similarity to the PDB structures, ((Berman et al., 2000), https://www.rcbs.org), but fail to identify the single best model, and their quality score is constrained by the number of available models. As an alternative, single-model methods are trained to extract inherent features of a model to estimate its quality, and the metrics they predict are good for sorting models based on their quality (Uziela et al., 2017; Kryshtafovych et al., 2016). QA methods are routinely trained and/or tested on datasets extracted from different editions of the Critical Assessment of Structure Prediction (CASP, https://predictioncenter.org/) and they exploit a vast suite of features and deep learning techniques (Kryshtafovych et al., 2016, and references therein).

Overall, our analysis indicates that when the PDB structure is known, both AlphaFold2 and ESMFold generate good quality models, with a slightly better performance of AlphaFold2 than ESMFold. When proteins lack structural information, the QA scores of both models progressively deteriorate at increasing divergence of the models. In this region, 49 % of ESMfold models are of higher quality than that of AlphaFold2. We also notice that when models are dissimilar, AlphaFold2 models are less compact than their ESMFold counterparts and that, consistently, their coil content is higher, with longer unstructured regions.

2. Materials and methods

2.1. Dataset

The dataset adopted for the analysis consists of 42,942 proteins extracted from the human Reference Proteome (UP000005640) available at UniProt (UniProt Consortium, 2023), (release 2023_January 03, 2023) and included in the Alpha&ESMhFolds database (Manfredi et al., 2024), with the respective AlphaFold2 and ESMFold models (for a total of 85,884 models). As previously described (Manfredi et al., 2024), we derived AlphaFold2 models from the AlphaFold database (https://alphafold.com) and computed in-house the ESMFold models (Manfredi et al., 2024). Both predicted models are full coverage with respect to the protein sequence as derived from UniProt. In the total human set, 2900 proteins are endowed with experimentally resolved 3D structures deposited in the PDB. We assess the reliability of the QA methods by comparing their predictions to the ground truth of the PDB structures. On the remaining 40,042 proteins, we carry out a large-scale analysis of model quality, based on the assessment of the three selected methods.

2.2. QA methods

We adopt three recently introduced QA methods: DeepAccNet (Hiranuma et al., 2021), QMEANDisCo (Studer et al., 2020) and QATEN (Zhang et al., 2023). As they all fall into the category of single-model methods, they are well-suited for our problem where we only have two models for each protein and we need to identify the best one. Importantly, they are also very fast, allowing us to obtain predictions for a total of 88,784 models (85,884 predicted and 2900 experimental) in less than a week. The size of our dataset was indeed a strong constraint, and it is the reason why methods such as ModFOLD9 (McGuffin and Alharbi, 2024), which require 24–48 h of computation for a single protein, were not considered.

DeepAccNet is trained to predict the LDDT score (pLDDT (Mariani et al., 2013)) for each residue, and it computes the average over the sequence as a metric of global quality (mean pLDDT). It is trained on a dataset derived from the PISCES (Wang and Dunbrack, 2003) server and tested on data from CASP13 (https://predictioncenter.org/). The architecture is based on Convolutional Neural Networks and it adopts a complex set of features as input, including distance maps, amino acid identities and properties, local atomic environments, backbone angles, residue angular orientations, Rosetta energy terms, and secondary structure information.

QMEANDisCo is also trained to predict the LDDT score for each residue, and it computes several global metrics (Studer et al., 2020). For consistency with DeepAccNet, we chose the average pLDDT. It is trained on data derived from the CAMEO dataset (Haas et al., 2018) and CASP12 targets, and tested on CAMEO benchmark sets and CASP13 targets. The architecture is based on fully connected feed-forward neural networks that combine the features adopted by the previous release of QMEAN (Benkert et al., 2008), including solvent accessibility, with a consensus-based distance constraint score (DisCo).

Finally, QATEN is trained to predict both the LDDT score for each residue and the GDT-TS (Zemla, 2003) as a global metric. For the sake of comparison with the other tools, we chose to compute the average pLDDT over the protein sequence for our analysis. The method is trained on a dataset generated by merging the training sets of DeepAccNet and GNNRefine (Jing and Xu, 2021), and it is tested on data from CASP14. The architecture is based on Graph Convolution and Attention Layers and it encodes the input model as a graph where nodes represent atoms, each encoded as an ensemble of distance-based features, orientational features, and bond-related features.

Ultimately, the chosen methods represent the most recent state-of-the-art regarding single-model quality assessment. They all perform extremely well in their respective benchmarks, and they can be executed in-house to predict scores for all 88,784 models included in our database. Seeing that their performances are comparable and that they produced different results on some of the models, we included all three in our analysis to allow us to generate a consensus.

2.3. Other methods

We run the DSSP program (Kabsch and Sander, 1983) on AlphaFold2 and ESMFold models for further validation of their quality. From the DSSP output, we specifically extracted the Accessible Surface Area (ASA) and the Secondary Structure (SS). We converted the ASA of each residue to Relative Surface Accessibility (RSA) values using the Sander and Rost scale (Rost and Sander, 1994). In order to distinguish unstructured from structured regions of the sequence, we converted the 8 SS classes to a binary distinction between COIL (T: Turn, B: Bend or None) and NON COIL (H: Alpha helix, G: 3–10 helix, I: Pi helix, E: Strand, B: Isolated beta-bridge residue).

3. Results and discussion

3.1. Validation of QA methods on PDB structures

The three chosen Quality Assessment methods (see Methods) are first validated on 2495 out of the 2900 PDB structures available in our dataset (405 structures were removed since one or more methods failed to produce a result due to problems in the PDB file; most notably, this never happens with the PDB files generated by AlphaFold2 and ESMFold). Since we routinely consider the experimental 3D structure as ground truth, an ideal QA method should attribute a high score to all the 2495 PDBs. Although this is not the case, our results are in line with the performances declared by the authors of the methods, and overall show their reliability. Fig. 1A reports boxplots of the predictions of the three QA methods, indicating that DeepAccNet computes values close to 1 more consistently and that QMEANDisCo is the one with the shorter tail, meaning that it produces bad predictions for fewer structures. Among the three, QATEN never scores a model higher than 0.9, a trend later observed on the whole dataset pointing to an inherent bias. However,

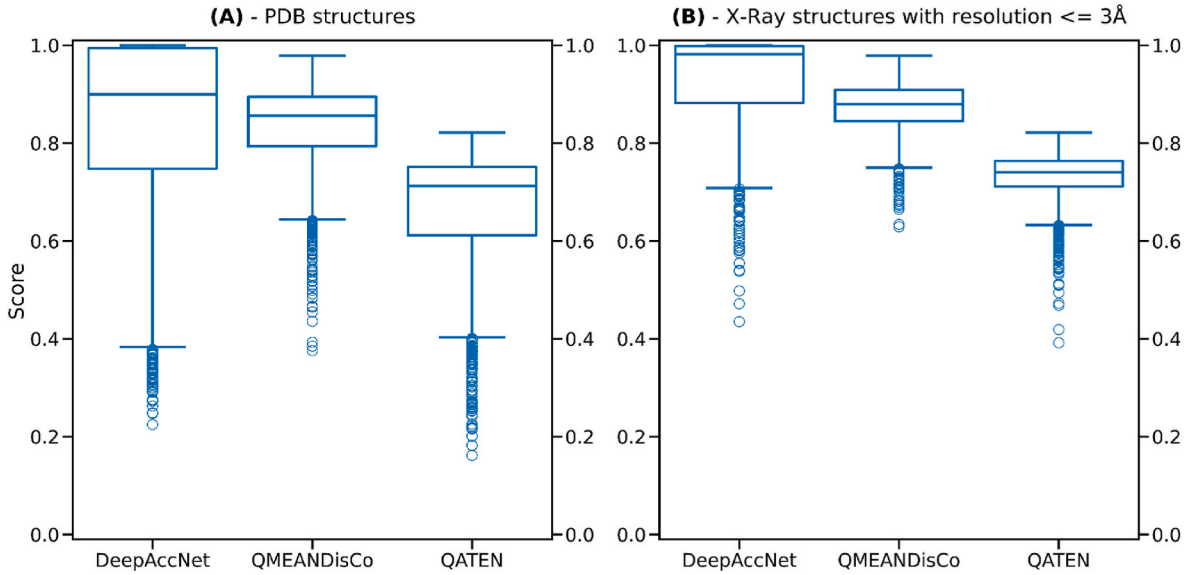


Fig. 1. Boxplots of the distributions of predictions of quality assessment with the three QA methods considered. DeepAccNet, QMEANDisCo, and QATEN compute the average of the predicted LDDT per protein (range (0–1)). The metrics measure how similar a predicted model would be to the experimental PDB structure and the predictors should assign high scores to all structures. From left to right, we report A) the distributions computed on all 2495 PDB structures for which the QA methods were able to produce a result. B) the distributions computed on a subset of 1575 PDB structures determined by X-ray diffraction with a resolution $\leq$ 3 Å.

being a consistent trend, it should not impact the comparison between two different models of the same protein. Fig. 1B shows similar boxplots obtained on a smaller subset of 1575 structures determined with X-ray diffraction methods with a resolution $\leq$ 3 Å. On this subset, the performances of QA methods increase considerably, confirming that the three-dimensional structure accuracy is important in the evaluation (Hiranuma et al., 2021; Studer et al., 2020; Zhang et al., 2023).

3.2. Validation of QA methods on predicted models of proteins with a PDB structure

We then analyse the scores of the QA methods when evaluating AlphaFold2 and ESMFold models of the 2900 proteins endowed with a PDB structure. For this subset, we have as a reference the TM-score between each predicted model and the corresponding PDB. The score estimates how good the computed model is with respect to the protein experimental structure (Manfredi et al., 2024; Zhang and Skolnick, 2004). Considering the TM-scores, we can cluster our models into two sets: one with the models with a score higher or equal to 0.6, indicating good fitting to the PDB structures; the other with the remaining models (TM-scores lower than 0.6), (Zhang and Skolnick, 2004). In the first cluster, we find the largest fraction (5,436) of the models generated by AlphaFold2 (2784) and by ESMFold (2,652). In the second cluster, 364 models derive from AlphaFold2 (116) and ESMFold (248), respectively. When the corresponding PDB structures are available, indeed both predictors give good models (Manfredi et al., 2024). Fig. 2 shows the boxplots of the scores of the three QA methods for both clusters. As expected, values are lower when models are worse.

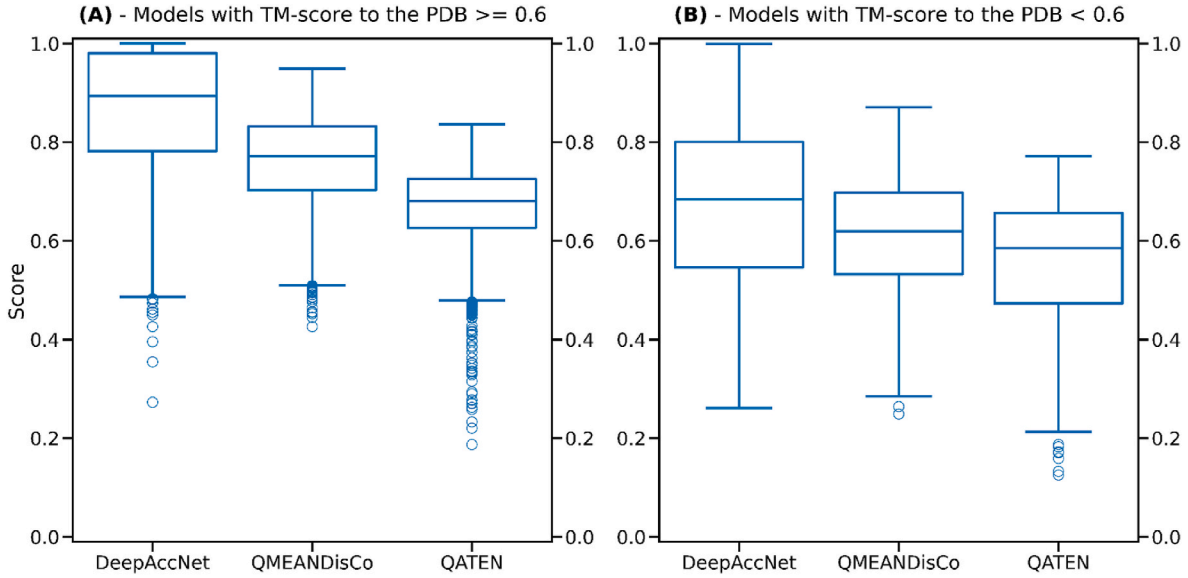


Fig. 2. Boxplots of the distributions of quality assessment of 5800 AlphaFold2 and ESMFold models (for 2900 human proteins) with the three QA methods considered. Each method computes the average per protein predicted LDDT (pLDDT). From left to right the figure shows: A) the distributions computed on 5436 models with TM-score to their corresponding PDB $\geq$ 0.6; B) the distributions computed on 364 models with TM-score to their corresponding PDB $<$ 0.6.

Overall, the results of our analysis confirm the reliability of the adopted Quality Assessment methods, giving us more confidence in their application to proteins lacking an experimental 3D structure.

3.3. Quality assessment of AlphaFold2 and ESMFold models without PDB structures

3.3.1. Validation of QA methods towards AlphaFold2 and ESMFold self-assessment

In our human dataset, 40,042 proteins lack a direct PDB structure. For assessing the superimposition of two protein models we rely again on the TM-score and adopt the threshold value of 0.6 to discriminate similar from dissimilar models (Manfredi et al., 2024; Zhang and Skolnick, 2004). Out of 80,084 models (2 for each protein), 33,204 have a TM-score higher than 0.6 and 46,880 lower than 0.6. Furthermore, both AlphaFold2 and ESMFold provide a per-residue predicted LDDT score (pLDDT) which estimates the reliability of the generated model (Jumper et al., 2021; Lin et al., 2023). In our database (Manfredi et al., 2024), we reported for each sequence the pairwise models and their TM- and pLDDT scores. These measures are self-assessed by each method, and for this reason, here we investigate how three external validation procedures qualify the models. We think that this is of great relevance, since the proteins lack experimental structure, and the adoption of QA tools should help the identification of the best model between the pair AlphaFold2 and ESMFold.

We compare the average pLDDT per sequence, as obtained from the three QA tools adopted, after analysing their consistency. The comparison among QA scores and AlphaFold2 and ESMFold average pLDDT

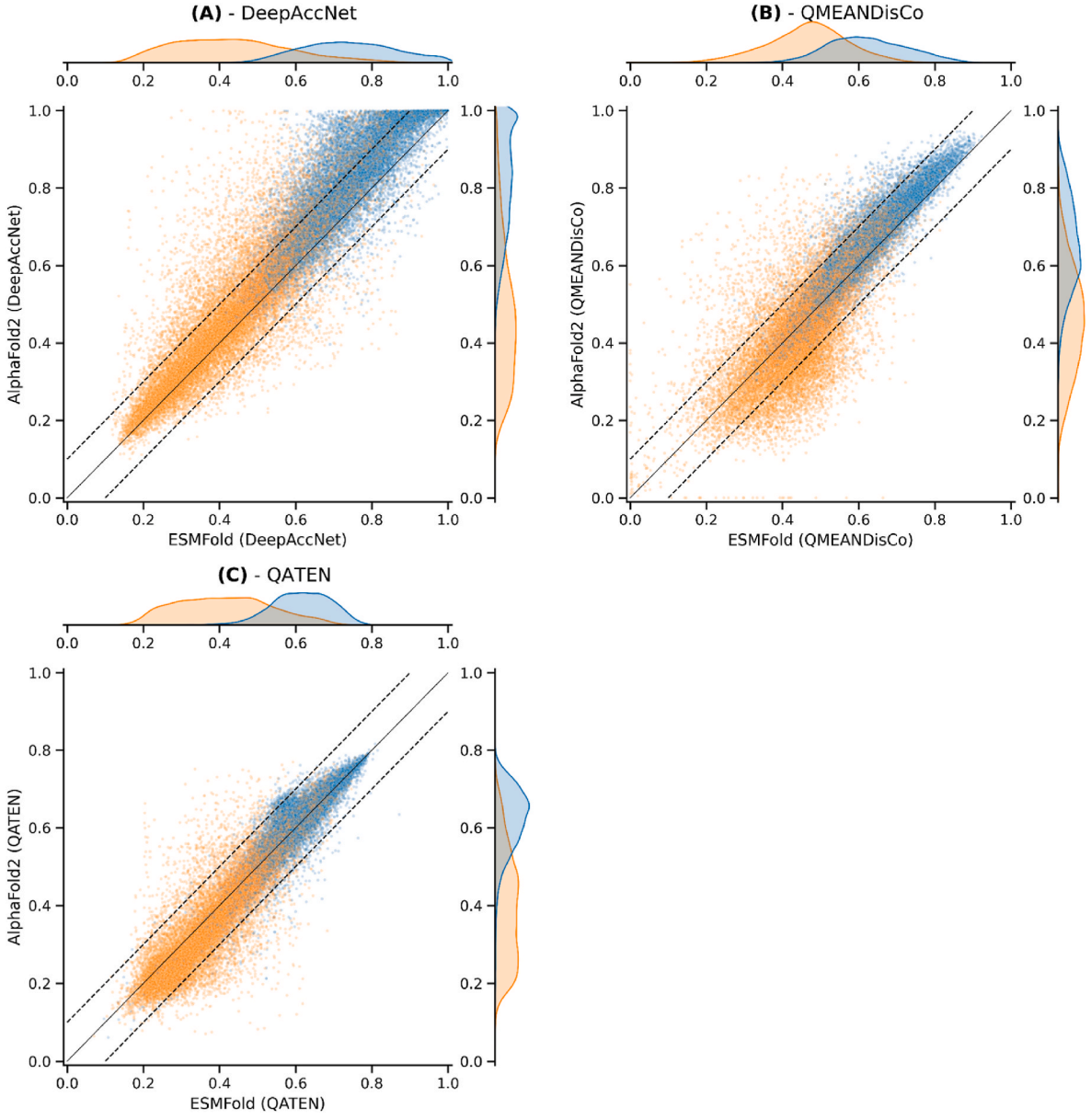


Fig. 3. Scatterplots showing one point for each protein in the dataset. From left to right, top to bottom, A) on the x-axis, we report the DeepAccNet score for the ESMFold model, against the DeepAccNet score for the AlphaFold2 model on the y-axis; B) on the x-axis, we report the QMEANDisCo score for the ESMFold model, against the QMEANDisCo score for the AlphaFold2 model on the y-axis; C) on the x-axis, we report the QATEN score for the ESMFold model, against the QATEN score for the AlphaFold2 model on the y-axis. In the three graphs, points above the diagonal correspond to proteins for which AlphaFold2 models have a higher score, and points below the diagonal correspond to proteins for which ESMFold models have a higher score. The two dotted lines divide points for which the absolute difference in average pLDDT is higher or lower than 0.1. Each protein is represented with a blue point when the two models have TM-score $\geq$ 0.6, with an orange one otherwise. The curves at the edges of the scatterplots are the distributions of the two classes as computed with kernel density estimation.

values indicates that their correlation holds for models with TM-scores higher or lower than 0.6 (Fig. S1 refers to AlphaFold2 models; Fig. S2 to ESMFold models).

In Fig. 3, the quality scores (average pLDDT) are computed with each QA tool for the two different sets of protein models, distinguished as similar or dissimilar. Points are colour-coded according to the pairwise TM-score for each protein, with blue points indicating models with TM-score $\geq$ 0.6 (similar), and orange points indicating models with TM-score $<$ 0.6 (dissimilar).

For the sake of simplification, we introduce a threshold of 0.1 average pLDDT for considering values predicted by the QA tools to be different. Our analysis highlights that DeepAccNet (Fig. 3A) scores AlphaFold2 models higher than those of ESMFold for 91 % of the proteins; QMEANDisCo (Fig. 3B) and QATEN (Fig. 3C) assign higher scores to AlphaFold2 models in 45 % and 38 % of the proteins, respectively. ESMFold models are scored better than AlphaFold2 ones by QMEANDisCo and QATEN in 55 % and 62 % of the cases, respectively (Table 1).

For the models above the 0.6 TM-score value, DeepAccNet and QMEANDisCo consistently attribute better performances to AlphaFold2 models (96 % and 94 %, Fig. 3A and B, and Table 1, respectively). Below the 0.6 TM-score value, two out of the three QA tools, score ESMFold higher.

3.3.2. Comparison of solvent accessibility in models generated by AlphaFold2 and ESMFold

To further characterise our models, we also compute the Residue Solvent Accessibility (RSA) values (Fig. 4) and the coil content of Secondary Structure (SS) (Fig. 5) of each model, both derived from DSSP. Models are grouped as above, considering a TM-score of 0.6 as a discriminative threshold among similar and dissimilar models, colour-coded blue and orange, respectively. In Fig. 4, each point corresponds to a different protein with RSA values computed for its ESMFold model (x-axis) and its AlphaFold2 model (y-axis).

It appears that proteins for which the two models agree are consistently characterised by lower RSA values (considering also their distribution). For diverging models, RSA values are higher suggesting an increasing loss of compactness. Here, it is evident that AlphaFold2 models are overall less packed than those of ESMFold. In the set of diverging models, some 81 % of AlphaFold2 models have a mean RSA value higher than that of the ESMFold ones.

3.3.3. The content of coils in models generated by AlphaFold2 and ESMFold

The coil content of our PDB set is 40.3 %. In Fig. 5, the scatterplot of the data indicates that the increase in coil content in the models is mainly due to those obtained with AlphaFold2. Indeed, by comparing the two methods, we observe that the overall content of coil structure is higher for AlphaFold2 in 68.8 % of cases, to be contrasted with the 26.2 % for ESMFold. The coil content of similar and dissimilar models is 41.4 % and 59.1 %, respectively.

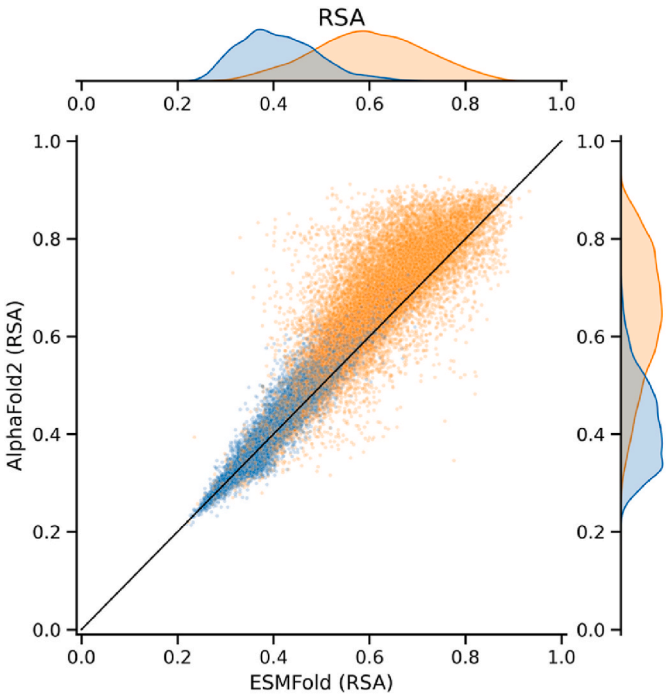


Fig. 4. Scatterplot showing one point for each protein in the dataset. On the x-axis, we report the mean RSA value for the ESMFold model, against the mean RSA value for the AlphaFold2 model on the y-axis. Each protein is represented with a blue point when the two models have TM-score $\geq$ 0.6, an orange one otherwise. The curves at the edges of the scatterplots are the distributions of the two classes as computed with kernel density estimation.

Table 2 indicates that AlphaFold2 consistently generates longer unstructured regions with respect to ESMFold. Indeed, although the total number of coil segments is similar between the two methods, AlphaFold2 coil segments reach lengths that are almost twice the length of the longer segments in ESMFold models, with a much higher number of segments longer than 300 residues.

4. Conclusions

AlphaFold2 models of proteins from different organisms, including humans, are available to the scientific community and have been proven, in CASP editions (CASP 14 and 15), to be of better quality than other predictors over a reduced set of proteins. ESMFold has been recently implemented (Lin et al., 2023) and, differently from AlphaFold2, relies on embedding representations of the proteins at hand. The different methodologies have different strengths and weaknesses and

Table 1
Statistics of QA scores on the set of models for proteins without PDB.

Dataset	Number proteins	QA method	Mean (std) score[a]	Number proteins diff $>$ 0.1[b]	Better AlphaFold2 models[c]	Better ESMFold models[c]
All	40,042	DeepAccNet	0.59 (0.22)	10,186	9266 (91 %)	920 (9 %)
		QMEANDisCo	0.53 (0.15)	5993	2708 (45 %)	3285 (55 %)
		QATEN	0.55 (0.12)	2490	935 (38 %)	1555 (62 %)
TM-score $\geq$ 0.6	16,602	DeepAccNet	0.77 (0.13)	5156	4973 (96 %)	183 (4 %)
		QMEANDisCo	0.64 (0.11)	962	907 (94 %)	55 (6 %)
		QATEN	0.64 (0.08)	406	201 (50 %)	205 (50 %)
TM-score $<$ 0.6	23,440	DeepAccNet	0.46 (0.17)	5030	4293 (85 %)	737 (15 %)
		QMEANDisCo	0.46 (0.13)	5031	1801 (36 %)	3230 (64 %)
		QATEN	0.49 (0.10)	2084	734 (35 %)	1350 (65 %)

[a] Mean and standard deviation of the average pLDDT scores computed by each QA method for both models of all the proteins.
[b] Number of proteins for which the absolute difference between QA score values attributed to AlphaFold2 and ESMFold is higher than 0.1 (outside dotted lines in Fig. 3).
[c] Number (and percentage) of proteins in which one model is better than the other considering again the set for which the pLDDT score difference is higher than 0.1.

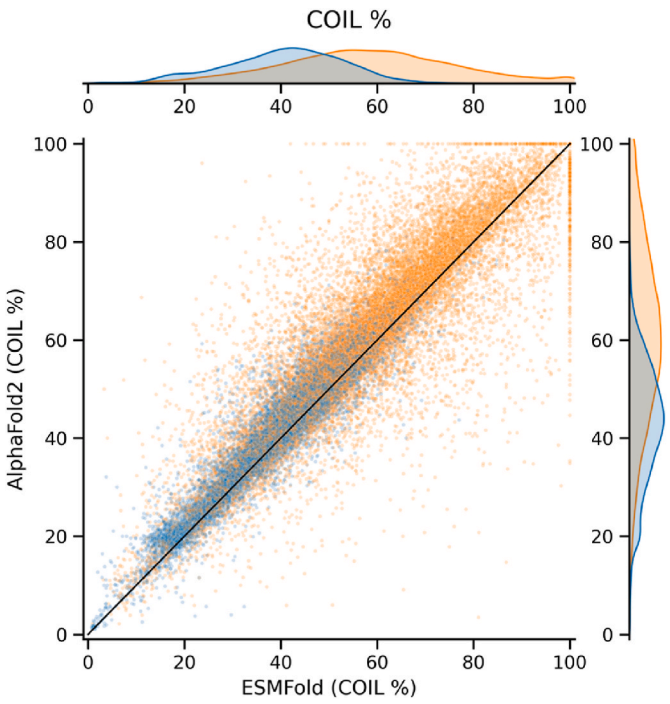


Fig. 5. Scatterplot showing one point for each protein in the dataset. On the x-axis, we report the percentage of residues in a coil state when observed in the ESMFold model, against the percentage of residues in a coiled state when observed in the AlphaFold2 model on the y-axis. Each protein is represented with a blue point when the two models have TM-score $\geq$ 0.6, an orange one otherwise. The curves at the edges of the scatterplots are the distributions of the two classes as computed with kernel density estimation.

Table 2
Statistics on the number and length distribution of coil segments in AlphaFold2 and ESMFold models.

	Model to model TM-score $<$ 0.6		Model to model TM-score $\geq$ 0.6	
	AlphaFold2	ESMFold	AlphaFold2	ESMFold
Coil segments (#)	446,024	470,141	500,328	499,429
Maximal length (# residues)	1489	755	935	596
Mean length (standard deviation)	13.2 (36.5)	12.0 (25.0)	6.2 (9.7)	6.0 (8.7)
Segments with length $\geq$100 (#)	10,683	7999	935	596
Segments with length $\geq$300 (#)	1400	314	7	3

= number of.

choosing the best available model for a protein of interest can improve downstream analysis. In the absence of experimental data, this is a complex task, but tools specialised to perform a Quality Assessment can help.

Recently, we released a public web server to access a database (Alpha&ESMhFolds (Manfredi et al., 2024)) that contains AlphaFold2 and ESMFold models for 42,942 human reference proteins. This resource can help researchers to visually superimpose and compare models generated by different tools for the same protein. Here, we perform the analysis of the paired models available from our database by computing scores generated by Quality Assessment methods: DeepAccNet, QMEANDisCo and QATEN. Furthermore, we evaluate the relative solvent accessibility and coil content of the paired models.

All the methods, including the ones predicting the protein structure from the sequence, compute the average predicted LDDT, as a final score for quality assessment. All the scoring values of the methods correlate, including the self-assessments performed by AlphaFold2 and ESMFold, supporting the notion that we can compare the evaluations. Nonetheless, each one reveals specific trends and we argue that their consensus is a safer measure of the confidence of each predicted model. In Fig. 6, two Venn diagrams contain the number of proteins in which the methods score higher AlphaFold2 (Fig. 6A) or ESMFold models (Fig. 6B). Among the 40,042 proteins of our database lacking a PDB structure, 63 % of them have a better AlphaFold2 model for at least two out of three QA methods, and 37 % of them have a better ESMFold model for at least two out of three QA methods. Overall, the three QA methods are in total agreement for 46 % of the proteins without PDB structure.

When AlphaFold2 and ESMFold generate similar models with high confidence, the first should routinely be preferred to the latter. Notably, 40 % of the proteins where the two models have a TM-score $\geq$ 0.6 have a good three-dimensional template available from other organisms, and AlphaFold2, better than ESMFold, integrates this information due to its implementation (Jumper et al., 2021; Lin et al., 2023). Conversely, on the 23,340 proteins in which AlphaFold2 and ESMFold diverge, ESMFold computes better models for half of the proteins. In this region, we highlight that ESMFold routinely computes more compact models, with coil structures that are less and much shorter than those of AlphaFold2. This may explain why QMEANDisCo and QATEN, contrary to DeepAccNet, compute higher scores for ESMFold models. Indeed, QMEANDisCo explicitly adopts RSA as an input feature and QATEN represents proteins as graphs that benefit from more compact structures. Summing up, our analysis supports the notion that AlphaFold2 models are of higher quality with respect to those of ESMFold when the two models are similar (TM-score higher than or equal to 0.6) (Manfredi et al., 2024); however, when the two models are dissimilar (TM-score lower than 0.6) nearly half of them have a better ESMFold model for at least two out of three QA methods. It is important to note that QA methods are machine learning-based tools trained on real structures and decoys (Hiranuma et al., 2021; Studer et al., 2020; Zhang et al., 2023). New experimental data will always be fundamental in order to increase our knowledge of protein structures. In the meantime, our results confirm that QA methods can qualify predicted models in the absence of structural information.

CRediT authorship contribution statement

Matteo Manfredi: Methodology, Software, Validation, Writing – review & editing. **Castrense Savojardo:** Conceptualization, Writing – review & editing. **Pier Luigi Martelli:** Conceptualization, Supervision, Writing – review & editing. **Rita Casadio:** Conceptualization, Supervision, Writing – review & editing.

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), "HEAL ITALIA" (Investment PNRR-M4C2-I1.3, Project PE_00000019, CUP J33C22002920006), and "National Centre for HPC, Big Data and Quantum Computing" (Investment PNRR-M4C2-I1.4, Project CN_00000013).

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.

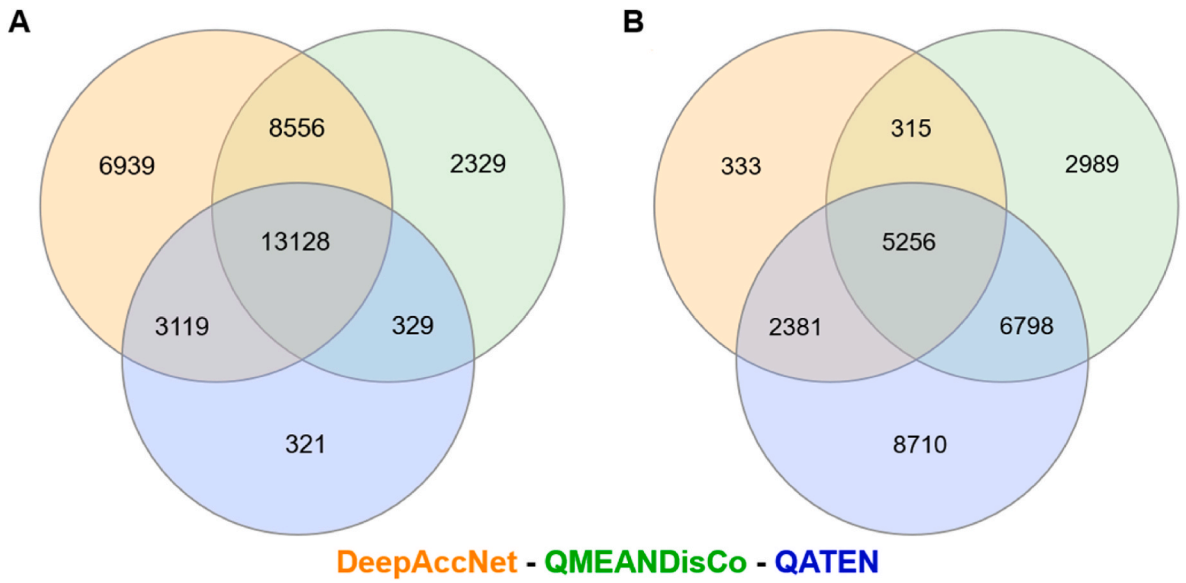


Fig. 6. Venn diagram showing the consensus of the three QA methods on the 40,042 proteins without a PDB structure. From left to right, A) we report the number of proteins for which the AlphaFold2 model receives a higher score than the ESMFold model; B) we report the number of proteins for which the ESMFold model receives a higher score than the AlphaFold2 model. In both diagrams, the orange set corresponds to DeepAccNet, the green one to QMEANDisCo, and the blue one to QATEN. Notably, the numbers do not sum to 40,042 as there are proteins for which the two predicted models receive an equal score from a QA method (15 proteins for DeepAccNet, 342 proteins for QMEANDisCo, 0 proteins for QATEN).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crstbi.2025.100167.

Data availability

The protein models analysed in the manuscript are freely available at https://alpha-esmhfolds.biocomp.unibo.it/

References

Benkert, P., Tosatto, S.C.E., Schomburg, D., 2008. QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71, 261–277.
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.
Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., Schwede, T., 2018. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins 86 (Suppl. 1), 387–398.
Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., Baker, D., 2021. Improved protein structure refinement guided by deep learning based accuracy estimation. Nat. Commun. 12, 1340.
Jing, X., Xu, J., 2021. Fast and effective protein model refinement using deep graph neural networks. Nat. Comput. Sci. 1, 462–469.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.
Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.
Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., Tramontano, A., 2016. Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. Proteins 84 (Suppl. 1), 349–369.
Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J., 2021. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins 89, 1607–1617.
Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J., 2023. Critical assessment of methods of protein structure prediction (CASP)-Round XV. Proteins 91, 1539–1549.
Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130.
Manfredi, M., Savojardo, C., Iardukhin, G., Salomoni, D., Costantini, A., Martelli, P.L., Casadio, R., 2024. Alpha&ESMhFolds: a web server for comparing AlphaFold2 and ESMFold models of the human reference proteome. J. Mol. Biol. 436, 168593.
Mariani, V., Biasini, M., Barbato, A., Schwede, T., 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722–2728.
McGuffin, L.J., Alharbi, S.M.A., 2024. ModFOLD9: a web server for independent estimates of 3D protein model quality. J. Mol. Biol. 436, 168531.
Ray, A., Lindahl, E., Wallner, B., 2012. Improved model quality assessment using ProQ2. BMC Bioinf. 13, 224.
Rost, B., Sander, C., 1994. Conservation and prediction of solvent accessibility in protein families. Proteins 20, 216–226.
Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, R., Haas, J., Schwede, T., 2020. QMEANDisCo-distance constraints applied on model quality estimation. Bioinformatics 36, 2647.
UniProt Consortium, 2023. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531.
Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., Elofsson, A., 2017. ProQ3D: improved model quality assessments using deep learning. Bioinformatics 33, 1578–1580.
Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.
Zemla, A., 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31, 3370–3374.
Zhang, Y., Skolnick, J., 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57, 702–710.
Zhang, P., Xia, C., Shen, H.-B., 2023. High-accuracy protein model quality assessment using attention graph neural networks. Briefings Bioinf. 24.