How AlphaFold shaped the structural coverage of the human transmembrane proteome

Márton A. Jambrich[1], Gabor E. Tusnady[1,2,*], Laszlo Dobson[1,3]
[1]Protein Bioinformatics Research Group, Institute of Enzymology, Research Centre for Natural Sciences, Magyar Tudósok körútja 2, H-1117 Budapest, Hungary
[2]Department of Bioinformatics, Semmelweis University, Tűzoltó u. 7, H-1094 Budapest, Hungary
[3]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany
*To whom correspondence should be addressed. Tel: +36 1 382 6709; Fax: +36 1 382 6295; Email: tusnady.gabor@ttk.hu

Abstract

AlphaFold2 (AF2) provides structures for every protein, opening up prospects for virtually every field in structural biology. However, transmembrane proteins pose a challenge for experimental scientists, resulting in a limited number of solved structures. Consequently, algorithms trained on this finite training set also face difficulties. To address this issue, we recently launched the TmAlphaFold database, where predicted AlphaFold2 structures are embedded into the membrane plane and a quality assessment is provided for each prediction using geometrical evaluation. In this paper, we analyze how AF2 has changed the structural coverage of membrane proteins compared to earlier years when only experimental structures were available, and high-throughput structure prediction was greatly limited. We also evaluate how AF2 can be used to search for (distant) homologs in highly diverse protein families. By combining quality assessment and homology search, we can pinpoint protein families where AF2 accuracy is limited, and experimental approaches are still desired.

Introduction

Transmembrane proteins (TMPs) are gatekeepers that regulate the movement of nutrients and drugs in and out of cells while also filtering or amplifying signals in neurotransmission and perception. After the Human Genome Project, it was determined that around 20-30% of the coded protein in the human genome encodes TMPs, which means that there are around 5-7 thousand TMPs in the human proteome [1]. The lipid bilayer surrounding TMPs exhibit them to three different phases and their proper embedding requires a multi-step folding process [2]. Despite the crucial roles that TMPs play in cellular processes, their structural characterization lags far behind that of globular proteins due to their natural dual environment [3], which makes their purification and crystallization tedious [4].

The challenging experimental conditions have given rise to various computational approaches, which range from topology prediction [5], which determines the exact location of transmembrane segments and the orientation of connecting loops relative to the membrane (e.g., cytosolic, extracellular), to 3D prediction of TMPs [6]. The latter has become a popular field, with approaches borrowed from the prediction of globular proteins and extended with

special conditions arising from the topology during contact map definition [7]. Despite tremendous effort, the accuracy of these topology and 3D predictions is limited, as the training set for tuning these methods is restricted by the number of experimentally solved structures. In the 2000's numerous structural genomic target selection projects started to pave the wave of membrane protein structure determination [8,9], with limited success. Notably, in the last 10 years, the situation has improved as cryo-electron microscopy (cryo-EM) has yielded hundreds of novel TMP structures.

Yet, the breakthrough seems to arrive from a computational tool: AlphaFold2 (AF2) [10], developed by the team at DeepMind, has revolutionized the field of structural biology by greatly improving the speed and accuracy of protein structure prediction. The software uses deep learning algorithms and a vast amount of genomic data to predict the 3D structure of a protein based solely on its amino acid sequence. It was quickly realized that AF2 can be utilized in many ways for structural biology research, including experimental model building, prediction of intrinsically disordered regions or assessing disease causing mutations [11]. The outstanding question however, how AF2 performs in a field of membrane proteins where the training data it could use was still highly limited.

Recently, we launched the TmAlphaFold database [12], which adds the possible localization of the lipid bilayer relative to the AF2 predicted helical transmembrane protein structures and evaluates the quality of each prediction. In this study, we analyze the human transmembrane proteome to identify bottlenecks where we encountered limitations with the AF2 structures. We made a quality assessment on the state of the human transmembrane proteome, we performed sequence and structure based clustering of TMPs and then highlighted proteins or protein families where there is still room for improvements.

Results

TmAlphaFold database is a good starting point to discriminate good and bad quality structures

The TmAlphaFold database [12] not only gives a possible localization of the lipid bilayer in predicted structures but also evaluates the quality of the predicted membrane embedded structures by applying filters, such as globular domains embedded in the membrane, conflicts with topology prediction, highlighting membrane regions folded outside the bilayer, and more (see Supplementary Material and TmalphaFold webpage for details). We constructed a non-redundant human TMP benchmark set consisting of proteins with I) solved 3D structures [3] and II) high-quality topology predictions from the Human Transmembrane Proteome (HTP) database [1] (Supplementary Table 1). First, we checked whether the different filters capture unique errors to rule out that they all detect the same problems but with a different approach (Figure 1A). Using the benchmark set, we calculated the correlation between the filters and noticed that most of the filters are diverse.

Next, we wanted to find out whether these filters really detect structures with errors. For proteins with solved 3D structures, we made a 2D projection to define their topography (location of transmembrane helices in the amino acid sequence). Then we compared the experimentally established or predicted topography to the topography we derived from the AF2 structures, after we modeled the location of the membrane bilayer using TmAlphaFold

structures. We defined the correct prediction for proteins where the topographies agreed (the number of transmembrane segments were equal and all helices were covered with at least 5 amino acids). We checked how different filters perform on correct and incorrect structures (Figure 1B). Around 80% of proteins pass each filter when the structure is correct, while for incorrect structures, this is always lower, ranging from 0-77%. TMAlphaFold uses a 5-level quality filter based on the filters, categorizing structures from failed to excellent. Proteins with excellent quality are indeed topographically correct in 97% of the cases, while in lower quality categories, the number of errors increases (Figure 1C). It is also visible that, considering the benchmark set, 40% of the proteins belong to the Excellent category, and below 10% of proteins have Poor or Failed categories.

For proteins where experimental 3D structure is available, we also calculated the Root Mean Square Deviation (RMSD) and TMScore using TMAlign [13]. When the experimental and predicted structure shows low RMSD (<1), almost 90% of the proteins have Good or Excellent quality, while the higher the RMSD, the lower is the proportion of Good and Excellent structures (Figure 1D). The same applies to TMScore (Figure 1E), where above the 0.5 threshold, 80% of the proteins have Good or Excellent quality, while for below the 0.5 threshold, this is only 59%.

AF2 also gives a probability value for each predicted residue, called the predicted Local Distance Difference Test (pLDDT): the distribution of these values for TM regions shows that, for higher quality predictions, AF2 is more certain (Figure 1F).

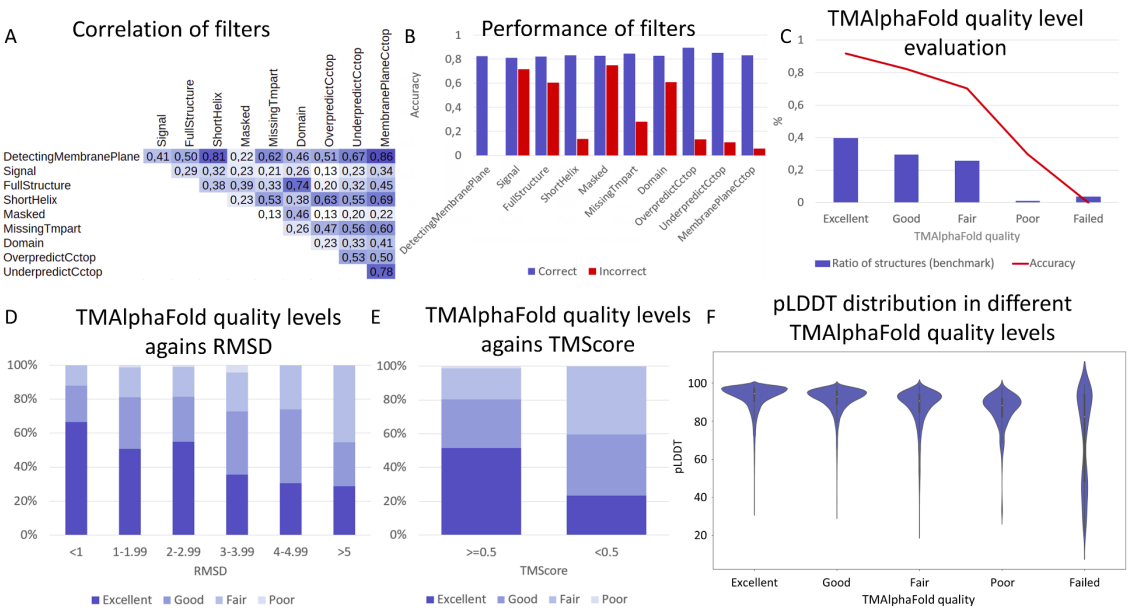

Figure 1: A) Correlation of different filters from TMAlphaFold on the benchmark set; B) Performance of different filters on correct (topographically correct) and incorrect structures in the benchmark set; C) Ratio of proteins and the measured accuracy (by checking topography) in different quality levels on the benchmark set; D) Distribution of TMAlphaFold quality levels in different RMSD intervals on the benchmark set; E) Distribution of TMAlphaFold quality levels in different TMScore intervals on the benchmark set; F) Distribution of pLDDT values in the transmembrane regions in different TMAlphaFold quality levels on the benchmark set.

How AlphaFold2 shaped the coverage of the human transmembrane proteome

Next, we extended our analysis to the entire human transmembrane proteome. Firstly, we assessed how the structural coverage of the transmembrane proteome changed after the release of AF2. To do this, we used HHBlits [14] to search for PDBTM structures where the deposited structure covered all transmembrane helices. For each year, we only incorporated PDB [15] structures that were available at that time. We categorized hits based on sequence identity and created a graph to visualize the data (see Figure 2A). In the early 2000s, the number of solved structures with high sequence identity (>80%) was below 100. Even when considering highly distant homologs (<20%), around 20% of the transmembrane proteome had a recognizable paralog. Notably, this also marked the upper limit of proteins where 3D structure prediction could be performed using homology-recognition or threading approaches. As more research groups accessed cryo-EM, the number of close (and distant) homologs increased, showing a promising trend. However, it would have taken decades to cover the entire transmembrane proteome with high-quality templates. If we add AF2 structures, using TmAlphaFold quality assessment, to this graph, the impact of AF2 becomes readily visible, with more than half of the transmembrane proteome belonging to the Good and Excellent category (see Figure 2A). It is interesting that the number of available experimentally solved structures in 2021 and the good/excellent quality AF2 structures in 2022 almost identical, however inspecting their overlap reveals that AF2 miss hundreds of cases where there is experimental result, but in turn provides good quality results for some that was not solved earlier (Supplementary Material, Supplementary Figure 1).

Although AF2 performs quite well, it is a supervised method, and this becomes evident when we check the quality of proteins where no homologous protein was available in the PDB and when there are 1-5 or more template structures already solved (see Figure 2B). The number of excellent structures is slightly higher when there are a few homologs, but AF2's performance is truly convincing when at least five homologs are available. AF2 was trained on X-ray structures, and therefore it performs better on proteins where there are available X-ray templates compared to cases when only NMR or cryo-EM homologs are available in the PDB (see Figure 2B).

The distribution of proteins in different quality categories on the benchmark set and the whole proteome are almost identical; therefore, we can assume that around 40% of the whole transmembrane proteome has 97% accuracy (considering topography), and on around 90% of the transmembrane proteome, it has 70% accuracy (see Figure 2C).

We also calculated the distribution of pLDDT values on (I) predicted transmembrane regions (see Figure 2D) and (II) within the modeled lipid bilayer according to TMAlphaFold (see Figure 2E). In both cases, the higher the structure quality, the higher the mean pLDDT value in membrane regions. In the case of the modeled lipid bilayer, on all quality levels except Excellent, the distribution is bimodal, likely reflecting the flexible protein segments violating the membrane. Since for the training of AF2, the lipid bilayer was not defined, disordered regions often cross it, resulting in incorrectly located domains.

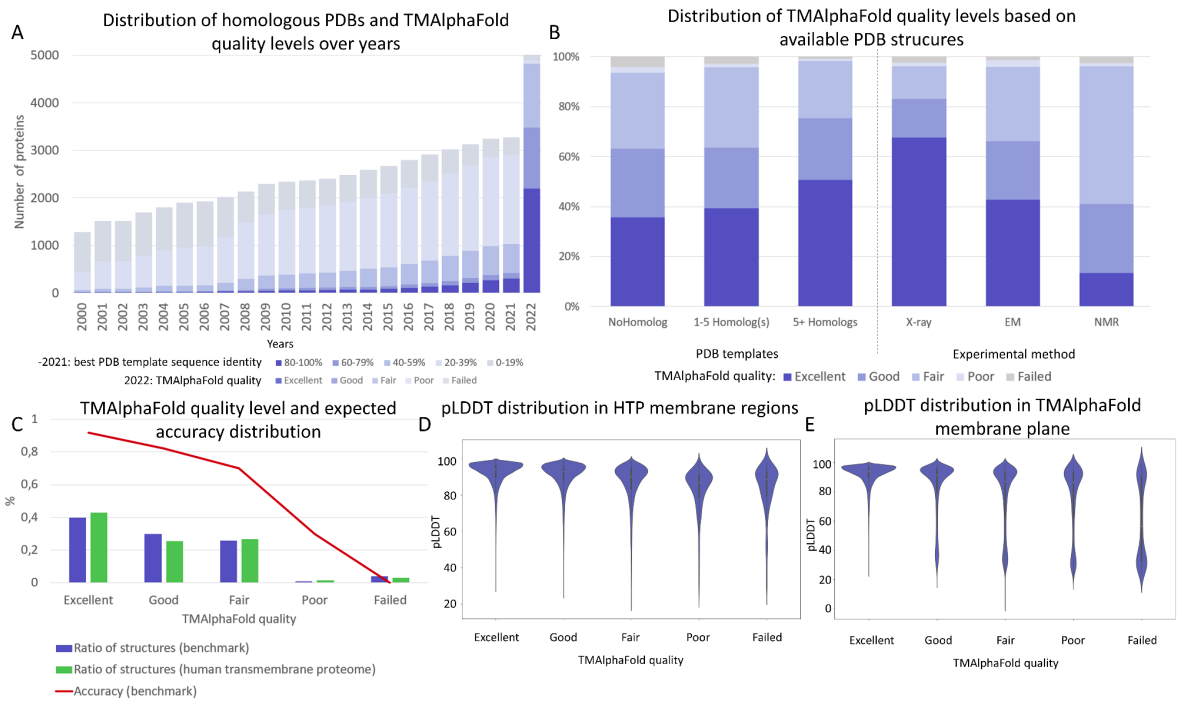

Figure 2: A) Coverage of PDB structures (from 2000-2021) according to their sequence identity and TMAlphaFold (2022) structures based on their quality level on the Human Transmembrane Proteome; B) Distribution of TMAlphaFold quality levels with and without homologous structures and based on experimental technique; C) Distribution the number of TMAlphaFold structures in different quality levels (benchmark set results are also displayed for comparison); D) Distribution of pLDDT values in the HTP predicted transmembrane *regions in different TMAlphaFold quality levels; E) Distribution of pLDDT values in the TMAlphaFold detected* transmembrane *regions in different TMAlphaFold quality levels.*

AlphaFold2 provides a fast and accurate searching and clustering compared to sequence based methods

Searching for distant homologs and clustering protein sets are commonly used tools to associate annotations or generate benchmark sets. In this study, we aimed to compare traditional homology searching tools with Foldseek clustering [16], based on AlphaFold2 structures. In line with this, we collected protein sets from various databases, including CATH [17], UniProt [18], Membranome [19], and Transporter Classification Database [20]. Next, we manually discarded proteins or merged groups to generate protein families where the number of transmembrane segments is the same within each group. This means that we have multiple groups for SLC proteins (with different numbers of transmembrane segments), but we also have distinct protein families with the same number of membrane regions (such as Aquaporins and a subset of ABC proteins). Our goal was to provide a dataset that poses a decent challenge for searching and clustering algorithms.

We first searched for homologs using HHBlits, BLAST [21], and Foldseek (using the full AF2 structures and also using only the TMAlphaFold predicted bilayer with +-5 Ångström). In both cases, we used two approaches: once all hits were accepted, and we also performed a search when only those hits were accepted that have the same number of transmembrane segments. For sequence-based approaches, we used the HTP database as a reference, while for AF2 structures, we used TMAlphaFold to define membrane regions. The sequence and structure-based searches yielded similar results (Table 1): in both cases, the best

sensitivity that could be achieved was around 0.88, and the best specificity was almost 1. Considering more universal metrics that take into account both positive and negative cases (Matthew's Correlation Coefficient: MCC and Balanced Accuracy: BACC), HHBlits and Foldseek provided the best results. We ran TMAlign to provide TMScores on paired structures and use the results to define a better cutoff, but this approach did not provide an advantage. We also used the union of the most sensitive sequence and structure based approach, and we achieved 0.934 sensitivity using this approach. Notably, this combination has the highest BACC and MCC too. Similarly, the section of the two most specific methods increased the specificity to 0.999, however this value was already high.

Table 1: Homology searching results using different approaches.)TP: True positive, FP: False positive, FN: False negative, TN: True negative, Sens: Sensitivity, Spec: Specificity, BACC: Balanced Accuracy, MCC: Matthew's Correlation Coefficient. seqID: sequence identity)

Approach	TP	FP	FN	TN	Sens	Spec	BACC	MCC	AUC/evalue	AUC/seqID	AUC/TMScore
HHBlits all	24456	4227	3424	73463	**0,877**	0,946	0,911	0,815	0,915	0,932	x
HHBlits Tms covered	22808	2456	5072	75234	0,818	0,968	0,893	0,813	0,894	0,905	x
BLAST all	21979	597	5901	77093	0,788	0,992	0,890	0,839	0,893	0,893	x
BLAST Tms covered	20039	60	7841	77630	0,719	**0,999**	0,859	0,806	0,859	0,859	x
Foldseek all	24406	3447	3474	74243	**0,875**	0,956	0,916	0,831	0,932	x	0,931
Foldseek Tms covered	21948	1522	5932	76168	0,787	0,980	0,884	0,814	0,891	x	0,891
Foldseek_mem all	21326	294	6554	77396	0,765	0,996	0,881	0,831	0,882	x	0,881
Foldseek_mem Tms covered	16884	165	10996	77525	0,606	**0,998**	0,802	0,723	0,802	x	0,802
Sensitive union	25818	4660	2062	73030	**0,934**	0,945	0,928	0,844	x	x	x
Specific section	15591	31	12289	77659	0,561	**0,999**	0,782	0,695	x	x	x

We also assessed how well different methods could find proteins belonging to the same families. Considering the number of pairs, the most sensitive approach (HHBlits) found most homologs in a group of protein families, such as Neurotransmitter ion-channel family, Calcium channels, Potassium channels, Aquaporins, Olfactory receptors, Rhodopsins, Phospholipid-transporting ATPases, but had lower performance on bitopic membrane proteins, ATP-binding cassette transporters (ABC), or Solute Carrier Family (SLC) members (Figure 3A, top line: ratios). The number of cross predictions between groups was higher within distinct ABC subfamilies (with different topology) or Olfactory receptors and Rhodopsins (which, however, belong to the superfamily of GPCR-s and share the same topology and high structural similarity).

Using the HHBlits results and DBScan [22,23] method, we also performed clustering (Figure 3B). While several families were merged (e.g., surprisingly bitopic and 4 transmembrane region proteins were clustered into the same group, also GPCRs and some ABC proteins with different topology), some families could be clearly clustered using this approach.

We also used the inbuilt clustering approach of Foldseek (Figure 3C) on the full structures: its performance was superior to sequential homology searching and DBScan clustering. Some protein families were split into two groups. Cross predictions mostly occurred for ABC proteins and GPCR family members. The results can be easily improved by adding topology filters that do not allow proteins with different topology into the same cluster.

Using Foldseek to search for homologs by only considering residues in and near the modeled lipid bilayer (Figure 3D) and then clustering the results with DBScan (Figure 3E)

provided the most specific results, but the number of homologs found was relatively low. It's worth noting that when we examined missing and found pairs, it became clear that the lower quality AF2 structures had a significant impact on the results (Figure 3F), as their structures could not be detected using structure-based searching algorithms.

A Most sensitive method (Hhblits) hits

		numTM:1		numTM:2	numTM: 4			numTM: 6			numTM: 7		numTM: 10	numTM: 11	numTM: 12		numTM: 15
		Single helix bin	Integrin	KCN_3	Eic/Glu me. domain	Neurotm ion-channel	CACN_2	ABC_2	Aquaporin	KCN_1	Rhodopsin	Olfactory	ATP_1	SLC_1	ABC_1	SLC_2	ABC_3
	Found (ratio)	0,00	0,58	0,71	0,33	1,00	1,00	0,67	1,00	1,00	0,88	0,98	1,00	0,36	0,71	0,49	0,48
	Single helix bin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
numTM:1	Integrin	0	348	0	0	0	0	0	0	0	0	0	0	0	0	0	0
numTM:2	KCN_3	0	0	216	0	0	0	0	0	0	0	0	0	0	0	0	0
	Eic/Glu me. Domain	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
	Neurotm ion-channel	0	0	0	0	42	0	0	0	0	0	0	0	0	0	0	0
numTM: 4	CACN_2	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0
	ABC_2	0	0	0	0	0	0	20	0	0	0	0	0	0	30	0	25
	Aquaporin	0	0	0	0	0	0	0	110	0	0	0	0	0	0	0	0
numTM: 6	KCN_1	0	0	0	0	0	0	0	0	240	0	0	0	0	0	0	0
	Rhodopsin	0	0	0	0	0	0	0	0	0	268	3614	0	0	0	0	0
numTM: 7	Olfactory	0	0	0	0	0	0	0	0	0	3614	45102	0	0	0	0	0
numTM: 10	ATP_1	0	0	0	0	0	0	0	0	0	0	0	380	0	0	0	0
numTM: 11	SLC_1	0	0	0	0	0	0	0	0	0	0	0	0	48	0	400	0
	ABC_1	0	0	0	0	0	0	30	0	0	0	0	0	0	30	0	30
numTM: 12	SLC_2	0	0	0	0	0	0	0	0	0	0	0	0	400	0	3668	0
numTM: 15	ABC_3	0	0	0	0	0	0	25	0	0	0	0	0	0	30	0	20

D Most specific (Foldseek, bilayer, overlapping TMs) hits

		numTM:1		numTM:2	numTM: 4			numTM: 6			numTM: 7		numTM: 10	numTM: 11	numTM: 12		numTM: 15
		Single helix bin	Integrin	KCN_3	Eic/Glu me. domain	Neurotm ion-channel	CACN_2	ABC_2	Aquaporin	KCN_1	Rhodopsin	Olfactory	ATP_1	SLC_1	ABC_1	SLC_2	ABC_3
	Found (ratio)	0,00	0,00	0,00	0,33	0,48	0,33	0,00	0,58	0,78	0,44	0,70	0,15	0,11	0,24	0,11	0,14
	Single helix bin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
numTM:1	Integrin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
numTM:2	KCN_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Eic/Glu me. Domain	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
	Neurotm ion-channel	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
numTM: 4	CACN_2	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
	ABC_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Aquaporin	0	0	0	0	0	0	0	64	0	0	0	0	0	0	0	0
numTM: 6	KCN_1	0	0	0	0	0	0	0	0	188	0	0	0	0	0	0	0
	Rhodopsin	0	0	0	0	0	0	0	0	0	134	148	0	0	0	0	0
numTM: 7	Olfactory	0	0	0	0	0	0	0	0	0	148	32406	0	0	0	0	0
numTM: 10	ATP_1	0	0	0	0	0	0	0	0	0	0	0	56	0	0	0	0
numTM: 11	SLC_1	0	0	0	0	0	0	0	0	0	0	0	0	14	0	17	0
	ABC_1	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
numTM: 12	SLC_2	0	0	0	0	0	0	0	0	0	0	0	0	17	0	858	0
numTM: 15	ABC_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6

B DBScan clustering based on the most sensitive method

	numTM:1		numTM:2	numTM: 4			numTM: 6			numTM: 7		numTM: 10	numTM: 11	numTM: 12		numTM: 15
	Single helix bin	Integrin	KCN_3	Eic/Glu me. domain	Neurotm ion-channel	CACN_2	ABC_2	Aquaporin	KCN_1	Rhodopsin	Olfactory	ATP_1	SLC_1	ABC_1	SLC_2	ABC_3
Cluster 1	2	25	3	3	7	6	1	0	0	1	2	0	0	1	7	2
Cluster 2	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 3	0	0	0	0	0	0	5	0	0	0	0	0	0	6	0	5
Cluster 4	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0
Cluster 5	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
Cluster 6	0	0	0	0	0	0	0	0	0	17	212	0	0	0	0	0
Cluster 7	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
Cluster 8	0	0	0	0	0	0	0	0	0	0	0	0	3	0	57	0
Cluster 9	0	0	0	0	0	0	0	0	0	0	0	0	6	0	21	0
Cluster 10	0	0	0	0	0	0	0	0	0	0	1	0	3	0	2	0

E DBScan clustering based on the most specific method

	numTM:1		numTM:2	numTM: 4			numTM: 6			numTM: 7		numTM: 10	numTM: 11	numTM: 12		numTM: 15
	Single helix bin	Integrin	KCN_3	Eic/Glu me. domain	Neurotm ion-channel	CACN_2	ABC_2	Aquaporin	KCN_1	Rhodopsin	Olfactory	ATP_1	SLC_1	ABC_1	SLC_2	ABC_3
Cluster 1	2	25	18	3	7	6	6	11	16	17	14	20	12	7	87	7
Cluster 2	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0
Cluster 3	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0
Cluster 4	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
Cluster 5	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0
Cluster 6	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Cluster 7	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
Cluster 8	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
Cluster 9	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
Cluster 10	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
Cluster 11	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
Cluster 12	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Cluster 13	0	0	0	0	0	0	0	0	0	1	76	0	0	0	0	0

C Best Foldseek clustering result

	numTM:1		numTM:2	numTM: 4			numTM: 6			numTM: 7		numTM: 10	numTM: 11	numTM: 12		numTM: 15
	Single helix bin	Integrin	KCN_3	Eic/Glu me. domain	Neurotm ion-channel	CACN_2	ABC_2	Aquaporin	KCN_1	Rhodopsin	Olfactory	ATP_1	SLC_1	ABC_1	SLC_2	ABC_3
Cluster 1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
Cluster 2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 3	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 4	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 5	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 6	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 7	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
Cluster 8	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0
Cluster 9	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
Cluster 10	0	0	0	0	0	0	4	0	0	0	0	0	0	6	0	5
Cluster 11	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0
Cluster 12	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
Cluster 13	0	3	0	0	0	0	2	0	0	18	215	0	0	1	1	2
Cluster 14	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
Cluster 15	0	0	0	0	0	0	0	0	0	0	0	0	6	0	59	0
Cluster 16	0	0	0	0	0	0	0	0	0	0	0	0	6	0	21	0
Cluster 17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0

F Ratio of TMAlphaFold quality levels in missed/found homologs

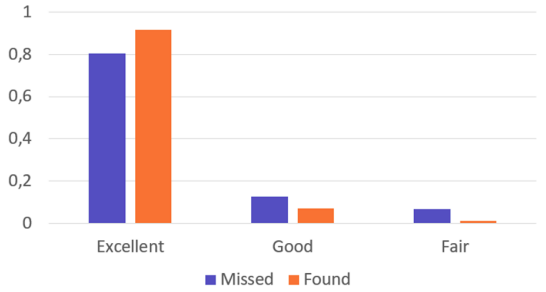

Figure 3: A) Top: ratio of proteins found of different protein families using HHBlits, below: number of pairs detected between and within protein families; B) Number of proteins in different clusters categorized by protein families, using HHBlits and DBscan. C) Number of proteins in different clusters categorized by protein families, using Foldseek clustering. D) Top: ratio of proteins found of different protein families using Foldseek, below: number of pairs detected between and within protein families; E) Number of proteins in different clusters categorized by protein families, using Foldseek and DBscan. F) Ratio of different TMAlphaFold quality structures in found and missing hits.

Areas where AlphaFold2 has limited accuracy

We also searched for protein groups for which AF2 provides confusing results and where there is no or only distant experimental PDB structure available. Foldseek is a good choice for defining clusters, but when we looked for low-quality structures, grouping them based on sequence search provided better results. We transformed the TMAlphaFold quality level to a scale of 1-5 (from failed to excellent) and searched for proteins with an average score below 4 (good). To avoid counting the same protein more than once, we used a greedy approach, discarding each selected protein from the list to avoid listing the same groups multiple times.

We noticed that many of these proteins have only low reliability topology predictions, so we did not expect the query and hit proteins to have overlapping transmembrane regions. This approach yielded 1600 proteins in 104 groups, where the largest group contained 598 proteins, and 24 groups included only 1 protein. Notably, there are also 107 proteins in this list that are FDA approved drug targets (Supplementary Material, Supplementary Figure 2). Nevertheless, from the 104 groups, we selected those where the number of good or excellent structures within the group was below 25%.

The first group we examined consisted of 15 members of the Spermatogenesis-associated proteins, which are expressed in spermatocytes and the retina, and some of its members are associated with retinal disease (Figure 4, top). HTP predicts them as bitopic membrane proteins, but the reliability of the predictions is always below 85%. TMDET [24]using the AlphaFold2 structures detects 1-5 transmembrane segments, and all of the structures have fair quality. Inspecting the structures reveals that some helical segments are present, but low pLDDT segments connecting them often cross the membrane and place the helical segments randomly. Low pLDDT values indicate flexible disordered regions, and in this case, MemDis [25] confirms that these proteins are highly disordered.

The second group we examined consisted of 13 endogenous retrovirus proteins that mediate receptor recognition and membrane fusion during early infection (Figure 4, middle). HTP largely predicts this group as a 2 transmembrane region protein, some of the members with high reliability values. For AF2, we observed a similar case as for Spermatogenesis-associated proteins, where flexible regions (with low pLDDT values) connect helical segments, and they randomly cross the bilayer, resulting in a different number of transmembrane segments among the structures. Notably, in this case, MemDis only predicts the N-terminal and a few other shorter segments as disordered, in contrast to pLDDT based disorder prediction (to be more precise, AF2 still predicts secondary structure for the majority of the protein, however helical elements mostly belong to the low confidence (<70 pLDDT) category).

The third group we examined consisted of reticulons, which inhibit amyloid precursor protein processing and play a role in trafficking of SLC family members (Figure 4, bottom). The number of predicted transmembrane segments varies from 2-4 according to HTP (with low reliability), while TmAlphaFold structures consist of 2-6 segments. Reticulons are also highly disordered, apart from their C-terminal membrane domain, according to MemDis. Notably, in this case, AF2 likely predicts actual transmembrane segments outside the bilayer in several group members.

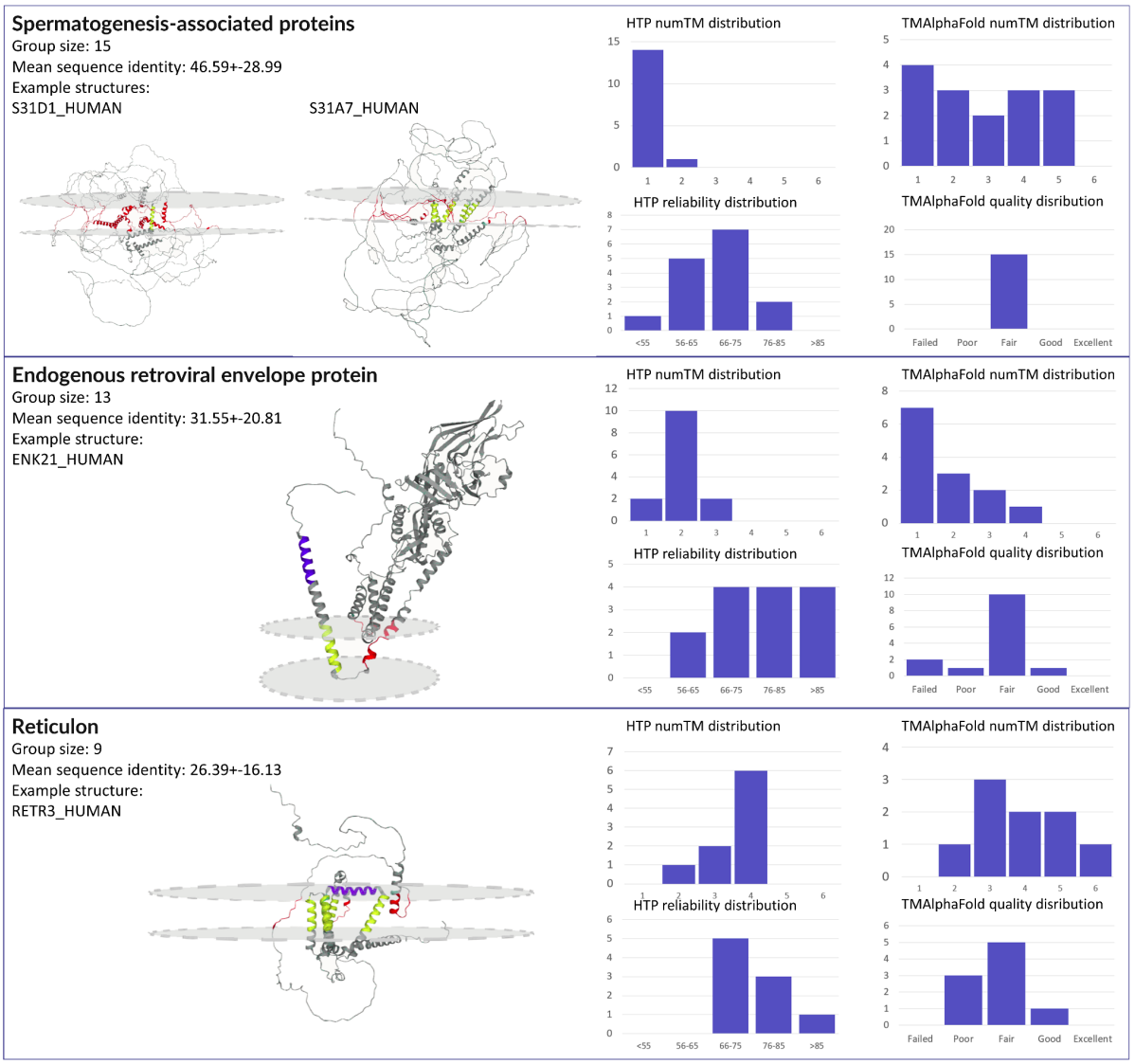

Figure 4: Example structures, transmembrane helix *distribution (based on HTP and TMAlphaFold), HTP reliability levels and TMAlphaFold quality levels in bad quality groups. Top: Spermatogenesis-associated proteins; Middle: Endogenous retrovirus proteins; Bottom: Reticulons*

Conclusion

In this study, we evaluated the quality of alpha-helical membrane protein structures predicted by AlphaFold2. Although AF2 has greatly contributed to the coverage of the human transmembrane proteome, there are still protein groups where its performance seems to be lower. Scientists often analyze their protein of interest or research focus, where AF2 might perform better (as these proteins are usually already under investigation for decades), for example in the case of ABC transporters [26]. However, it is interesting to note that around 30% of the proteome still lacks structures with convincing quality. One area where AF2 definitely has limitations is multimeric proteins, where the correct fold highly depends on the interaction. Although AF2 was quickly tailored by the scientific community to accept multiple protein sequences [27], this is not yet reflected in the AlphaFold Protein Structure database [28].

Defining protein clusters is also a challenging area, as the sequence identity greatly varies between different homology groups. For generating non-redundant datasets, highly sensitive methods are better, as they can more reliably ensure that two homologous proteins will not be placed in the same group and cause a bias in method development. For modeling or analyzing function by taking conservation into account, high sensitivity is only useful until not too many false positive hits are present that provide noise. The search mode also has to be optimized for global and local searching purposes (e.g. we can find the transmembrane domain of ABC transporters, or we can look for subfamilies with identical topologies). Unfortunately, it seems like the membrane fold predicted by AlphaFold alone is not sufficient for structural homology search or clustering and the globular domains in the proteins aided the search.

It is also visible that AF2 currently heavily relies on PDB, as the quality of predicted structures drastically dropped when we considered proteins where no template structure could be found. The most frequent problems we found are caused by flexible regions, which incorrectly cross the lipid bilayer, often placing globular domains and other fragments into the membrane and violating the topology of the protein. Another frequent problem is when AF2 tries to fold helical regions into the membrane bilayer, even if they do not belong there.

Methods

Performance of AlphaFold on the Human Transmembrane Proteome

First we prepared a dataset of human proteins where reasonably high-quality structural information was available from PDBTM structures [3] or topology predictions (reliability>95% in HTP [1]). First we used BLAST with e-value 10^{-5}, 40% sequence identity to collect homologous structures for proteins deposited into the HTP database. We only accepted hits where all transmembrane segments were covered. We expanded this set with proteins that have high reliability (above>95%) topology prediction in the HTP. This dataset was then filtered for redundancy using CD-HIT [29], by decrementing identity (90, 70, 50 ,40) and word length (5,4,3,2). The list of collected proteins is available in Supplementary Table 1.

We used TMAlign to compare predicted and experimentally solved structures (Supplementary Table 2).

Next we compared the predicted topology from HTP and the projected topology from AF2 structures, as the membrane bilayer was defined by TMAlphaFold [12]. Correct prediction means that all transmembrane segments overlap with at least 5 residues. TMAlphaFold filter results are also visible here. (Supplementary Table 3).

Coverage of structure on Human Transmembrane Proteins

We downloaded the full UniProt sequence of PDBTM alpha-helical structures using SIFTS [30]. Using these sequences we searched for homologous hits from HTP using HHBlits, where the mapped PDBTM membrane regions have overlap with HTP sequence in the alignment - however as some topology predictions might not be accurate, the predicted transmembrane segments from HTP were not taken into consideration (Supplementary Table 4).

TMAlphaFold quality levels (Supplementary Table 5) and the mapped structures were used to produce Figure 2A (data: Supplementary Table 6).

Searching for homologous proteins

We used the UniProt[18], CATH [17], Membranome [19] and the Transporter Classification Database [20] to search for protein families (Supplementary Table 7). We manually processed results to construct a dataset, discarded proteins where there is little or no annotation available, or merged groups if two databases had overlapping results (Supplementary Table 8).

We performed multiple homology search strategies (Supplementary Table 9): BLAST, HHBlits, Foldseek (full AF2 structures and only residues in and around the bilayer detected by TMAlphaFold). We also performed the same search, however we only accepted the result if the topology of the query and the hit protein matched (i.e. all transmembrane regions overlapped with at least 5 residues).

We used multiple clustering techniques too (Supplementary Table 10): as protein sequence identity and structural similarity highly varies between families, we used DBScan [22] to cluster the results (as it is a density-based clustering method) using the similarity matrix provided by the searching tools. We iterated through different epsilon values and selected the one where the number of clusters was between 10 and 20 (considering 16 original protein families), and selected the one where the most families had 85% of their members in a separate cluster. We also used Foldseek clustering on full and partial structures (with residues in and around the bilayer), using 0.2, 0.4, 0.6, 0.8 thresholds. The selection criteria was the same as for DBScan.

Searching for groups with low quality AlphaFold structures

We used HHBlits to search for homologous proteins, using the settings that resulted in the most sensitive search. We searched for proteins where the average TMAlphaFold quality was below good. To avoid listing the same proteins multiple times, we used a greedy algorithm that did not allow us to select the same protein multiple times (Supplementary Table 11). We also counted the number of good/excellent quality structures, as well as distant PDB relatives within the groups (Supplementary Table 12), and in the results we showed the result of 3 most populated clusters, where the ratio of solved structures or good/excellent predictions were below 25%. The list of FDA approved drug targets were obtained from Human Protein Atlas (Supplementary Table 12).

Acknowledgment

Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund [K132522].

References

1. Dobson L, Reményi I, Tusnády GE. The human transmembrane proteome. Biol Direct. 2015;10: 31.

2. Bowie JU. Solving the membrane protein folding problem. Nature. 2005;438: 581–589.

3. Kozma D, Simon I, Tusnády GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res. 2013;41: D524–9.

4. Varga JK, Tusnády GE. The TMCrys server for supporting crystallization of

transmembrane proteins. Bioinformatics. 2019;35: 4203–4204.

5. Dobson L, Reményi I, Tusnády GE. CCTOP: a Consensus Constrained TOPology prediction web server. Nucleic Acids Res. 2015;43: W408–12.

6. Kozma D, Tusnády GE. TMFoldRec: a statistical potential-based transmembrane protein fold recognition tool. BMC Bioinformatics. 2015;16: 201.

7. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012;149: 1607–1621.

8. Punta M, Love J, Handelman S, Hunt JF, Shapiro L, Hendrickson WA, et al. Structural genomics target selection for the New York consortium on membrane protein structure. J Struct Funct Genomics. 2009;10. doi:10.1007/s10969-009-9071-1

9. Varga J, Dobson L, Reményi I, Tusnády GE. TSTMP: target selection for structural genomics of human transmembrane proteins. Nucleic Acids Res. 2017;45: D325–D330.

10. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

11. Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, et al. A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol. 2022;29: 1056–1067.

12. Dobson L, Szekeres LI, Gerdán C, Langó T, Zeke A, Tusnády GE. TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. Nucleic Acids Res. 2023;51: D517–D522.

13. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33: 2302–2309.

14. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9: 173–175.

15. Bittrich S, Bhikadiya C, Bi C, Chao H, Duarte JM, Dutta S, et al. RCSB Protein Data Bank: Efficient Searching and Simultaneous Access to One Million Computed Structure Models Alongside the PDB Structures Enabled by Architectural Advances. J Mol Biol. 2023; 167994.

16. Kim H, Mirdita M, Steinegger M. Foldcomp: a library and format for compressing and indexing large protein structure sets. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad153

17. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2021;49: D266–D273.

18. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51: D523–D531.

19. Lomize AL, Hage JM, Pogozheva ID. Membranome 2.0: database for proteome-wide profiling of bitopic proteins and their dimers. Bioinformatics. 2018;34: 1061–1062.

20. Saier MH, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, et al. The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 2021;49: D461–D467.

21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410.

22. SanderJörg. DBSCAN Revisited, Revisited. ACM Transactions on Database Systems (TODS). 2017 [cited 14 Apr 2023]. doi:10.1145/3068335

23. VaroquauxGaël, DuchesnayÉdouard. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011 [cited 14 Apr 2023]. doi:10.5555/1953048.2078195

24. Tusnády GE, Dosztányi Z, Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics. 2005;21: 1276–1277.

25. Dobson L, Tusnády GE. MemDis: Predicting Disordered Regions in Transmembrane Proteins. Int J Mol Sci. 2021;22. doi:10.3390/ijms222212270

26. Hegedűs T, Geisler M, Lukács GL, Farkas B. Ins and outs of AlphaFold2 transmembrane protein structure predictions. Cell Mol Life Sci. 2022;79: 73.

27. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2022. p. 2021.10.04.463034. doi:10.1101/2021.10.04.463034

28. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50: D439–D444.

29. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.

30. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res. 2019;47: D482–D489.