





## RESEARCH ARTICLE

# REVISED In silico prediction of structure and function for a large family of transmembrane proteins that includes human Tmem41b [version 2; peer review: 2 approved, 1 approved with reservations]

Shahram Mesdaghi, David L. Murphy, Filomeno Sánchez Rodríguez ,  
J. Javier Burgos-Mármol , Daniel J. Rigden

Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK

**V2** First published: 03 Dec 2020, 9:1395  
<https://doi.org/10.12688/f1000research.27676.1>  
 Latest published: 25 Mar 2021, 9:1395  
<https://doi.org/10.12688/f1000research.27676.2>

## Abstract

**Background:** Recent strides in computational structural biology have opened up an opportunity to understand previously uncharacterised proteins. The under-representation of transmembrane proteins in the Protein Data Bank highlights the need to apply new and advanced bioinformatics methods to shed light on their structure and function. This study focuses on a family of transmembrane proteins containing the Pfam domain PF09335 ('SNARE\_ASSOC'/ 'VTT '/ 'Tvp38'/ 'DedA'). One prominent member, Tmem41b, has been shown to be involved in early stages of autophagosome formation and is vital in mouse embryonic development as well as being identified as a viral host factor of SARS-CoV-2.






**Methods:** We used evolutionary covariance-derived information to construct and validate *ab initio* models, make domain boundary predictions and infer local structural features.


**Results:** The results from the structural bioinformatics analysis of Tmem41b and its homologues showed that they contain a tandem repeat that is clearly visible in evolutionary covariance data but much less so by sequence analysis. Furthermore, cross-referencing of other prediction data with covariance analysis showed that the internal repeat features two-fold rotational symmetry. *Ab initio* modelling of Tmem41b and homologues reinforces these structural predictions. Local structural features predicted to be present in Tmem41b were also present in Cl<sup>-</sup>/H<sup>+</sup> antiporters.

**Conclusions:** The results of this study strongly point to Tmem41b and its homologues being transporters for an as-yet uncharacterised substrate and possibly using H<sup>+</sup> antiporter activity as its mechanism for transport.

## Open Peer Review

Reviewer Status   

	Invited Reviewers		
	1	2	3
<b>version 2</b>			
(revision)			
25 Mar 2021		report	report
		↑	↑
<b>version 1</b>			
03 Dec 2020	report	report	report

- Pradip Panta**, Louisiana State University, Baton Rouge, USA  
**William T. Doerrler**, Louisiana State University, Baton Rouge, USA
- Gábor Tusnády** , Research Centre for Natural Sciences, Budapest, Hungary  
**László Dobson**, Research Centre for Natural Sciences, Budapest, Hungary
- Claudio Bassot** , Stockholm University, Stockholm, Sweden

Any reports and responses or comments on the article can be found at the end of the article.

**Keywords**

ab initio modelling, bioinformatics, autophagy, contact predictions, evolutionary covariance, DedA, SARS-CoV-2, Tmem41b, VTT domain

**Corresponding author:** Daniel J. Rigden ([drigden@liverpool.ac.uk](mailto:drigden@liverpool.ac.uk))

**Author roles:** **Mesdaghi S:** Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Murphy DL:** Investigation, Writing – Review & Editing; **Sánchez Rodríguez F:** Investigation, Writing – Review & Editing; **Burgos-Mármol JJ:** Investigation, Writing – Review & Editing; **Rigden DJ:** Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The author(s) declared that no grants were involved in supporting this work.

**Copyright:** © 2021 Mesdaghi S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Mesdaghi S, Murphy DL, Sánchez Rodríguez F *et al.* **In silico prediction of structure and function for a large family of transmembrane proteins that includes human Tmem41b [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2021, 9:1395 <https://doi.org/10.12688/f1000research.27676.2>

**First published:** 03 Dec 2020, 9:1395 <https://doi.org/10.12688/f1000research.27676.1>

**REVISED Amendments from Version 1**

Input from the referees led to the conclusion that the re-entrant PDBTM screen needed to be reimplemented; the use of re-entrant loop sequences in order to perform the screen may not be appropriate due to the poor sequence similarity between the re-entrant loops with a view that a structural comparison being more informative. Subsequently, pdb structures of the loops were used for the clustering exercise. The boundaries for the experimentally determined structures were extracted from the PDBTM and the boundaries for the models were predicted using the OMP server. As this investigation focused on re-entrant loops that are immediately preceded by a TMhelix that is packed with the re-entrant loop, all re-entrant loops in addition to the preceding 30 residues were extracted from a non-redundant re-entrant loop containing subset of the PDB. The resulting 193 library entries, supplemented with the re-entrant loop features from the ab initio models, underwent an all-against-all structural alignment utilising Dali. The Z-scores for these alignments were then used to cluster all the structures. The reimplemented screen resulted in the query re-entrant loop feature structures clustering with the re-entrant loop features of Cl<sup>-</sup>/H<sup>+</sup> antiporters; this was a similar result to the original sequence-based clustering.

An additional figure has been added to the manuscript showing a multiple sequence alignment for a selection of DedA domain proteins. The alignment has been annotated to highlight the relative positions of the DedA and the PF09665 domains as well as the re-entrant loop positions for the example DedA proteins that were modelled.

The amphipathic helix prediction test paragraph in the results section has been re-written for the purpose of clarity.

Finally, in addition to the correction of typographical errors, the citations have been updated as recommended by the referees as well as to reflect the changes in the experimental procedure.

**Any further responses from the reviewers can be found at the end of the article**

**Introduction**

A protein's structural information is crucial to understand its function and evolution. Currently, there is only experimental structural data for a tiny fraction of proteins (Khafizov *et al.*, 2014). For instance, membrane proteins are encoded by 30% of the protein-coding genes of the human genome (Almén *et al.*, 2009), but they only have a 3.3% representation in the Protein Data Bank (PDB) (5785 membrane proteins out of 174507 PDB entries). Membrane protein families are particularly poorly understood due to experimental difficulties, such as over-expression, which can result in toxicity to host cells (Grisshammer & Tateu, 1995), as well as difficulty in finding a suitable membrane mimetic to reconstitute the protein. Additionally, membrane proteins are much less conserved across species compared to water-soluble proteins (Sojo *et al.*, 2016), making sequence-based homologue identification a challenge, and in turn rendering homology modelling of these proteins more difficult. Membrane proteins can be grouped according to their interaction with various cell membranes: integral membrane proteins (IMPs) are permanently anchored whereas peripheral membrane proteins transiently adhere to cell membranes. IMPs that span the membrane are known as transmembrane

proteins (TMEMs) as opposed to IMPs that adhere to one side of the membrane (Fowler & Coveney, 2006). Membrane proteins also include various lipid-modified proteins (Resh, 2016).

One IMP protein family is Tmem41, which has two human representatives, namely Tmem41a and Tmem41b; both share the PF09335 ('SNARE\_ASSOC'/'VTT '/'Tvp38'/'DedA') Pfam (El-Gebali *et al.*, 2019) domain. The profile of Tmem41b has recently risen due to experimental evidence pointing to its involvement in macroautophagy regulation (making it a possible Atg protein, i.e. an autophagy related protein) and lipid mobilisation (Moretti *et al.*, 2018). Other studies identify Tmem41b to be involved in motor circuit function, with TMEM41B-knockout *Drosophila* showing neuromuscular junction defects and aberrant motor neuron development in knockout zebrafish (Lotti *et al.*, 2012). Also, it has been reported that in TMEM41B-knockout HeLa cells there is an inhibition of Zika virus replication (Scaturro *et al.*, 2018). Tmem41b has also been identified as a host cell factor for SARS-CoV-2 (Schneider *et al.*, 2020). Tmem41b is the only common host cell factor identified for flaviviruses and coronaviruses and is the only autophagy-related protein identified as a viral host factor (Hoffmann *et al.*, 2021).

Additionally, Tmem41b has been shown to be essential for mouse embryonic development: homozygous knockout mice embryos suffer early termination of their development after 7–8 weeks (Van Alstyne *et al.*, 2018). Tmem41b is a structurally uncharacterised 291-residue protein found in the endoplasmic reticulum (ER) localising at the mitochondria-associated ER membranes (Moretti *et al.*, 2018). Disruption of the PF09335 domain by various residue substitutions (Tábara *et al.*, 2019) or its removal (Morita *et al.*, 2018) results in inhibition of autophagosome formation and impaired lipid mobilisation in human embryonic kidney (HEK) cells.

Tmem41b homologues, hereafter referred to as DedA proteins (Morita *et al.*, 2019), are present in all domains of life (Keller & Schneider, 2013). The Pfam PF09335 domain was first identified in the *Saccharomyces cerevisiae* protein Tvp38 (Inadome *et al.*, 2007), and the authors concluded that Tvp38 associates with the tSNAREs in Tlg2-containing compartments, suggesting a role in membrane transport. Investigations into the bacterial and archaeal prevalence of these proteins showed that 90% of bacterial species and 70% of archaeal species encoded proteins with the PF09335 domain (Doerrler *et al.*, 2013). Bacterial and archaeal PF09335-containing proteins are collectively known as the DedA family (Doerrler *et al.*, 2013; Nonet *et al.*, 1987). Detailed studies of the *Escherichia coli* DedA proteins have indicated that there are eight *E. coli* representatives of the DedA family (YqjA, YghB, YabI, YohD, DedA, YdjX, YdjZ, and YqaA) with overlapping functions (Doerrler *et al.*, 2013; Keller & Schneider, 2013), with YdjX and YdjZ being the most closely related to human Tmem41b in terms of sequence similarity (Doerrler *et al.*, 2013). Phenotypically, DedA knock-out *E. coli* cells display increased temperature sensitivity, cell division defects, activation envelope stress pathways, compromised proton motive force, sensitivity to alkaline pH and increased antibiotic susceptibility

(Doerfler *et al.*, 2013; Keller *et al.*, 2014). As *E. coli* expresses multiple DedA homologues, lethal effects are not observed as long as at least one DedA is expressed (Kumar & Doerfler, 2014; Thompkins *et al.*, 2008). *Borrelia burgdorferi* contains only one DedA protein in its genome and knockout cells display the same phenotype as the *E. coli* knockout strains. The *B. burgdorferi* homologue is indeed essential (Liang *et al.*, 2010). Interestingly, *E. coli* knockout cells can be rescued with the *B. burgdorferi* homologue that shows only 19% sequence identity with YqjA. The functions of DedA have also been studied in the pathogen *Burkholderia thailandensis* where one family member was found to be required for resistance to polymyxin (Panta *et al.*, 2019).

Until the structure of poorly characterised protein families such as Pfam family PF09335 can be elucidated experimentally, *ab initio* protein modelling can be used to predict a fold allowing for structure-based function inferences (Rigden *et al.*, 2017). Such methods have made significant strides recently due to the availability of contact predictions (Kinch *et al.*, 2016). Prediction of residue-residue contacts relies on the fact that each pair of contacting residues covaries during evolution. The process of co-variation occurs as the properties of the two residues complement each other in order to maintain structural integrity of that local region and, consequently, its original functionality. Therefore, if one residue from the pair is replaced, the other must also change to compensate the physical chemical variation and hence preserve the original structure (Lapedes *et al.*, 1999). The link between two residues can be then reliably detected in multiple sequence alignments by using direct coupling analysis (Morcos *et al.*, 2011) as well as machine learning algorithms (Wu *et al.*, 2020). The predicted contacts can be used for a range of analyses such as the identification of domain boundaries (Rigden, 2002; Simkovic *et al.*, 2017a), but their main application is for contact-based modelling methods which can address larger targets than conventional fragment-assembly-based *ab initio* methods (Yang *et al.*, 2020). Contact-based modelling methods have been proven successful previously in modelling membrane proteins (Hopf *et al.*, 2012).

In the current study, we first linked the Pfam PF09335 family to the PF06695 family and chose a conveniently small Archaeal sequence and then utilised state of the art methods to make structural predictions for not only the Archaeal sequence but also for two prominent members of the Pfam family PF09335 (Tmem41b and YqjA) by exploiting data derived from sequence, evolutionary covariance and *ab initio* modelling. We are able to predict that both PF09335 homologues (DedA proteins) and PF06695 homologues contain re-entrant loops (stretches of protein that enter the bilayer but exit on the same side of the membrane) as well as a pseudo-inverted repeat topology. The predicted presence of both of these structural features strongly suggests that DedA proteins are secondary active transporters for an uncharacterised substrate.

## Methods

### Multiple Sequence Alignment

A multiple sequence alignment was generated using PSI/TM-COFFEE variant (RRID:SCR\_019024) with default settings (Floden *et al.*, 2016).

### Pfam database screening

Searches using the sequences of DedA domain proteins Tmem41b, YqjA, YdjX, Ydjz, Tvp38 and Mt2055 were made against the Pfam-A\_v32.0 (RRID:SCR\_004726) (El-Gebali *et al.*, 2019) database using the HHPred (RRID:SCR\_010276) v3.0 server (Zimmermann *et al.*, 2018) with default parameters (-p 20 -Z 10000 -loc -z 1 -b 1 -B 10000 -ssm 2 -sc 1 -seq 1 -dbstrlen 10000 -norealign -maxres 32000 -contxt /cluster/toolkit/production/bioprog/tools/hh-suite-build-new/data/context\_data.crf) and eight iterations for MSA generation in the HHblits (Remmert *et al.*, 2012) stage.

### Contact map predictions

The DeepMetapsicov v1.0 server (Kandathil *et al.*, 2019) was used to generate contact predictions with ConKit v0.12 (Simkovic *et al.*, 2017b) utilised to visualise the contact maps. ConPlot (RRID:SCR\_019216) was used to overlay additional prediction data (Sánchez Rodríguez *et al.*, 2021).

### Other prediction data

Transmembrane helical topology predictions were obtained from the Topcons server (Tsirigos *et al.*, 2015). Secondary structure predictions were made employing a local installation of PSIPRED (RRID:SCR\_010246) v4.0 (McGuffin *et al.*, 2000). ConKit was also used to predict and visualise potential structural domain boundaries (Rigden, 2002; Simkovic *et al.*, 2017a). Residue analysis of putative amphipathic regions were performed using HELIQUEST (Gautier *et al.*, 2008) to determine the presence, direction and magnitude of any hydrophobic moment. Residue conservation was determined using the ConSurf server (Ashkenazy *et al.*, 2016).

### Dataset for custom re-entrant database

A library of re-entrant loop pdb structures together with the putative re-entrant loop structures from the query protein models were clustered on their structural similarity. The library was built by obtaining a non-redundant (removing redundancy with a 40% sequence identity threshold) set of 125 chains from the PDBTM (RRID:SCR\_011962) (Kozma *et al.*, 2013) that contain at least one re-entrant loop. As this investigation focuses on re-entrant loops that are immediately preceded by a TM helix that is packed against the loop, all re-entrant loops (boundaries defined by PDBTM) in addition to the preceding 30 residues were extracted. The resulting 193 library entries ([https://figshare.com/articles/dataset/repository\\_zip/14055212](https://figshare.com/articles/dataset/repository_zip/14055212)), supplemented with the re-entrant loop features (defined by the OMP server (Lomize *et al.*, 2012) and accompanied by the preceding 30 residues) from the *ab initio* modelling underwent an all-against-all structural alignment using a local installation of Dali v4.0 (Holm & Laakso, 2016). The Z-scores for these alignments were then used for clustering with CLANS v1.0 (Frickey & Lupas, 2004) with a Z-score of 4.5 used as the cut-off threshold.

### Model building

*Ab initio* models were built using the trRosetta (Yang *et al.*, 2020) server with default settings. Conservation was mapped on to the models using the ConSurf server (Ashkenazy *et al.*, 2016). Visualisation of models was achieved using PyMOL (RRID:SCR\_000305) v2.3.0 (DeLano, 2002).

## Structural alignments

Dali (RRID:SCR\_013433) v4.0 (Holm & Laakso, 2016) was used to structurally align the output models and to query against the PDBTM (Kozma *et al.*, 2013).

An earlier version of this article can be found on bioRxiv (doi: <https://doi.org/10.1101/2020.06.27.174763>)

## Results and discussion

### Sequence comparisons suggest Pfam families PF09335 and PF06695 are related

HHpred (Zimmermann *et al.*, 2018) was used to screen a selection of DedA proteins against the Pfam database (El-Gebali *et al.*, 2019). Hits were observed in the same region against both PF09335 and the Pfam domain PF06695 ('Sm\_multidrug\_ex') which is strongly indicative of homology: a probability of 99.4% with an E-value of 9E-17 for the PF09335 hit and 98.3% and 2E-10 respectively for PF06695. A HHpred search against the Pfam database using a member of PF06695 - the short archaeal sequence Mt2055 (UniProt code W9DY28) (Apweiler *et al.*, 2004) - returned similar results (Table 1). Figure 1 shows the MSA for the same sequences along with the matched regions of the two Pfam domains under investigation. The Mt2055 sequence originates from the unpublished draft genome of the archaeobacterium *Methanobolus tindarius* DSM 2278. For many of the subsequent analyses, the shorter archaeal sequence was used initially but the clear homology among this set of proteins means that inferences can be drawn across the group.

There are no known experimental protein structures representing PF09335 or PF06695, but both Gremlin and DMPfold have constructed *ab initio* models for these Pfam domains (Greener *et al.*, 2019; Ovchinnikov *et al.*, 2017).

### The predicted Pfam domains are inconsistent with a structural domain

Analysis of the HHpred results obtained for the archaeal protein Mt2055 revealed the presence of additional hits for both

PF06695 and PF09335 Pfam domains, in which the C-terminal half of the domains aligned with the N-terminal half of the Archaea protein. For example, residues 1-69 of the archaeal protein aligned with residues 52-117 of the Pfam PF09335 profile with a probability of 74.15%. Interestingly, contact density analysis (Rigden, 2002; Sadowski, 2013) supported the existence of a domain boundary around residue 60, in broad agreement with the HHpred results (Figure 2). Both the HHpred and contact density results therefore pointed to a specific domain structure being present.

### Sequence & contact prediction map analysis indicate that PF06695 is made up of a tandem repeat

When the Mt2055 sequence was split at residue 60-61, the resulting N-terminal region of 60 residues and the C-terminal section of 79 residues could be aligned using HHalgn (Soding, 2005) with a 78% probability and an E-value of 1.9E-3. Examination of the map of predicted contacts for Mt2055 reveals features that are present in both the N- and C-terminal halves of the protein (Figure 2c). Taken together, these data strongly support the existence of a tandem repeat within the Mt2055 protein and hence across the PF06695 and PF09335 protein families.

Interestingly, an equivalent sequence analysis with HHpred of other PF09335 homologues including Tmem41b itself does not reveal a repeat. However, inspection of their corresponding predicted contact maps does reveal features repeated when N- and C-halves of the protein are compared (Figure 3). Apparently, evolutionary divergence has removed all trace of the repeat sequence signal in bacterial and eukaryotic proteins, although the feature remains visible by evolutionary covariance analysis.

### *Ab initio* modelling of Mt2055 reveals an unusual topology

Several authors have deposited structures of uncharacterised Pfam families in databases (El-Gebali *et al.*, 2019); however, Pfam domain boundaries for PF09335/PF06695, which define the limits of these previous modelling exercises, do

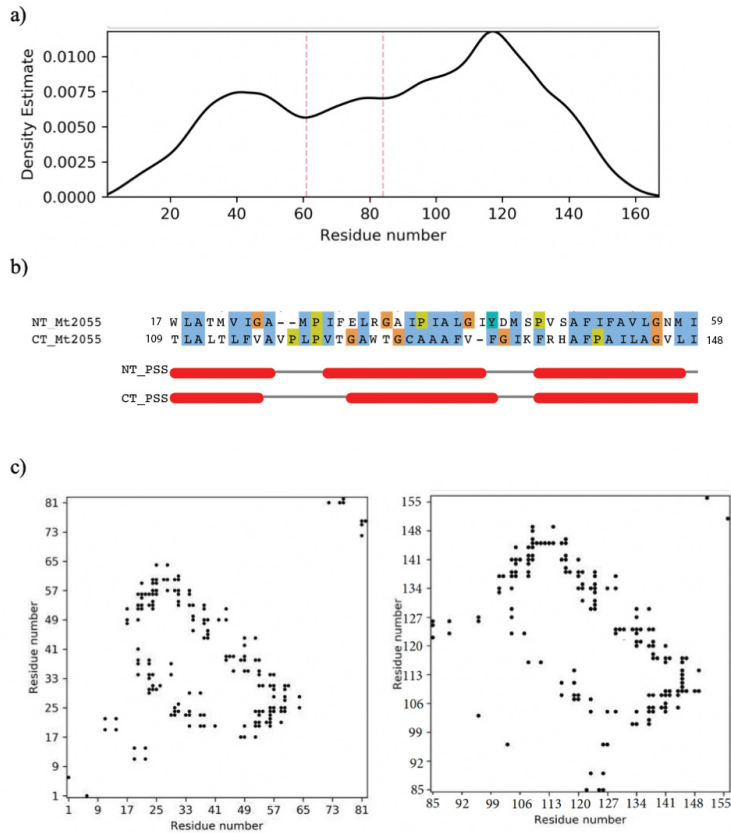
**Table 1.** HHpred results for Tmem41b and homologues demonstrate homology between Pfam families PF09335 and PF06695.

	Species	UniProt Code	Length	PF09335 'SNARE_ASSOC'/'VTT'/'Tvp38' /DedA		PF06695 'Sm_multidrug_ex'	
				Probability	E-Value	Probability	E-Value
Tmem41b	<i>Homo sapiens</i>	Q5BJD5	291	99.4	9E-17	98.3	2E-10
YdjX	<i>Escherichia coli</i>	P76219	236	99.6	2.1E-17	99.1	9.9E-13
Ydjz	<i>Escherichia coli</i>	P76221	235	99.6	1.1E-17	99.0	4.5E-16
YqjA	<i>Escherichia coli</i>	P0AA63	220	99.62	5.6E-15	99.41	1.3E-12
Tvp38	<i>Saccharomyces cerevisiae</i>	P36164	337	99.4	7.9E-15	98.7	2.7E-10
Mt2055	<i>Methanobolus tindarius</i>	W9DY28	168	99.0	2.4E-10	99.8	1.8E-20

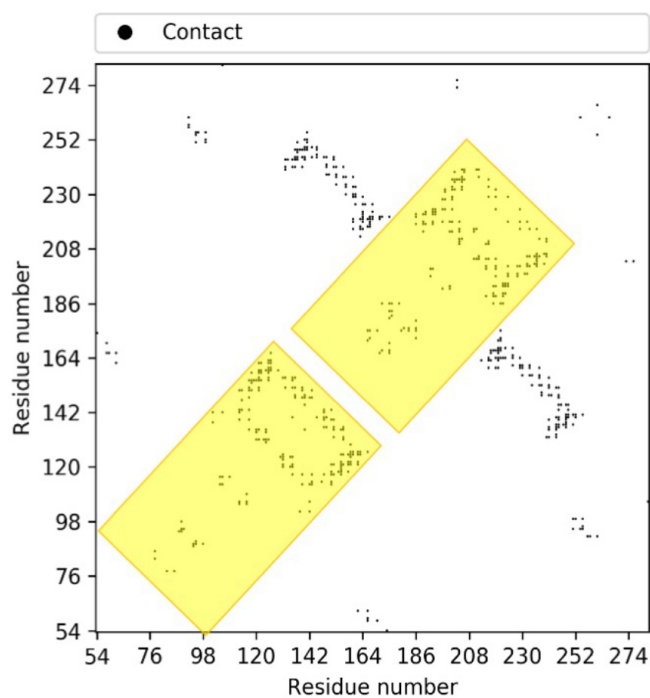




**Figure 1. Multiple Sequence Alignment for query protein selection listed in Table 1.** Magenta highlights the regions matched by HHpred to the PF09665 Pfam domain while purple is used for additional residues included in the PF09335 Pfam domain matches. The black boxed regions represent the locations of the putative re-entrant loops as identified by the modeling of the respective proteins. The secondary structure for the archaeal W9DY29 sequence (Mt2055) is also depicted with the relative positions of alpha helices shown as red blocks.



**Figure 2. Mt2055 domain analysis.** (a) Contact density profile constructed by ConKit (Simkovic et al., 2017b) utilising DeepMetaPSICOV contact prediction. Solid black line represents contact density and dotted red lines mark density minima corresponding to possible domain boundaries. (b) HAlign alignments for the N-terminal and C-terminal Mt2055 halves, formatted using Jalview (Waterhouse et al., 2009) and coloured according to the ClustalX scheme. Red bars represent helical secondary structure. (c) Maps of predicted contacts generated by DeepMetaPSICOV and plotted using ConKit; left is N-terminal half (residues 1-84) and right is C-terminal half (residues 85-168). Black points represent predicted intramolecular contacts.



**Figure 3. Tmem41b Contact map constructed using DeepMetaPSICOV and plotted using Conkit.** The highlighted areas represent repeat units that have been revealed through evolutionary covariance analysis.

not reflect the conserved structural domain that we predict. Given the fact that the available *ab initio* models were inconsistent with the transmembrane helix, secondary structure and contact predictions, we constructed our own models of Mt2055 as well as Tmem41b and YqjA with trRosetta. ([https://figshare.com/articles/dataset/repository\\_zip/14055212](https://figshare.com/articles/dataset/repository_zip/14055212))

The Mt2055, Tmem41b and YqjA models had estimated TM scores from the trRosetta server of 0.633, 0.624 and 0.635 respectively, suggesting that they were likely to have captured the native fold of the family. All-against-all pairwise structural superposition of the models with DALI gave a mean Z-score of 11.9 confirming their strong similarity. We also used satisfaction of predicted contacts to validate the models (Figure 4) (Simkovic *et al.*, 2017a). This showed that 80% of the top  $L$  predicted contacts (where  $L$  is the length of the protein) are satisfied by the model contacts for both Mt2055 and YqjA and a value of 60% was achieved for Tmem41b suggestive of good quality models (de Oliveira *et al.*, 2017).

The models (Figure 3) contained interesting features: two inversely symmetrical repeated units each possessing a helix lying parallel to the membrane surface (green) and a re-entrant loop (orange) packed with a TM helix (red).

The presence of a re-entrant loop packed against each TM helix can also be seen on predicted contact maps for these proteins (Figure 4b). Interestingly, each of the re-entrant helices are

predicted as a single transmembrane region in the TopCons predictions. When cross-referenced with the PSIPRED secondary structure prediction it is noted that there is a predicted two-residue region of coil around the mid-point of the first TM helix prediction. A similar observation can be made for the fourth TM helix prediction with the equivalent coil region being six residues in length (see the diagonal of Figure 4b) Such a prediction would more obviously be treated as indicative of some kind of kink in the helix (Law *et al.*, 2016) but the explanation here is that these regions form re-entrant helices. Similar contact map features, indicative of re-entrant loops packing against TM helices, can be seen clearly on the contact maps of other DedA proteins (data not shown). The MSA in Figure 1 shows the relative positions of the re-entrant loops in their respective sequences.

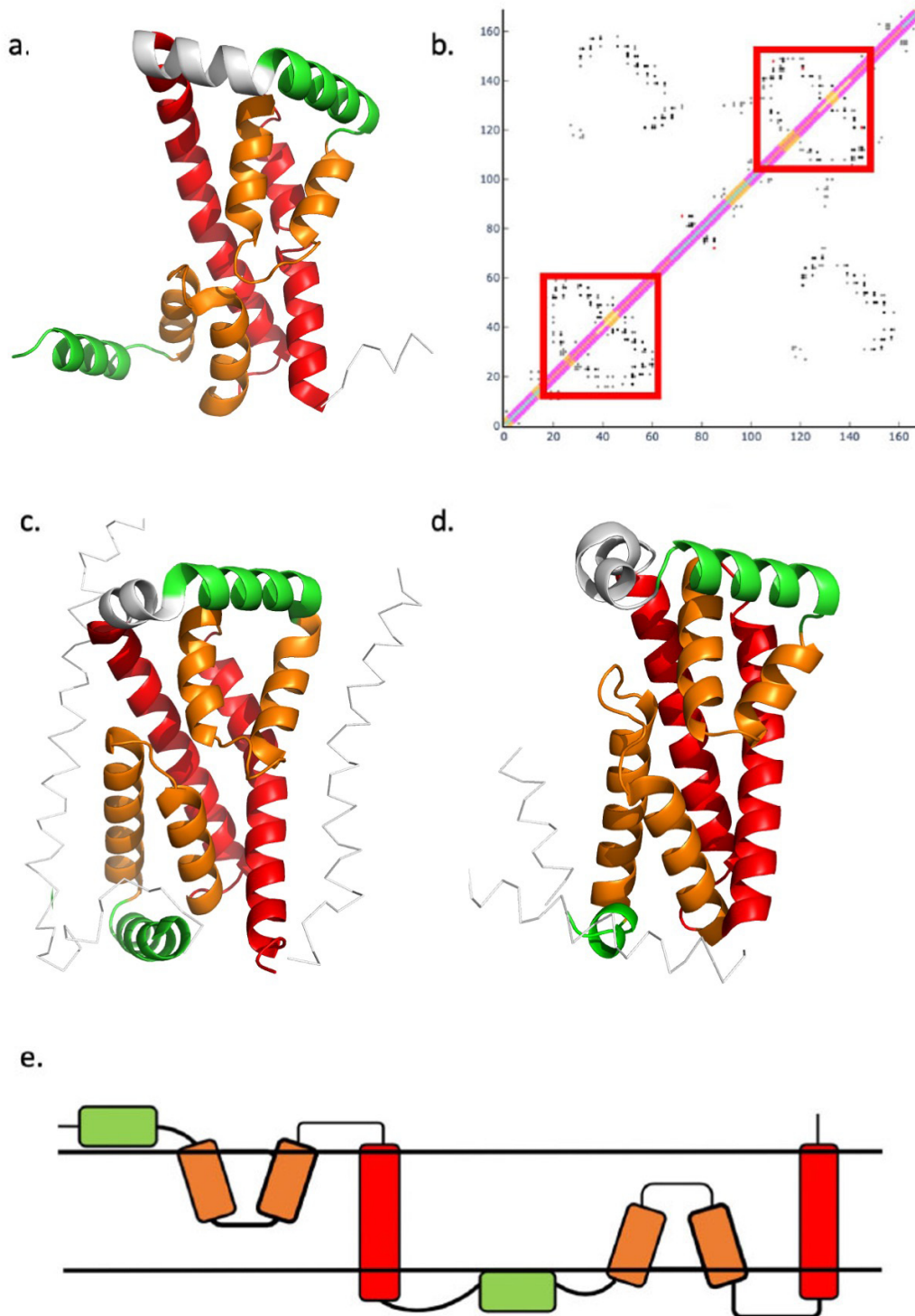
In order to test for test whether the membrane-parallel helices (green in Figure 3) were amphipathic, an analysis of helical wheel diagrams for the fifteen residues preceding the putative re-entrant loops was performed with HELIQUEST (Gautier *et al.*, 2008). The quantitative measures of the hydrophobic moment for the regions being analysed (Figure 5) support that they are indeed amphipathic helices. The hydrophobic moments ranged from 0.298 to 0.546.

The predicted presence of the *amphipathic-re-entrant loop-TM helix* features in DedA domain proteins prompted a desire to map sequence conservation on to the *ab initio* models. Using the ConSurf server to perform the mapping of sequence conservation onto the query models, it revealed that the re-entrant loop sequences are highly conserved. The high sequence conservation of re-entrant loops indicate that they are likely to be functionally and/or structurally important (Figure 6).

### Re-entrant loops are also present in Cl/H<sup>+</sup> Antiporters

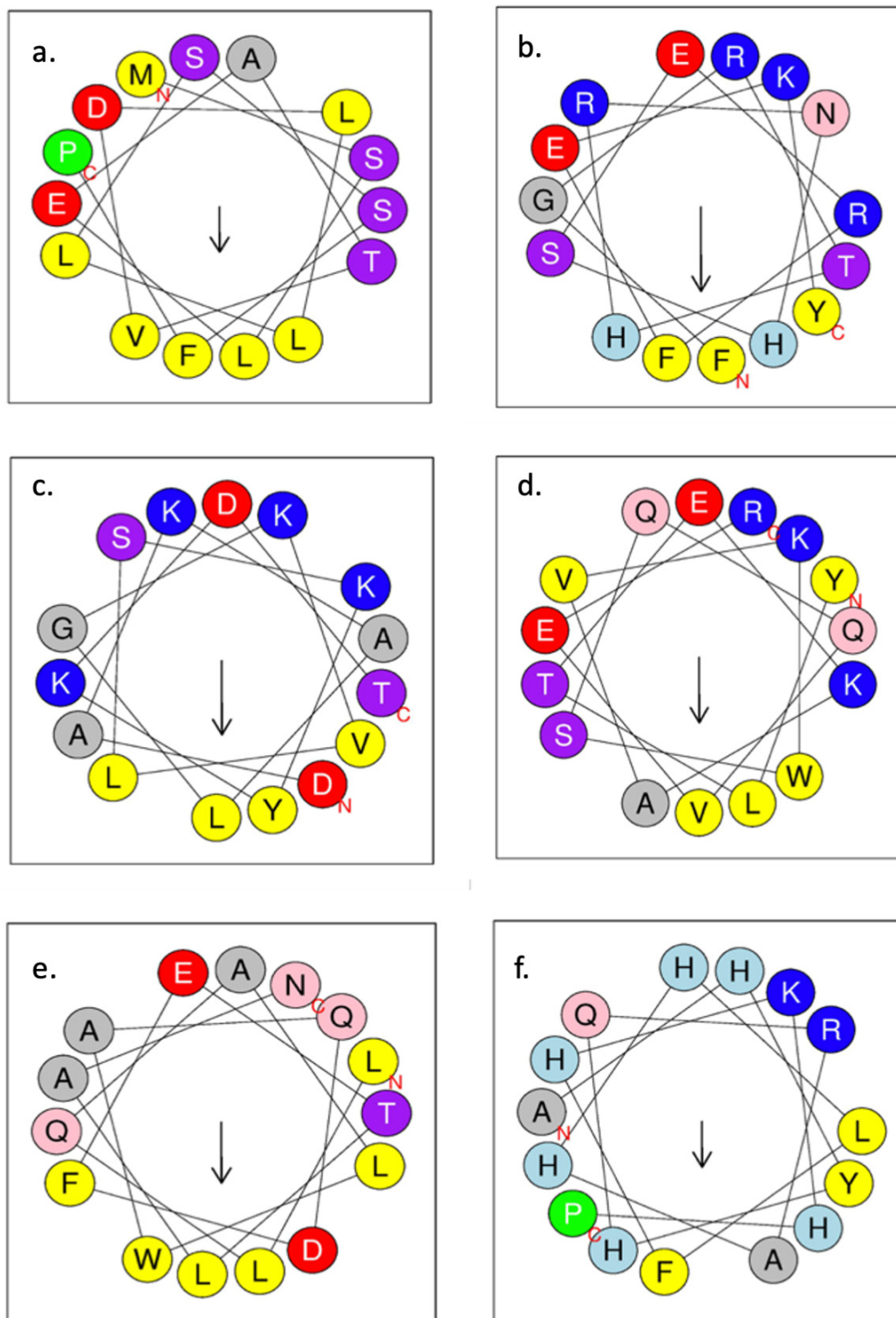
The presence of re-entrant loops and the high density of conserved residues within them caused us to examine experimentally characterised re-entrant loops in the PDBTM database. A total of 193 non-redundant re-entrant helices were identified (see *Methods*). All 193 were clustered with the putative re-entrant loops from Mt2055, Tmem41b and YqjA relative z-scores derived from an all-against-all DALI run and subsequently clustered in CLANS (Frickey & Lupas, 2004) with a z-score cut-off of 4.5.

As expected all six re-entrant structures from the query models clustered together. The CLC transporter re-entrant structures of 3orgA (re-entrant 1 and re-entrant 2), 7bxu and 5tqq also clustered with the queries. Additionally, the re-entrant structure from an Undecaprenyl pyrophosphate phosphatase (UppP) (6cb2) also clustered with the queries. UppP is an integral membrane protein that recycles lipid and has structural similarities to CLC transporters (Workman *et al.*, 2018). Contact maps derived from the pdb files of CLC and UppP structures show the contact map signature corresponding to the re-entrant/TM helix structural feature. Interestingly, the UppP is more similar to the query proteins being only 271 residues in length and having only 6 TM helices.

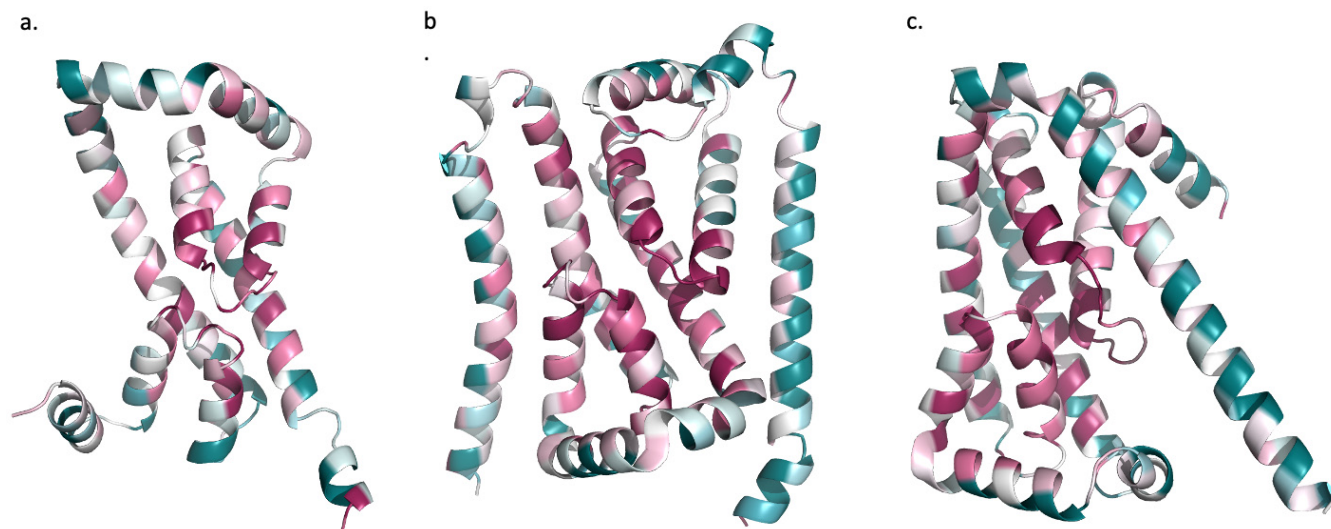


**Figure 4.** (a) trRosetta model of MT2055 - amphipathic helix (green) and a re-entrant loop (orange) packed with a TM helix (red) (b) Superposition of DMP predicted contact map for Mt2055 and contacts from the Mt2055 model. Black points are matching contacts, red are mismatches and grey are contacts predicted but not present in the model. Diagonal is a visual representation of transmembrane helix and secondary structure prediction - central diagonal is the visualisation of the TopCons transmembrane prediction (orange being a TM helix) and the outer diagonals are the visual representation of the PSIPRED secondary structure prediction (pink - alpha helix and yellow - coil). Red boxes highlight the re-entrant loop and TM helix packing contact map signature. (c) trRosetta model of Tmem41b only showing the conserved structural domain (residues 39-217) (d) trRosetta model of YqjA only showing the conserved structural domain (residues 14-176). (e) Proposed topology for (extended) DedA domain.





**Figure 5. Helical wheel diagrams generated using the HELIQUEST server.** Hydrophobic residues are shown in yellow, serine and threonine in purple, basic residues in dark blue, acidic residues in red, asparagine and glutamine in pink, alanine and glycine in grey, histidine in light blue and proline in green circles. Arrows represent direction and magnitude of the hydrophobic moment and residue marked with 'N' is the N-terminal end of the putative amphipathic helix with the residue marked 'C' being the C-terminal end. (a) Mt2055 putative amphipathic helix 1 (hydrophobic moment of 0.298). (b) Mt2055 putative amphipathic helix 2 (hydrophobic moment of 0.546). (c) Tmem41b putative amphipathic helix 1 (hydrophobic moment of 0.471). (d) Tmem41b putative amphipathic helix 2 (hydrophobic moment of 0.420). (e) YqjA putative amphipathic helix 1 (hydrophobic moment of 0.295). (f) YqjA putative amphipathic helix 2 (hydrophobic moment of 0.396).



**Figure 6.** trRosetta models with ConSurf conservation mapping for (a) Mt2055 (b) Tmem41b (c) YqjA. Conservation is shown as a spectrum from purple (highly conserved) to blue (not conserved).

Analysis of the Cl<sup>-</sup>/H<sup>+</sup> antiporter structures show that they contain a similar inverted repeat as we infer for the DedA homologues, resulting in pseudo-2-fold axis of symmetry running along the membrane (Duran & Meiler, 2013). Again similarly, the Cl<sup>-</sup>/H<sup>+</sup> antiporter 3orgA also contains the amphipathic helices on the N-terminal side of the re-entrant loops. The fact that the presence of the amphipathic helices is restricted only to 3orgA and not found in all homologues suggest that these features are not essential for function (Figure 7).

#### A possible antiporter role for DedA proteins

The presence of re-entrant loops in a transmembrane protein strongly indicates a transporter or pore functionality since this structural feature has, hitherto, only been found in proteins of this kind (Yan & Luo, 2010). The structural similarities between the DedA proteins and the Cl<sup>-</sup>/H<sup>+</sup> antiporters raise the possibility that the families studied here are, in fact, unsuspected distant homologues having this putative pore feature in common. In that regard it is relevant to recall a hypothesis that DedA proteins are H<sup>+</sup> antiporters resulting from site directed mutagenesis (SDM) experiments (Kumar & Doerrler, 2014; Kumar *et al.*, 2016).

A recent study has identified key residues (Figure 8) in the *E. coli* DedA protein YqjA that, when replaced in site directed mutagenesis experiments, resulted in properly folded (membrane localized) but non-functional proteins unable to complement alkaline pH sensitivity of *E. coli* YqjA mutant and antibiotic sensitivity of YqjA/YghB double mutant (Panta *et al.*, 2019). Highlighting the essential residues (E39, D51, R130 and R136) on the YqjA model show that they come together in three-dimensional space with the N-terminal side of the first re-entrant possessing E39 and the C-terminal side possessing D51. R130 and R136 are similarly positioned on the second

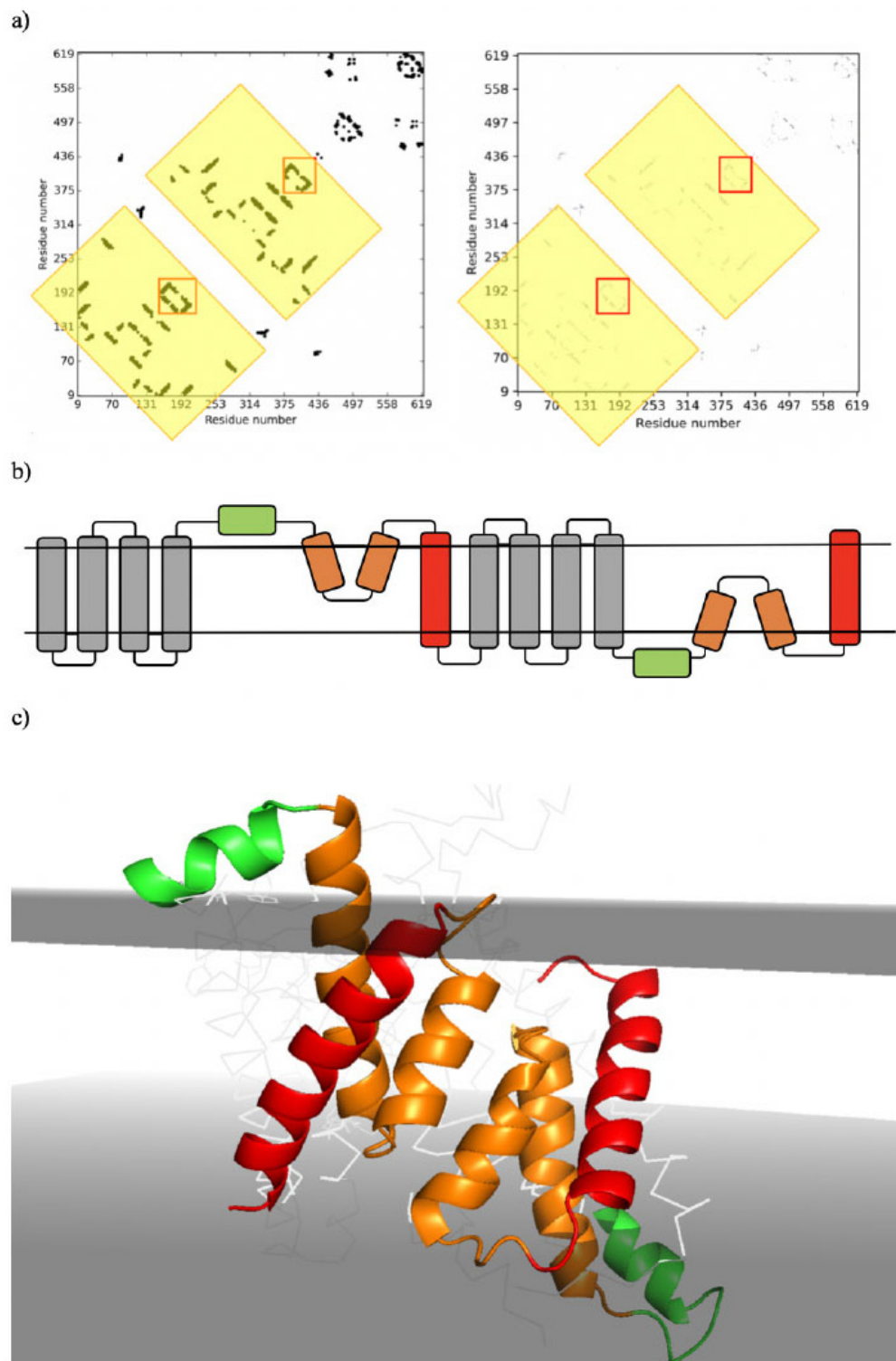
re-entrant loop (Figure 8). Re-entrant loops are known to form pores and here we have two proton-titratable residues (E39, D51) in close proximity to essential basic residues (R130 and R136) within a putative pore. This three-dimensional arrangement of key residues could serve a role in the coupling of the protonation status with the binding of a yet to be characterised substrate as is postulated for the multi-drug H<sup>+</sup> antiporter MdfA (Heng *et al.*, 2015) where these same residues are located inside a central cavity.

#### Conclusions

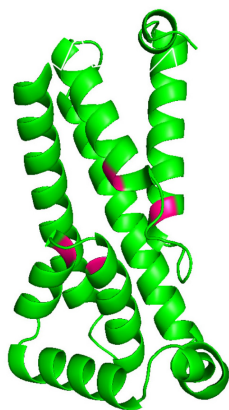
This study demonstrates how covariance prediction data have multiple roles in modern structural bioinformatics: not just by acting as restraints for model making and serving for validation of the final models but by predicting domain boundaries and revealing the presence of cryptic internal repeats not evidenced by sequence analysis. Furthermore, we characterised a contact map feature characteristic of a re-entrant helix which may in future allow detection of this feature in other protein families.

Sequence, co-variance and *ab initio* modelling analyses show that the Pfam PF09335 and PF06695 domains are distantly homologous. These domains contain a structural core composed of a pseudo-inverse repeat of an amphipathic helix, a re-entrant loop and a TM helix. All PF09335 homologues contain this central core with additional TM-helices flanking either side.

Querying the models against the PDB using Dali did not yield any significant hits. However, analysis of the prediction data revealed two features of DedA proteins that independently suggest that they are secondary transporters: both an inverted repeat architecture and the presence of a re-entrant loop, which are both independently and strongly associated with transporter function (Duran & Meiler, 2013; Yan & Luo, 2010).



**Figure 7.** (a) Left - Predicted Contact map with repeating units highlighted in yellow boxes, contact map signature of re-entrant loop packed with TM helix in red boxes.; Right - The Experimental Contact map obtained from the PDB structure with repeating units highlighted in yellow boxes, contact map signature of re-entrant loop packed with TM helix in red boxes. (b) Actual 3orgA topology; grey: TM Helices that are additional to the core; red: TM helices contributing to the formation of the core; orange; re-entrant loops contributing to the formation of the core; green: amphipathic helices contributing to the formation of the core. (c) The 2-fold pseudo symmetry of the amphipathic/re-entrant loop/TM helix core inverted repeat structure of 3orgA with membrane positions shown as grey planes obtained from PDBTM.



**Figure 8.** Essential residues determined by SDM experiments highlighted in pink on a truncated YqjA model.

Additionally, the fact that DedA proteins show structural similarities with H<sup>+</sup> antiporters indicate that these proteins may also couple substrate transport with an opposing H<sup>+</sup> current. Indeed, the YqjA homologue also contains strategically placed residues known to be involved in H<sup>+</sup> antiporter activity. The *ab initio* models show that the essential residues come together in the region that would be buried in the membrane potentially forming a substrate chamber consistent with the transport of a specific substrate. Further research needs to be carried out to determine what this substrate is and confirm the mechanism of transport.

### Data availability

Figshare: Final models and a list of PDB structures used for the clustering exercise <https://doi.org/10.6084/m9.figshare.14055212.v1> (Mesdaghi, 2021)

### Reference

- Almén MS, Nordström KJ, Fredriksson R, *et al.*: Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.* 2009; **7**: 50.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Apweiler R, Bairoch A, Wu CH, *et al.*: UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 2004; **32**(Database issue): 115D–119.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ashkenazy H, Abadi S, Martz E, *et al.*: ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* 2016; **44**(W1): W344–W350.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- de Oliveira SHP, Shi J, Deane CM: Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinformatics.* 2017; **33**(3): 373–381.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- DeLano W: The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, California, USA. 2002.
- Doerfler WT, Sikdar R, Kumar S, *et al.*: New Functions for the Ancient DedA Membrane Protein Family. *J Bacteriol.* 2013; **195**(1): 3–11.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Duran AM, Meiler J: Inverted topologies in membrane proteins: a mini-review. *Comput Struct Biotechnol J.* 2013; **8**(11): e201308004.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- El-Gebali S, Mistry J, Bateman A, *et al.*: The Pfam protein families database in 2019. *Nucleic Acids Res.* 2019; **47**(D1): D427–D432.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Floden EW, Tommaso PD, Chatzou M: PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic Acids Res.* 2016; **44**(W1): W339–W343.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fowler PW, Coveney PV: A computational protocol for the integration of the monotopic protein prostaglandin H2 synthase into a phospholipid bilayer. *Biophys J.* 2006; **91**(2): 401–410.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Frickey T, Lupas A: CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics.* 2004; **20**(18): 3702–3704.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gautier R, Douguet D, Antony B, *et al.*: HELIQUEST: a web server to screen sequences with specific alpha-helical properties. *Bioinformatics.* 2008; **24**(18): 2101–2102.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Greener JG, Kandathil SM, Jones DT: Deep learning extends *de novo* protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun.* 2019; **10**(1): 3977.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grisshammer R, Tateu CG: Overexpression of integral membrane proteins for structural studies. *Q Rev Biophys.* 1995; **28**(03): 315–422.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Heng J, Zhao Y, Liu M, *et al.*: Substrate-bound structure of the *E. coli* multidrug resistance transporter MdfA. *Cell Res.* 2015; **25**(9): 1060–1073.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hoffmann HH, Schneider WM, Rozen-Gagnon K, *et al.*: TMEM41B Is a Pan-flavivirus Host Factor. *Cell.* 2021; **184**(1): 133–148. e20.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Holm L, Laakso LM: Dali server update. *Nucleic Acids Res.* 2016; **44**(W1): W351–W355.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hopf TA, Colwell LJ, Sheridan R, *et al.*: Three-dimensional structures of membrane proteins from genomic sequencing. *Cell.* 2012; **149**(7): 1607–1621.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Inadome H, Noda Y, Kamimura Y, *et al.*: Tvp38, Tvp23, Tvp18 and Tvp15: Novel membrane proteins in the Tlg2-containing Golgi/endosome compartments of *Saccharomyces cerevisiae*. *Exp Cell Res.* 2007; **313**(4): 688–697.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kandathil S, Greener J, Jones D: Prediction of inter-residue contacts with DeepMetaPSICOV in CASP13. *BioRxiv.* 2019; 586800.  
[Publisher Full Text](#)
- Keller R, Schneider D: Homologs of the yeast Tvp38 vesicle-associated protein are conserved in chloroplasts and cyanobacteria. *Front Plant Sci.* 2013; **4**: 467.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Keller R, Ziegler C, Schneider D: When two turn into one: evolution of membrane transporters from half modules. *Biol Chem.* 2014; **395**(12): 1379–1388.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Khafizov K, Madrid-Aliste C, Almo SC, *et al.*: Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc Natl Acad Sci U S A.* 2014; **111**(10): 3733–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kinch LN, Li W, Monastyrskyy B, *et al.*: Assessment of CASP11 contact-assisted predictions. *Proteins.* 2016; **84** Suppl 1(Suppl 1): 164–180.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kozma D, Simon I, Tusnády GE: PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* 2013; **41**(Database issue): D524–D529.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kumar S, Bradley CL, Mukashyaka P, *et al.*: Identification of essential arginine residues of *Escherichia coli* DedA/Tvp38 family membrane proteins YqjA and YghB. *FEMS Microbiol Lett.* 2016; **363**(13): fnw133.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kumar S, Doerfler WT: Members of the conserved DedA family are likely membrane transporters and are required for drug resistance in *Escherichia coli*. *Antimicrob Agents Chemother.* 2014; **58**(2): 923–930.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lapedes A, Giraud B, Liu L, *et al.*: Correlated mutations in models of protein sequences: phylogenetic and structural effects. *JSTOR.* 1999; **33**: 236–256.  
[Reference Source](#)



- Law EC, Wilman HR, Kelm S, *et al.*: **Examining the Conservation of Kinks in Alpha Helices.** *PLoS One.* 2016; **11**(6): e0157553.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Liang FT, Xu Q, Sikdar R, *et al.*: **BB0250 of *Borrelia burgdorferi* is a conserved and essential inner membrane protein required for cell division.** *J Bacteriol.* 2010; **192**(23): 6105–6115.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lomize MA, Pogozheva ID, Joo H, *et al.*: **OPM database and PPM web server: Resources for positioning of proteins in membranes.** *Nucleic Acids Res.* 2012; **40**(Database issue): D370–6.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lotti F, Imlach WL, Saieva L, *et al.*: **An SMN-Dependent U12 Splicing Event Essential for Motor Circuit Function.** *Cell.* 2012; **151**(2): 440–454.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McGuffin LJ, Bryson K, Jones DT: **The PSPRED protein structure prediction server.** *Bioinformatics.* 2000; **16**(4): 404–405.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mesdaghi S: **repository.zip.** *figshare.* Dataset. 2021.  
<http://www.doi.org/10.6084/m9.figshare.14055212.v1>
- Morcos F, Pagnani A, Lunt B, *et al.*: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families.** *Proc Natl Acad Sci U S A.* 2011; **108**(49): E1293–301.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Moretti F, Bergman P, Dodgson S, *et al.*: **TMEM41B is a novel regulator of autophagy and lipid mobilization.** *EMBO Rep.* 2018; **19**(9): e45889.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Morita K, Hama Y, Izume T, *et al.*: **Genome-wide CRISPR screen identifies *TMEM41B* as a gene required for autophagosome formation.** *J Cell Biol.* 2018; **217**(11): 3817–3828.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Morita K, Hama Y, Mizushima N: **TMEM41B functions with VMP1 in autophagosome formation.** *Autophagy.* 2019; **15**(5): 922–923.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nonet ML, Marvelj CC, Tolanlt DR: **The *hist-purF* Region of the *Escherichia coli* K-12 Chromosome. IDENTIFICATION OF ADDITIONAL GENES OF THE *hist* AND *purF* OPERONS.** *J Biol Chem.* 1987; **262**(25): 12209–17.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ovchinnikov S, Park H, Varghese N, *et al.*: **Protein structure determination using metagenome sequence data.** *Science.* 2017; **355**(6322): 294–298.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Panta PR, Kumar S, Stafford CF, *et al.*: **A DedA Family Membrane Protein Is Required for *Burkholderia thailandensis* Colistin Resistance.** *Front Microbiol.* 2019; **10**: 2532.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Remmert M, Biegert A, Hauser A, *et al.*: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nat Methods.* 2012; **9**(2): 173–5.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Resh MD: **Lipid Modification of Proteins.** *Biochemistry of Lipids, Lipoproteins and Membranes.* 2016; 391–414.  
[Publisher Full Text](#)
- Rigden DJ: **Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments.** *Protein Eng.* 2002; **15**(2): 65–77.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rigden DJ, Cymerman IA, Bujnicki JM: **Prediction of protein function from theoretical models.** In *From Protein Structure to Function with Bioinformatics: Second Edition.* Springer Netherlands. 2017; 467–498.  
[Publisher Full Text](#)
- Sadowski MI: **Prediction of protein domain boundaries from inverse covariances.** *Proteins.* 2013; **81**(2): 253–260.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sánchez Rodríguez F, Mesdaghi S, Simpkin AJ, *et al.*: **ConPlot: Web-based application for the visualisation of protein contact maps integrated with other data.** *Bioinformatics.* 2021; btab049.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Scaturro P, Stukalov A, Haas DA, *et al.*: **An orthogonal proteomic survey uncovers novel Zika virus host factors.** *Nature.* 2018; **561**(7722): 253–257.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Schneider WM, Luna JM, Hoffmann HH, *et al.*: **Genome-scale identification of SARS-CoV-2 and pan-coronavirus host factor networks.** *bioRxiv.* 2020; 2020.10.07.326462.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Simkovic F, Ovchinnikov S, Baker D, *et al.*: **Applications of contact predictions to structural biology.** *IUCr.* 2017a; **4**(Pt 3): 291–300.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Simkovic F, Thomas JMH, Rigden DJ: **ConKit: a python interface to contact predictions.** *Bioinformatics.* 2017b; **33**(14): 2209–2211.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Soding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics.* 2005; **21**(7): 951–960.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sojo V, Dessimoz C, Pomiankowski A, *et al.*: **Membrane Proteins Are Dramatically Less Conserved than Water-Soluble Proteins across the Tree of Life.** *Mol Biol Evol.* 2016; **33**(11): 2874–2884.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tåbara LC, Vincent O, Escalante R: **Evidence for an evolutionary relationship between Vmp1 and bacterial DedA proteins.** *Int J Dev Biol.* 2019; **63**(1–2): 67–71.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Thompkins K, Chattopadhyay B, Xiao Y, *et al.*: **Temperature sensitivity and cell division defects in an *Escherichia coli* strain with mutations in yghB and yqjA, encoding related and conserved inner membrane proteins.** *J Bacteriol.* 2008; **190**(13): 4489–4500.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tsirigos KD, Peters C, Shu N, *et al.*: **The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides.** *Nucleic Acids Res.* 2015; **43**(W1): W401–7.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Van Alstyne M, Lotti F, Dal Mas A, *et al.*: **Stasimon/Tmem41b localizes to mitochondria-associated ER membranes and is essential for mouse embryonic development.** *Biochem Biophys Res Commun.* 2018; **506**(3): 463–470.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Waterhouse AM, Procter JB, Martin DMA, *et al.*: **Jalview Version 2—a multiple sequence alignment editor and analysis workbench.** *Bioinformatics.* 2009; **25**(9): 1189–1191.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Workman SD, Worrall LJ, Strynadka NCJ: **Crystal structure of an intramembranal phosphatase central to bacterial cell-wall peptidoglycan biosynthesis and lipid recycling.** *Nat Commun.* 2018; **9**(1).  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wu T, Hou J, Adhikari B, *et al.*: **Analysis of several key factors influencing deep learning-based inter-residue contact prediction.** *Bioinformatics.* 2020; **36**(4): 1091–1098.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yan C, Luo J: **An Analysis of Reentrant Loops.** *Protein J.* 2010; **29**(5): 350–354.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Yang J, Anishchenko I, Park H, *et al.*: **Improved protein structure prediction using predicted interresidue orientations.** *Proc Natl Acad Sci U S A.* 2020; **117**(3): 1496–1503.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zimmermann L, Stephens A, Nam SZ, *et al.*: **A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.** *J Mol Biol.* 2018; **430**(15): 2237–2243.  
[PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:   

---

## Version 2

Reviewer Report 09 April 2021

<https://doi.org/10.5256/f1000research.55392.r82225>

© 2021 Bassot C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Claudio Bassot** 

Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

Thank you to the authors for address the points I raised in my review. However, there is still a small typo in the first correction.

*The figure mentioned should be Figure 4, not Figure 3.*

*"The models ( Figure 3) contained interesting features: two inversely symmetrical repeated units each possessing an amphipathic helix lying parallel to the membrane surface (green) and a re-entrant loop (orange) packed with a TM helix (red)."*

.....

*"In order to test whether the membrane-parallel helices (green in Figure 3) were amphipathic, an analysis of helical wheel diagrams for the fifteen residues preceding the putative re-entrant loops was performed by with HELIQUEST ( Gautier et al., 2008). The quantitative measures of the hydrophobic moment for the regions being analysed ( Figure 5) support that they are indeed amphipathic helices. The hydrophobic moments ranged from 0.298 to 0.546."*

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Modelling of transmembrane proteins.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 07 April 2021

<https://doi.org/10.5256/f1000research.55392.r82226>

© 2021 Tuszányi G et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Gábor Tuszányi**

Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

**László Dobson**

Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

All but one of our previous comments were responded. We accepted all responses, but authors should responded to this point too:

"The most serious one: As it can be seen on Fig7b-c, 3org contains additional helices that surround the interfacial helix - re-entrant loop - tm structure. Indeed the protein are dimer, where the dimer interface are formed by the re-entrant loops and the additional transmembrane helices surround this core. This arrangement ensure the lipid embedded structure is energetically stable. In the proposed model, re-entrant loops are not wrapped by other helices thus lipids may interact them. This is energetically unfavorable and does not prefer for the suggested function too. The validity of the model should be further investigate by molecular dynamic simulations of lipid embedded structures."

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Topology and structure prediction of transmembrane proteins.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

---

### Version 1

Reviewer Report 20 January 2021

<https://doi.org/10.5256/f1000research.30592.r75805>

© 2021 Bassot C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Claudio Bassot**

Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

The authors model ab-initio Tmem41b and homologues, characterizing them as secondary transporters. The models are reliable and the study is scientifically robust and worthy of indexing.

However, two points are unclear in the text and need to be clarified, plus, some minor changes will improve the readability of the manuscript. Finally, the models could be made available to ensure full reproducibility.

Major:

- The sentence: "The analysis was performed by HELIQUEST (Gautier *et al.*, 2008) which constructed helical wheel diagrams and provided a quantitative measure of the hydrophobic moment for the region being analysed (Figure 4)." is out of context. In that paragraph are described the reentrant helices, shouldn't the sentence (and the figure) be in the paragraph before where are mentioned the amphipathic helices? The figure discussion in the text should be extended.
- The role of Figure 7 is not clear. It is mentioned in the context of the description of the putative active residues but these are not shown in the figure. Moreover, the only reference to structure 3orgA (shown in the figure) is in the previous paragraph but it's not related to Figure 7. The authors should describe Figure 7 better.

Minor

Introduction:

- "but they only have a 2% representation in the Protein Data Bank (PDB) (Koehler Leman *et al.*, 2015)". The number of transmembrane proteins has grown significantly in the past few years. From the statistics of PDB and PDBTM the ratio of membrane proteins appears close to 4% now.

Results:

- Figure 1b could be clearer with the residues numbering on the sequences.
- Figure 5. The colours are misleading because they are the opposite of the standard consurf colouration (blue not conserved, purple conserved). The standard colouration would allow a faster understanding of the figure.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

No

**Are the conclusions drawn adequately supported by the results?**

Yes



**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Modelling of transmembrane proteins.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 10 Mar 2021

**Daniel Rigden**, University of Liverpool, UK

**The authors model ab-initio Tmem41b and homologues, characterizing them as secondary transporters. The models are reliable and the study is scientifically robust and worthy of indexing.**

**However, two points are unclear in the text and need to be clarified, plus, some minor changes will improve the readability of the manuscript. Finally, the models could be made available to ensure full reproducibility.**

Models are now available as now mentioned in the text;

[https://figshare.com/articles/dataset/repository\\_zip/14055212](https://figshare.com/articles/dataset/repository_zip/14055212)

**Major:**

- **The sentence: “The analysis was performed by HELIQUEST (Gautier *et al.*, 2008) which constructed helical wheel diagrams and provided a quantitative measure of the hydrophobic moment for the region being analysed (Figure 4).” is out of context. In that paragraph are described the reentrant helices, shouldn't the sentence (and the figure) be in the paragraph before where are mentioned the amphipathic helices? The figure discussion in the text should be extended.**

Yes, we are in agreement with you, the paragraph did seem out of place as well as unfinished. In response we have re-worked the paragraph and noted the helices as merely membrane-parallel at first mention, deferring the question of amphipathicity.

*“The models ( Figure 3) contained interesting features: two inversely symmetrical repeated units each possessing an amphipathic helix lying parallel to the membrane surface (green) and a re-entrant loop (orange) packed with a TM helix (red).”*

.....

*“In order to test whether the membrane-parallel helices (green in Figure 3) were amphipathic, an analysis of helical wheel diagrams for the fifteen residues preceding the putative re-entrant loops was performed by with HELIQUEST ( Gautier *et al.*, 2008). The quantitative measures of the hydrophobic moment for the regions being analysed ( Figure 5) support that they are indeed amphipathic helices. The hydrophobic moments ranged from 0.298 to 0.546.”*

- **The role of Figure 7 is not clear. It is mentioned in the context of the description of the putative active residues but these are not shown in the figure. Moreover, the only reference to structure 3orgA (shown in the figure) is in the previous paragraph but it's not related to Figure7. The authors should describe Figure 7 better.**

Thank you for pointing out the ambiguity of Figure 7. There was an error in the figure numbering resulting in Figure 7 not being cited in the main text. This has now been amended and the relevance of figure 7 is highlighted in the following paragraph; *"Analysis of the Cl<sup>-</sup>/H<sup>+</sup> antiporter structures show that they contain a similar inverted repeat as we infer for the DedA homologues, resulting in pseudo-2-fold axis of symmetry running along the membrane (Duran & Meiler, 2013). Again similarly, the Cl<sup>-</sup>/H<sup>+</sup> antiporter 3orgA also contains the amphipathic helices on the N-terminal side of the re-entrant loops. The fact that the presence of the amphipathic helices is restricted only to 3orgA and not found in all homologues suggest that these features are not essential for function (Figure 7)."*

### Minor

#### Introduction:

- **"but they only have a 2% representation in the Protein Data Bank (PDB) (Koehler Leman et al., 2015)". The number of transmembrane proteins has grown significantly in the past few years. From the statistics of PDB and PDBTM the ratio of membrane proteins appears close to 4% now.**

As suggested using the PDB & PDBTM stats we can see the ratio of membrane proteins is above the 2% previously quoted. The PDBTM has 5785 entries with a total of 174507 entries for the PDB. I have updated the introduction accordingly;

*"For instance, membrane proteins are encoded by 30% of the protein-coding genes of the human genome (Almén et al., 2009), but they only have a 3.3% representation in the Protein Data Bank (PDB) (5785 membrane proteins out of 174507 PDB entries)."*

#### Results:

- **Figure 1b could be clearer with the residues numbering on the sequences.**  
Yes agreed. Numbering has been added to figure 1b (now 2b, due to the inclusion of an additional figure in the revised manuscript). It can be seen that the additional detail of the numbering makes it easier to cross reference the images that make up this figure. The new figure is shown;
- **Figure 5. The colours are misleading because they are the opposite of the standard consurf colouration (blue not conserved, purple conserved). The standard colouration would allow a faster understanding of the figure.**

We agree. Colouring on the B-factor column directly produces the results we show but the new blue-purple spectrum does seem to be well-adopted. We have therefore replaced the figure and updated the legend to read

*"trRosetta models with Consurf conservation mapping for (a) Mt2055 (b) Tmem41b (c) Yqja. Conservation is shown as a spectrum from purple (highly conserved) to blue (not conserved)."*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 07 January 2021

<https://doi.org/10.5256/f1000research.30592.r75806>

© 2021 Tuszányi G et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### László Dobson

Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

### Gábor Tuszányi

Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

In this manuscript Mesdaghi *et al.* describe the *in silico* structure modeling of three homologous integral membrane proteins Mt2055, Yqja and human Tmem41b. Structure determination of transmembrane proteins lacks behind globular ones for several reasons, giving space for computational tools. The proper use of these tools may unveil important structural aspects of transmembrane proteins but interpretation of results of such analysis should be done carefully. While the generated models in the manuscript are interesting and might be fully or partly true, the sequence analysis and interpretation of the results are problematic.

Major:

- The authors should be more specific about the exact boundaries of Pfam domains in different proteins as well as the sequence relations of proteins presented in Table 1. Please provide multiple sequence alignment for these proteins indicating the localization of the two pfam domains and the proposed re-entrant loops/transmembrane regions in the sequences.
- The authors propose Mt2055 contains a tandem repeat and suggest the duplication is present in Tmem41b and Yqja structure as well even if it is undetectable from sequence analysis. The proposed domain boundary in Figure 1a and arguments for tandem duplication does not seem convincing. The e-value of 1.9E-3 is quite large for the alignment. The authors should rule out that results in their paper may occur purely by chance. Please test the statistical significance of this value by generating pairwise alignments of transmembrane regions of unrelated transmembrane proteins with similar length. Moreover, contact maps for Mt2055 and Tmem41b were generated from the same multiple alignment, and therefore they must be identical/similar. Thus the similarities does not prove the tandem duplication occurred in Tmem41b too.
- Structure modeling of membrane proteins is somewhat different from globular ones for several reasons. It is highly recommended to use specific software for this task or argue why used a non-specific one. On one hand, in general, topology prediction is more accurate than structure modeling and should be used as an input to aid the modeling. The reviewer is not sure the result of a standard *ab initio* structure modeling program is sufficient to question topology prediction results. On the other hand, topology prediction results are different for Tmem41b (6 TM helix) and Mt2055 (4 TM helix). Notably, other consensus topology method (CCTOP) have a similar result for Mt2055 (4 helix), but different for Tmem41b (6 helix). Using a third method (Octopus) a re-entrant loop is predicted. The authors should elaborate on such results instead of picking one method and running it on only one of the sequences.
- Authors state: "For many of the subsequent analyses, the shorter archaeal sequence was used initially but the clear homology among this set of proteins means that inferences can be drawn

across the group.” - Please provide the used multiple sequence alignment with pairwise similarities to support this statement.

- It is not clear how helical wheels and hydrophobic moments support the manuscript - please provide a better description or omit these results.

- Problems/Validation of re-entrant loops:

- The authors selected 56 sequence regions from PDBTM database and run an all-against-all Blast search and create clusters based on the search results. Since the sequence complexity of membrane regions are lowest than regions of globular proteins, the analysis should be repeated on randomly selected transmembrane segments. Please provide the list of the selected 56 re-entrant loops together with the results of the repeated analysis.
- Authors state: “The presence of a re-entrant loop packed against each TM helix can also be seen on predicted contact maps for these proteins (Figure 3b).” Re-entrant loops cannot be seen on contact map, only parallel and anti-parallel structures. A similar contact map can be easily generated from 3 transmembrane helices (1 parallel pair and two anti-parallel ones).
- The authors filtered removing any sequences of less than 10 residues and more than 20. Although the exact sequence localisation and length of the predicted re-entrant loop are not provided, the regions indicated as the “sign” of re-entrant loops on Figure 3b is larger than 20 residues and on the structures the orange regions contain 7 turns, thus the sequence length of them should be more than 20 residues ( $7 \times 3.5 = 24.5$ ).
- The most serious one: As it can be seen on Fig7b-c, 3org contains additional helices that surround the interfacial helix - re-entrant loop - tm structure. Indeed the protein are dimer, where the dimer interface are formed by the re-entrant loops and the additional transmembrane helices surround this core. This arrangement ensure the lipid embedded structure is energetically stable. In the proposed model, re-entrant loops are not wrapped by other helices thus lipids may interact them. This is energetically unfavorable and does not prefer for the suggested function too. The validity of the model should be further investigate by molecular dynamic simulations of lipid embedded structures.

- “The analysis was performed by HELIQUEST (Gautier *et al.*, 2008) which constructed helical wheel diagrams and provided a quantitative measure of the hydrophobic moment for the region being analysed (Figure 4).” - This sentence and Figure 4 are pointless, containing data not used in the validation of the results.

Minor:

- Abstract/Results: “The results from the structural bioinformatics analysis of Tmem41b and its homologues showed that they contain a tandem repeat that is clearly visible in evolutionary covariance data but much less so by sequence analysis.”  
- As I showed above, this statement might not be true. Moreover evolutionary covariance data is the results of sequence analysis, so this sentence is void of sense. Please rephrase.
- Introduction: “there are eight E. coli representatives of the DedA family (YqjA, YghB, YabI, Yoh, DedA, YdjX, YdjZ, and YqaA)”  
- Character D is missing in Yoh.



- Introduction: "In the current study, we utilised state of the art methods to make structural predictions for two prominent members of the Pfam family PF09335 (Tmem41b and Yqja) by exploiting data derived from sequence, evolutionary covariance and ab initio modelling."  
- The most part of the manuscript deal with the sequence analysis of Mt2055, please rephrase this sentence in order to mirror this fact.
- "Interestingly, each of the re-entrant helices is predicted as a single transmembrane region in the TopCons predictions (see the diagonal of Figure 3b) with a two-residue region of coil in the centre."  
- TOPCONS does not predict coils and such details cannot be seen on the figure - please clarify this sentence.
- The authors should provide the generated PDB files as Supplementary Material.
- Contact map on Figure 7a left is the same that on right (numbering, dots). They should be different if one based on prediction and the other based on experimental data.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

No

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

No

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Topology and structure prediction of transmembrane proteins.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 10 Mar 2021

**Daniel Rigden**, University of Liverpool, UK

**Major:**

**- The authors should be more specific about the exact boundaries of Pfam domains in different proteins as well as the sequence relations of proteins presented in Table 1. Please provide multiple sequence alignment for these proteins indicating the localization of the two pfam domains and the proposed re-entrant loops/transmembrane regions in the sequences.**

*Yes agreed, we agree that a MSA is useful. An MSA generated using PSI/TM-COFFEE has been added as Figure 1 to the manuscript. The Pfam domains in questions well as the putative re-entrant loops for the modelled proteins have been highlighted to illustrate their relative positions.*

**- The authors propose Mt2055 contains a tandem repeat and suggest the duplication is present in Tmem41b and Yqja structure as well even if it is undetectable from sequence analysis. The proposed domain boundary in Figure1a and arguments for tandem duplication does not seem convincing. The e-value of 1.9E-3 is quite large for the alignment. The authors should rule out that results in their paper may occur purely by chance. Please test the statistical significance of this value by generating pairwise alignments of transmembrane regions of unrelated transmembrane proteins with similar length.**

*Utilisation of HHalign does result in an e-value of 1.9E-3 which on its own is not compelling. However, as highlighted, HHalign also expressed a probability score (a measure of statistical significance) of 78% which the software developers argue is a better indicator of significance than the e-value alone.*

*Additionally (as explained earlier in the text) 'Analysis of the HHpred results (against Pfam database) obtained for the archaeal protein Mt2055 revealed the presence of additional secondary hits for both PF06695 and PF09335 Pfam domains, in which the C-terminal half of the domains aligned with the N-terminal half of the Archaea protein. For example, residues 1-69 of the archaeal protein aligned with residues 52-117 of the Pfam PF09335 profile with a probability of 74.15%.'*

*It is important to remember that generating and scoring alignments with unrelated TM proteins is an intrinsic part of the database search. We have estimated that in Pfam there are currently around 1377 Pfam domains containing two or more TM-helices. This is based on analysis of the Phobius predictions that are part of the current 33.1 release. In the search above, only three of these domains - PF08566, PF09835, PF13571 - scored comparably (probabilities of 73.3-76.6%) with the secondary hits for PF06695 and PF09335. These results clearly place the secondary PF06695 and PF09335 matches at the extreme end of the score distribution for TM-helical Pfam domains, supporting their significance.*

*Arguably the above findings alone do not provide absolutely conclusive evidence of the presence of a repeat. However, reinforcing these findings we have the repeat that is revealed by the plotting of the predicted contacts and, consequently, the inverse repeat that is witnessed by the modelling.*

**Moreover, contact maps for Mt2055 and Tmem41b were generated from the same multiple alignment, and therefore they must be identical/similar. Thus the similarities does not prove the tandem duplication occurred in Tmem41b too.**

*Interesting point: however, we can confirm that the MSAs used to generate the contacts maps for Mt2055 and Tmem41b were not identical. The MSAs constructed by DMP were constructed independently using HHblits against the Uniprot database. The manuscript used the predicted contacts from the server and the MSAs generated to make the contact predictions are not made available to download with the results. However, performing the contact prediction locally and utilising the same HHblits settings as the DMP server generates MSAs with 5000 sequences for each of the query proteins. The predicted contact maps are very similar to those presented in the paper yet analysis reveals that the MSAs had only 1010 sequences in common.*

**- Structure modeling of membrane proteins is somewhat different from globular ones for several reasons. It is highly recommended to use specific software for this task or argue why used a non-specific one.**

*At the beginning of this project we had similar thoughts to you, therefore initially Rosetta membrane was utilised to build the models. However, the membrane protocol 'forced' TM helices where it was later clear from contact map analysis that re-entrant loops should be present. Therefore, it was decided that contact restrained modelling software with proven success in regard to ab initio modelling of membrane proteins was used. Both DMPfold (local & server) as well as the trRosetta (server) models were constructed and similar folds were observed. We note that the DMPfold paper benchmarked using transmembrane protein as explicitly says it 'works just as well for transmembrane proteins.' (<https://www.nature.com/articles/s41467-019-11994-0>). The trRosetta method was benchmarked against CASP13 targets which included transmembrane proteins (<https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25775>). The use of these covariance-based methods for membrane proteins has a long history so the following citation has been included in the revised manuscript;*

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., & Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7), 1607–1621.

**On one hand, in general, topology prediction is more accurate than structure modelling and should be used as an input to aid the modelling. The reviewer is not sure the result of a standard ab initio structure modelling program is sufficient to question topology prediction results. On the other hand, topology prediction results are different for Tmem41b (6 TM helix) and Mt2055 (4 TM helix). Notably, other consensus topology method (CCTOP) have a similar result for Mt2055 (4 helix), but different for Tmem41b (6 helix). Using a third method (Octopus) a re-entrant loop is predicted. The authors should elaborate on such results instead of picking one method and running it on only one of the sequences.**

*The different membrane topology prediction tools were used initially to predict the TMhelix boundaries for the query proteins. We observed the same between the results of the different methods as yourself. It was the variability of the topology predictions in addition to the contact*

map features that led to the conclusion that something other than straightforward TM helices is present in the Pfam domains in question. Indeed, TMHMM does show lower probability TMhelix predictions for the regions that the contact prediction and model making predict to be re-entrant loops.

To investigate further, visual representations of the membrane topology from TopCons and the psipred secondary structure prediction were plotted along the diagonal of the contact prediction for the query proteins. This clearly highlights that the N and C halves of the predicted TM helices in question are making contact with each other (by a length of around 10 residues). Additionally, the secondary structure plot shows an interruption at the halfway point of the predicted TM helices which would account for the abrupt change in direction of helix in the membrane.

Additionally, we have identified a crystal structure that is comparable in terms of size (293 residues) and has common structural features (inverted repeat with 2 re-entrant/TMhelix structures) to our query proteins; 6cb2. For this protein, the TopCons topology prediction was compared to the actual topology of the crystal structure.

The above figure shows actual contacts for 6cb2 (black points) and a visual representation of the TopCons topology prediction (green -outside, red - TM helix, yellow-inside, yellow boxes are the re-entrant loop-TM-helix 'signature'). Cross-referencing the first re-entrant contact map feature with the TopCons topology prediction it is clear that the TopCons topology must be wrong; the first TopCons predicted TM helix cannot be making contact with a region out-side of the membrane. Indeed, examination of the crystal structure reveals that the contact feature highlighted does in fact result from a re-entrant loop packed with a TMhelix .

Furthermore, constructing *ab initio* models of 6cb2 using both trRosetta (server) and DMPfold (server) yielded models that correctly fold the re-entrant loop in question. Performing structural alignments of the *ab initio* models against the crystal structure using Dali (server) give Z scores of 35 (trRosetta) and 27.5 (DMPfold). These scores leave no doubt that the correct fold has been modelled. The image below shows the crystal structure in green and the trRosetta model in magenta. The second image highlights the re-entrant feature that we are interested in.

**- Authors state: "For many of the subsequent analyses, the shorter archaeal sequence was used initially but the clear homology among this set of proteins means that inferences can be drawn across the group." - Please provide the used multiple sequence alignment with pairwise similarities to support this statement.**

*A multiple sequence alignment is now provided as Figure 1 and we note that all query sequences share the same Pfam domains so their homology is assured.*

**- It is not clear how helical wheels and hydrophobic moments support the manuscript - please provide a better description or omit these results.**

**- "The analysis was performed by HELIQUEST (Gautier et al., 2008) which constructed helical wheel diagrams and provided a quantitative measure of the hydrophobic moment for the region being analysed (Figure 4)." - This sentence and Figure 4 are**

**pointless, containing data not used in the validation of the results.**

Yes, we are in agreement with you, the paragraph did seem out of place as well as unfinished. In response we have re-worked the paragraph in question providing more clarity and analysis for the amphipathic analysis of the queries;

*"In order to test for the presence of the amphipathic helices, an analysis of helical wheel diagrams for the fifteen residues preceding the putative re-entrant loops was performed with HELIQUEST (Gautier et al., 2008). The quantitative measures of the hydrophobic moment for the regions being analysed (Figure 5) support that they are indeed amphipathic helices. The hydrophobic moments ranged from 0.298 to 0.546."*

To clarify; from the helical wheel figures the amphipathic nature of the approximately 15 residues preceding the putative re-entrant loops is clear. The importance of this finding is explained in relation to the structural comparison with the Cl<sup>-</sup>/H<sup>+</sup> antiporter 3org which also possesses the same structural features that we predict for the DedA proteins.

**- Problems/Validation of re-entrant loops:**

- **The authors selected 56 sequence regions from PDBTM database and run an all-against-all Blast search and create clusters based on the search results. Since the sequence complexity of membrane regions are lowest than regions of globular proteins, the analysis should be repeated on randomly selected transmembrane segments. Please provide the list of the selected 56 re-entrant loops together with the results of the repeated analysis.**
- **The authors filtered removing any sequences of less than 10 residues and more than 20. Although the exact sequence localisation and length of the predicted re-entrant loop are not provided, the regions indicated as the "sign" of re-entrant loops on Figure 3b is larger than 20 residues and on the structures the orange regions contain 7 turns, thus the sequence length of them should be more than 20 residues (7\*3.5=24.5).**
- **The most serious one: As it can be seen on Fig7b-c, 3org contains additional helices that surround the interfacial helix - re-entrant loop - tm structure. Indeed the protein are dimer, where the dimer interface are formed by the re-entrant loops and the additional transmembrane helices surround this core. This arrangement ensure the lipid embedded structure is energetically stable. In the proposed model, re-entrant loops are not wrapped by other helices thus lipids may interact them. This is energetically unfavorable and does not prefer for the suggested function too. The validity of the model should be further investigate by molecular dynamic simulations of lipid embedded structures.**

*Yes we are in agreement with you; the imposition of the re-entrant loop boundaries for the ab initio models were relatively arbitrary. Therefore, the re-entrant loop screen against the PDBTM has been completely re-implemented. Membrane boundaries for the models have now been predicted using the OMP server. These boundaries provide the lengths of the putative re-entrant loops. Consequently it is now recognised that the 20 residue 'typical length' of re-entrant loops may not be valid for the query models and the filtering of the larger loops for the clustering stage of this research had weak justification.*

*The clustering Methods text now reads*



*"A library of re-entrant loop pdb structures together with the putative re-entrant loop structures from the query protein models were clustered on their structural similarity. The library was built by obtaining a non-redundant (removing redundancy with a 40% sequence identity threshold) set of 125 chains from the PDBTM (RRID:SCR\_011962) (Kozma et al., 2013) that contain at least one re-entrant loop. As this investigation focuses on re-entrant loops that are immediately preceded by a TM helix that is packed against the loop, all re-entrant loops (boundaries defined by PDBTM) in addition to the preceding 30 residues were extracted. The resulting 193 library entries ([https://figshare.com/articles/dataset/repository\\_zip/14055212](https://figshare.com/articles/dataset/repository_zip/14055212)), supplemented with the re-entrant loop features (defined by the OMP server (Lomize, Pogozheva, Joo, Mosberg, & Lomize, 2012) and accompanied by the preceding 30 residues) from the ab initio modelling underwent an all-against-all structural alignment using a local installation of Dali v4.0 (Holm & Laakso, 2016). The Z-scores for these alignments were then used for clustering with CLANS v1.0 (Frickey & Lupas, 2004) with a Z-score of 4.5 used as the cut-off threshold."*

*The Results of that protocol are now reported as follows*

*"The presence of re-entrant loops and the high density of conserved residues within them caused us to examine experimentally characterised re-entrant loops in the PDBTM database. A total of 193 non-redundant re-entrant helices were identified (see Methods). All 193 were clustered with the putative re-entrant loops from Mt2055, Tmem41b and YqjA using relative z-scores derived from an all-against-all DALI run and subsequently clustered in CLANS (Frickey & Lupas, 2004) with a z-score cut-off of 4.5.*

*The as expected all six re-entrant structures from the query models clustered together. The CLC transporter re-entrant structures of 3orgA (re-entrant 1 and re-entrant 2), 7bxu and 5tqq also clustered with the queries. Additionally, the re-entrant structure from an Undecaprenyl pyrophosphate phosphatase (UppP) (6cb2) also clustered with the queries. UppP is an integral membrane protein that recycles lipid and has structural similarities to CLC transporters (Workman, Worrall, & Strynadka, 2018). Contact maps derived from the pdb files of CLC and UppP structures show the contact map signature corresponding to the re-entrant/TM helix structural feature. Interestingly, the UppP is more similar to the query proteins being only 271 residues in length and having only 6 TM helices."*

*A list of the 125 chains from which the re-entrant structures were extracted from will be made available in a repository.*

- **Authors state: "The presence of a re-entrant loop packed against each TM helix can also be seen on predicted contact maps for these proteins (Figure 3b)." Re-entrant loops cannot be seen on contact map, only parallel and anti-parallel structures. A similar contact map can be easily generated from 3 transmembrane helices (1 parallel pair and two anti-parallel ones).**

*Yes, a similar contact map feature can be easily generated from 3 transmembrane helices, however, this would result in a box feature of around 20x20 residues (and obviously reflected in the diagonal). Since the re-entrant loop is making contact with itself this can only result in an approximately 10 residue antiparallel feature on the contact map. Only approximately half of the TM helix that is packed with the re-entrant helix will be making contact with the re-entrant loop, therefore, this would result in an additional 10 residue antiparallel feature in addition to a 10-residue parallel feature. Together with the diagonal these will display an approximately 10x10 box feature (also reflected in the diagonal) on the contact map rather than the 20x20 box*

*feature that three transmembrane helices (1 parallel pair and two anti-parallel ones) would produce. This can be seen below;*

**Minor:**

- **Abstract/Results: "The results from the structural bioinformatics analysis of Tmem41b and its homologues showed that they contain a tandem repeat that is clearly visible in evolutionary covariance data but much less so by sequence analysis."**  
- As I showed above, this statement might not be true. Moreover evolutionary covariance data is the results of sequence analysis, so this sentence is void of sense. **Please rephrase.**

*We do not agree with this statement as sequence comparisons and co-variance comparisons are alternative methods to identify tandem repeats. Yes, co-variance is derived from sequence analysis; however, co-variance data contains information that may not be present in data acquired from conventional sequence analysis.*

- **Introduction: "there are eight E. coli representatives of the DedA family (YqjA, YghB, YabI, Yoh, DedA, YdjX, YdjZ, and YqaA)"**  
- Character D is missing in Yoh.

*Thank you, corrected.*

- **Introduction: "In the current study, we utilised state of the art methods to make structural predictions for two prominent members of the Pfam family PF09335 (Tmem41b and Yqja) by exploiting data derived from sequence, evolutionary covariance and ab initio modelling."**  
- The most part of the manuscript deal with the sequence analysis of Mt2055, please rephrase this sentence in order to mirror this fact.

*Thank you, the introduction has been updated to reflect the emphasis on the PF09665 Pfam domain and its representative Mt2055;*

*'In the current study, we first linked the Pfam PF09335 family to the PF06695 family and chose a conveniently small Archaeal sequence and then utilised state of the art methods to make structural predictions for not only the Archaeal sequence but also for two prominent members of the Pfam family PF09335 (Tmem41b and Yqja) by exploiting data derived from sequence, evolutionary covariance and ab initio modelling. We are able to predict that both PF09335 homologues (VTT proteins) and PF06995 homologues contain re-entrant loops (stretches of protein that enter the bilayer but exit on the same side of the membrane) as well as a pseudo-inverted repeat topology. The predicted presence of both of these structural features strongly suggests that VTT proteins are secondary active transporters for an uncharacterised substrate.'*

- **"Interestingly, each of the re-entrant helices is predicted as a single transmembrane region in the TopCons predictions (see the diagonal of Figure 3b) with a two-residue region of coil in the centre."**  
- TOPCONS does not predict coils and such details cannot be seen on the figure - please

**clarify this sentence.**

*It was not our intention to suggest that Topcons predicts secondary structure. We have changed the paragraph in question clarifying our intention;*

*"Interestingly, each of the re-entrant helices is predicted as a single transmembrane region in the TopCons predictions. When cross-referenced with the PSIPRED secondary structure prediction it is noted that there is a predicted two-residue region of coil region of coil around the mid-point of the first TM helix prediction. A similar observation can be made for the fourth TM helix prediction with the equivalent coil region being six residues in length (see the diagonal of Figure 4b)"*

- **The authors should provide the generated PDB files as Supplementary Material.**

Thank you for pointing out this important omission. Since this journal does not allow Supplementary Material we have deposited the models in a repository now mentioned in the paper.

[https://figshare.com/articles/dataset/repository\\_zip/14055212](https://figshare.com/articles/dataset/repository_zip/14055212)

- **Contact map on Figure 7a left is the same that on right (numbering, dots). They should be different if one based on prediction and the other based on experimental data.**

*The left image was generated using predictions and the right with pdb file. They are very similar, but this is to be expected. CIC transporters are a large family and therefore the co-variance-derived predictions will be very accurate.*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 17 December 2020

<https://doi.org/10.5256/f1000research.30592.r75807>

© 2020 Doerrler W et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Pradip Panta**

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA

**William T. Doerrler**

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA

This work describes the computational structural modeling of a conserved membrane protein family that includes human TMEM41B, a protein with a number of reported functions. Membrane proteins are poorly represented in the structural database and computational methods are increasingly valuable for understanding structure and function. Here, they use a method using evolutionary amino acid contact co-variation to predict a structure that supports a proposed function as a proton dependent antiporter. While I am not fluent in the computational methods

used, their prediction do align with published experimental work. The manuscript is well written and informative, but with a number of factual errors. I also suggest additional citations.

I would like to begin with nomenclature. I received an email from Dr. Noburo Mizushima several months ago. He has published work on the TMEM41B protein. Also included on the email was Lucy Forrest, Dirk Schneider, and Rebecca Keller. It was Dr. Mizushima's suggestion to name this protein family the "DedA superfamily" that includes both prokaryotic and eukaryotic proteins (DedA, VMP, and TMEM41 families). Accordingly, the shared domain will be called "DedA domain" and "VTT" domain would no longer be used. All recipients of this email agreed to using this nomenclature moving forward. Therefore, to avoid confusion, I would like the authors to adopt this nomenclature. I can forward the email upon request.

Since the manuscript contains no line numbers, I will list the suggested corrections by paragraph:

**Introduction:**

Paragraph 1: Formally, "membrane proteins" also include various lipid-modified proteins of both prokaryotes and eukaryotes in addition to integral and peripheral membrane proteins.

Paragraph 4: "DedA" does not stand for "death effector domain". It was named in a 1987 paper<sup>1</sup>. See page 12213 of that article. I would like to see this article cited as well for historical purposes.

The sentence that begins with "Phenotypically, DedA knockout E. coli..." should instead read "Phenotypically, E. coli lacking both *yqjA* and *yghB* (encoding proteins with 60% amino acid identity and partially overlapping functions)...." This paragraph should also cite<sup>2</sup>.

The sentence that reads "As *E. coli* expresses multiple DedA homologues, the redundancy protects the cells from the phenotypical effects of single or multiple knock-outs as long as at least one DedA is expressed" should read "As *E. coli* expresses multiple DedA homologues, lethal effects are not observed as long as at least one DedA is expressed". Cite the following article<sup>3</sup>.

You may also point out that the sole DedA family gene in *Borrelia burgdorferi* is indeed essential<sup>4</sup>.

YqjA is misspelled "YdjA"

The sentence "Attempts to rescue...." Should be removed, as it does not make sense.

The final sentence about *Pseudomonas* cites a non-peer reviewed proceeding abstract. I would like all citations to "Justice *et al.* 2016" removed from this article. This sentence can be replaced with the equally effective "The functions of DedA have also been studied in the pathogen *Burkholderia thailandensis* where one family member was found to be required for resistance to polymyxin"<sup>5</sup>.

Paragraph 6: "YqjA" is spelled "Yqja" here and throughout the manuscript and should be corrected. This includes in Table 1.

**Methods:**

Paragraph 1: Please spell "Ydjx" and other bacterial proteins as "YdjX" with the final letter capitalized.

**Results and discussion:**

Paragraph 5: first sentence, remove "however".

Paragraph 14: "A possible role for VTT proteins" final sentence remove "Justice *et al.*" and instead cite <sup>6,7</sup>.

Also, in this sentence, define "SDM" as "site directed mutagenesis".

Paragraph 15, first sentence. This statement is incorrect. Mutation of D51, E39, R130 or R136 in YqjA resulted in properly folded (membrane localized) but nonfunctional proteins unable to complement alkaline pH sensitivity of E. coli YqjA mutant and antibiotic sensitivity of YqjA/YghB double mutant.

Finally, another interesting example of a membrane protein antiporter with re-entrant helices is the undecaprenyl pyrophosphate phosphatase UppP. It is up to the authors if they would like to cite these articles <sup>8,9</sup>.

**References**

1. Nonet ML, Marvel CC, Tolan DR: The *hisT-purF* region of the Escherichia coli K-12 chromosome. Identification of additional genes of the *hisT* and *purF* operons. *J Biol Chem.* 1987; **262** (25): 12209-17 [PubMed Abstract](#)
2. Thompkins K, Chattopadhyay B, Xiao Y, Henk MC, et al.: Temperature sensitivity and cell division defects in an Escherichia coli strain with mutations in *yghB* and *yqjA*, encoding related and conserved inner membrane proteins. *J Bacteriol.* 2008; **190** (13): 4489-500 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Boughner LA, Doerrler WT: Multiple deletions reveal the essentiality of the DedA membrane protein family in Escherichia coli. *Microbiology (Reading).* 2012; **158** (Pt 5): 1162-1171 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Liang FT, Xu Q, Sikdar R, Xiao Y, et al.: BB0250 of *Borrelia burgdorferi* is a conserved and essential inner membrane protein required for cell division. *J Bacteriol.* 2010; **192** (23): 6105-15 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Panta PR, Kumar S, Stafford CF, Billiot CE, et al.: A DedA Family Membrane Protein Is Required for *Burkholderia thailandensis* Colistin Resistance. *Front Microbiol.* 2019; **10**: 2532 [PubMed Abstract](#) | [Publisher Full Text](#)
6. Kumar S, Doerrler WT: Members of the conserved DedA family are likely membrane transporters and are required for drug resistance in Escherichia coli. *Antimicrob Agents Chemother.* 2014; **58** (2): 923-30 [PubMed Abstract](#) | [Publisher Full Text](#)
7. Kumar S, Bradley CL, Mukashyaka P, Doerrler WT: Identification of essential arginine residues of Escherichia coli DedA/Tvp38 family membrane proteins YqjA and YghB. *FEMS Microbiol Lett.* **363** (13). [PubMed Abstract](#) | [Publisher Full Text](#)
8. Workman S, Worrall L, Strynadka N: Crystal structure of an intramembranal phosphatase central to bacterial cell-wall peptidoglycan biosynthesis and lipid recycling. *Nature Communications.* 2018; **9** (1). [Publisher Full Text](#)
9. El Ghachi M, Howe N, Huang C, Olieric V, et al.: Crystal structure of undecaprenyl-pyrophosphate phosphatase and its role in peptidoglycan biosynthesis. *Nature Communications.* 2018; **9** (1). [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**



Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bacterial genetics with interests in membrane proteins and antibiotic resistance mechanisms.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 10 Mar 2021

**Daniel Rigden**, University of Liverpool, UK

**I would like to begin with nomenclature. I received an email from Dr. Noburo Mizushima several months ago. He has published work on the TMEM41B protein. Also included on the email was Lucy Forrest, Dirk Schneider, and Rebecca Keller. It was Dr. Mizushima's suggestion to name this protein family the "DedA superfamily" that includes both prokaryotic and eukaryotic proteins (DedA, VMP, and TMEM41 families). Accordingly, the shared domain will be called "DedA domain" and "VTT" domain would no longer be used. All recipients of this email agreed to using this nomenclature moving forward. Therefore, to avoid confusion, I would like the authors to adopt this nomenclature. I can forward the email upon request.**

*Thanks very much for this helpful suggestion. We have changed the naming in the manuscript throughout to "DedA superfamily"*

**Introduction:**

**Paragraph 1: Formally, "membrane proteins" also include various lipid-modified proteins of both prokaryotes and eukaryotes in addition to integral and peripheral membrane proteins.**

*Thank you, this omission has been rectified;*

*'Membrane proteins can be grouped according to their interaction with various cell membranes: integral membrane proteins (IMPs) are permanently anchored whereas peripheral membrane proteins transiently adhere to cell membranes. IMPs that span the membrane are known as transmembrane proteins (TMEMs) as opposed to IMPs that adhere to one side of the membrane (Fowler & Coveney, 2006). Membrane proteins also include various lipid-modified proteins (Resh, 2016).'*

**Paragraph 4: "DedA" does not stand for "death effector domain". It was named in a 1987 paper<sup>1</sup>. See page 12213 of that article. I would like to see this article cited as well for historical purposes.**

*Your clarification on this nomenclature is appreciated. We have amended the manuscript to reflect this error and included the reference for its historical importance.*

- **The sentence that begins with "Phenotypically, DedA knockout E. coli..." should instead read "Phenotypically, E. coli lacking both *yqjA* and *yghB* (encoding proteins with 60% amino acid identity and partially overlapping functions)...." This paragraph should also cite<sup>2</sup>.**
- **The sentence that reads "As *E. coli* expresses multiple DedA homologues, the redundancy protects the cells from the phenotypical effects of single or multiple knock-outs as long as at least one DedA is expressed" should read "As *E. coli* expresses multiple DedA homologues, lethal effects are not observed as long as at least one DedA is expressed". Cite the following article<sup>3</sup>.**
- **Paragraph 14: "A possible role for VTT proteins" final sentence remove "Justice *et al.*" and instead cite<sup>6,7</sup>.**

*The additional references for previous experimental studies of the DedA proteins that you have suggested have been added to the manuscript. Thanks for your suggestions which certainly make the manuscript more comprehensively cite previous studies.*

- **You may also point out that the sole DedA family gene in *Borrelia burgdorferi* is indeed essential<sup>4</sup>.**
- **The final sentence about *Pseudomonas* cites a non-peer reviewed proceeding abstract. I would like all citations to "Justice *et al.* 2016" removed from this article. This sentence can be replaced with the equally effective "The functions of DedA have also been studied in the pathogen *Burkholderia thailandensis* where one family member was found to be required for resistance to polymyxin"<sup>5</sup>.**

*Thanks for these helpful suggestions. The inclusion of these points brings more clarity to the paragraph in question;*

*"*Borrelia burgdorferi* contains only one DedA protein in its genome and knockout cells display the same phenotype as the *E. coli* knockout strains. The *B. burgdorferi* homologue is indeed essential (Liang *et al.*, 2010). Interestingly, *E. coli* knockout cells can be rescued with the *B. burgdorferi* homologue that shows only 19% sequence identity with YqjA. The functions of DedA have also been studied in the pathogen *Burkholderia thailandensis* where one family member was found to be required for resistance to polymyxin (Panta *et al.*, 2019)."*

- **YqjA is misspelled "YdjA"**
- **Paragraph 6: "YqjA" is spelled "Yqja" here and throughout the manuscript and**

**should be corrected. This includes in Table 1.**

- **Paragraph 1: Please spell “Ydjx” and other bacterial proteins as “YdjX” with the final letter capitalized.**

*Corrections made. Thank you for pointing out these important errors.*

**The sentence “Attempts to rescue....” Should be removed, as it does not make sense.**

Sentence removed, thank you.

**Results and discussion:**

**Paragraph 5: first sentence, remove “however”.**

*Amendment made, thank you.*

**Also, in this sentence, define “SDM” as “site directed mutagenesis”.**

*Amendment made, thank you.*

**Paragraph 15, first sentence. This statement is incorrect. Mutation of D51, E39, R130 or R136 in YqjA resulted in properly folded (membrane localized) but nonfunctional proteins unable to complement alkaline pH sensitivity of E. coli YqjA mutant and antibiotic sensitivity of YqjA/YghB double mutant.**

*Thank you for pointing this out to us. We can see our description of the results of the study was incorrect. This has been amended with your direction;*

*“A recent study has identified key residues ( Figure 8) in the E. coli DedA protein YqjA that, when replaced in site directed mutagenesis experiments, resulted in properly folded (membrane localized) but non-functional proteins unable to complement alkaline pH sensitivity of E. coli YqjA mutant and antibiotic sensitivity of YqjA/YghB double mutant ( Panta et al., 2019).”*

**Finally, another interesting example of a membrane protein antiporter with re-entrant helices is the undecaprenyl pyrophosphate phosphatase UppP. It is up the authors if they would like to cite these articles [8](#),[9](#).**

*Yes, this is interesting; for the revision of the manuscript we re-implemented the re-entrant loop screen against the PDBTM. We found that Gcb2 (UppP) re-entrant loop structures were structurally very similar to the re-entrant models for the DedA domains.*

**Competing Interests:** No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**