Computational and Structural Biotechnology Journal 23 (2024) 3730–3740

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

Prediction of conformational states in a coronavirus channel using Alphafold-2 and DeepMSA2: Strengths and limitations

Jaume Torres [*], Konstantin Pervushin, Wahyu Surya

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore

ARTICLE INFO

Keywords:
Coronavirus
Envelope protein
SARS
IBV
AlphaFold-2
Ion channel
Deep MSA

ABSTRACT

The envelope (E) protein is present in all coronavirus genera. This protein can form pentameric oligomers with ion channel activity which have been proposed as a possible therapeutic target. However, high resolution structures of E channels are limited to those of the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), responsible for the recent COVID-19 pandemic. In the present work, we used Alphafold-2 (AF2), in ColabFold without templates, to predict the transmembrane domain (TMD) structure of six E-channels representative of genera alpha-, beta- and gamma-coronaviruses in the *Coronaviridae* family. High-confidence models were produced in all cases when combining multiple sequence alignments (MSAs) obtained from DeepMSA2. Overall, AF2 predicted at least two possible orientations of the α-helices in E-TMD channels: one where a conserved polar residue (Asn-15 in the SARS sequence) is oriented towards the center of the channel, 'polar-in', and one where this residue is in an interhelical orientation 'polar-inter'. For the SARS models, the comparison with the two experimental models 'closed' (PDB: 7K3G) and 'open' (PDB: 8SUZ) is described, and suggests a ~60° α-helix rotation mechanism involving either the full TMD or only its N-terminal half, to allow the passage of ions. While the results obtained are not identical to the two high resolution models available, they suggest various conformational states with striking similarities to those models. We believe these results can be further optimized by means of MSA subsampling, and guide future high resolution structural studies in these and other viral channels.

1. Introduction

Coronaviruses (CoVs) have the largest genomes among vertebrate animal RNA viruses. CoVs belong to the order nidovirales and the family *coronaviridae* [1] and are distributed into four genera: alpha-, beta-, gamma- and delta-CoV. CoVs cause a variety of diseases, e.g., respiratory, gastrointestinal and neurological, in mammalian and avian species. Of the seven human CoVs (hCoVs) identified, four are endemic and cause ~20 % of cases of common cold [2], whereas three beta-CoV are highly pathogenic: Severe Acute Respiratory Syndrome-CoV (SARS-CoV), Middle East Respiratory Syndrome-CoV (MERS-CoV) [3–5] and SARS-CoV-2 [6], appearing in 2003, 2012 and 2019, respectively. CoVs also cause severe and fatal diseases in animals such as alpha-CoV swine transmissible gastroenteritis virus (TGEV), feline CoV (FCoV) and gamma-CoV avian infectious bronchitis virus (IBV)[7]. TGEV and IBV are recognised as pathogens of international importance in the World Organization for Animal Health (WOAH) list of diseases [8].

CoVs are enveloped, non-segmented, single-stranded positive-sense RNA viruses with genomes 25–32 kb in length. They have six conserved open reading frames (ORF). The first two-thirds of the genome contains ORF1a and ORF1b, encoding the replicase/transcriptase proteins [1]. These are synthesized as two large polyproteins, pp1a and pp1ab, which are processed by viral proteases to generate 15 or 16 nsps in CoVs, which form the membrane-bound replication and transcription complex [9]. The remaining one-third of the genome encodes mainly four structural proteins: spike (S), envelope (E), membrane (M), nucleocapsid (N), and species-specific accessory proteins [10].

CoV envelope E proteins are about 100 residues long (e.g., 75 in SARS-CoV-2 and 109 in IBV), have a single TM domain and form oligomeric channels [11–25] that are potential antiviral drug targets [26]. In SARS, E protein localizes to the endoplasmic reticulum–Golgi intermediate compartment (ERGIC) of infected cells [23]. In SARS-2 and SARS, the single transmembrane domain (TMD) has identical sequences, and the N- and C-terminal tails are exposed luminally and cytoplasmically, respectively. Other E proteins have a similar topology [23,27] although in TGEV (group alpha) the topology is the opposite in infected cells [28–30].

In the better-studied SARS-CoV E protein (SARS E hereafter),

[*] Corresponding author.
E-mail address: jtorres@ntu.edu.sg (J. Torres).

https://doi.org/10.1016/j.csbj.2024.10.021
Received 15 June 2024; Received in revised form 1 October 2024; Accepted 15 October 2024
Available online 16 October 2024

channel activity is a virulence factor, as shown by the effect of channel-inactivating mutations, e.g., N15A, on a mouse-adapted SARS virus, which increased survival and reduced lung edema and proinflammatory cytokine levels [17]. Enhanced membrane permeability has been observed in bacterial and mammalian cells, or in in vitro systems, for MERS E [31], MHV E [32], SARS-CoV E [33], or human coronavirus 229E and IBV [12,13].

The SARS E-TMD forms pentameric oligomers in perfluorooctanoic acid (PFO) gels [34] and in ERGIC-like lipid membranes, as shown using ^{19}F spin diffusion solid-state NMR [35], and also in C14-betaine detergent [14,36–38]. More recently, the mass difference between lipidic nanodiscs reconstituted with full-length (FL) SARS-2 E protein and empty ones was reported to be consistent with an E protein pentameric oligomer [39].

The structure of the TMD channel of SARS-2 E protein (SARS E-TMD) was first obtained at high resolution in lipid membranes using solid-state NMR (PDB: 73KG) [40]. In this model, the channel lumen was reported to be dehydrated and described as being in a 'closed' conformation, with a radius of 2 Å in the narrowest section. The orientation of a pore-facing asparagine (Asn-15) was in agreement with previous structural models obtained in detergent and from lower-resolution techniques [21,41,42] as well as with the electrophysiological findings that mutation N15A completely abolishes channel activity [11,19]. Another structure for the SARS E-TMD channel was obtained later at a relatively low pH and high calcium concentration (PDB: 8SUZ), which was supposed to mimic the ERGIC and lysosomal environment experienced by the E protein in the cell. The data was consistent with a more 'open' conformation of the channel [43,44], with Asn-15 facing the lipid phase. However, we note that black lipid membrane (BLM) measurements show that no acidification is required for the channel of both SARS-2 E and SARS E proteins to be functional [38]. In the closed conformation, Phe-20 and Phe-26 interacted with residues Leu-19 and Leu-27, respectively, of neighbouring helices, whereas in the open conformation, the three Phe residues in ETM were oriented towards the lipid phase.

Since AF2 has been used successfully to predict the structure of other membrane proteins [45], given the lack of structural information for other CoV E proteins in other coronavirus groups, in the present paper we have used AlphaFold-2 (AF2) to predict the structure of the coronavirus E channel. In preliminary runs to predict oligomers, we observed that only the TMD showed relatively high confidence scores (not shown), therefore we focused on this part of the protein only. We used the TMD sequences equivalent to E7-K38 (32 residues) in the SARS E protein sequence, with the same stretch and sequence length for the TMD of other CoV E proteins (see Fig. 1). Prediction was helped by extracting large multiple sequence alignments (MSAs) using DeepMSA2, and combining them before assuming that in all cases the channels share a common backbone structure. Following this, the minimum set of sequences necessary to produce reliable models was obtained. The results suggest that in most cases at least two conformational states are predicted which seem to be common to all channels tested. Comparison with the two SARS high resolution models available is discused.

FCoV	7-**DD**HGMVVSVFFWLLLIIILILFSIALLNVI**K**L-38
TGEV	11-**DD**NGMVISIIFWFLLIIILILLSIALLNII**K**L-42
MERS	7-**ER**IGLFIVNFFIFTVVCAITLLVCMAFLTAT**R**-38
SARS	7-**EE**TGTLIVNSVLLFLAFMVFLLVTLAILTAL**R**-38
MHV	7-T**D**TVWYVGQIIFIFAVCLMVTIIVVAFLASI**K**-38
IBV	10-**EE**NGSFLTALYIIVGFLALYLLG**R**ALQAFVQA-41

Fig. 1. Representative CoV E TMD sequences used for AF2 prediction. Acidic residues (red), basic residues (blue) and a conserved polar residue (highlighted yellow) are indicated, to visualize the similarity between the sequences: E proteins tend to have negatively charged residues at the N-terminal side, and basic residues on the C-terminal side at juxtamembrane positions. It is also noted that one of the two mutations that completely inactivate the channel is Asn-15 in SARS, and a polar residue at this or equivalent position is conserved in most E proteins. This residue has been used as indicator to preliminarily visualise the orientation of the bundles (see next sections).

2. Materials and methods

2.1. Sequences of TMDs used

For the AF2 predictions, we used representative TMDs for the genus alpha, species alphacoronavirus 1 (FCoV and TGEV), three TMDs from the genus beta (species MHV, MERS and SARS) and one from the genus gamma (species IBV) (Fig. 1). From the alpha coronavirus genus, we used TGEV E (82 residues, ABG89336.1) [46] and FCoV E (82 residues, ASU62503.1). From beta-coronaviruses, we used the subgenus Sarbecovirus SARS-CoV E (76 residues, AAP51230.1[47]), which shares the same TMD with SARS-CoV-2 E (75 residues), and subgenus Merbecovirus MERS-CoV E, 82 residues, YP_007188584.1)[48] and MHV E (83 residues, AAC36596.1)[49]. Lastly, from the gamma-coronaviruses, we used IBV E (109 residues, AAO33465.1)[50]. These same E sequences were used as a query for DeepMSA2 [51], before extracting only the residues encompassing the TMD, equivalent to 7–38 in SARS-CoV E (Fig. 1).

2.2. Structure prediction

Structure prediction of pentameric channels encompassing the TMD was obtained in both a local AF2 installation and a ColabFold notebook. However, since no significant differences were observed between the results obtained with the two methods, we tested the effect of custom MSAs only on ColabFold.

2.2.1. Local AF2 installation

The TMD channel structure using the sequences in Fig. 1 was predicted with a local installation of AF2 (commit 7c9114c, 10 August 2023) [52] using the multimer mode [53] with full dataset and 5 seeds, selecting the top-ranked model.

2.2.2. ColabFold notebook

ColabFold (ColabFold v1.5.5: AlphaFold2 [52,54]) used the parameters, unless otherwise specified: no templates, 6 recycles (forced to complete 6 with 'recycle_early_stop_tolerance = 0') and 4 seeds (5 models each), which resulted in a total of 20 models for each prediction. The last model after each 6 recycles was used. For each prediction run, three scores were extracted from each model using a custom Perl script: (i) predicted local distance difference test (pLDDT), (ii) predicted TM-score (pTM) and (iii) interface predicted template modelling (ipTM). The latter measures the accuracy of the predicted interface between the subunits of a complex. Since we observed a high correlation between pLDDT and the average between pTM and ipTM ($r^2 > 0.95$, in all cases) (Supplementary Fig. S1A), the quality of the models was more conveniently visualized with a single score ('score') obtained by combining the three individual scores: [pLDDT/100 + (pTM + ipTM)/2]/2. As a guideline, we established an arbitrary cut-off of 80 % for this score, above which models were considered 'reliable'. With equal value for the three parameters, this is achieved with pLDDT of 80 %, and pTM and ipTM of 0.8.

2.2.3. MSA-1

For ColabFold, we initially used the default MMseqs2 multiple sequence alignment (MSA) with the TMD sequences in Fig. 1 as input (the resulting MSA is referred to as 'MSA-1'). In some cases, the number of sequences in the MSA was too small for a reliable prediction (<30) [55], but in other cases the number of sequences was sufficiently high (e.g., 90 sequences in the MSA for IBV TMD) (see Fig. 2).

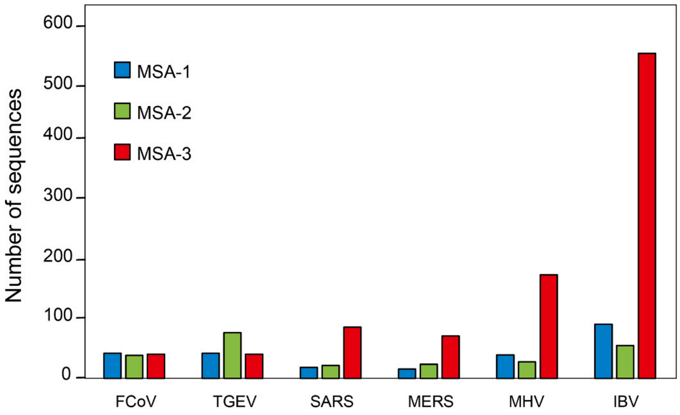

Fig. 2. Number of sequences in MSAs obtained by three different methods. MSA-1 (blue), MSA-2 (green) and MSA-3 (red).

2.2.4. MSA-2

The default MMseqs2 in ColabFold was also used to generate a custom MSA using E-FL as input. Afterwards, the resulting E-FL MSAs were cropped with Jalview to include just the TMD [56]. Alignments were polished by removing redundant sequences and sequences having excessive gaps and shorter fragments accounting for less than 50 % of the TMD, after Muscle [57] or ClustalWS [58], used with defaults. This method that used E-FL as input for ColabFold produced slightly longer MSAs than MSA-1, and is referred to as 'MSA-2' (Fig. 2).

2.2.5. MSA-3

A third type of MSA was obtained from DeepMSA (version 2) [51], a hierarchical approach that creates high-quality MSAs using both genomes and metagenomes [51,59]. Here, we used the E-FL sequences as a query. As with MSA-2 (2.2.2b), the MSAs were trimmed to include just the TMD after further alignment and removal of redundancies and truncated TMD sequences. This method resulted in longer MSAs than MSA-1 or MSA-2, and is referred to as 'MSA-3'(Fig. 2).

2.2.6. Finally

other forms of MSA were obtained by combining the individual MSAs obtained for each of the six sequences (SARS, TGEV, MHV, FCoV, MERS and IBV), assuming the backbone structure of their channels is the same. For example, combining the six individual MSA-1 with the six individual MSA-3. This was referred to as 'MSA-4'.

2.2.3 For each prediction, the best models were sorted by the predicted template modelling (pTM) score (0−1). pTM is based on a superposition of the predicted structure and an hypothetical true structure, where pTM > 0.5 means high similarity. The pLDDT (predicted local distance difference test) is a *per-residue* confidence score (>90 = high confidence, and >50 = low confidence)[60,61]. Regions with pLDDT > 90 are expected to be modelled with high accuracy, whereas regions with pLDDT < 50 may represent an unstructured region or only structured as part of a complex. The predicted aligned error (PAE) (measured in Ångströms and capped at 31.75 Å) indicates the expected positional error at residue x if the predicted and actual structures are aligned on residue y. Thus, low PAE values (colored generally in blue in a PAE plot) between two domains or subunits represent well-defined relative positions and orientations of these two bodies. Each model was energy minimized by OpenMM/Amber (relax_amber.ipynb), using default values 2000 max_iterations, tolerance 2.39 and stiffness 10 [54]. Graphical representation was performed in Chimera X [62,63] and channel lumenal volume and residue accessibility was obtained with the program HOLE [64]. From the HOLE output, a list of atoms closest to the pore centre were extracted and the corresponding residues were interpreted as hole-facing after manual confirmation.

2.3. Evolutionary tree

An evolutionary tree of TMD sequences in MSA-4 was calculated in Jalview using neighbor joining and BLOSUM62. This tree was used to identify the smaller subsets of sequences that would produce optimal results for AF2 structure prediction in ColabFold.

2.4. Helix rotational orientation comparisons

The orientation of the different amino acids respect to the centre of the helical bundle (pore) was obtained using a custom Perl script. First, PDB files were reoriented so that the central axis of the α-helical bundle was fully aligned with the Z axis (coordinates 0, 0, z), where all C=O bonds point down. After parsing the file and recording all α-carbon (**CA**) coordinates, the orientation ω for a specific residue X with coordinate **CA** was obtained as follows. A line was obtained between the **CA** geometrical average of four residues before residue X and the following four residues, forming a 'local' helix axis. A point **H** along this line, sharing the same Z coordinate as the **CA** of residue X was defined as the centre of the local helix segment, whereas a point **C** along (0, 0, z) with the same Z coordinate, was defined as the center of the bundle. Angle ω was calculated between vectors **H-CA** and **H-C**. Residues with **CA** facing exactly the centre of the bundle were defined to have ω = 0° whereas residues facing away from it were defined to have ω = 180°. A positive rotation was defined as being clockwise from 0° (see schematic in Fig. 5C). The angles corresponding to the same residue in the five helices were averaged.

3. RESULTS

3.1. Structure prediction using AF2

The structure of channels formed by E protein TMDs were predicted first with ColabFold. Using each representative TMD (Fig. 1) as input, a default MSA (MSA-1) for each sequence was generated by MMSeqs2 (MSA_mode: mmseqs2_uniref_env) with MSA lengths shown in Fig. 2. However, with this approach, the quality scores for the obtained models were only barely acceptable for the two sequences corresponding to alpha-CoV (TGEV and FCoV, see Fig. 3A), and only one model (FCoV) attained a score higher than the 0.8 cut-off. When we used ColabFold with a custom MSA (MSA-2, which is derived from E-FL)(see Methods section), MSA length improved significantly only for TGEV (Fig. 2) but model quality scores only improved for FCoV and TGEV (Fig. 3B). Results were not better when a local AF2 installation was used (Fig. 3C).

In an attempt to improve the confidence of the predictions, we assumed that all the pentameric channels formed by the sequences in Fig. 1 share a similar backbone structure, and thus we combined MSA-1 for all six sequences. Using this strategy, only a marginal improvement was observed in MERS and MHV, but except in the case of FCoV and TGEV, no models reached the cut-off level of 0.8 (Fig. 3D). Nevertheless, this marginal improvement suggested that a combination of MSAs might produce optimal results.

3.2. DeepMSA2

We then used DeepMSA2 [51,59] in an attempt to use even more diverse sequences. In most cases, these MSAs (MSA-3) contained more sequences than those obtained from MMseqs2, i.e., MSA-1 or MSA-2 (Fig. 2). Despite of this, results only improved significantly for IBV, with about half of the models scoring above the cut-off (Fig. 3E), consistent with the much larger number of sequences in MSA-3 for this sequence (Fig. 2). We then, as in Fig. 3D, assumed a common backbone structure for the channel model of all six sequences, and combined all six MSA-1 and all six MSA-3 to form a new MSA (MSA-4) that contained almost 800 sequences after removing redundancies. This combined MSA, common to all six sequences, produced excellent results for SARS,

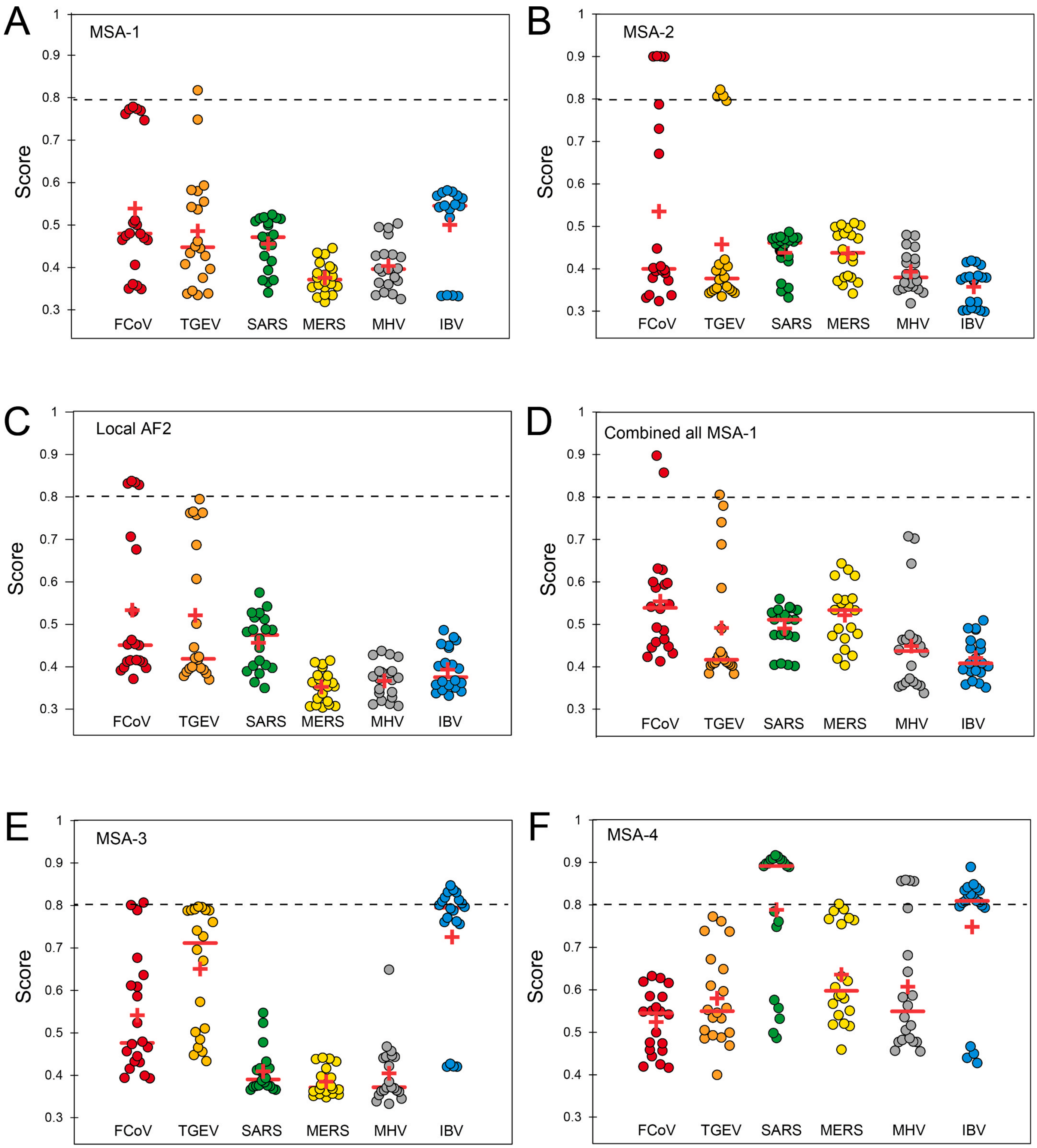

Fig. 3. Scores results obtained using various MSAs. (A) individual MSA-1; (B) individual MSA-2; (C) local AF2; (D) combined MSA-1; (E) individual MSA-3; (F) combined all MSA-1 and MSA-3 (MSA-4).

MERS, MHV and IBV, especially for SARS and IBV where more than half of the models scored above the cut-off (Fig. 3F). However, results worsened for the alpha-CoV sequences (FCoV and TGEV, especially for the former). We speculate that the different behavior of these two sequences may be related to their proposed opposed topology [28–30].

3.3. *Minimal set of sequences in the MSA*

Interestingly, a substantial fraction (~80 %) of the MSA-3 sequences derived from DeepMSA-2 were not related to CoV E proteins, or even to viral proteins. To test if using only sequences related to E protein would improve prediction quality, we extracted all E-related sequences in the six MSA-1 (215 sequences) and in the six MSA-3 (130 sequences).

Neither using only E-related sequences found in MSA-3 or all E-related sequences found in MSA-4 resulted in satisfactory results (Fig. S2). Since (i) a high percentage of the MSA-3 sequences are not related to E proteins and (ii) we obtained an optimal result using the combination MSA-4 (788 sequences) (Fig. 3F), we tried to identify the minimal number of sequences in the MSA-4 required to obtain optimal results. Thus, we clustered the sequences in MSA-4 based on closeness to CoV-E sequences, and we started considering the branch of the tree containing

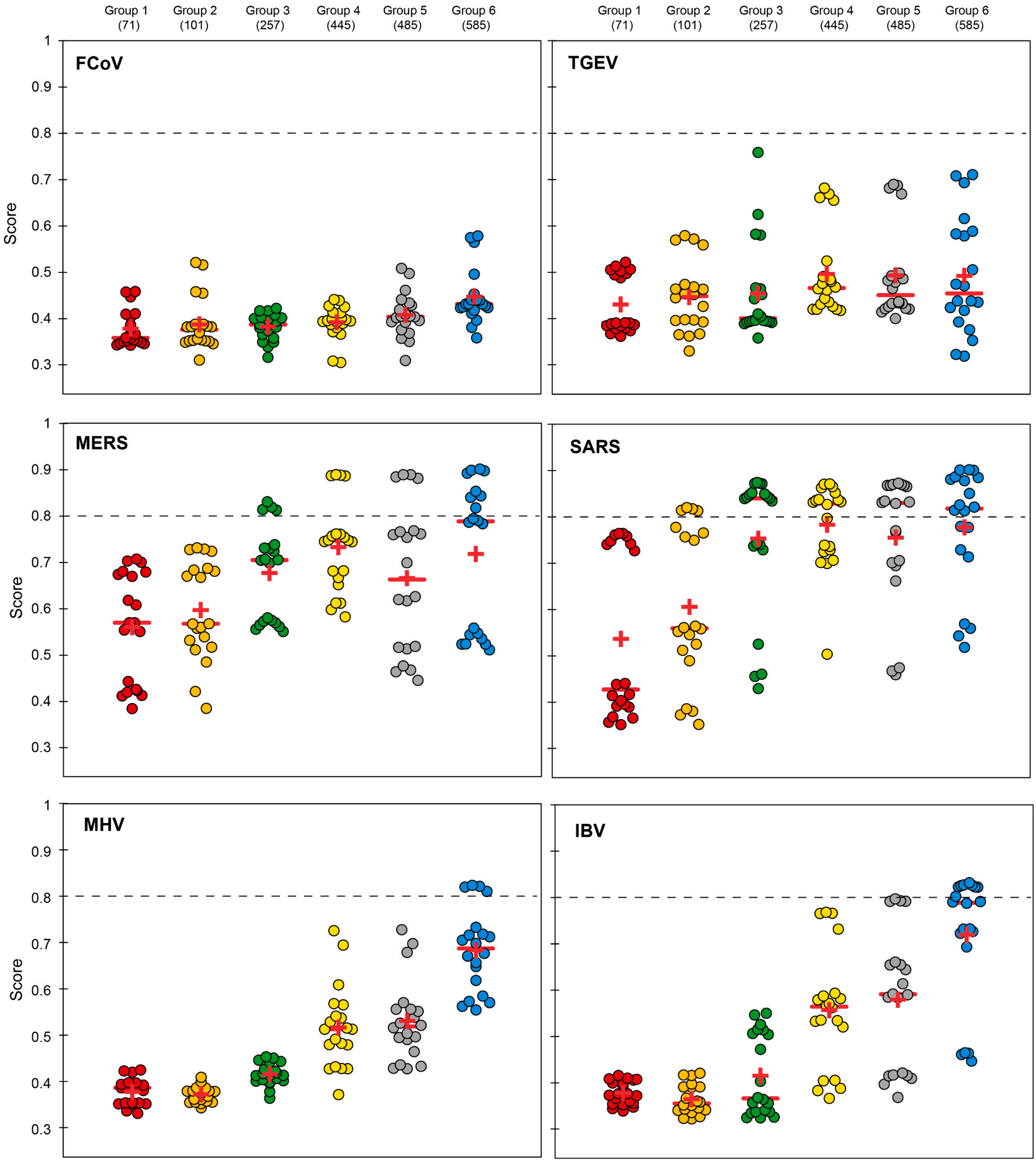

Fig. 4. Prediction for increasing number of sequences in the combined MSA-4. Scores corresponding to the 20 models obtained using MSA-4 that contained an increasing number of sequences; from 71 (group 1, including only E sequences), and after addition of branches of the evolutionary tree that were progressively farther from the E sequences, from group 2 (101 sequences) to group 6 (585 sequences) for the TMDs of the representative viruses indicated in each panel. Red cross and bar represent average and median scores, respectively.

only CoV E sequences (71 sequences, referred to as ‘group 1’). This was followed by the inclusion in that group of the closest branch to the E sequences in the tree. This produced a total 101 sequences, referred to as ‘group 2’. From there, we increased the MSA progressively including farther branches from the E sequences, to form ‘group 3’ (257 sequences), ‘group 4’ (445 sequences), ‘group 5’ (485 sequences) and ‘group 6’ (585 sequences).

The structure prediction was then run again using these grouped custom MSAs as input (Fig. 4). For alpha-CoVs (FCoV and TGEV), none of these partial MSAs produced results above the cutoff, but all other sequences produced optimal results for the largest (group 6) MSA. For SARS and MERS, some models were above the cut-off with just 101 sequences (group 2) or 257 sequences (group 3), respectively. For MHV and IBV this required 585 sequences (group 6). Results did not improve further after the addition of more branches that diverged more from the E sequences in the evolutionary tree (not shown).

3.4. Best model above the cut-off

For the best five models of each of the six sequences (Fig. 4) we obtained quality plots showing the corresponding MSA, pLDDT (but the scores for Figs. 3–4 used also pTM and ipTM) and PAE (Supplementary Figs. 3–5). The best model for each sequence is schematically shown in Fig. 5A. For all the sequences, the conserved polar residue (in SARS, Asn-15, see Fig. 1 highlighted) is oriented luminally (‘polar-in’ orientation), except for the MERS sequence, where this residue faces the neighboring helix (‘polar-inter’ orientation) (Fig. 5B). A more quantitative comparison between these models was obtained using the rotational orientation of the residues (see Methods section and Fig. 5C). The rotational orientation of ‘polar-in’ models (Fig. 5D) shows a very similar orientation along the length of the helix and the conserved polar residue is exposed directly to the lumen of the channel (Fig. 5E, blue stripe). The IBV E model was excluded from these ‘polar-in’ plots because it deviates from these orientations in the C-terminal half (results for IBV are

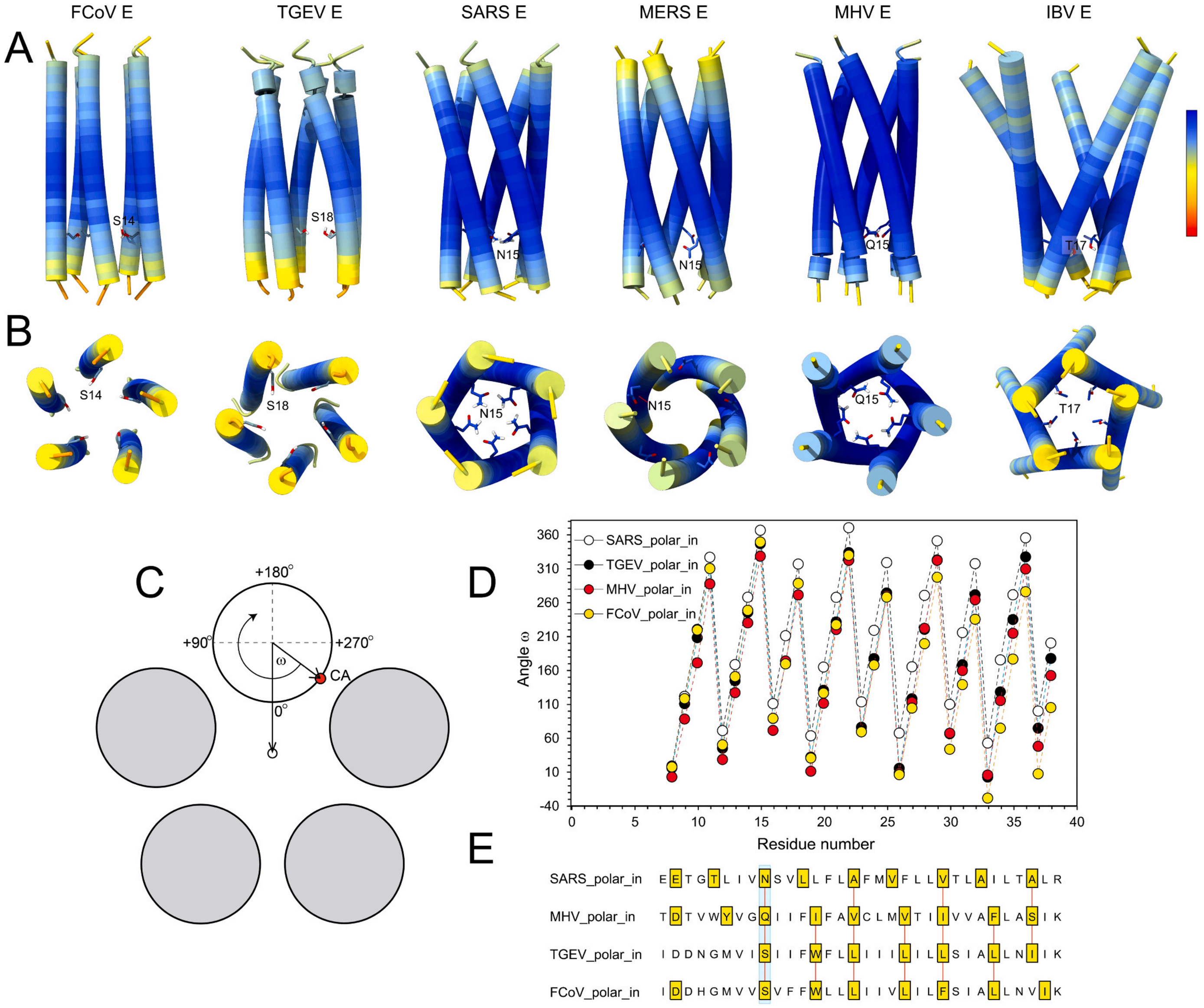

Fig. 5. Representative best models obtained for each of the TMD sequences. The pentameric models are represented in a side view (top row) and a view from the N-terminus (bottom row). To guide the eye, the side chain of the fully conserved polar residue in each sequence is shown as sticks. The models are colored according to pLDDT in AF2; (C) schematic representation of the way in which the rotational orientation per residue was calculated; one of the monomers in the pentamer shows two vectors, both coming from the centre the helix. One is directed towards the centre of the channel and the other towards the α-carbon (CA) of a residue. The angle between these two vectors, ω, was obtained and represented; (D) rotational orientation ω of the residues in the four ‘polar in’ orientations indicated; (E) residues directly exposed to the lumen of the channel (yellow) according to HOLE.

summarized in Supplementary Figs. S6, S7).

3.5. Multiple conformational states above the cut-off

After examination of all the models above the cut-off for each sequence, structural heterogeneity was observed in most cases, suggesting possible conformational intermediates.

3.5.1. SARS sequence

For the SARS sequence, the ten good models obtained clustered in two groups (almost equally populated) corresponding to either ‘polar-in’ or ‘polar-inter’ orientations (Fig. 6A-C). The lumen in the two clusters appears to have a constriction at two different locations along the channel (Fig. 6C), which suggests these may be conformational states. The same figure (Fig. 6D-E) shows the two high-resolution SARS models obtained in lipid membranes by ssNMR: a ‘closed’ pentamer (PDB: 7K3G) [40] and a proposed alternative ‘open’ form (PDB: 8SUZ) obtained using lower pH and high Ca^{2+} concentration [44]. The similarity between the AF2-predicted ‘polar-in’ model and the ‘closed’ model 7K3G is obvious comparing the orientation of Asn-15 in the N-terminal half of the channel (panels B and E). However, superposition of these models and calculation of RMSD was not possible because in the AF2-predicted models helices are continuous, tilted by $\sim 20^\circ$ an the bundles are left-handed, whereas the experimentaly-derived models have kinks and bends around Phe-20 and almost no tilt (Fig. 6F). Thus, a quantitative comparison of these models was made using the rotational orientation of the residues relative to the channel center (Fig. 6G-H). The two AF2-predicted models are separated by an almost uniform $\sim -60^\circ$ rotation. Interestingly, whereas the ‘polar-in’ model is very different from model 8SUZ (Fig. 6G), the latter and ‘polar-inter’ have a very similar orientation, especially in the C-terminal half (Fig. 6H) and share the same four residues exposed to the channel lumen (Fig. 6I). Overall, 7K3G, 8SUZ and ‘polar-inter’ share a similar rotational orientation in the C-terminal half (Fig. 5G). In contrast, the ‘polar-in’ model differs from these three models in the C-terminal half, but is very similar to 7K3G in the N-terminal half (where Asn-15 is located). This is also supported by the luminally exposed residues in each case (Fig. 6I).

3.5.2. MHV sequence

For the MHV E sequence, three clusters were observed: ‘polar-in’ (Fig. 7, left), ‘polar-inter’ (Fig. 7, right) and a third one that seems to represent an intermediate structure with a larger helix tilt (Fig. 7, middle). In this case, the ‘polar-inter’ model channel is completely blocked by Trp-11 (not shown), as shown by the channel volumes (Fig. 7 C).

3.5.3. IBV sequence

The best models for the IBV sequence formed a single cluster in a ‘polar-in’ orientation, with tilted helices resembling an inverted tepee (Fig. 5A). However, this appears to be context-dependent: a prediction that used a longer version of IBV E (residues 1 to 90), or one that removed all ‘non E’ sequences in MSA-4, produced channel structures with a similar low helix tilt to other sequences. This low helix tilt model was also ‘polar-in’(see comparison between these models in Fig. S7).

3.5.4. FCoV, TGEV and MERS sequences

For FCoV ad TGEV, the predominant model was also ‘polar-in’ although ‘polar-inter’ was also found in some models above the cut-off (not shown). Finally, for MERS, all the best models (above cut-off) had a ‘polar-inter’ orientation.

4. DISCUSSION

In this paper, we attempted to use AF2 to obtain model structures for E TMD pentameric oligomers of various representative coronaviruses. Full-length E sequences did not produce good models (scores typically $<$ 0.5), either using ColabFold or a local AF2 installation, although the TMD region is usually the one showing the highest prediction confidence (not shown). Thus, we focused our efforts in this part of the E protein. Validation of the resulting models, at least in the case of the SARS sequence, was helped by the fact that (i) two structures of the SARS TMD channel are available obtained by solid-state NMR and (ii) the oligomeric size of this channel has been reasonably proven to be pentameric, although other forms may also exist.

However, using ColabFold and MMseqs2 to produce MSAs was only adequate for the two alphacoronavirus sequences, FCoV and TGEV. This particularity may be related to the distinct reported topology for E protein in alpha-coronaviruses, e.g., TGEV, with $C_{exo}N_{endo}$, where the C-terminus is lumenal and the N-terminus is facing the cytoplasm [28–30]. The quality of the predictions improved when we used MSAs from DeepMSA2, especially for IBV E, where more than 500 sequences were available in MSA-3. We then reasoned that while the sequences are different, the backbone structure of the channel may be common to all these sequences, and a combination of all the individual MSA-3 alignments, together with the MSA-1 alignments, could produce a better result. This was indeed the case, again except for the alpha-CoV sequences FCoV and TGEV which in fact showed much worse results. In some cases (SARS, MERS, MHV and IBV E), more than half of the 20 models had a combined score of > 80, which implies high confidence.

Incidentally, a similar assumption regarding the similarity of the backbone for E TMD channel in CoV E proteins mas made in a paper published by one of us (J.T.) about 20 years ago that used evolutionary conservation data [21], where the correct pentameric model could not be decided among two types of of bundle related by a $\sim 45^\circ$ rotation. Similar models are found by AF2 herein, which we term ‘polar-in’ and polar-inter’. Since the AF2-predicted models do not have a break in the TMD (found in the NMR experimental model), RMSD or TM-score cannot be used for comparison. Instead, we used a quantitative measure based on the rotational orientation of each residue relative to the center of the channel. The ‘polar-in’ model is consistent with N-terminal half of 7K3G, the ‘closed’ channel structure reported previously in ERGIC lipids by solid-state NMR [40], as both show Asn-15 facing the lumen of the channel. However, in the C-terminal half, the two models are separated by a $\sim 60^\circ$ rotation. Interestingly, the ‘polar-inter’ model is almost identical in the C-terminal region to both experimental models 7K3G and 8SUZ, another experimental model obtained at lower pH and a high calcium concentration [43,44]. In the N-terminal half, the ‘polar-inter’ model is closer to 8SUZ, although the latter shows a larger rotation to the point that Asn-15 points to the lipid phase. We note that, as we have noted previously [38], the vast majority of electrophysiological experiments with the SARS E channel do not require either acidification nor calcium to produce open channels [11,13,16,18,19,34], and both low pH and addition of calcium over salts of monovalent cations have been shown to *reduce* channel conductance [65]. In any case, although we did not find such model using the techniques presented here, the ‘polar-inter’ model seems to be close to that ‘open’ conformation.

In summary, we found that in SARS the two AF2-predicted models are related by an almost constant $\sim 60^\circ$ rotation of the helices, whereas the two experimental models 7K3G and 8SUZ share a similar C-terminal half and differ by a $> 100^\circ$ rotation in the N-terminal half. Whether these differences are caused by the inability of AF2 to reproduce the experimentally observed bends or kinks in the TMD helices, or whether the more straight helices predicted represent true conformational intermediates requires further experimental validation. For several of the other sequences, we also obtained models that pertain to at least two different conformations, whereas in other sequences only one conformation was found (MERS and IBV, ‘polar-inter’ and ‘polar-in’, respectively).

Lastly, the conformational intermediates suggested by our study may be further defined by manipulating the MSAs. This is in a way similar to the strategy used here to improve reliability of the models, and would

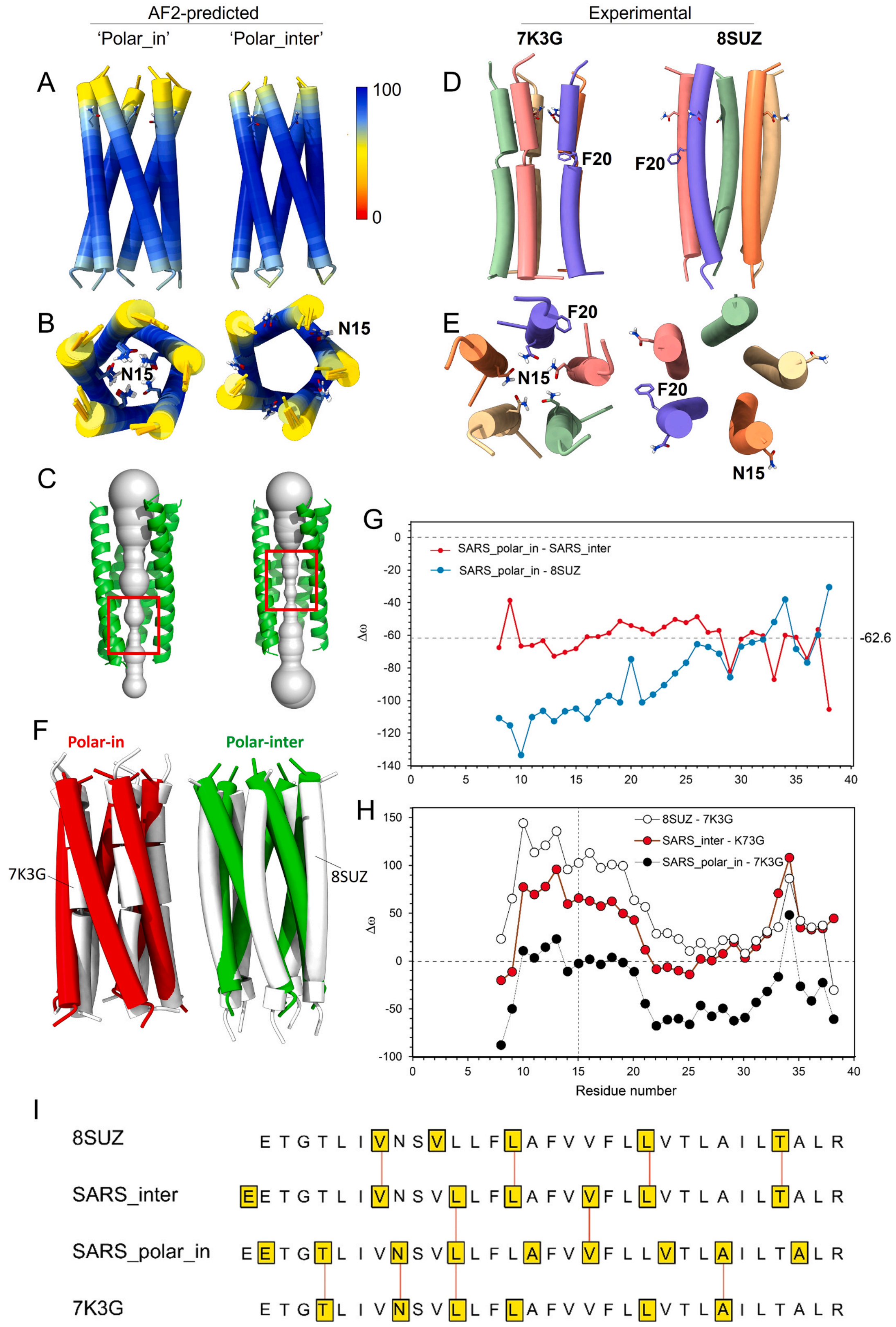

Fig. 6. Comparison of AF2-predicted and experimental SARS channels. (A) AF2-predicted 'polar_in' and 'polar-inter' models colored according to the pLDDT score; (B) top view from the N-terminus of the models in (A), showing the orientation of the polar residue Asn15; (C) inner volume of the channel for the models in (A) where the most constricted region is indicated by a red rectangle; (D-E) experimental models in side (D) and top (E) views; (F) discrepancy between the AF2-predicted and experimental models; (G) calculated difference between ω angles of SARS 'polar-in' minus 'SARS-inter' or minus 8SUZ. The horizontal dotted line shows the average ω difference between the two predicted SARS models (−62.6°); (H) same as (G) for the differences respect to model 7K3G; (I) residues in these four models directly in contact with the lumen of the channel (highlighted in yellow). A more complete description of the luminal orientation in these and other models is shown in Fig. S6.

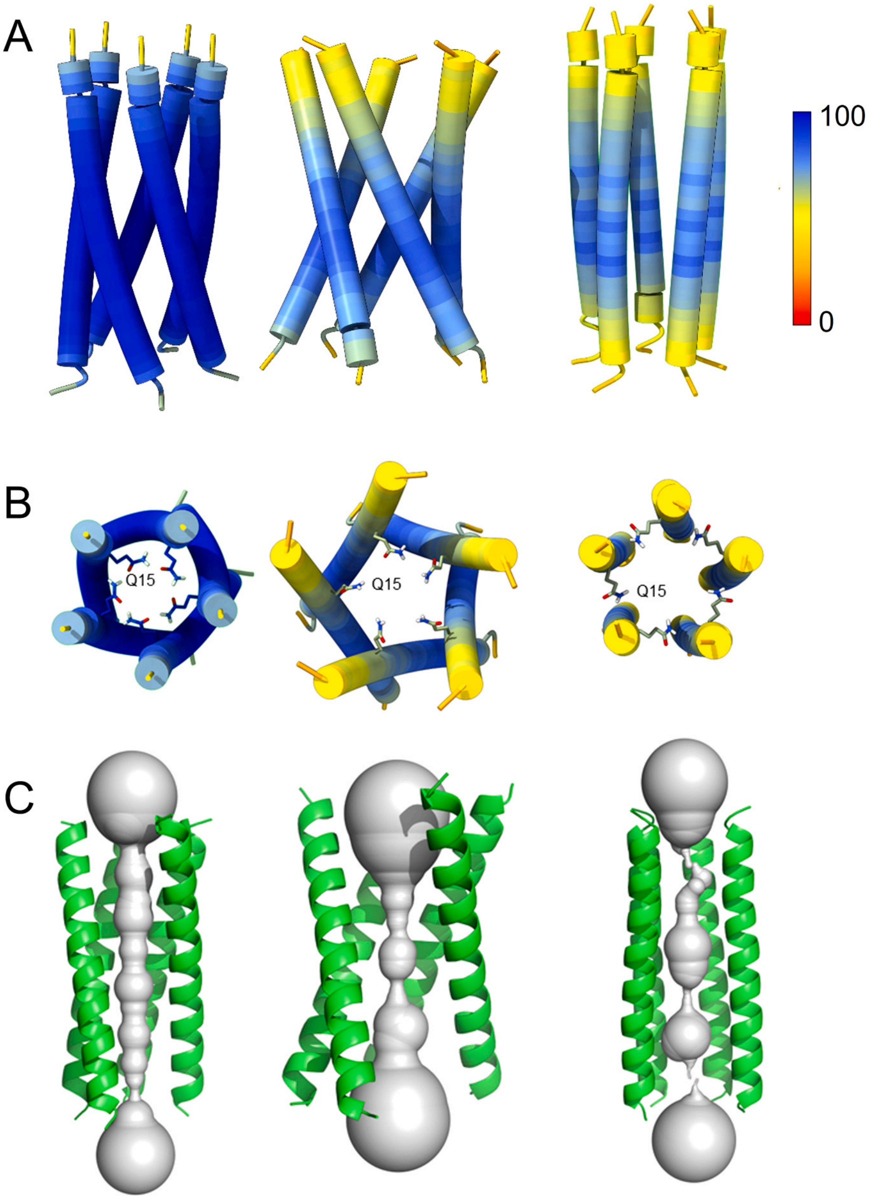

Fig. 7. Representative three clusters of models above the cut-off for MHV ETM pentamers. (A) Side view, (B) top view from the N-terminus, showing the orientation of the polar residue Q15; (C) lumenal space obtained with HOLE [64]. Models in (A-B) are colored according to the pLDDT score. Residues exposed to the channel are shown in Supplementary Fig. S6.

involve using only a subset of MSA sequences with specific features (or randomly selected), avoiding templates and extensive recycling. This method was employed recently [66] to obtain alternative conformational states of transporters and receptors. Alternative AF2-based methods that involve masking rows or columns in the MSA to mask coevolution information, subsampling, bias, and structure evaluation have been discussed in detail elsewhere [67].

Overall, we have shown that by restricting the sequence to the TMD and by using a larger dataset for MSA, very high-confidence models (as defined by AF2) can be obtained for the E protein TMD channel of coronaviruses. These may unveil or hint at mechanistic aspects that are hidden in experimental models.

CRediT authorship contribution statement

Konstantin Pervushin: Writing – review & editing, Project administration, Funding acquisition. **Wahyu Surya:** Writing – review & editing, Formal analysis. **Jaume Torres:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors have declared no conflict of interest.

Acknowledgements

J.T. and K.P. acknowledge the generous funding of the Ministry of Education of Singapore MOE Tier2 grant #021410-00001.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.10.021.

References

[1] Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. Nidovirales: evolving the largest RNA virus genome. Virus Res 2006;117:17–37.

[2] Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and sources of endemic human coronaviruses. Adv Virus Res 2018;100:163–88.

[3] Drosten C, Günther S, Preiser W, Van der Werf S, Brodt HR, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 2003;348:1967–76.

[4] Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, et al. A novel coronavirus associated with severe acute respiratory syndrome. New Engl J Med 2003;348:1953–66.

[5] Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New Engl J Med 2012;367:1814–20.

[6] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727–33.

[7] Saif LJ. Animal coronaviruses: What can they teach us about the severe acute respiratory syndrome? *OIE*. Rev Sci Et Tech 2004;23:643–60.

[8] (2024) WOAH, 2023. World Organisation for Animal Health Animal Diseases.

[9] Knoops K, Kikkert M, Van Den, Worm SHE, Zevenhoven-Dobbe JC, Van Der Meer Y, et al. SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. PLoS Biol 2008;6:1957–74.

[10] Masters PS. The molecular biology of coronaviruses. Adv Virus Res 2006;66: 193–292.

[11] Torres J, Maheswari U, Parthasarathy K, Ng L, Ding XL, Gong X. Conductance and amantadine binding of a pore formed by a lysine-flanked transmembrane domain of SARS coronavirus envelope protein. Protein Sci 2007;16:2065–71.

[12] Wilson L, Gage P, Ewart G. Hexamethylene amiloride blocks E protein ion channels and inhibits coronavirus replication. Virology 2006;353:294–306.

[13] Wilson L, McKinlay C, Gage P, Ewart G. SARS coronavirus E protein forms cation-selective ion channels. Virology 2004;330:322–31.

[14] Li Y, Surya W, Claudine S, Torres J. Structure of a conserved Golgi complex-targeting signal in coronavirus envelope proteins. J Biol Chem 2014;289: 12535–49.

[15] Parthasarathy K, Lu H, Surya W, Vararattanavech A, Pervushin K, Torres J. Expression and purification of coronavirus envelope proteins using a modified beta-barrel construct. Protein Expr Purif 2012;85:133–41.

[16] Nieto-Torres JL, Verdia-Baguena C, Jimenez-Guardeno JM, Regla-Nava JA, Castano-Rodriguez C, Fernandez-Delgado R, et al. Severe acute respiratory syndrome coronavirus E protein transports calcium ions and activates the NLRP3 inflammasome. Virology 2015;485:330–9.

[17] Nieto-Torres JL, DeDiego ML, Verdia-Baguena C, Jimenez-Guardeno JM, Regla-Nava JA, Fernandez-Delgado R, et al. Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. PLoS Pathog 2014;10:e1004077.

[18] Verdia-Baguena C, Nieto-Torres JL, Alcaraz A, Dediego ML, Enjuanes L, Aguilella VM. Analysis of SARS-CoV E protein ion channel activity by tuning the protein and lipid charge. Biochim Biophys Acta 2013;1828:2026–31.

[19] Verdia-Baguena C, Nieto-Torres JL, Alcaraz A, Dediego ML, Torres J, Aguilella VM, et al. Coronavirus E protein forms ion channels with functionally and structurally-involved membrane lipids. Virology 2012;432:485–94.

[20] Torres J, Parthasarathy K, Lin X, Saravanan R, Kukol A, Liu DX. Model of a putative pore: the pentameric α-helical bundle of SARS coronavirus E protein in lipid bilayers. Biophys J 2006;91:938–47.

[21] Torres J, Wang J, Parthasarathy K, Liu DX. The transmembrane oligomers of coronavirus protein E. Biophys J 2005;88:1283–90.

[22] Pervushin K, Tan E, Parthasarathy K, Xin L, Jiang FL, Yu D, et al. Structure and inhibition of the SARS coronavirus envelope protein ion channel. PLoS Pathog 2009;5.

[23] Nieto-Torres JL, Dediego ML, Alvarez E, Jimenez-Guardeno JM, Regla-Nava JA, Llorente M, et al. Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein. Virology 2011;415:69–82.

[24] Li Y, Surya W, Claudine S, Torres J. Structure of a conserved golgi complex-targeting signal in coronavirus envelope proteins. J Biol Chem 2014;289: 12535–49.

[25] Parthasarathy K, Ng L, Lin X, Liu DX, Pervushin K, Gong X, et al. Structural flexibility of the pentameric SARS coronavirus envelope protein ion channel. Biophys J 2008;95:L39–41.

[26] To J, Surya W, Torres J. Targeting the channel activity of viroporins. Adv Protein Chem Struct Biol 2016;104:307–55.

[27] Corse E, Machamer CE. Infectious bronchitis virus E protein is targeted to the Golgi complex and directs release of virus-like particles. J Virol 2000;74:4319–26.

[28] Maeda J, Repass JF, Maeda A, Makino S. Membrane topology of coronavirus E protein. Virology 2001;281:163–9.

[29] Godet M, L'Haridon R, Vautherot JF, Laude H. TGEV corona virus ORF4 encodes a membrane protein that is incorporated into virions. Virology 1992;188:666–75.

[30] Corse E, Machamer CE. Infectious bronchitis virus E protein is targeted to the Golgi complex and directs release of virus-like particles. J Virol 2000;74:4319–26.

[31] Surya W, Li Y, Verdià-Bàguena C, Aguilella VM, Torres J. MERS coronavirus envelope protein has a single transmembrane domain that forms pentameric ion channels. Virus Res 2015;201:61–6.

[32] Madan V, Garcia Mde J, Sanz MA, Carrasco L. Viroporin activity of murine hepatitis virus E protein. FEBS Lett 2005;579:3607–12.

[33] Liao Y, Yuan Q, Torres J, Tam JP, Liu DX. Biochemical and functional characterization of the membrane association and membrane permeabilizing activity of the severe acute respiratory syndrome coronavirus envelope protein. Virology 2006;349:264–75.

[34] Parthasarathy K, Ng L, Lin X, Ding XL, Pervushin K, Gong X, et al. Structural flexibility of the pentameric SARS coronavirus envelope protein ion channel. Biophys J 2008;95:L39–41.

[35] Somberg NH, Wu WW, Medeiros-Silva J, Dregni AJ, Jo H, DeGrado WF, et al. SARS-CoV-2 envelope protein forms clustered pentamers in lipid bilayers. *Biochem* Online Print 2022.

[36] Parthasarathy K, Lu H, Surya W, Vararattanavech A, Pervushin K, Torres J. Expression and purification of coronavirus envelope proteins using a modified β-barrel construct. Protein Expr Purif 2012;85:133–41.

[37] Surya W, Torres J. Oligomerization-dependent beta-structure formation in SARS-CoV-2 envelope protein. Int J Mol Sci 2022;23:13285.

[38] Surya W, Tavares-Neto E, Sanchis A, Queralt-Martín M, Alcaraz A, Torres J, et al. The complex proteolipidic behavior of the SARS-CoV-2 envelope protein channel: weak selectivity and heterogeneous oligomerization. Int J Mol Sci 2023;24:12454.

[39] Poggio, E. , Vallese, F. , Hartel, A.J.W. , Morgenstern, T.J. , Kanner, S.A. , Rauh, O. , , , , , , and et al. (2023) Perturbation of the host cell Ca(2+) homeostasis and ER-mitochondria contact sites by the SARS-CoV-2 structural proteins E and M. 14, 297.

[40] Mandala VS, McKay MJ, Shcherbakov AA, Dregni AJ, Kolocouris A, Hong M. Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. Nat Struct Mol Biol 2020;27:1202–8.

[41] Torres J, Parthasarathy K, Lin X, Saravanan R, Kukol A, Ding XL. Model of a putative pore: the pentameric α-helical bundle of SARS coronavirus E protein in lipid bilayers. Biophys J 2006;91:938–47.

[42] Pervushin K, Tan E, Parthasarathy K, Lin X, Jiang FL, Yu D, et al. Structure and Inhibition of the SARS Coronavirus Envelope Protein Ion Channel. PLoS Pathog 2009;5.

[43] Medeiros-Silva J, Somberg NH, Wang HK, McKay MJ, Mandala VS, Dregni AJ, et al. pH- and calcium-dependent aromatic network in the SARS-CoV-2 envelope protein. J Am Chem Soc 2022;144:6839–50.

[44] Medeiros-Silva J, Dregni AJ, Somberg NH, Duan P, Hong M. Atomic structure of the open SARS-CoV-2 E viroporin. Sci Adv 2023;9:eadi9007.

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

[46] Zhang X, Hasoksuz M, Spiro D, Halpin R, Wang S, Stollar S, et al. Complete genomic sequences, a key residue in the spike protein and deletions in nonstructural protein 3b of US strains of the virulent and attenuated coronaviruses, transmissible gastroenteritis virus and porcine respiratory coronavirus. Virology 2007;358:424–35.

[47] Wu Q, Zhang Y, Lü H, Wang J, He X, Liu Y, et al. The E protein is a multifunctional membrane protein of SARS-CoV. Genom Proteom Bioinforma 2003;1:131–44.

[48] de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, et al. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol 2013;87:7790–2.

[49] Fischer F, Stegen CF, Masters PS, Samsonoff WA. Analysis of constructed E gene mutants of mouse hepatitis virus confirms a pivotal role for E protein in coronavirus assembly. J Virol 1998;72:7885–94.

[50] Brooks JE, Rainer AC, Parr RL, Woolcock P, Hoerr F, Collisson EW. Comparisons of envelope through 5B sequences of infectious bronchitis coronaviruses indicates recombination occurs in the envelope and membrane genes. Virus Res 2004;100: 191–8.

[51] Zheng W, Wuyun Q, Li Y, Zhang C, Freddolino PL, Zhang Y. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. Nat Methods 2024;21:279–89.

[52] Jumper J, E R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.

[53] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv 2022:463034. 2021.2010.2004.

[54] Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods 2022;19:679–82.

[55] 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.

[56] Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009; 25:1189–91.

[57] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.

[58] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–8.

[59] Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2019;36:2105–12.

[60] 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.

[61] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596: 590–6.

[62] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–12.

[63] Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. Protein Sci 2021;30:70–82.

[64] Smart OS, Goodfellow JM, Wallace BA. The pore dimensions of gramicidin A. Biophys J 1993;65:2455–60.

[65] Verdiá-Báguena C, Aguilella VM, Queralt-Martín M, Alcaraz A. Transport mechanisms of SARS-CoV-E viroporin in calcium solutions: lipid-dependent Anomalous Mole Fraction Effect and regulation of pore conductance. Biochim Et Biophys Acta (BBA)-Biomembr 2021;1863:183590.

[66] del Alamo D, Sala D, McHaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. eLife 2022; 11:e75751.

[67] Sala D, Engelberger F, McHaourab HS, Meiler J. Modeling conformational states of proteins with AlphaFold. Curr Opin Struct Biol 2023;81:102645.