

1
2
3
4
5
6
7
8
9
10
11
12

**Remote homolog detection places insect chemoreceptors in a cryptic
protein superfamily spanning the tree of life**

13

Nathaniel J. Himmel^{1,*}, David Moi² and Richard Benton^{1,*}

14

¹Center for Integrative Genomics

15

²Department of Computational Biology

16

Faculty of Biology and Medicine

17

University of Lausanne

18

CH-1015

19

Lausanne

20

Switzerland

21

22

*Corresponding authors:

23

nathanieljohn.himmel@unil.ch

24

richard.benton@unil.ch

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46 **Summary**

47

48 Many proteins exist in the so-called “twilight zone” of sequence alignment, where
49 low pairwise sequence identity makes it difficult to determine homology and
50 phylogeny^{1,2}. As protein tertiary structure is often more conserved³, recent
51 advances in *ab initio* protein folding have made structure-based identification of
52 putative homologs feasible^{4–6}. However, structural screening and phylogenetics
53 are in their infancy, particularly for twilight zone proteins. We present a pipeline for
54 the identification and characterization of distant homologs, and apply it to 7-
55 transmembrane domain ion channels (7TMICs), a protein group founded by insect
56 Odorant and Gustatory receptors. Previous sequence and limited structure-based
57 searches identified putatively-related proteins, mainly in other animals and plants
58^{7–10}. However, very few 7TMICs have been identified in non-animal, non-plant taxa.
59 Moreover, these proteins’ remarkable sequence dissimilarity made it uncertain if
60 disparate 7TMIC types (Gr/Or, Grl, GRL, DUF3537, PHTF and GrlHz) are
61 homologous or convergent, leaving their evolutionary history unresolved. Our
62 pipeline identified thousands of new 7TMICs in archaea, bacteria and unicellular
63 eukaryotes. Using graph-based analyses and protein language models to extract
64 family-wide signatures, we demonstrate that 7TMICs have structure and sequence
65 similarity, supporting homology. Through sequence and structure-based
66 phylogenetics, we classify eukaryotic 7TMICs into two families (Class-A and Class-
67 B), which are the result of a gene duplication predating the split(s) leading to
68 Amorphea (animals, fungi and allies) and Diaphoretickes (plants and allies). Our
69 work reveals 7TMICs as a cryptic superfamily with origins close to the evolution of
70 cellular life. More generally, this study serves as a methodological proof of principle
71 for the identification of extremely distant protein homologs.

72

73

74 Results and Discussion

75

76 Insect Odorant receptors (Ors) and Gustatory receptors (Grs) are 7-
77 transmembrane domain ion channels (7TMICs) critical for the behavior and
78 evolution of insects^{7,11,12}. Although originally thought to be insect-specific¹³⁻¹⁸, the
79 genomic revolution enabled sequence-based searches to identify putative
80 homologs in animals (Gustatory receptor-like proteins; Grls), plants (DUF3537
81 proteins) and single-celled eukaryotes (GRLs)^{7-9,19}. However, the representation
82 of 7TMICs across taxa remained sparse, recognized in only a small number of
83 unicellular eukaryotes (17 proteins from 7 species), and missing from several
84 holozoan lineages, including chordates, choanoflagellates, comb jellies and
85 sponges^{7-9,19}.

86 The best-characterized 7TMICs are insect Ors, which function as odor-
87 gated heterotetrameric (or in some cases homotetrameric) ion channels²⁰⁻²³. A
88 substantial breakthrough came from two Or cryo-electron microscopy structures:
89 the fig wasp *Apocryta bakeri* Or co-receptor Orco²⁰ and the jumping bristletail
90 *Machilis hrabei* Or5²¹ (**Figure 1A**). Or monomers have several notable structural
91 features, including: (i) 7 transmembrane alpha helices with a characteristic packing
92 pattern; (ii) an intracellular N-terminus and extracellular C-terminus; (iii) shorter
93 extracellular than intracellular loops; (iv) long TM4, TM5, and TM6 helices that
94 extend into the intracellular space, forming the “anchor domain,” where most inter-
95 subunit interactions occur; (v) an unusual “split” TM7 helix, composed of an
96 intracellular TM7a (part of the anchor domain) and a transmembrane-spanning
97 TM7b (which lines the pore of the ion channel); and (vi) an N-terminal re-entrant
98 loop (TM0)^{20,21,24,25}. These tertiary structural features are remarkably highly-
99 conserved despite low primary sequence conservation; for example, the two
100 experimental structures have virtually indistinguishable folds while having only 19%
101 amino acid sequence identity (**Figure 1B**). Importantly, these structures can be
102 accurately predicted *in silico* by several algorithms^{8,25}, notably AlphaFold (**Figure**
103 **1C**)^{4,10}.

104 Recently, we took advantage of the structural similarity of 7TMICs to perform
105 structure-based screens for putative homologs that had not been identified by
106 sequence-based screening. These screens identified several proteins adopting the
107 7TMIC fold, including: fly-specific Gustatory receptor-like proteins (Grls); a highly-
108 conserved lineage of eukaryotic proteins (PHTFs, an acronym for the misnomer
109 Putative Homeodomain Transcription Factor); a holozoan-specific Grl lineage
110 (GrlHz); and trypanosome 7TMICs¹⁰. However, these searches were limited by the
111 high computational requirements of the structural alignment tool—Dali^{26,27}—and
112 only ~564,000 AlphaFold models from 48 species were screened. Thus, large
113 taxonomic gaps still exist: fewer than 50 proteins have been identified outside of
114 animals and plants, and none have been identified in prokaryotes (despite
115 screening 17 prokaryotic proteomes¹⁰). Beyond the technical limitations leading to
116 sparse taxonomic sampling, the PHTF, GrlHz, Gr/Or, DUF3537 and various
117 unicellular eukaryotic 7TMIC proteins share little to no recognizable sequence
118 similarity. It is thus unclear how many 7TMICs exist across taxa and if 7TMICs form
119 a single or many homologous protein families. We thus sought to build a new
120 pipeline for remote homolog detection, validation, and sequence/structure analysis,
121 aiming to resolve the evolutionary history of 7TMICs, be they homologous or
122 convergent.

123

124 **Insect Ors and Grs have high structural similarity despite exceptional
125 sequence dissimilarity**

126 Comparisons of the AlphaFold models of *Drosophila melanogaster* Ors and Grs
127 exemplifies the discordance between sequence and structure similarity: pairwise
128 comparisons average only ~13% pairwise amino acid sequence identity (**Figure
129 1D**, y-axis), placing these proteins at the border of the so-called “twilight zone” (10-
130 40% sequence identity)¹ and “midnight zone” (<10% sequence identity)² of
131 sequence alignment. By contrast, pairwise comparisons of the corresponding
132 AlphaFold structures—using Dali Z-scores, a widely-used metric of fold similarity
133^{26,27}—reveals that all pairwise comparisons fall within the “safe zone” of structural
134 alignments, indicating high statistical confidence in their similarity (**Figure 1D**, x-
135 axis). When visualized as a sequence similarity network (produced by all-to-all
136 BLASTP searches), Ors and Grs—together with other *D. melanogaster* 7TMICs,
137 i.e. Grls and Phtf¹⁰—segregate into several non-contiguous clusters (**Figure S1A**).
138 This analysis demonstrates that no single receptor protein can be used to identify
139 all others via simple sequence-based searches. By contrast, structure-based
140 search strategies (e.g. Dali, **Figure S1B**) are capable of densely networking these
141 proteins. As *D. melanogaster* Ors and Grs are just a very small subset of 7TMICs
142 that likely had a single common ancestor²⁸, these observations emphasise how
143 structure-based screens are a greatly superior way to search for distant homologs
144 across more phylogenetically diverse species³.
145

146 **A pipeline for identifying extremely distant protein homologs**
147

148 Foldseek—a recently released tool for structure-based protein comparisons—
149 operates orders of magnitude faster than Dali and other structural alignment tools,
150 making large protein homolog screens feasible⁵. We first benchmarked Foldseek
151 on *D. melanogaster* 7TMICs. When forced to compare the AlphaFold model of Orco
152 to all other 7TMICs of this species, Foldseek produced structural similarity scores
153 that correlate with Dali Z-scores (**Figure S1C**). As proof of concept for screening,
154 we used the *D. melanogaster* Orco AlphaFold model to survey the AlphaFold
155 structural proteome of *D. melanogaster* (**Figure 1E**). Foldseek was able to recover
156 all *D. melanogaster* 7TMICs except Phtf: thus, the method can result in false
157 negatives. However, with the most permissive settings—which would allow the
158 most sensitive homolog detection—Foldseek also had an extremely high false
159 positive rate (73.3%), and the most divergent relatives (e.g. Grls) had higher E-
160 values and/or lower percent sequence identity than false positives (**Figure 1E**). As
161 we were interested in screening for distant and divergent 7TMIC homologs across
162 much longer evolutionary distances than only within *D. melanogaster*, we
163 recognized that neither E-value nor sequence identity could serve as an effective
164 threshold. These benchmarks illustrated the need for additional search and
165 validation steps to minimise both false positive and false negative results in our
166 screen.
167

168 We therefore implemented Foldseek as part of a screening and validation
169 pipeline, with the goal of determining the presence or absence of 7TMICs across
170 the tree of life (**Figure 1F**). This pipeline first uses Foldseek to search for
171 structurally similar models in the AlphaFold Protein Structure Database, which
172 currently consists of ~200,000,000 models from >1,200,000 species (see Methods
173 for details on exclusions). After structural validation, it employs PSI-BLAST in a

174 sequence-based screen, providing structurally-informed access to >400,000,000
175 sequences—with diverse transcriptomic, proteomic, genomic, and metagenomic
176 origins—that might not have a corresponding protein model. This second step also
177 allows for the identification of proteins with models that were missed in the first
178 structure-based screen, which we expected to occur due to the occurrence of false
179 negatives at hypothetically vast evolutionary distances (e.g. Orco to Phtf (**Figure**
180 **1E**)).

181 As false positives can have high scores, and as some public data can be
182 incomplete or of low quality, we implemented several verification steps to extract
183 true hits. For proteins identified by structural model, we: (i) curated proteins based
184 on membrane topology as predicted by the protein language model DeepTMHMM
185²⁹; (ii) validated structural alignments using Dali; and (iii) visually inspected putative
186 hits for the previously-described 7TMIC features. Proteins identified through
187 sequence similarity were curated based on membrane topology (DeepTMHMM).
188

189 **7TMICs are present across the tree of life**

190

191 This screen recovered thousands of previously unidentified 7TMICs spanning the
192 tree of life (**Figure 1G-H** and **Figure S1E**). These hits not only include new
193 eukaryotic 7TMICs (hereafter, Euk7TMICs), but also sequences from all major
194 branches of bacteria (Bac7TMICs) and archaea (Arch7TMICs) (see “Protein
195 nomenclature” section in the Methods). These proteins come from several
196 obviously monophyletic clades, apparent as clusters in a network representing all-
197 to-all BLASTP searches (**Figure 1H**). However, they can exhibit very little pairwise
198 sequence similarity, represented by few edges between clusters in the BLASTP
199 network (**Figure 1H**).

200 Euk7TMICs could be visually sorted into two types of structure:
201 Or/Gr/Grl/GRL/DUF3537-like (**Figure 2A**), having the canonical insect Or-like fold;
202 or PHTF-like (**Figure 2B**), having the same core structure, but with a long first
203 intracellular loop (IL1). While the various prokaryotic 7TMICs have a striking degree
204 of structural similarity to Euk7TMICs (**Figure 2C-E**), we observed that they
205 generally had shorter TM4 and TM5 helices, which constitute a component of the
206 anchor domain in insect Ors (**Figure 1A**). Heimdallarchaeota 7TMICs (**Figure 2E**)
207 were an exception: their overall tertiary structure appeared eukaryote-like. This
208 qualitative similarity (supported by subsequent quantitative analyses, described
209 below) is notable, as Heimdallarchaeota are proposed to be the most closely
210 related extant archaea to eukaryotes³⁰⁻³⁴. In addition, a small number of
211 metagenomically-identified prokaryotic 7TMICs have Euk7TMIC-like folds (**Figure**
212 **2F**). Notably, these show high sequence similarity to Euk7TMICs (green nodes in
213 the eukaryotic PHTF-like cluster, **Figure 1H**), suggesting that these sequences are
214 the result of eukaryote-to-prokaryote horizontal gene transfer(s), a hypothesis
215 further supported phylogenetically (see below).
216

217 **7TMICs have a shared tertiary structure and amino acid sequence profile, 218 supporting homology**

219

220 While we observed structural similarities between the proteins our screen
221 identified, it remained unclear if these sequences are homologous, or if they
222 represent cases of structural convergence. To address this fundamental issue, we
223 adapted established protein comparison tools into a graph-based approach for

224 determining homology based on both structure and sequence. For protein
225 structures, we calculated all-to-all template modelling (TM) scores, where those
226 >0.5 indicate high statistical confidence of fold similarity³⁵. For protein sequences,
227 we performed all-to-all PSI-BLAST searches; PSI-BLAST builds iterative multiple
228 sequence alignments, thereby identifying distant homologs by family-wise
229 sequence profiles, rather than by simple pairwise sequence similarities³⁶. In
230 essence, PSI-BLAST networking is equivalent to performing PSI-BLAST homolog
231 searches starting with every structurally-validated 7TMIC as a query (see
232 Methods). For both methods, one expects homologous proteins to form bi-
233 directional connections between each other (i.e. that pairs will be reciprocal hits),
234 and that homologous families will be highly interconnected, thereby collapsing into
235 visually identifiable clusters in structure- and sequence-space. We performed these
236 analyses with Type-I and Type-II opsins as control groups, as these large families
237 are 7-transmembrane domain proteins (unrelated to 7TMICs) that adopt highly
238 similar folds to one another, despite no recognized sequence similarity³⁷.

239 In the structural similarity network, 7TMICs formed a densely connected
240 linkage cluster, disconnected from a unified opsin linkage cluster (**Figure 3A**).
241 7TMICs also clustered in sequence space – after 3 PSI-BLAST iterations, 7TMICs
242 collapsed into a single, highly connected community structure (**Figure 3B** and
243 **Figure S2A**). In stark contrast, the opsins separated into distinct Type-I and Type-
244 II community structures, demonstrating that structure and sequence are not
245 necessarily linked (**Figure 3B**). While there were connections between 7TMICs
246 and the opsins, in the third iteration these constituted only 18 of the 1,117,609
247 connections (0.0016%), almost certainly representing spurious similarity. A small
248 minority of 7TMICs (33/2421 representative sequences) from diverse eukaryotic
249 taxa showed no connectivity to the core 7TMIC cluster in the second iteration and
250 weak connectivity in the third; these may be extremely rapidly evolving proteins
251 and/or cases of independent structural convergence.

252 We next sought to determine which, if any, regions of 7TMICs are more
253 conserved. It was previously observed that insect Ors display the highest
254 conservation in the anchor domain and pore-forming region, with greater
255 divergence in the N-terminal region that forms the odor-binding pocket^{20,21}. We
256 calculated sequence embedding-based conservation scores, which identify sites
257 that are evolutionarily constrained³⁸. This analysis elucidated a similar
258 conservation pattern for newly identified 7TMICs: while absolute amino acid
259 sequence identity is low (averaging 15% across sites, **Figure 3C** and **Figure S2B**),
260 embedding-based conservation analysis revealed that the most highly conserved
261 regions are in three locations: the hypothetical anchor domain (intracellular
262 sequences spanning TM4-TM5 and TM6-TM7a), the hypothetical pore (TM7b), and
263 TM5-TM6, which form lateral ion permeation conduits in Ors^{20,21} (**Figure 3D, 3F**).

264 We next used the protein language model PeSTo³⁹ to predict protein-
265 protein interactions in 7TMICs, revealing two conserved regions (**Figure 3E, 3F**).
266 The first was N-terminal, corresponding to the re-entrant loop (TM0); this region
267 has an important, albeit poorly-understood, function in Orco²⁵. The second region
268 was in the hypothetical anchor domain and pore, in the same regions as the highest
269 peaks of sequence conservation.

270 These findings are not biased by the inclusion of proteins previously
271 determined to be homologous (insect Ors/Grs and animal Grls); on the contrary,
272 removing these sequences improved average conservation (and interaction)
273 scores in these regions (**Figure 3D-E** and **Figure S2C**).

274 While we cannot know *a priori* whether these proteins form tetramers like
275 insect Ors, these patterns of conservation and predicted protein-protein
276 interactions suggests they may assemble as multimers using the same domains.
277 Consistent with this idea, using AlphaFold-multimer^{40–43} to predict complexes of
278 tetramers of newly-identified 7TMICs, the vast majority of resulting quaternary
279 structures had striking similarity to experimentally-derived Or structures (**Figure 3G**
280 and **Figure S2D**). In these models, the hypothetical anchor domain (particularly
281 TM7a) contains the closest protein-protein interactions and TM7b lines the putative
282 pore.

283 These results quantitatively demonstrate that 7TMICs have a common
284 structure, a shared sequence profile, and similar patterns of sequence
285 conservation. Thus, the most parsimonious hypothesis is that 7TMICs are a
286 homologous protein superfamily.
287

288 The evolutionary history of 7TMICs

289 Having obtained evidence for the homology of 7TMICs, we next sought to elucidate
290 the evolutionary history of the superfamily. As we expected pairwise sequence
291 dissimilarity would make multiple sequence alignments difficult, we performed
292 sequence-based phylogenetics on an ensemble of alignments, thus resulting in a
293 “forest” of phylogenetic trees (**Figure S3A-E**), from which we extracted the median
294 sample tree (**Figure 4A**). These analyses suggested that there are two main
295 Euk7TMIC families, hereafter termed Class-A and Class-B Euk7TMICs. While
296 Class-A Euk7TMICs appear to be monophyletic, the monophyly of Class-B is
297 uncertain. Class-A Euk7TMICs include insect Ors/Grs, animal Grls, plant
298 DUF3537, holozoan GrlHz, and various unicellular eukaryotic 7TMICs. Class-B
299 Euk7TMICs are PHTF-like proteins from diverse taxa, including a small number of
300 bacterial and archaeal proteins (see green nodes in the PHTF-like cluster the
301 BLASTP network (**Figure 1H**) and an example structure (**Figure 2F**)). The
302 phylogenetic separation of these from other prokaryotic 7TMICs suggests they
303 arose through horizontal gene transfer(s). This analysis also suggests that
304 kinetoplastid 7TMICs (Kineto7TMICs) branch more proximally to prokaryotic
305 7TMICs, consistent with the hypothesis that kinetoplastids (and allies; collectively
306 Discoba) split early in eukaryotic evolution⁴⁴. The median sampled tree (**Figure**
307 **4A**) generally represents this diverse tree space: Kineto7TMICs branch proximally
308 to Arch/Bac7TMICs (here deeply, but with low branch support; 0.79 and 0.409 for
309 the two most proximal branches); Class-A is monophyletic, with modestly strong
310 branch support (0.91); and Class-B is paraphyletic, but with extremely low branch
311 support on the relevant branch (0.22) (**Figure 4A**).

312 To complement the sequence-based phylogenetic approach, we also
313 employed a recently-developed structure-based phylogenetic method (fold_tree),
314 which infers a minimum evolution tree from a matrix of Foldseek-derived structural
315 alignments⁴⁵. We made three notable observations of the resulting tree (**Figure**
316 **4B**), which shared many similarities to the sequence-based phylogenies (**Figure**
317 **4A**). First, the prokaryotic branch most proximal to the Euk7TMICs included
318 Heimdallarchaeota 7TMICs (**Figure 4B**, asterisk), consistent with their proposed
319 relation to eukaryotes, and suggesting that the shared Euk7TMIC structure (i.e.
320 longer TM4 and TM5 (**Figure 2**)) emerged just before eukaryogenesis. Second,
321 Kineto7TMICs were placed as the sister clade to all other eukaryotic 7TMICs,

323 consistent with their presumed early branching⁴⁴. Third, Class-B Euk7TMICs were
324 essentially monophyletic.

325 The most parsimonious interpretation of these data is that the Class-
326 A/Class-B split is the result of a gene duplication which occurred after
327 eukaryogenesis, but before the speciation event(s) leading to Amorphea and
328 Diaphoretickes (**Figure 4C-D**).

329

330 **Concluding remarks**

331

332 We have described a structure- and sequence-based screening strategy for
333 identifying extremely distant transmembrane protein homologs, revealing that
334 7TMICs are present across the tree of life, including novel discoveries of
335 representatives in Bacteria and Archaea. We have also shown that, despite
336 substantial pairwise sequence dissimilarity, 7TMICs have extremely high structural
337 similarity and identifiable family-wise sequence similarity. Together our results
338 provide the first strong evidence that these disparate proteins form a single,
339 homologous superfamily. This finding contrasts with the Type-I and Type-II opsins,
340 whose structural similarity to each other might represent a case of convergent
341 evolution^{37,46}. Despite the phylogenetic breadth and conserved structure of
342 7TMICs, our knowledge of their function is almost entirely restricted to a subset of
343 insect proteins^{47,48}, which represents only a single, insect-specific lineage of this
344 family. Our work lays a foundation for the analysis of the presumably diverse
345 functions of 7TMICs across a wide range of species. Moreover, we suspect this
346 ancient and cryptic superfamily is only one of many that wait to be discovered in
347 the depths of the twilight zone of sequence space.

348

349 **Acknowledgements**

350

351 We thank Matteo Dal Peraro, Lucien Krapp and members of the Benton laboratory
352 for comments on the manuscript, and Christophe Dessimoz for support of this work.
353 Silhouettes in Figures 1 and 4 were sourced from PhyloPic (www.phylopic.org/).
354 N.J.H. is supported by a Human Frontier Science Program Long-Term Postdoctoral
355 Fellowship (LT-0003/2022L). D.M. is supported by a Swiss National Science
356 Foundation grant (216623) to Christophe Dessimoz. Research in R.B.'s laboratory
357 is supported by the University of Lausanne, an ERC Advanced Grant (833548) and
358 the Swiss National Science Foundation (310030B-185377).

359

360 **Author Contributions**

361

362 Conceptualization, NJH and RB; Methodology, NJH; Software, NJH and DM;
363 Validation, NJH and DM; Formal analysis, NJH and DM; Investigation, NJH;
364 Resources, NJH; Data Curation, NJH and DM; Writing – Original Draft, NJH;
365 Writing – Review & Editing, NJH, DM, and RB; Supervision, RB; Project
366 administration, NJH and RB; Funding acquisition, NJH and RB.

367 DM designed and performed the fold_tree analysis. NJH performed all other
368 formal analyses.

369

370 **Declaration of interests**

371

372 The authors declare no competing interests.

373

374 **Figure Legends**

375

376 **Figure 1. A structure- and sequence-based screen for the identification and**
377 **validation of extremely distant 7TMIC homologs.**

378 (A) Left: top and side views of the cryo-EM structure of the *A. bakeri* Orco
379 homotetramer (PDB 6C70), with one subunit colored ²⁰. Right: transmembrane
380 prediction of *A. bakeri* Orco by DeepTMHMM and Phobius illustrating the
381 characteristic membrane topology of 7TMICs. A cartoon representation of 7TMIC
382 membrane topology is shown below.

383 (B) Aligned cryo-EM structures of *A. bakeri* Orco and *M. hrabei* Or5 (PDB 7LIC).

384 (C) Aligned cryo-EM and AlphaFold structures of *A. bakeri* Orco.

385 (D) Sequence identity versus structural similarity for all pairwise comparisons of
386 AlphaFold models of *D. melanogaster* Ors and Grs, using Dali. The cluster of dots
387 at the top right are self-to-self comparisons and isoforms of the same gene.

388 (E) Proof of principle Foldseek screen of the AlphaFold structural proteome of *D.*
389 *melanogaster*, with results of the screen plotted by Foldseek-derived percent amino
390 acid sequence identity and E-values.

391 (F) Outline of the screen and validation pipeline.

392 (G) Cladogram of taxa in which 7TMICs were identified.

393 (H) All-to-all BLASTP network of 7TMICs (each represented by a dot), which
394 visualizes only pairwise sequence similarity. Several clusters form, suggesting
395 monophyly within clusters (annotated manually based on CLANS clustering). At
396 presumed longer evolutionary distances, 7TMICs show little-to-no pairwise
397 sequence identity, represented by the weak connectivity (i.e. few edges) between
398 most clusters.

399

400 **Figure 2. Examples of newly-identified 7TMICs.**

401 Transmembrane predictions, and top and side views of the AlphaFold structure of
402 newly-identified 7TMICs.

403 (A) Representative example of a Gr- and DUF3537-like Euk7TMIC, subsequently
404 phylogenetically classified as Class-A (**Figure 4**). These proteins have all the
405 stereotyped 7TMIC features.

406 (B) Representative example of a PHTF-like Class-B 7TMIC (**Figure 4**). These
407 proteins have stereotyped 7TMIC features, with the addition of a long intracellular
408 loop between TM2 and TM3 (IL1).

409 (C-E) Representative examples of a Bac7TMIC and Arch7TMICs. When clustered
410 by Foldseek at 90% coverage (data not shown), prokaryotic 7TMICs form three
411 structure clusters (represented here in (C), (D), and (E)), although they cannot be
412 easily distinguished visually. These proteins all share the stereotyped 7TMIC
413 features but, with the exception of Heimdallarchaeota 7TMICs (E), have shorter
414 TM4 and TM5.

415 (F) Representative example of the small number of bacterial and archaeal proteins
416 with extreme fold and high sequence-similarity to Euk7TMICs, which are presumed
417 to have arisen through horizontal gene transfer(s) (HGT) (**Figure 4**).

418

419 **Figure 3. Evidence for 7TMIC homology through structural and sequence**
420 **similarity.**

421 (A) Structural similarity network of 7TMICs derived from all-to-all TM-scores
422 (schematized at the top).

423 (B) Sequence similarity network of 7TMICs produced by all-to-all PSI-BLAST

424 searches (schematized at the top; iteration 3 is shown), providing evidence that
425 7TMICs have a family-wide sequence profile. This pattern of strong, bidirectional
426 linkages became apparent already in PSI-BLAST iteration 2, while subsequent
427 iterations resembled iteration 3 (**Figure S2A** and Supplemental Material).

428 **(C)** Amino acid sequence identity derived from a query-centered Foldseek
429 alignment for the centermost node in the structural similarity network (*Symbiodium*
430 *natans* A0A812K102). Transmembrane predictions are from Phobius. TM7b was
431 annotated manually. TM0 (the re-entrant loop) is indicated with a dashed line, as it
432 is inconsistently predicted by DeepTMHMM and Phobius, and often predicted with
433 low confidence region in AlphaFold models. Query-centered alignments for all
434 7TMIC models analyzed here are available in the supplemental data.

435 **(D)** Average sequence embedding-based conservation scores for 7TMICs, with the
436 curves interpolated to match the length of A0A812K102. Column conservation
437 scores are significantly correlated with column sequence identity for A0A812K102
438 (**Figure S2B**). The location and strength of conservation likely varies by 7TMIC
439 family and subfamily, and excluding the well-established 7TMICs (insect Ors/Grs
440 and animal Grls) led to overall increased conservation scores (light blue line, also
441 **Figure S2C**). Embedding-based conservation score for all 7TMIC models analyzed
442 here are available in the supplemental data.

443 **(E)** Average PeSTo protein-protein interaction predictions. The region with the
444 most consistently-predicted protein-protein interactions is near the C-terminus,
445 correlating with the site of highest sequence conservation; again, exclusion of
446 Ors/Grs/Grts led to higher prediction scores (light orange line). PeSTo predictions
447 for all 7TMIC models analyzed here are available in the supplemental data.

448 **(F)** Sequence-embedding based conservation scores (blue) and PeSTo-derived
449 protein-protein interaction scores (orange) mapped onto the AlphaFold model of
450 A0A812K102.

451 **(G)** Top: top (presumed extracellular) and bottom (presumed intracellular) views of
452 a hypothetical tetramer of A0A812K102 (predicted by AlphaFold-Multimer),
453 showing that individual subunits have their closest interactions in the pore and
454 anchor domains, similar to Ors^{20,21}. Bottom: side view of the A0A812K102
455 tetramer, with two subunits masked for clarity, and the presumed anchor and pore
456 regions colored on the visualized subunits. In total, we modelled 85 tetramers; 83
457 of these were Or-like, in that they displayed rotational symmetry, with the closest
458 interactions in the hypothetical anchor and pore regions (further examples in
459 **Figure S2D**).

460

461 **Figure 4. A model for the evolution of the 7TMIC superfamily.**

462 **(A)** The median phylogenetic tree of 7TMICs sampled from the Robinson-Foulds-
463 based tree space of 48 sequence-based phylogenetic trees. For visualization
464 purposes, the tree is arbitrarily rooted in the last common ancestor of all
465 Arch/Bac7TMICs (which are highly reticulated); the true root is likely at the
466 unidentified location of the last universal common ancestor (LUCA) within the
467 prokaryotic branch. Branch lengths are derived from the average number of
468 substitutions per site. Tree space is visualized in **Figure S2C-E**, and all trees and
469 alignments are available in the Supplemental Data.

470 **(B)** TM-score based structural tree of 7TMICs derived from fold_tree, automatically
471 rooted using the MAD method. As in (A), the true root is likely at the unidentified
472 location of LUCA. Branch lengths are derived from the underlying distance matrix
473 of TM-scores. The asterisk marks the branch containing Heimdallarchaeota

474 7TMICs; a fully annotated tree is available in the supplemental data.
475 (C) Collapsed version of the tree in (B) highlighting the major branching patterns.
476 Kinetoplastid 7TMICs likely branched early, while the Class-A/Class-B split
477 occurred after the emergence of the last eukaryotic common ancestor (LECA), but
478 before the split(s) leading to Amorphea and Diaphoretickes.
479 (D) Summary of the results of the screen and evolutionary analyses. The left tree
480 shows assumed relationships between the various taxa in which 7TMICs were
481 identified, while the top tree shows the evolutionary history of 7TMICs themselves.
482 The colored dots represent the presence or absence of 7TMIC families. At the
483 subfamily level, many Class-B PHTF-like proteins may be the result of horizontal
484 gene transfer(s), as there is broad but sparse taxonomic diversity within this
485 putative subfamily. Note that this screen did not recover previously identified
486 7TMICs from Amoebozoa or Chytridiomycota, which were inferred to be Class-A
487 based on previously described sequence similarity to insect Grs/Ors⁸; in addition,
488 here, “Fungi” only refers to Chytridiomycota and Blastocladiomycota.
489

490 **STAR Methods**

491 **Resource availability**

494 **Lead contact**

495 Further information and requests for resources and reagents should be directed to
496 and will be fulfilled by the lead contacts, Nathaniel Himmel
497 (nathanieljohn.himmel@unil.ch) and Richard Benton (richard.benton@unil.ch).
498

500 **Materials availability**

501 This study did not generate new unique reagents.
502

504 **Data availability**

505 All data have been deposited in Dryad (<https://doi.org/10.5061/dryad.fqz612jz9>)
506 and are publicly available as of the date of the publication.
507

508
509

510 **Method details**

511

512 **Structural screen and validation**

513

514 Proof-of-concept screens were carried out using local implementations of Foldseek
515 ⁵ and DaliLite^{26,27}. Subsequent structure-based screens of the AlphaFold Protein
516 Structure Database (<https://alphafold.ebi.ac.uk/>)^{4,6} were performed on the
517 Foldseek server (<https://search.foldseek.com/>), using the following query
518 structures/models: *A. bakeri* Orco (6C70); *M. hrabei* Or5 (7LIC); *D. melanogaster*
519 Gr1Hz (Q9W1W8); *B. belcheri* Gr1Hz (A0A6P5ACQ6); *T. adhaerens* Gr1Hz
520 (B3RTY0); *Z. mays* DUF3537 (A0A1D6LEW8, B4FJ88, and B6SUZ0); *P. patens*
521 DUF3537 (A0A2K1ICX7, A0A2K1JKU0, and A0A2K1L324); *D. melanogaster* Phtf
522 (Q9V9A8); *H. sapiens* PHTF1 (Q9UMS5) and PHTF2 (Q8N3S3); *P. halstedii* PHTF
523 (A0A0P1B782); *L. infantum* GRL1 (A4HWQ9); and *T. brucei* brucei GRL1
524 (Q57U78) (**Figure S1D**). For the initial screen, we masked the WD40 repeats in
525 trypanosome GRL1 and the long intracellular loop 1 in PHTF, thus restricting the
526 search to the core 7TMIC domain. We did not set a statistical threshold (E-value)
527 for putative homolog identification. For eukaryotic hits, we initially considered all
528 hits from the screen. For archaeal and bacterial hits, we took the more stringent
529 approach of only further analyzing those that were hits for all the query groups
530 (annotated in **Figure S3A**). We did not formally screen animal or vascular land
531 plant species because we considered that these taxa have been sufficiently
532 screened^{7–10,19}, and we were most interested in the very early evolution of 7TMICs.
533 Indeed, preliminary Foldseek screens did not elucidate any obvious new plant-
534 and/or animal-specific 7TMICs (data not shown).

535 Subsequent validation was performed in several steps. First, transmembrane topology was predicted using DeepTMHMM (the BioLib
536 implementation at <https://dtu.biolib.com/DeepTMHMM/> and a local
537 implementation)²⁹. For putative eukaryotic homologs, we assessed these
538 predictions alongside structural models (visualized in PyMol), looking for: (i) 7
539 predicted transmembrane alpha helices; (ii) shorter extracellular than intracellular
540 loops; (iii) an intracellular N-terminus and extracellular C-terminus; (iv) longer TM4,
541 TM5, and TM6 helices; and (v) the exceptional “split” TM7 helix^{20,21,24,25}. We did
542 not consider the re-entrant loop (TM0) as a criterion, as it is inconsistently predicted
543 by transmembrane prediction methods^{8,10}. For archaea and bacteria, we only
544 further assessed hits with exactly 7 predicted transmembrane segments in the
545 stereotyped architecture. We also used Phobius (<https://phobius.sbc.su.se/>)^{49,50}
546 and the transformer model PeSTo (<https://pesto.epfl.ch/>)³⁹ to predict
547 transmembrane topology; both were used for visualization, but neither was used to
548 curate sequences. Finally, we used a local implementation of DaliLite to compare
549 all remaining hits with the original query structures and three negative controls. For
550 the negative controls we selected an Adiponectin receptor (*Homo sapiens* ADPR1;
551 5LXG) and a channelrhodopsin (*Chlamydomonas reinhardtii* Channelrhodopsin-2;
552 6EID) in advance of the screen, as both have 7 transmembrane domains but are
553 unrelated to 7TMICs; we added the ABC transporter permease (*Escherichia coli*
554 A0A061Y968) *post hoc*, as many of the screen hits were errantly annotated as ABC
555 transporters. Only hits with Dali Z-scores >8 as compared to 7TMIC queries were
556 further analyzed. This threshold is based on Holm’s criteria⁵¹, where Z-scores >20
557 indicate definite homology, 8–20 probable homology, 2–8 a “gray area” (here,
558 “twilight zone”) and <2 non-significant (here, “midnight zone”). We conceptualized
559

560 these scores as “protein fold similarity” in place of “homology,” as we infer
561 homology based on a holistic view of sequence, structure, and taxonomic features.
562 Pearson’s correlation analysis and the Bayesian equivalent were performed in
563 JASP (<https://jasp-stats.org/>).

564

565 Sequence-based homolog identification

566

567 For putative eukaryotic homologs, the results of the Foldseek screen were used to
568 select query sequences. CLANS was used to generate an all-to-all BLASTP
569 network (E-value cutoff 0.01), which was subsequently clustered by the global
570 network clustering option^{52–54}. PSI-BLAST homolog searches were carried out
571 using all singlets and a representative sequence from each cluster (the node with
572 the highest neighborhood connectivity). Searches were run on the NCBI server
573 (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) against the clustered non-redundant
574 (clustered_nr) sequence database, until convergence. PSI-BLAST searches were
575 performed with an E-value cutoff of 0.05, but final candidates were selected only if
576 they had a minimum coverage of 50% (with coverage of the transmembrane region)
577 and a final E-value at or below 10⁻¹⁰. For searches recovering canonical animal
578 Grs/Ors/Grls, the PSI-BLAST searches were stopped when the top 1000 hits were
579 recovered, as these searches quickly converged on tens of thousands of
580 predominately insect sequences, which was computationally time-consuming and
581 methodologically unnecessary for this study.

582

583 For Arch7TMIC homologs, sequence databases were likewise assembled
584 using PSI-BLAST, using each of the structural screen hits as a query sequence.
585 Compared to the eukaryote-based searches, we took a more stringent approach,
586 setting an E-value cutoff of 10⁻¹⁰ for both the PSI-BLAST search and final hit
587 selection. Query sequences that were orphans, or which had very few sequence-
588 based homologs (<10), were excluded from further analyses. These searches
589 recovered the Bac7TMICs, so efforts were not repeated using Bac7TMIC queries.

590

591 After preliminary homolog identification, DeepTMHMM was used to predict
592 transmembrane topology. For all sequences identified via PSI-BLAST, we kept
593 sequences with >6 (rather than 7) TM segments, as DeepTMHMM had previously
594 failed to predict TM7 despite the presence of TM7 helices in the associated
595 structural models¹⁰. Finally, to reduce redundancy, and thus simplify computation
596 and presentation, CD-HIT (<https://cd-hit.org>)^{55,56} was used to cluster sequences—
597 first by 70% for the initial BLASTP sequence similarity network (**Figure 1H**), then
598 by 50% for all subsequent analyses—keeping the longest sequence as the cluster
599 representative.

600

601 A notable limitation of this approach is the use of metagenomics for the
602 identification of some prokaryotic 7TMICs. As these data are assembled from
603 environmental samples, these sequences could be misidentified. While this
604 possibility cannot be completely discounted, it is not a compelling problem, as most
605 of the metagenomically identified sequences described herein correspond to tens
606 to hundreds of homologous proteins in closed prokaryotic genomes. The only
607 obvious exceptions are the small number of archaeal and bacterial sequences most
608 closely resembling Class-B Euk7TMICs.

609

610

611

612

613

610 ***Ab initio* protein folding and structural analyses**

611

612 All monomer models were downloaded from the AlphaFold Protein Structure
613 Database. Protein multimers (5 models each) were generated for *A. bakeri* Orco,
614 the example 7TMICs in **Figure 2**, *Symbiodinium natans* A0A812K102 (the most
615 central node in the structural network, **Figure 3A**), and 7 additional structures
616 derived from Foldseek clustering of 7TMICs by 50% alignment coverage (thus
617 representing nearly the entire 7TMIC fold space, **Figure S2D** and supplemental
618 data). Predictions were performed in Google Colaboratory
619 (<https://research.google.com/colaboratory>) using AlphaFold2+MMSeqs2 as
620 implemented by Colabfold (<https://github.com/sokrypton/ColabFold>)⁴⁰⁻⁴³. These
621 models were not interpreted as accurate predictions of protein stoichiometry, but
622 rather as hypothetical tetramers and as indirect predictions of protein-protein
623 interactions. We also generated hypothetical dimers, trimers, and pentamers for *A.
624 bakeri* Orco (available in the supplemental data) and observed that the protein
625 subunits assembled in a globally similar way – i.e. closest contact at the anchor
626 domain(s). Transmembrane prediction was performed using DeepTMHMM,
627 Phobius, and PeSTo web servers, as described above. Protein-protein interactions
628 were predicted using a local implementation of PeSTo (<https://github.com/LBM-EPFL/PeSTo>). All proteins were visualized in PyMol. Visualized structural
629 alignments were generated using Coot⁵⁷.

630

631 **Network and conservation analyses**

632

633 We used graph-based strategies for visualizing relatedness among proteins⁵⁸.
634 Structure-based networks were generated from the results of all-to-all DaliLite or
635 Foldseek searches, where connections are derived from Z-scores >8 or TM scores
636 >0.5, respectively. BLASTP sequence-based networking was performed using the
637 CLANS webserver (<https://toolkit.tuebingen.mpg.de/tools/clans>) and a local
638 implementation of CLANS⁵²⁻⁵⁴, using attraction values derived from E-values
639 <0.01; clusters were identified using the built-in network clustering algorithm with
640 the global averages option.

641

642 PSI-BLAST networking was performed via all-to-all PSI-BLAST searches
643 using a local implementation of BLAST+³⁶. First, BLAST databases were prepared
644 from the sequences databases described above. Insect Ors were excluded, as they
645 are an insect-specific radiation^{28,59}; their removal thus reduced the likelihood of
646 spurious connectivity between distantly related 7TMICs, as demonstrated by their
647 relatively high connectivity in the BLASTP network (see **Figure 1H**). In other words,
648 the removal of Ors hypothetically weakened network connectivity overall, but
649 increased our confidence in homology between linked sequences. BLAST+ was
650 then used to perform all-to-all PSI-BLAST searches, stopping at either
651 convergence or 10 iterations. PSSMs were generated with an E-value cutoff of 0.01
652 and the final network was assembled from hits where the PSSM query coverage
653 was >70%. For any query-to-subject relationship, only the first significant PSI-
654 BLAST hit was kept, corresponding to the weakest significant connection (as
655 connections tend to strengthen in subsequent PSI-BLAST iterations), thus
656 providing the most conservative interpretation of the network. The opsin
657 control/outgroup databases were from previous studies^{60,61}.

658

659 All networks were visualized, annotated, and quantitatively analyzed in
CLANS, CytoScape⁶² and Adobe Illustrator.

660 For conservation analyses, query-centered sequence alignments were first
661 produced by Foldseek; in the figures, we visualized the alignment from the model
662 with the highest closeness centrality (i.e. the centermost model; A0A812K102) from
663 the structural similarity network. Amino acid sequence identity scores were
664 calculated in Jalview. Embedding-based conservation scores were calculated
665 using the esm2_t33_650M_UNR50D protein language model⁶³, via the methods
666 and scripts described by³⁸ (<https://github.com/esbgkannan/kibby>). The mean
667 conservation scores were calculated by spline interpolating each individual data
668 series (corresponding to each protein) to match the length of A0A812K102, then
669 averaging those values; as such, family- and subfamily-specific conservation
670 patterns are likely not represented in the average curve. The embedding-based
671 conservation scores, PeSTo predictions, and query-centered multiple sequence
672 alignments for all representative models are available in the supplemental data.
673 Pearson's correlation analysis (and the Bayesian equivalent) was performed in
674 JASP.

675
676 **Phylogenetics**
677

678 7TMIC GenBank accession numbers from our 50% clustered sequence database
679 were matched to UniProt and 1947 AlphaFold-derived protein models were
680 downloaded from the AlphaFold Protein Structure Database. All subsequent
681 phylogenetic analyses were carried out on these 1947 representative proteins.

682 Muscle5 was used to generate the ensemble of multiple sequence
683 alignments (MSAs)⁶⁴. Because the alignments were extremely long and gap rich
684 (**Table S1**), MSAs were trimmed using trimal with the -gappyout option⁶⁵. Each
685 trimmed MSA was then used to generate phylogenetic trees using FastTree2⁶⁶,
686 using 3 different amino acid substitution models (JTT, WAG, and LG), and with
687 branch lengths rescaled to optimize the Gamma20 likelihood. The initial MSAs had
688 extremely high dispersion and extreme lack of consensus (**Figure S3B**) indicating
689 widespread alignment errors⁶⁴. These errors resulted in non-trivial topological
690 differences in the phylogenetic trees, resulting in extreme non-consensus (even for
691 obviously monophyletic clades, such as the insect Ors). This suggested a high
692 degree of phylogenetic instability, likely due to both alignment errors (from low PID)
693 and phylogenetic errors (e.g. long branch attraction).

694 To minimize alignment and phylogenetic errors, we repeated MSA and tree
695 inference after identifying and removing rogue taxa (i.e. the most unstable leaves
696 in the previous ensemble analysis) via RogueNaRok⁶⁷. Although the resulting
697 ensemble of MSAs still had high dispersion, the resulting phylogenetic trees were
698 more consistent in the assignment of the various subfamilies as monophyletic
699 clades (**Figure S3C**). These trees were used for subsequent analysis.

700 The structural phylogeny was generated using fold_tree⁴⁵. Here, we
701 emphasize the tree derived from all-against-all TM-scores, thereby sampling
702 structural space based on pairwise global rigid structural comparisons, mirroring
703 our network-based analysis, as described above. Structural trees derived from
704 pairwise distances based on the Foldseek structural alphabet (**Figure S3F**) or
705 pairwise IDDT scores (**Figure S3G**) produced radically different topologies; neither
706 has obviously high congruence with the sequence-based phylogenetics, nor with
707 the presumed taxonomy of the species included in this analysis. All trees were
708 analyzed using the ape⁶⁸, phytools⁶⁹, and treespace⁷⁰ R packages. Tree topology
709 space was explored by principal coordinate analysis of the Robison-Folds

710 distances between the unrooted phylogenies. Trees were visualized and annotated
711 using R, iTol (<https://itol.embl.de/>)⁷¹, and Adobe Illustrator.

712

713 Protein nomenclature

714

715 Most previous naming conventions have not been evolutionarily informed. Terms
716 such as Gustatory receptor-like (Grl and GRL) do not refer to monophyletic clades,
717 but instead correspond to many taxon-specific 7TMIC branches. For animal Grls
718 and unicellular eukaryotic GRLs, the terms were chosen because they resembled
719 insect Grs in either amino acid sequence and/or tertiary structure^{8,9,19}; by contrast,
720 insect Grls were named based on the *absence* of sequence similarity to Grs despite
721 the presence of structural similarity¹⁰. While these terms are useful in situational
722 contexts, they are uninformative at long evolutionary scales. We propose that the
723 7TMIC superfamily be split into domain-specific families; for eukaryotes, these are
724 Class-A and Class-B. We suggest that the more complex nomenclature of previous
725 work (e.g. Or, Gr, GrlHz) should be reserved for taxon-specific contexts. Relatedly,
726 the evolution of Arch7TMICs and Bac7TMICs is highly reticulated. Although we
727 saw proximity between Heimdallarchaeota 7TMICs and Euk7TMICs in our
728 structure-based phylogeny, there were no other clear recapitulations of
729 Asgard/Eukaryota monophyly or the Archaea-Bacteria split. Therefore,
730 “Arch7TMIC” and “Bac7TMIC” serve only as terms of convenience, and we strongly
731 caution that they do not refer to monophyletic clades.

732 Supplemental Figures

733

734 **Figure S1. All-to-all pairwise protein similarity networks of *D. melanogaster*** 735 **7TMICs, Foldseek benchmarking, and summary of the Foldseek screen.**

736 (A) All-to-all BLASTP network of *D. melanogaster* 7TMICs; consistent with their low
737 pairwise sequence similarity, this analysis fails to link every 7TMIC to all others.
738 Rather, the major *D. melanogaster* classes (Ors and Grs) are separated into two
739 identifiable community structures, with sparse connectivity among the Grs, and
740 between the Grs and Ors. Other 7TMICs—including Grls, GrlHz, Phtf and two
741 Grs—form singlets, indicating an inability to identify hypothetical homologs using
742 BLASTP.

743 (B) All-to-all Dali network of *D. melanogaster* 7TMICs. In contrast to (A), structural
744 comparisons result in a “hairball” structure, wherein nearly all proteins are linked to
745 all others, excepting Phtf, which is presumed to be the most distantly related.

746 (C) Plots of structural similarity scores between Orco and other *D. melanogaster*
747 7TMICs, comparing Dali to Foldseek-derived scores. Foldseek generates Orco-to-
748 all E-values that tightly correlate with the rapidly generated 3Di+AA-derived E-
749 values (top) and the slowly generated TM-align-derived TM-scores (bottom).

750 (D) Protein models used in the Foldseek screen, and negative controls used for
751 subsequent Dali-based validation, with a clustering dendrogram based on all-to-all
752 Dali comparisons between the queries and negative controls. The dendrogram is
753 derived from the Dali Z-score distance matrix. The heatmap shows all-to-all Dali Z-
754 scores and TM-scores.

755 (E) Stacked density plot showing the frequency distribution of the hits of the
756 Foldseek screen, by E-value, with the inset pie-chart showing the proportion of true
757 positives to false positives. Most true positives had relatively poor E-values, with
758 similar or worse scores than many false positives, demonstrating the need for
759 structural validation in a Foldseek screen.

760

761 **Figure S2. Initial iterations of the PSI-BLAST sequence similarity networks,** 762 **7TMIC sequence conservation analysis, and predicted quaternary structures** 763 **of select, newly-identified 7TMICs.**

764 (A) Sequence similarity networks were generated by all-to-all PSI-BLAST searches
765 of a 50% clustered sequence database of 7TMICs, alongside databases of Type-I
766 and Type-II opsins. Iterations 1 and 2 are visualized here. Subsequent iterations
767 resemble the clustering pattern of iteration 3, as visualized in **Figure 3**, albeit with
768 strengthening community structures. Left: PSI-BLAST iteration 1. In this network,
769 sequences formed several non-contiguous clusters, and failed to cluster together
770 7TMICs and Type-I opsins, which is expected given the substantial sequence
771 dissimilarity of 7TMICs. Right: PSI-BLAST iteration 2. Surprisingly, PSI-BLAST
772 networking produced bidirectional linking of the majority of 7TMICs, although
773 presumed spurious linkages to outgroups began to form (which did not greatly
774 multiply in subsequent iterations), and a small number of 7TMICs do not form links
775 to the core 7TMIC cluster(s) (although all join a 7TMIC community structure by
776 iteration 3 (**Figure 3B**)).

777 (B) Embedding-based conservation scores weakly but significantly correlate with
778 column sequence identity from the A0A812K102-centered sequence alignment.

779 (C) Average embedding-based conservation scores for different subsets of
780 7TMICs, demonstrating that, while family-specific patterns exist, the conservation
781 of anchor domain and pore regions is consistent. The TM and domain labels are

782 derived from A0A812K102, as visualized in **Figure 3**.
783 (D) Predicted tetramers for select 7TMICs. Top: top (presumed extracellular) and
784 side views of the tetrameric arrangement of 7TMICs predicted by AlphaFold-
785 Multimer, showing the formation of a hypothetical pore along TM7b, similar to *A.
bakeri* Orco (far-left). Bottom: local Distance Difference Test (IDDT) scores (used
786 to assess model confidence), plotted for each of the 5 replicate models generated.
787 Each color represents a different replicate. Vertical black lines separate each of the
788 modelled subunits. Generally, the transmembrane-spanning alpha helices are the
789 most confidently predicted, leading to the similar pattern of IDDT peaks and troughs
790 across models.
791

792

793 **Figure S3. Phylogenetic and tree space analysis.**

794 (A) Pipeline for sequence-based phylogenetic analysis. First, an ensemble of 16
795 multiple sequence alignments (MSAs) are made by perturbing the guide tree and
796 the Hidden Markov model's pseudorandom number generator (HMM PRNG).
797 Second, phylogenetic trees are generated for each of the MSAs, using 3 different
798 amino acid substitution models, resulting in 48 trees. Finally, differences in the
799 topology of the 48 trees are calculated by pairwise Robinson-Foulds distances; the
800 resulting distance matrix is subsequently visualized in two dimensions by principal
801 coordinate analysis (PCoA).

802 (B) Majority consensus tree for the 48 phylogenetic trees based on alignments of
803 the representative 7TMIC sequences. 7TMIC clades/colors were assigned
804 manually based on visual inspection of a CLANS-based clustering analysis. Branch
805 colors indicate the percent consensus. There is essentially no clear consensus
806 among these 48 initial trees; obviously monophyletic clades—such as insect Ors—
807 are not reliably predicted, suggesting substantial alignment/phylogenetic errors (as
808 expected for this highly divergent superfamily).

809 (C) Majority consensus tree for the 48 phylogenetic trees based on alignments of
810 a 7TMIC dataset where rogue taxa (i.e. the most phylogenetically unstable leaves)
811 have been removed (with colors matching (B)). While there is still no greatly
812 informative majority consensus topology, this analysis better recapitulates more
813 obvious monophyletic clades, with higher branch consensus, indicating that errors
814 have been minimized (but not eliminated, which we did not expect to occur at these
815 levels of sequence dissimilarity).

816 (D) PCoA of Robinson-Foulds tree space for trees from (C). Trees form 6 topology
817 clusters.

818 (E) Majority consensus trees for each of the 6 clusters, with colors matching (C)
819 and (D). Five of these clusters agree that Kinet7TMICs branch proximally to
820 prokaryotic 7TMICs, consistent with the hypothesis that kinetoplastids (and allies:
821 Discoba) split early in eukaryotic evolution⁴⁴. Clusters 1 and 4 do not have majority
822 consensus on deep 7TMIC branching. The remaining clusters suggest there are at
823 least two Euk7TMIC families, termed Class-A and Class-B Euk7TMICs, but do not
824 agree on the monophyly of Class-B Euk7TMICs. Clusters 4-6 suggest Class-B
825 monophyly, while clusters 1-3 suggest that many proteins are basally branching
826 (and thus, paraphyly). Given that structure-based phylogenetics suggest a
827 monophyletic Class-B, this discordance may be the result of lingering long branch
828 attraction or other errors resulting from the inclusion of rapidly evolved, horizontally-
829 transferred, or structurally-convergent proteins.

830 (F) Structural phylogeny derived from pairwise distances used the Foldseek 3Di
831 structural alphabet, with colors matching the panels above. This tree is presented

832 as rooted, but as in Figure 4, the true root is likely within the prokaryotic 7TMICs,
833 at the location of the Last Universal Common Ancestor.

834 (G) Structural phylogeny derived from pairwise IDDT scores, with colors matching
835 the panels above. As in (F), the true root is likely at location of the Last Universal
836 Common Ancestor.

837

838 **Supplemental Table**

839

840 **Table S1. Muscle5 multiple sequence alignment analysis.**

841 Column confidence is a measure of the reproducibility of each column, where 0
842 indicates the column is never found, and 1 indicates it is found across all
843 alignments. Dispersion is measured as the median dispersion of aligned letter pairs
844 over the ensemble (D_LP), and the median dispersion of columns over the
845 ensemble (D_Cols) (Robert Edgar, personal communication, 10 May 2023), where
846 0 is all the same and 1 is all different. Dispersion was extremely high. For the initial
847 set of alignments: D_LP=0.5836 D_Cols=1.0000. After removal of rogue taxa:
848 D_LP=0.5855 D_Cols=1.0000.

849

MSA Replicate	MSA Perturbations		All Sequences		No Rogue Taxa	
	Guide Tree	PRNG	Columns	Column Confidence	Columns	Column Confidence
1	abc	0	38154	0.315	36865	0.311
2	abc	1	38357	0.297	34833	0.295
3	abc	2	36970	0.314	35376	0.325
4	abc	3	37064	0.314	31247	0.304
5	acb	0	38375	0.31	35627	0.304
6	acb	1	40342	0.298	34754	0.302
7	acb	2	37080	0.317	34994	0.321
8	acb	3	35735	0.313	31345	0.305
9	bca	0	38391	0.315	35914	0.314
10	bca	1	39831	0.3	35813	0.294
11	bca	2	37647	0.316	34847	0.322
12	bca	3	36515	0.305	31406	0.303
13	none	0	38700	0.3	35842	0.309
14	none	1	40114	0.306	34808	0.293
15	none	2	37443	0.308	34932	0.309
16	none	3	37831	0.309	32689	0.298

850

851 **References**

- 852
- 853 1. Doolittle, R.F. (1986). Of Urfs And Orfs: A Primer on How to Analyze Derived
854 Amino Acid Sequences (University Science Books).
- 855 2. Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and
856 Design* 2, S19–S24. 10.1016/S1359-0278(97)00059-X.
- 857 3. Illergård, K., Ardell, D.H., and Elofsson, A. (2009). Structure is three to ten
858 times more conserved than sequence—A study of structural response in
859 protein cores. *Proteins: Structure, Function, and Bioinformatics* 77, 499–508.
860 10.1002/prot.22458.
- 861 4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
862 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021).
863 Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–
864 589. 10.1038/s41586-021-03819-2.
- 865 5. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist,
866 C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein
867 structure search with Foldseek. *Nat Biotechnol*, 1–4. 10.1038/s41587-023-
868 01773-0.
- 869 6. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova,
870 G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein
871 Structure Database: massively expanding the structural coverage of protein-
872 sequence space with high-accuracy models. *Nucleic Acids Research* 50,
873 D439–D444. 10.1093/nar/gkab1061.
- 874 7. Benton, R. (2015). Multigene Family Evolution: Perspectives from Insect
875 Chemoreceptors. *Trends in Ecology & Evolution* 30, 590–600.
876 10.1016/j.tree.2015.07.009.
- 877 8. Benton, R., Dessimoz, C., and Moi, D. (2020). A putative origin of the insect
878 chemosensory receptor superfamily in the last common eukaryotic ancestor.
879 *eLife* 9, e62507. 10.7554/eLife.62507.
- 880 9. Robertson, H.M. (2015). The Insect Chemoreceptor Superfamily Is Ancient in
881 Animals. *Chemical Senses* 40, 609–614. 10.1093/chemse/bjv046.
- 882 10. Benton, R., and Himmel, N.J. (2023). Structural screens identify candidate
883 human homologs of insect chemoreceptors and cryptic *Drosophila* gustatory
884 receptor-like proteins. *eLife* 12, e85537. 10.7554/eLife.85537.
- 885 11. Joseph, R.M., and Carlson, J.R. (2015). *Drosophila* Chemoreceptors: A
886 Molecular Interface Between the Chemical World and the Brain. *Trends in
887 Genetics* 31, 683–695. 10.1016/j.tig.2015.09.005.

- 888 12. Robertson, H.M. (2019). Molecular Evolution of the Major Arthropod
889 Chemoreceptor Gene Families. *Annual Review of Entomology* *64*, 227–242.
890 10.1146/annurev-ento-020117-043322.
- 891 13. Clyne, P.J., Warr, C.G., Freeman, M.R., Lessing, D., Kim, J., and Carlson,
892 J.R. (1999). A novel family of divergent seven-transmembrane proteins:
893 candidate odorant receptors in *Drosophila*. *Neuron* *22*, 327–338.
894 10.1016/s0896-6273(00)81093-4.
- 895 14. Clyne, P.J., Warr, C.G., and Carlson, J.R. (2000). Candidate taste receptors
896 in *Drosophila*. *Science* *287*, 1830–1834. 10.1126/science.287.5459.1830.
- 897 15. Gao, Q., and Chess, A. (1999). Identification of candidate *Drosophila* olfactory
898 receptors from genomic DNA sequence. *Genomics* *60*, 31–39.
899 10.1006/geno.1999.5894.
- 900 16. Scott, K., Brady, R., Cravchik, A., Morozov, P., Rzhetsky, A., Zuker, C., and
901 Axel, R. (2001). A chemosensory gene family encoding candidate gustatory
902 and olfactory receptors in *Drosophila*. *Cell* *104*, 661–673. 10.1016/s0092-
903 8674(01)00263-x.
- 904 17. Vosshall, L.B., Amrein, H., Morozov, P.S., Rzhetsky, A., and Axel, R. (1999).
905 A spatial map of olfactory receptor expression in the *Drosophila* antenna. *Cell*
906 *96*, 725–736. 10.1016/s0092-8674(00)80582-6.
- 907 18. Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley,
908 R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., et
909 al. (2002). Comparative Genome and Proteome Analysis of *Anopheles*
910 *gambiae* and *Drosophila melanogaster*. *Science* *298*, 149–159.
911 10.1126/science.1077061.
- 912 19. Saina, M., Busengdal, H., Sinigaglia, C., Petrone, L., Oliveri, P., Rentzsch, F.,
913 and Benton, R. (2015). A cnidarian homologue of an insect gustatory receptor
914 functions in developmental body patterning. *Nat Commun* *6*, 6243.
915 10.1038/ncomms7243.
- 916 20. Butterwick, J.A., del Mármol, J., Kim, K.H., Kahlson, M.A., Rogow, J.A., Walz,
917 T., and Ruta, V. (2018). Cryo-EM structure of the insect olfactory receptor
918 Orco. *Nature* *560*, 447–452. 10.1038/s41586-018-0420-8.
- 919 21. del Mármol, J., Yedlin, M.A., and Ruta, V. (2021). The structural basis of
920 odorant recognition in insect olfactory receptors. *Nature* *597*, 126–131.
921 10.1038/s41586-021-03794-8.
- 922 22. Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L.B., and
923 Touhara, K. (2008). Insect olfactory receptors are heteromeric ligand-gated
924 ion channels. *Nature* *452*, 1002–1006. 10.1038/nature06850.
- 925 23. Wicher, D., Schäfer, R., Bauernfeind, R., Stensmyr, M.C., Heller, R.,
926 Heinemann, S.H., and Hansson, B.S. (2008). *Drosophila* odorant receptors
927 are both ligand-gated and cyclic-nucleotide-activated cation channels. *Nature*
928 *452*, 1007–1011. 10.1038/nature06861.

- 929 24. Benton, R., Sachse, S., Michnick, S.W., and Vosshall, L.B. (2006). Atypical
930 Membrane Topology and Heteromeric Function of *Drosophila* Odorant
931 Receptors In Vivo. *PLOS Biology* 4, e20. 10.1371/journal.pbio.0040020.
- 932 25. Hopf, T.A., Morinaga, S., Ihara, S., Touhara, K., Marks, D.S., and Benton, R.
933 (2015). Amino acid coevolution reveals three-dimensional structure and
934 functional domains of insect odorant receptors. *Nat Commun* 6, 6077.
935 10.1038/ncomms7077.
- 936 26. Holm, L., Kääriäinen, S., Rosenström, P., and Schenkel, A. (2008). Searching
937 protein structure databases with DaliLite v.3. *Bioinformatics* 24, 2780–2781.
938 10.1093/bioinformatics/btn507.
- 939 27. Holm, L., and Park, J. (2000). DaliLite workbench for protein structure
940 comparison. *Bioinformatics* 16, 566–567. 10.1093/bioinformatics/16.6.566.
- 941 28. Robertson, H.M., Warr, C.G., and Carlson, J.R. (2003). Molecular evolution of
942 the insect chemoreceptor gene superfamily in *Drosophila melanogaster*.
943 *Proceedings of the National Academy of Sciences* 100, 14537–14542.
944 10.1073/pnas.2335847100.
- 945 29. Hallgren, J., Tsirigos, K.D., Pedersen, M.D., Armenteros, J.J.A., Marcatili, P.,
946 Nielsen, H., Krogh, A., and Winther, O. (2022). DeepTMHMM predicts alpha
947 and beta transmembrane proteins using deep neural networks.
948 2022.04.08.487609. 10.1101/2022.04.08.487609.
- 949 30. Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., and Embley, T.M. (2008). The
950 archaebacterial origin of eukaryotes. *Proceedings of the National Academy of
951 Sciences* 105, 20356–20361. 10.1073/pnas.0810647105.
- 952 31. Eme, L., Tamarit, D., Caceres, E.F., Stairs, C.W., De Anda, V., Schön, M.E.,
953 Seitz, K.W., Dombrowski, N., Lewis, W.H., Homa, F., et al. (2023). Inference
954 and reconstruction of the heimdallarchaeal ancestry of eukaryotes. *Nature*
955 618, 992–999. 10.1038/s41586-023-06186-2.
- 956 32. Liu, Y., Makarova, K.S., Huang, W.-C., Wolf, Y.I., Nikolskaya, A.N., Zhang, X.,
957 Cai, M., Zhang, C.-J., Xu, W., Luo, Z., et al. (2021). Expanded diversity of
958 Asgard archaea and their relationships with eukaryotes. *Nature* 593, 553–557.
959 10.1038/s41586-021-03494-3.
- 960 33. Spang, A., Eme, L., Saw, J.H., Caceres, E.F., Zaremba-Niedzwiedzka, K.,
961 Lombard, J., Guy, L., and Ettema, T.J.G. (2018). Asgard archaea are the
962 closest prokaryotic relatives of eukaryotes. *PLOS Genetics* 14, e1007080.
963 10.1371/journal.pgen.1007080.
- 964 34. Williams, T.A., Foster, P.G., Cox, C.J., and Embley, T.M. (2013). An archaeal
965 origin of eukaryotes supports only two primary domains of life. *Nature* 504,
966 231–236. 10.1038/nature12779.
- 967 35. Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity
968 with TM-score = 0.5? *Bioinformatics* 26, 889–895.
969 10.1093/bioinformatics/btq066.

- 970 36. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W.,
971 and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation
972 of protein database search programs. *Nucleic Acids Res* 25, 3389–3402.
- 973 37. Rozenberg, A., Inoue, K., Kandori, H., and Béjà, O. (2021). Microbial
974 Rhodopsins: The Last Two Decades. *Annual Review of Microbiology* 75, 427–
975 447. 10.1146/annurev-micro-031721-020452.
- 976 38. Yeung, W., Zhou, Z., Li, S., and Kannan, N. (2023). Alignment-free estimation
977 of sequence conservation for identifying functional sites using protein
978 sequence embeddings. *Briefings in Bioinformatics* 24, bbac599.
979 10.1093/bib/bbac599.
- 980 39. Krapp, L.F., Abriata, L.A., Cortés Rodriguez, F., and Dal Peraro, M. (2023).
981 PeSTo: parameter-free geometric deep learning for accurate prediction of
982 protein binding interfaces. *Nat Commun* 14, 2175. 10.1038/s41467-023-
983 37701-8.
- 984 40. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., S"oding, J., and
985 Steinegger, M. (2017). UniClust databases of clustered and deeply annotated
986 protein sequences and alignments. *Nucleic Acids Res.* 45, D170–D176.
987 10.1093/nar/gkw1081.
- 988 41. Mirdita, M., Steinegger, M., and S"oding, J. (2019). MMseqs2 desktop and
989 local web server app for fast, interactive sequence searches. *Bioinformatics*
990 35, 2856–2858. 10.1093/bioinformatics/bty1057.
- 991 42. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and
992 Steinegger, M. (2022). ColabFold: Making Protein folding accessible to all.
993 *Nature Methods*. 10.1038/s41592-022-01488-1.
- 994 43. Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane,
995 G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., et al. (2019).
996 MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*
997 10.1093/nar/gkz1035.
- 998 44. Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The New
999 Tree of Eukaryotes. *Trends in Ecology & Evolution* 35, 43–55.
1000 10.1016/j.tree.2019.08.008.
- 1001 45. Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., and Dessimoz,
1002 C. (2023). Structural phylogenetics unravels the evolutionary diversification of
1003 communication systems in gram-positive bacteria and their viruses.
1004 manuscript in preparation.
- 1005 46. Larusso, N.D., Ruttenberg, B.E., Singh, A.K., and Oakley, T.H. (2008). Type II
1006 Opsins: Evolutionary Origin by Internal Domain Duplication? *J Mol Evol* 66,
1007 417–423. 10.1007/s00239-008-9076-6.
- 1008 47. Benton, R. (2022). Drosophila olfaction: past, present and future. *Proceedings
1009 of the Royal Society B: Biological Sciences* 289, 20222054.
1010 10.1098/rspb.2022.2054.

- 1011 48. Chen, Y.-C.D., and Dahanukar, A. (2020). Recent advances in the genetic
1012 basis of taste detection in *Drosophila*. *Cell. Mol. Life Sci.* *77*, 1087–1101.
1013 10.1007/s00018-019-03320-0.
- 1014 49. Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2004). A Combined
1015 Transmembrane Topology and Signal Peptide Prediction Method. *Journal of*
1016 *Molecular Biology* *338*, 1027–1036. 10.1016/j.jmb.2004.03.016.
- 1017 50. Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined
1018 transmembrane topology and signal peptide prediction—the Phobius web
1019 server. *Nucleic Acids Research* *35*, W429–W432. 10.1093/nar/gkm256.
- 1020 51. Holm, L. (2020). Using Dali for Protein Structure Comparison. In *Structural*
1021 *Bioinformatics Methods in Molecular Biology.*, Z. Gáspári, ed. (Springer US),
1022 pp. 29–42. 10.1007/978-1-0716-0270-6_3.
- 1023 52. Frickey, T., and Lupas, A. (2004). CLANS: a Java application for visualizing
1024 protein families based on pairwise similarity. *Bioinformatics* *20*, 3702–3704.
1025 10.1093/bioinformatics/bth444.
- 1026 53. Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas,
1027 A.N., and Alva, V. (2020). Protein Sequence Analysis Using the MPI
1028 Bioinformatics Toolkit. *Current Protocols in Bioinformatics* *72*, e108.
1029 10.1002/cpb1.108.
- 1030 54. Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M.,
1031 Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A Completely
1032 Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its
1033 Core. *Journal of Molecular Biology* *430*, 2237–2243.
1034 10.1016/j.jmb.2017.12.007.
- 1035 55. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for
1036 clustering the next-generation sequencing data. *Bioinformatics* *28*, 3150–
1037 3152. 10.1093/bioinformatics/bts565.
- 1038 56. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and
1039 comparing large sets of protein or nucleotide sequences. *Bioinformatics* *22*,
1040 1658–1659. 10.1093/bioinformatics/btl158.
- 1041 57. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and
1042 development of Coot. *Acta Cryst D* *66*, 486–501.
1043 10.1107/S0907444910007493.
- 1044 58. Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009). Using
1045 Sequence Similarity Networks for Visualization of Relationships Across
1046 Diverse Protein Superfamilies. *PLOS ONE* *4*, e4345.
1047 10.1371/journal.pone.0004345.
- 1048 59. Brand, P., Robertson, H.M., Lin, W., Pothula, R., Klingeman, W.E., Jurat-
1049 Fuentes, J.L., and Johnson, B.R. (2018). The origin of the odorant receptor
1050 gene family in insects. *eLife* *7*, e38340. 10.7554/eLife.38340.

- 1051 60. Bulzu, P.-A., Andrei, A.-Ş., Salcher, M.M., Mehrshad, M., Inoue, K., Kandori,
1052 H., Beja, O., Ghai, R., and Banciu, H.L. (2019). Casting light on
1053 Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat Microbiol* 4,
1054 1129–1137. 10.1038/s41564-019-0404-y.
- 1055 61. Ramirez, M.D., Pairett, A.N., Pankey, M.S., Serb, J.M., Speiser, D.I.,
1056 Swafford, A.J., and Oakley, T.H. (2016). The Last Common Ancestor of Most
1057 Bilaterian Animals Possessed at Least Nine Opsins. *Genome Biology and*
1058 *Evolution* 8, 3640–3652. 10.1093/gbe/evw248.
- 1059 62. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D.,
1060 Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software
1061 Environment for Integrated Models of Biomolecular Interaction Networks.
1062 *Genome Res* 13, 2498–2504. 10.1101/gr.1239303.
- 1063 63. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R.,
1064 Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-
1065 level protein structure with a language model. *Science* 379, 1123–1130.
1066 10.1126/science.adc2574.
- 1067 64. Edgar, R.C. (2022). Muscle5: High-accuracy alignment ensembles enable
1068 unbiased assessments of sequence homology and phylogeny. *Nat Commun*
1069 13, 6968. 10.1038/s41467-022-34630-w.
- 1070 65. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a
1071 tool for automated alignment trimming in large-scale phylogenetic analyses.
1072 *Bioinformatics* 25, 1972–1973. 10.1093/bioinformatics/btp348.
- 1073 66. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately
1074 Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* 5, e9490.
1075 10.1371/journal.pone.0009490.
- 1076 67. Aberer, A.J., Krompass, D., and Stamatakis, A. (2013). Pruning Rogue Taxa
1077 Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice.
1078 *Systematic Biology* 62, 162–166. 10.1093/sysbio/sys078.
- 1079 68. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern
1080 phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528.
1081 10.1093/bioinformatics/bty633.
- 1082 69. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative
1083 biology (and other things). *Methods in Ecology and Evolution* 3, 217–223.
1084 10.1111/j.2041-210X.2011.00169.x.
- 1085 70. Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace:
1086 Statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology*
1087 *Resources* 17, 1385–1392. 10.1111/1755-0998.12676.
- 1088 71. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online
1089 tool for phylogenetic tree display and annotation. *Nucleic Acids Research* 49,
1090 W293–W296. 10.1093/nar/gkab301.

Figure 1

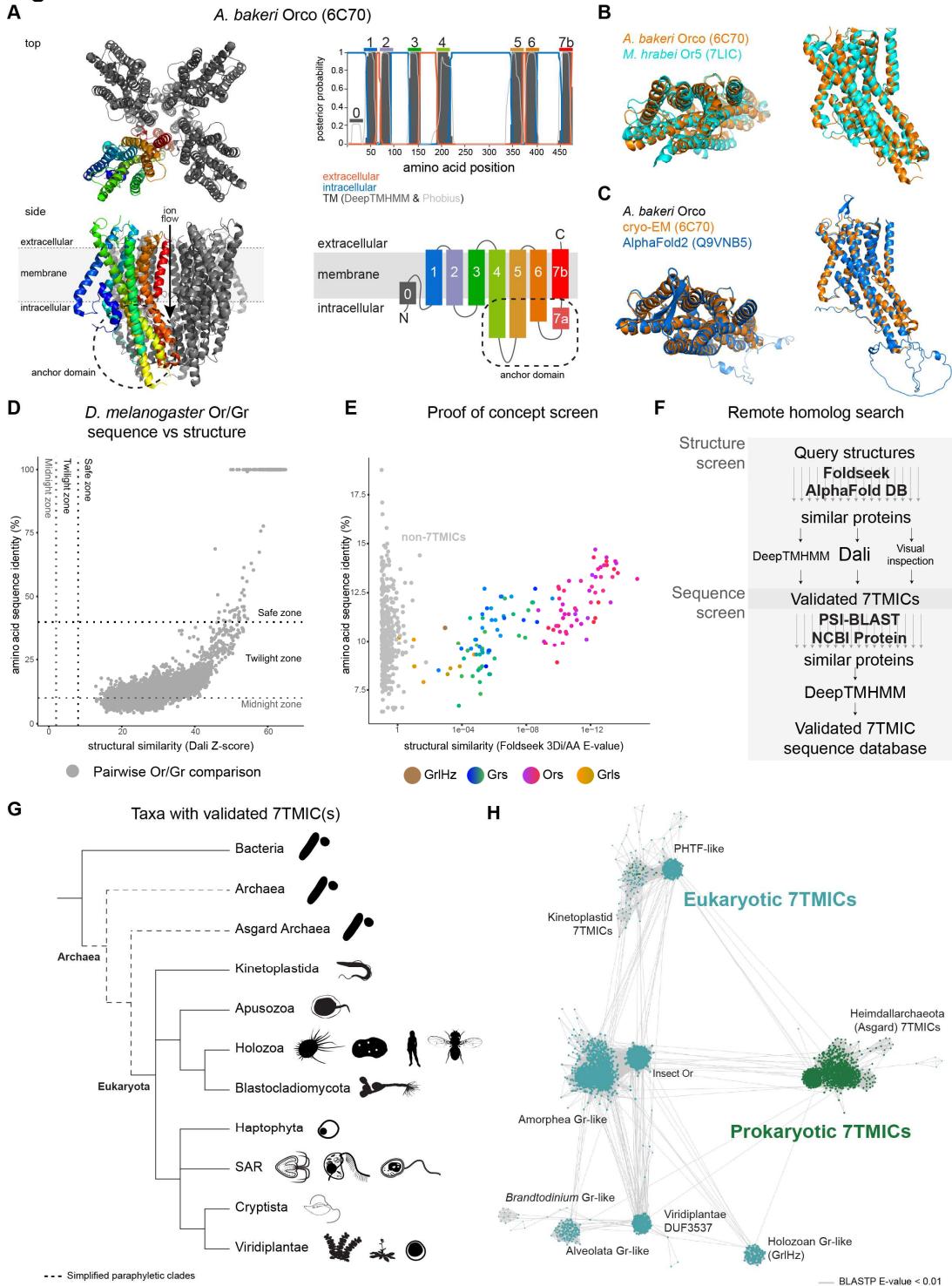


Figure 2

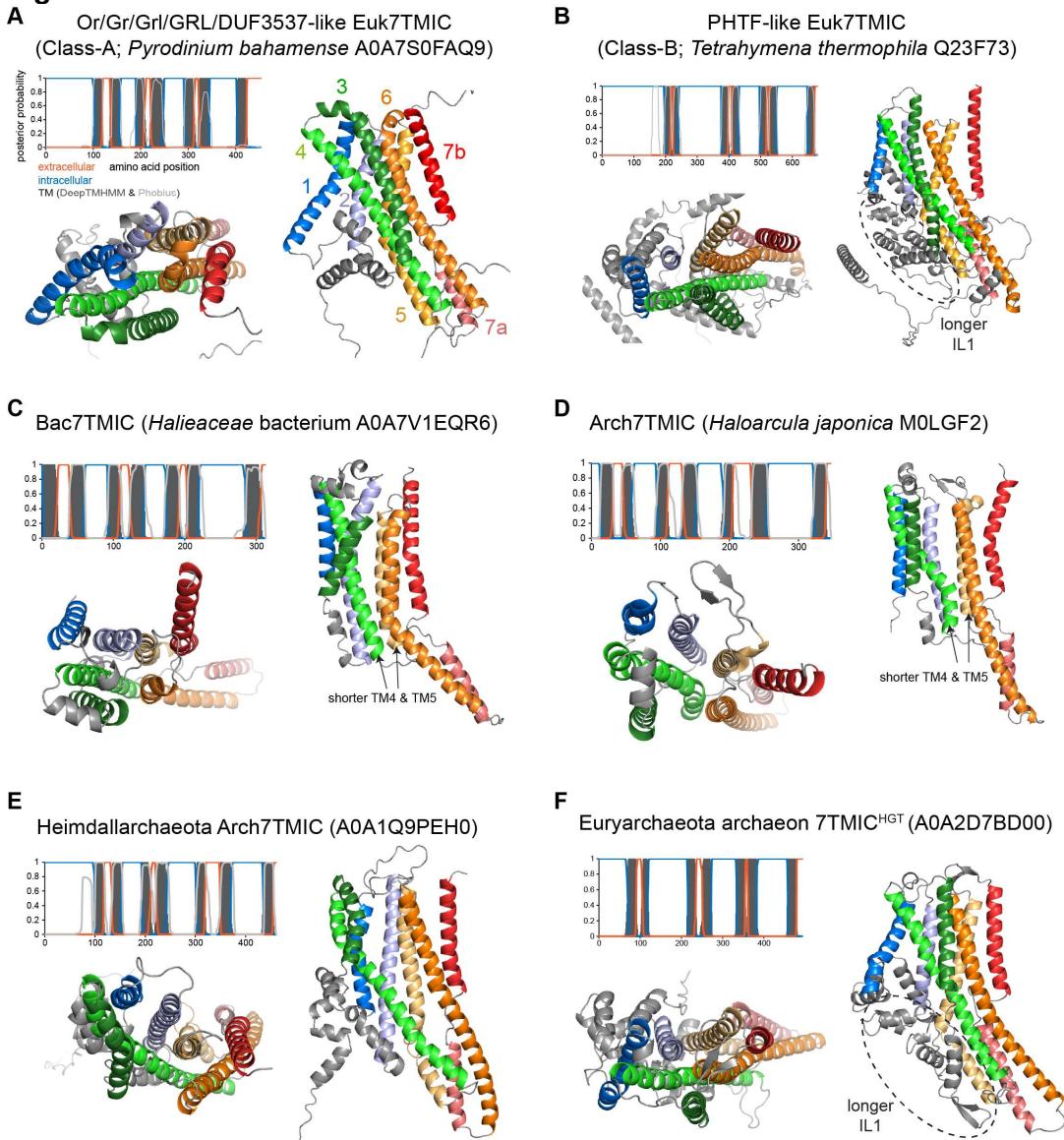


Figure 3

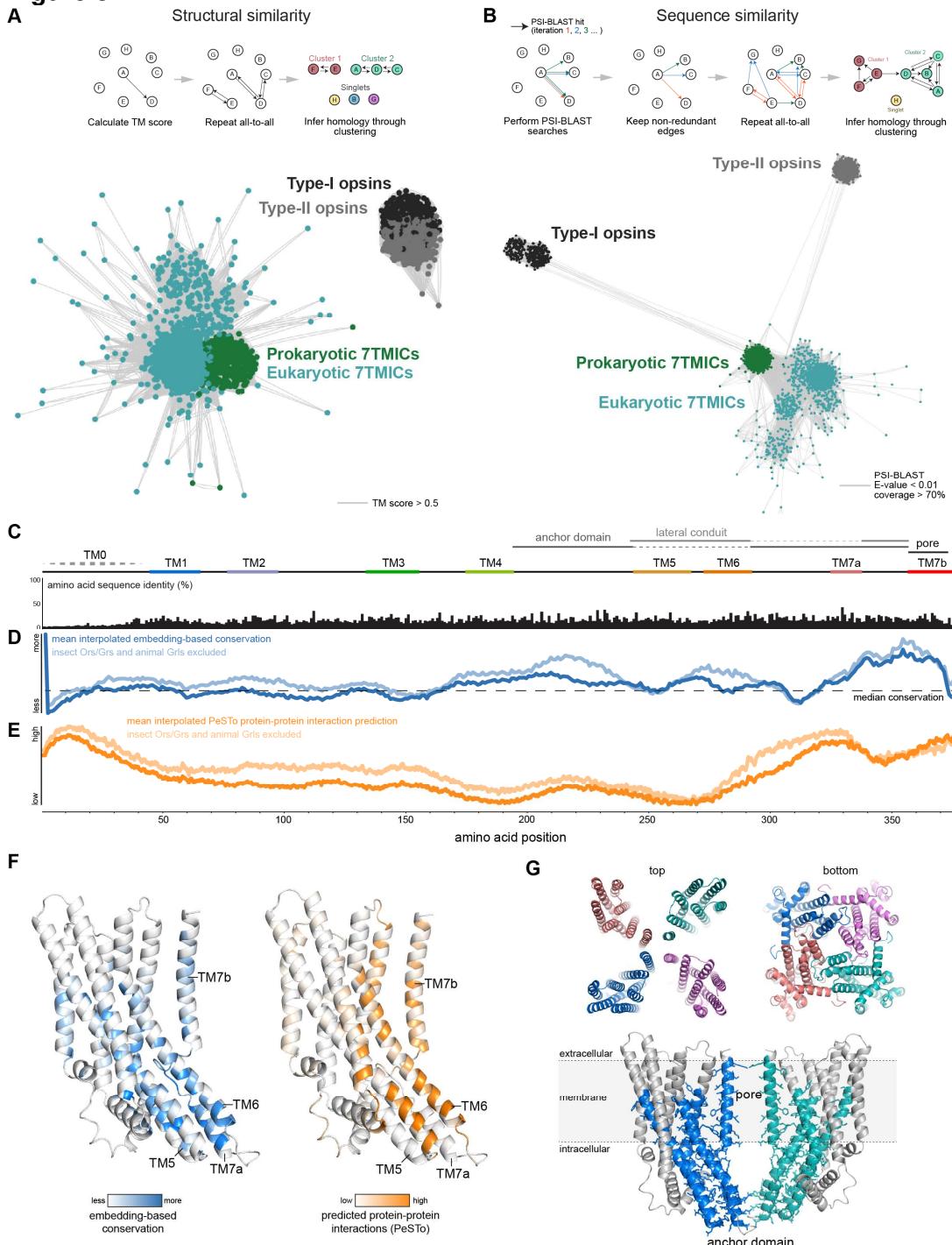


Figure 4

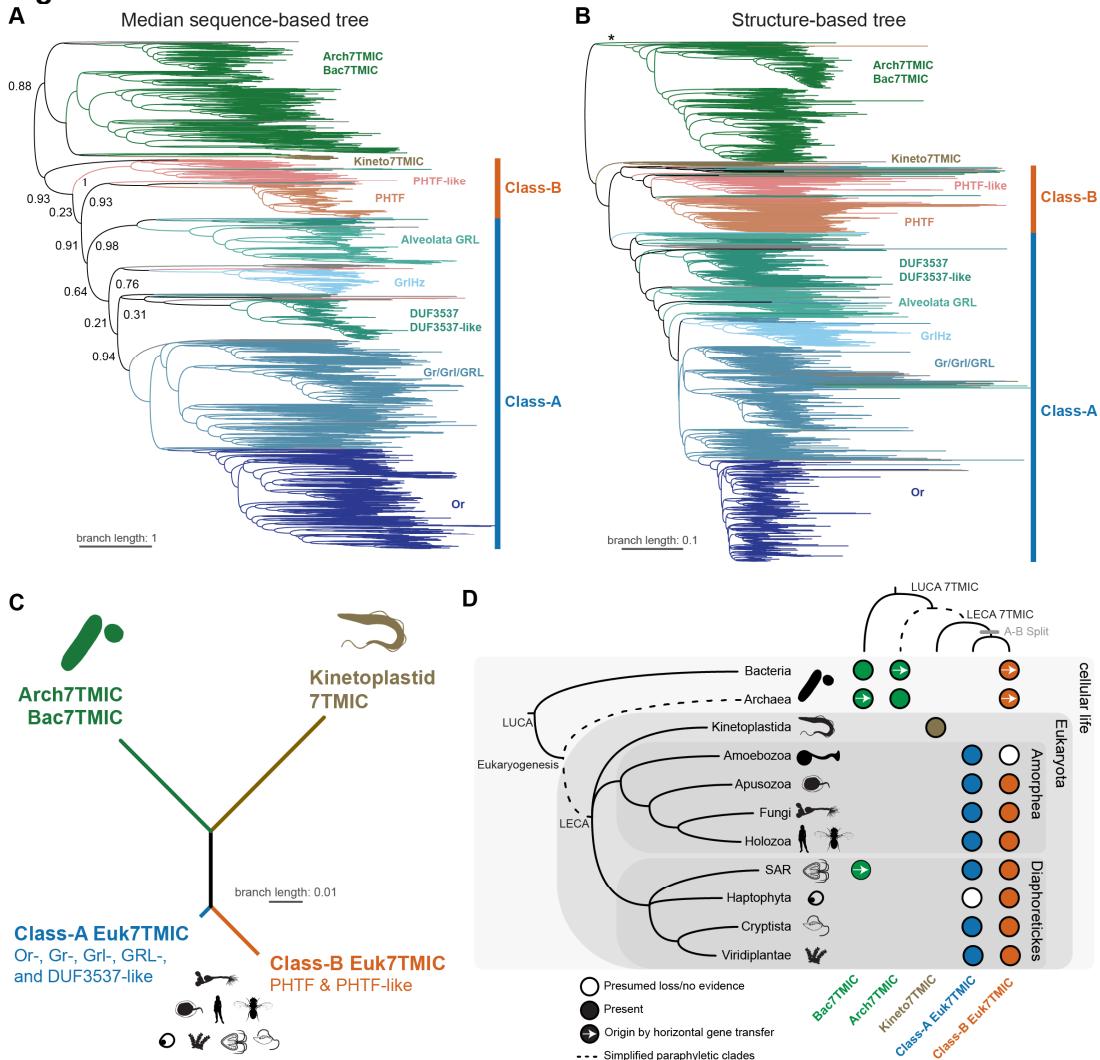


Figure S1

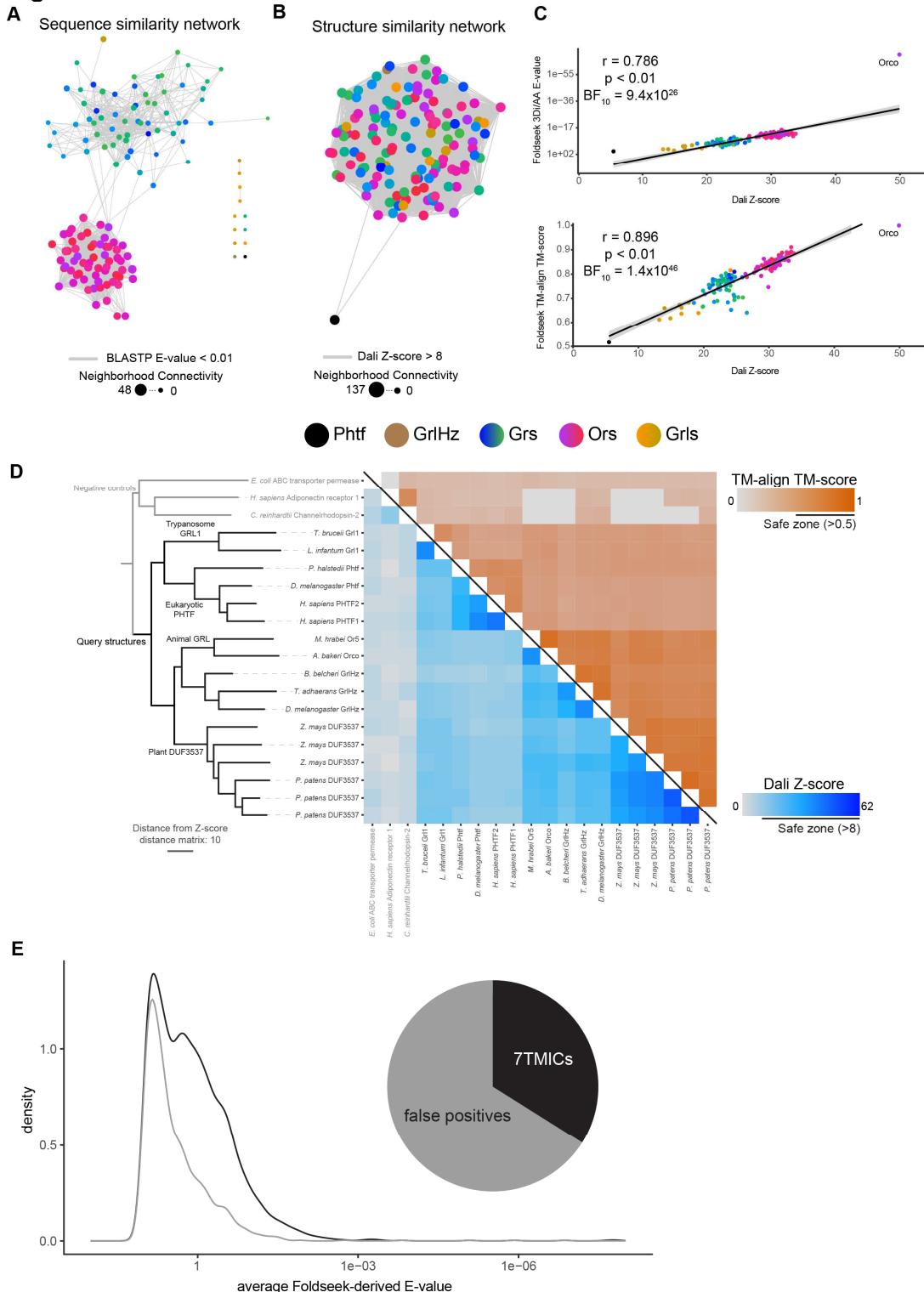
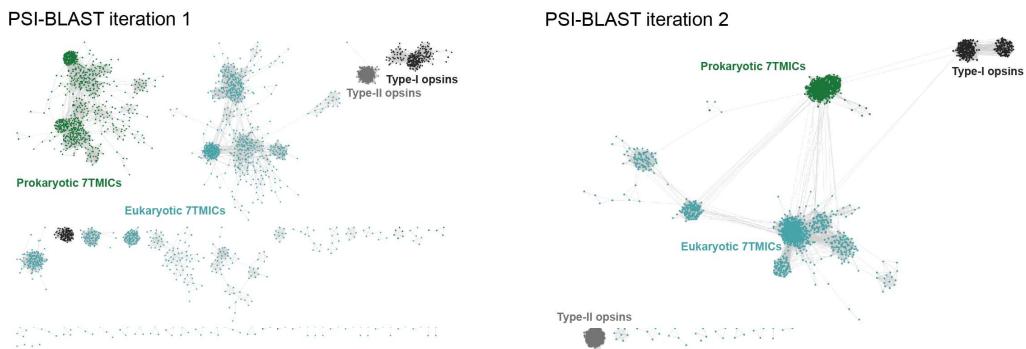
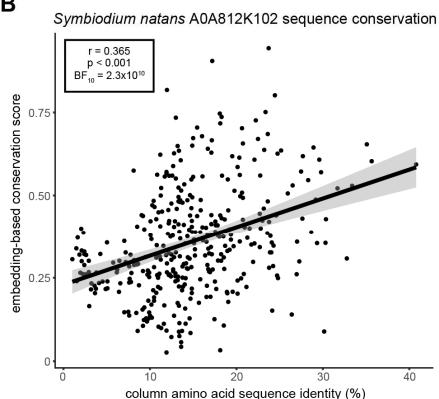


Figure S2

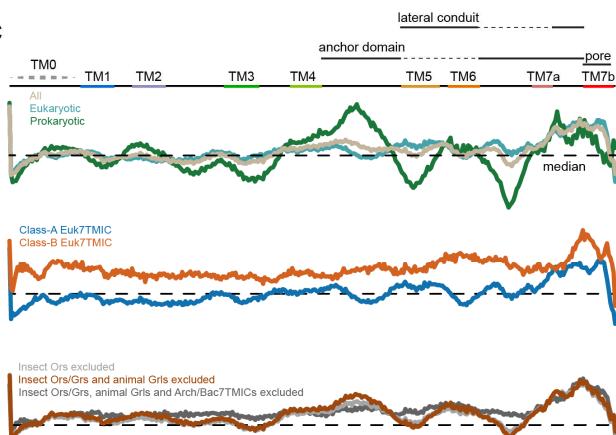
A



B



C



D

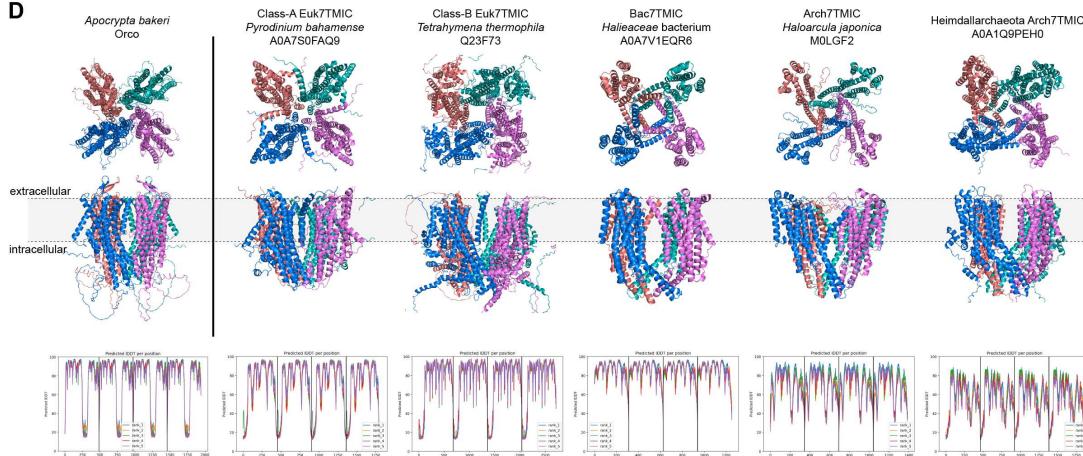
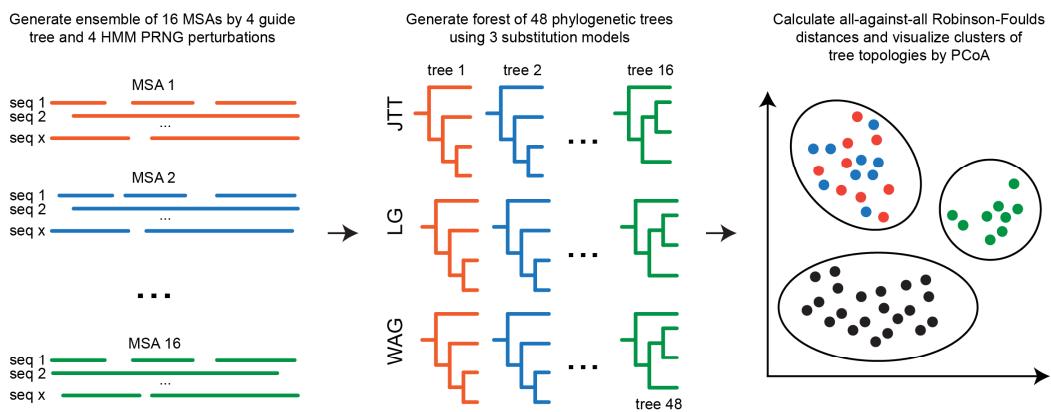
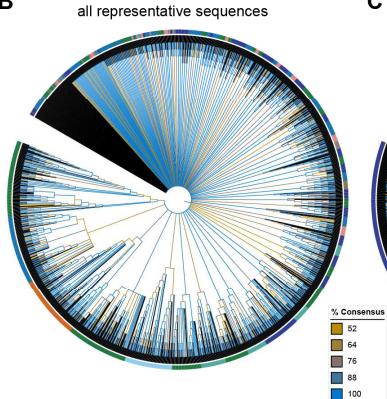


Figure S3

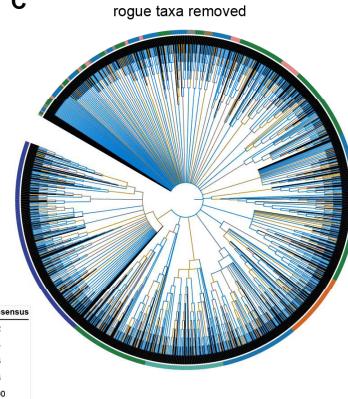
A



B

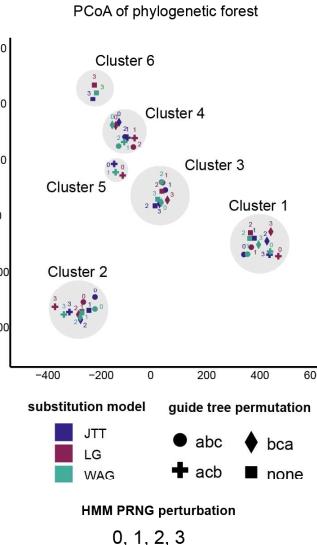


C

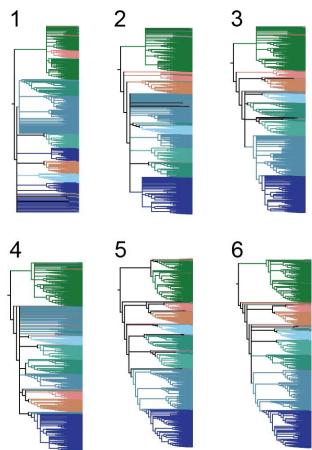


Arch7TMIC Kineto7TMIC PHTF PHTF-like DUF3537
Bac7TMIC Alveolata GRL GrHz Gr/Grl/GRL Or incertae sedis
DUF3537-like

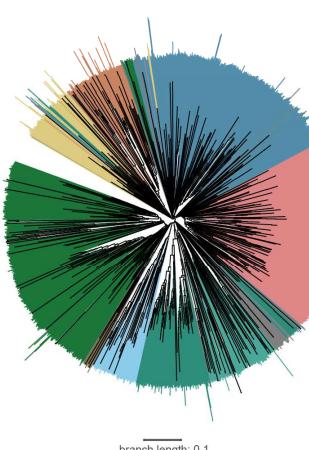
D



E



F



G

