Biophysical Journal, Volume *122*

Supplemental information

Interpreting the molecular mechanisms of disease variants in human transmembrane proteins

Johanna Katarina Sofie Tiemann, Henrike Zschach, Kresten Lindorff-Larsen, and Amelie Stein

Article

SUPPORTING MATERIAL: Interpreting the molecular mechanisms of disease variants in human transmembrane proteins

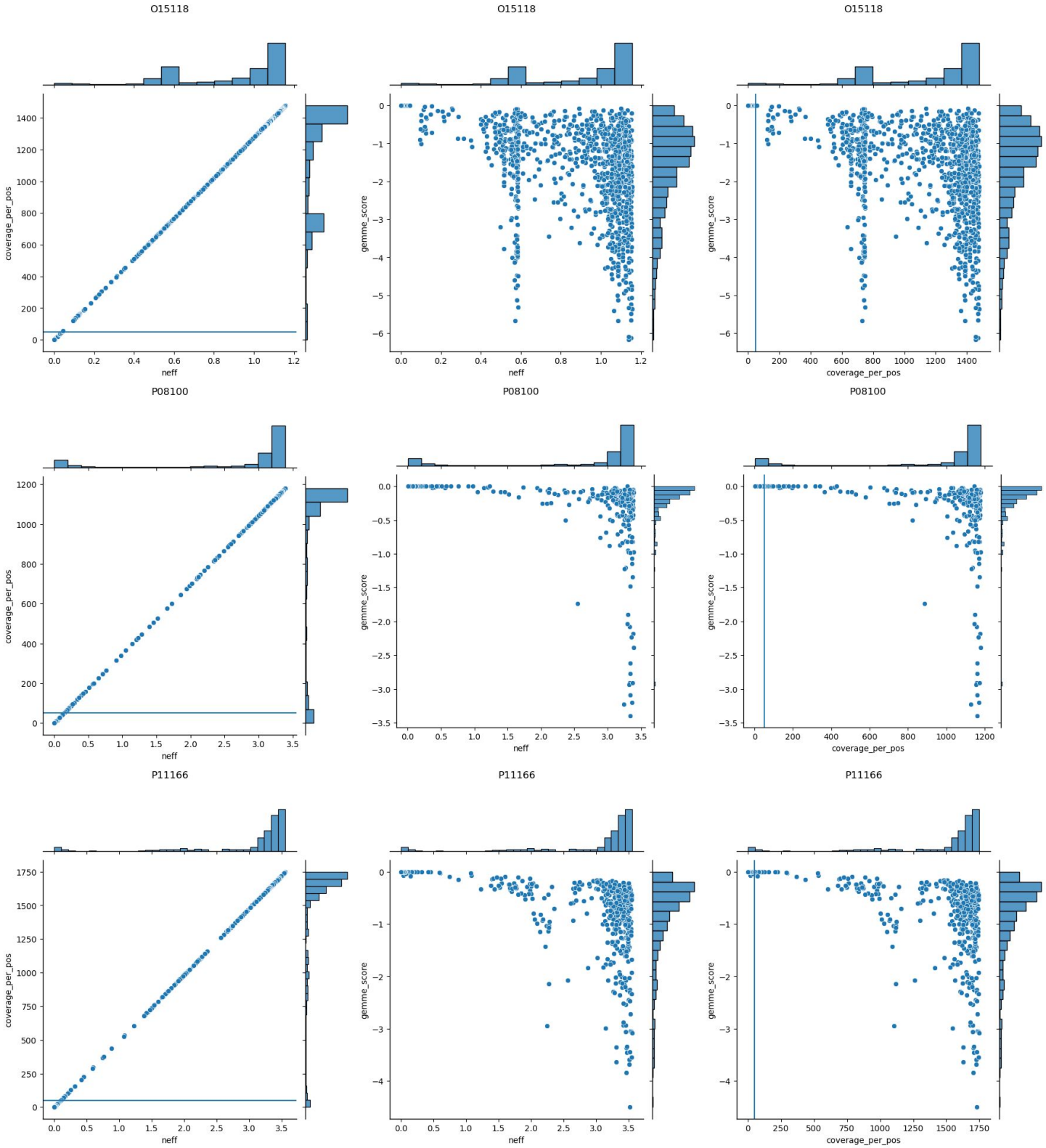
Johanna Katarina Sofie Tiemann[1,2], Henrike Zschach[1,2], Kresten Lindorff-Larsen[1*], and Amelie Stein[2*]

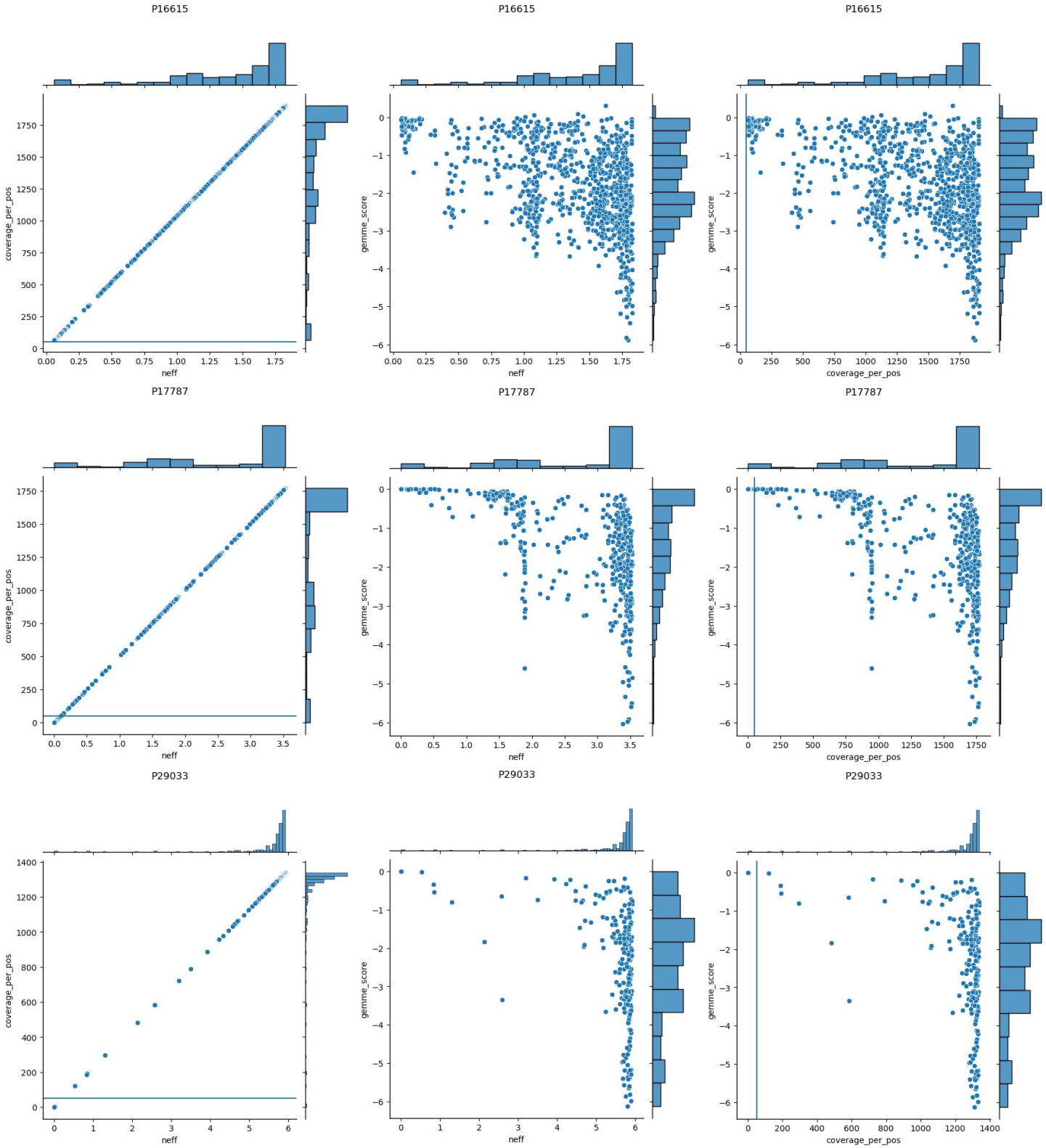
[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark.
[2]Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark.
[*]Correspondence: lindorff@bio.ku.dk (KL-L)
[*]Correspondence: amelie.stein@bio.ku.dk (AS)

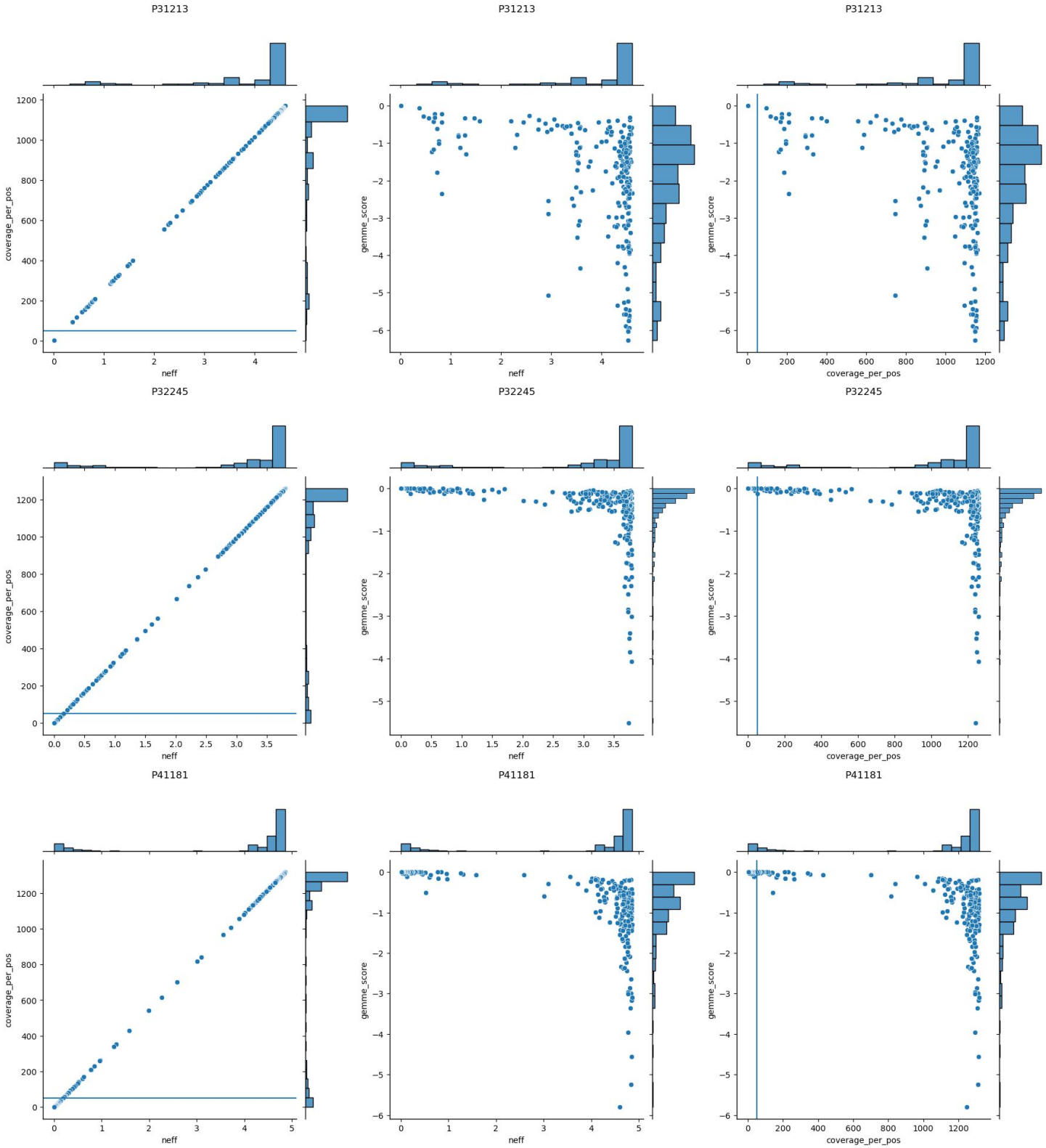
SUPPLEMENTARY MATERIAL

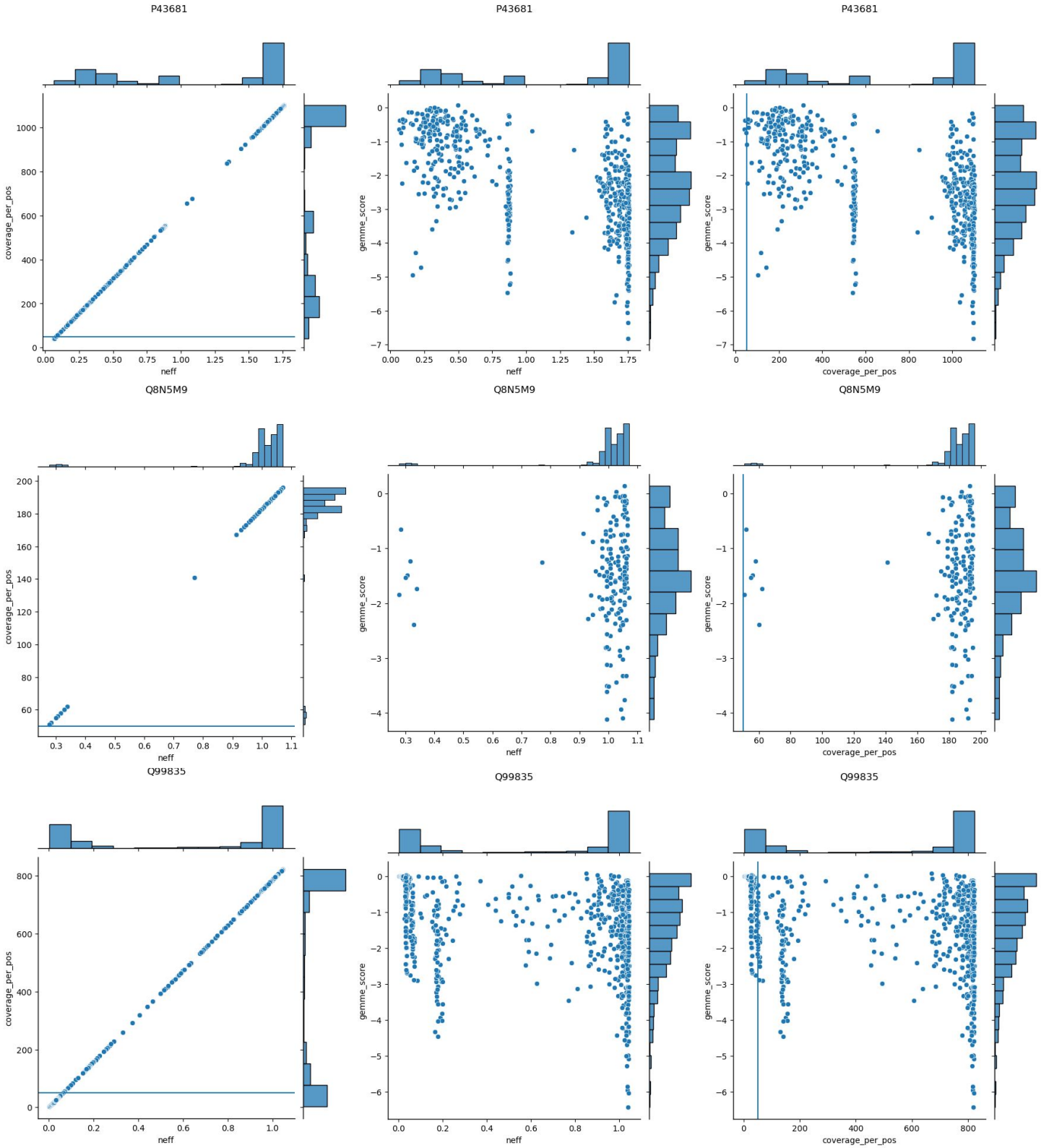
Supplementary Figures

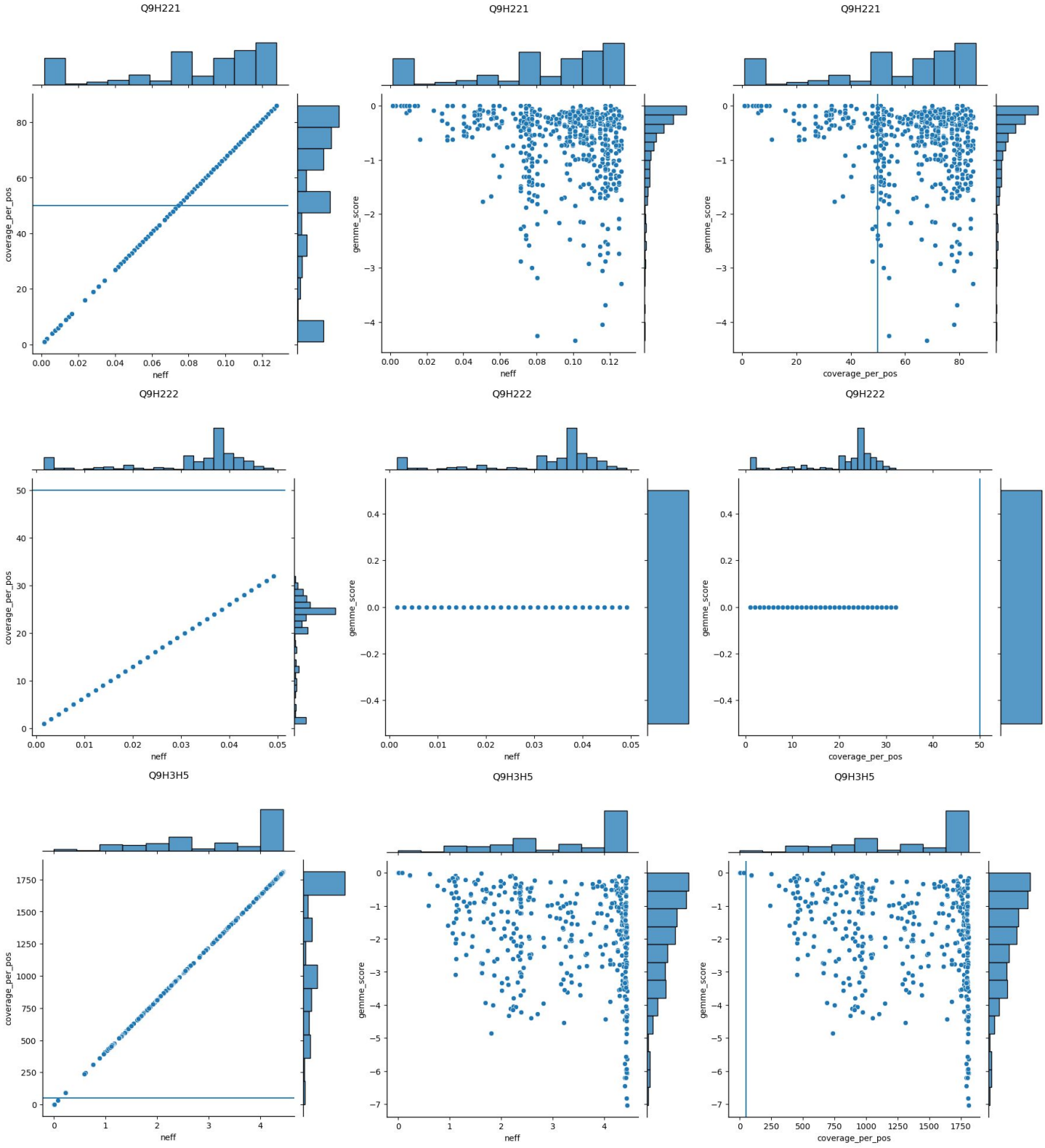
[1]Present Address: Center for Health Data Science, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen N., Denmark.


O15118
O15118
O15118
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos
P08100
P08100
P08100
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos
P11166
P11166
P11166
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos


P16615
P16615
P16615
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos
P17787
P17787
P17787
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos
P29033
P29033
P29033
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos


P31213
P31213
P31213
P32245
P32245
P32245
P41181
P41181
P41181
coverage_per_pos
gemme_score
neff
coverage_per_pos


P43681
P43681
P43681
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos
Q8N5M9
Q8N5M9
Q8N5M9
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos
Q99835
Q99835
Q99835
coverage_per_pos
neff
gemme_score
neff
gemme_score
coverage_per_pos


Q9H221
coverage_per_pos
neff
Q9H221
gemme_score
neff
Q9H221
gemme_score
coverage_per_pos
Q9H222
coverage_per_pos
neff
Q9H222
gemme_score
neff
Q9H222
gemme_score
coverage_per_pos
Q9H3H5
coverage_per_pos
neff
Q9H3H5
gemme_score
neff
Q9H3H5
gemme_score
coverage_per_pos

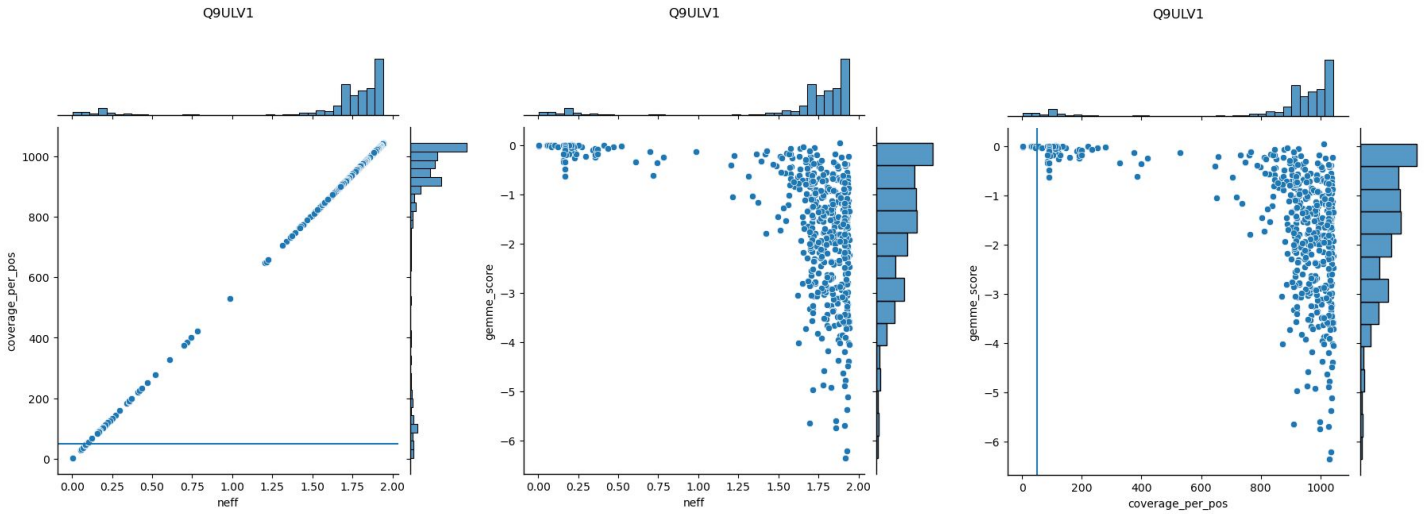

Q9ULV1
coverage_per_pos
neff
Q9ULV1
gemme_score
neff
Q9ULV1
gemme_score
coverage_per_pos

Figure S1: Neff values evaluated against GEMME coevolutionary score and sequence coverage per position: Per protein in the X-ray subset three plots are shown: first, the sequence coverage at each residue position vs. neff; second, the GEMME score (not normalized) vs. neff; third, the GEMME score vs. MSA sequence coverage. For the sequence coverage, our chosen threshold line is drawn at 50.

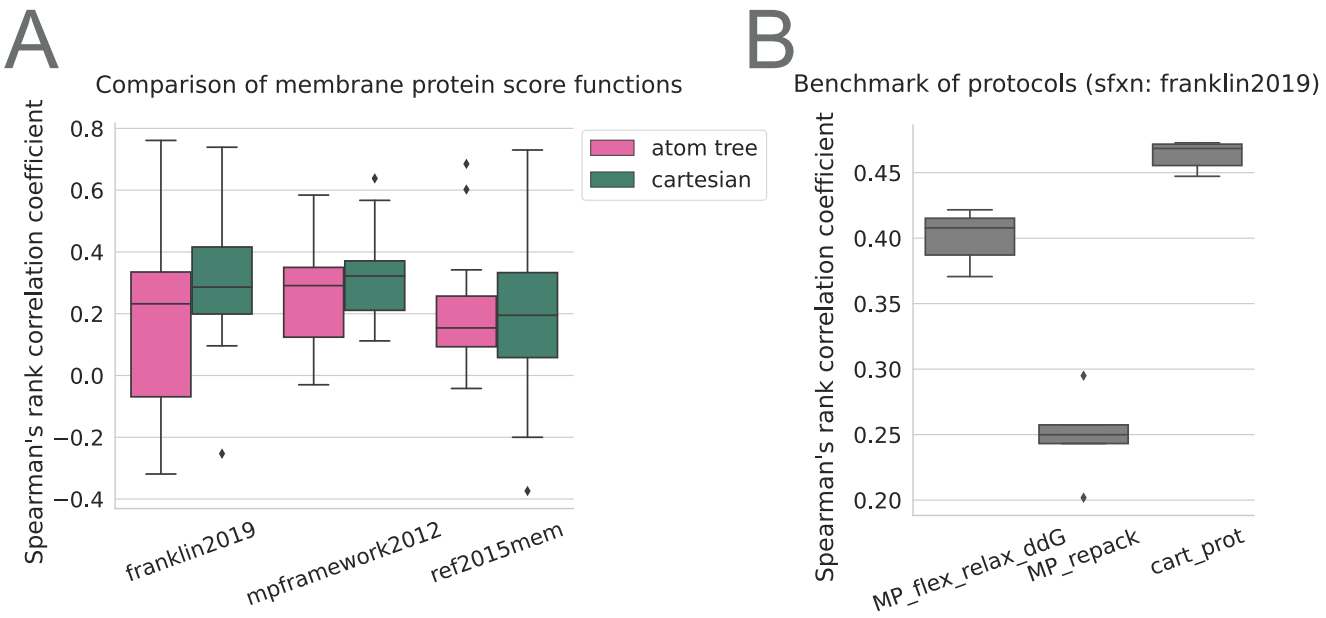


Figure S2: Benchmark of Rosetta ΔΔG calculations for MPs. (A) Comparison of accuracy of stability calculations performed with different membrane protein score functions but using the same protocol and data set. Data was extracted from (1). (B) Comparison using different protocols but the same score function (franklin2019; (2)) and was conducted on the benchmark set described in table 1 in the main manuscript.

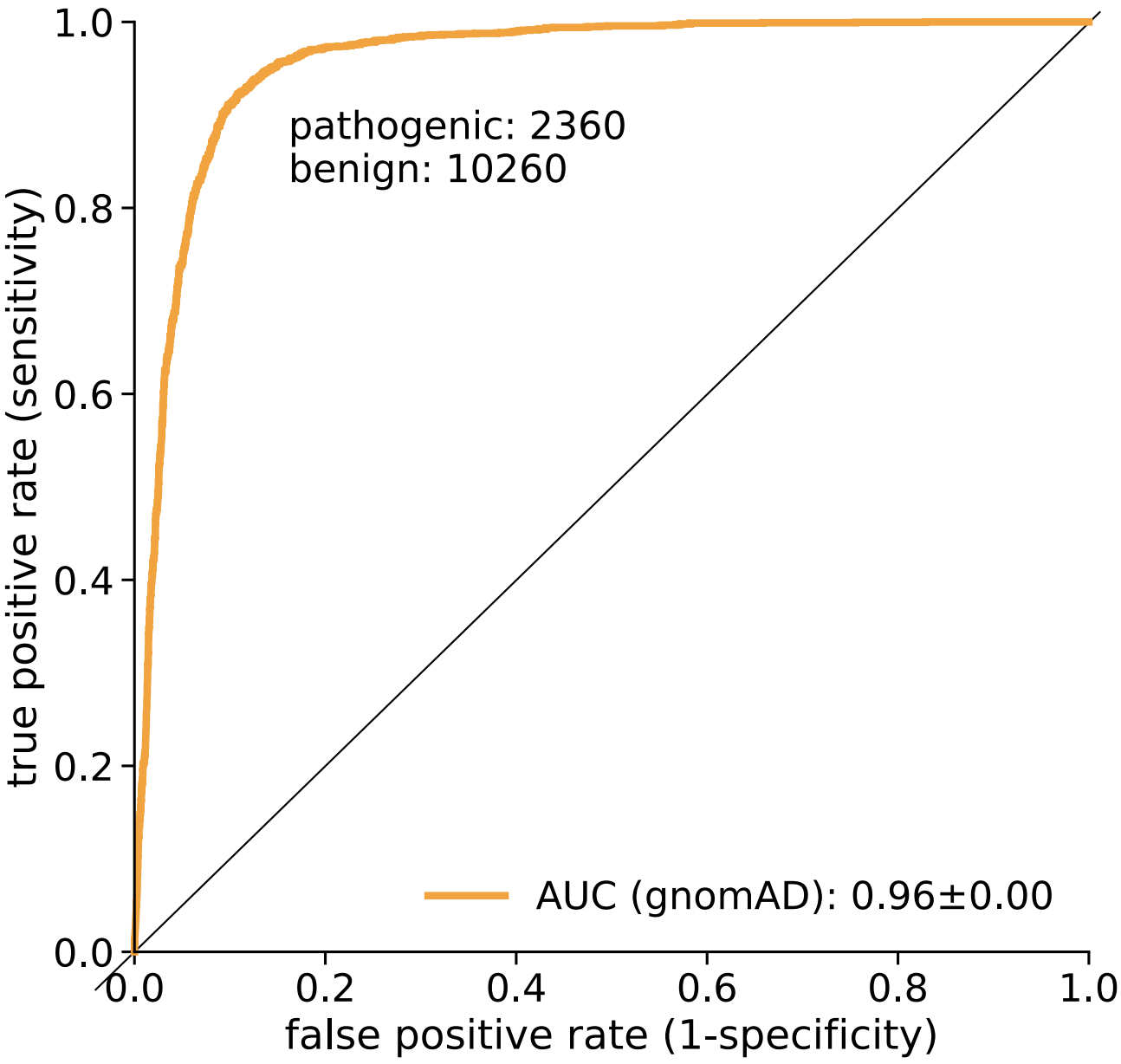


Figure S3: ROC analysis for gnomAD allele frequencies

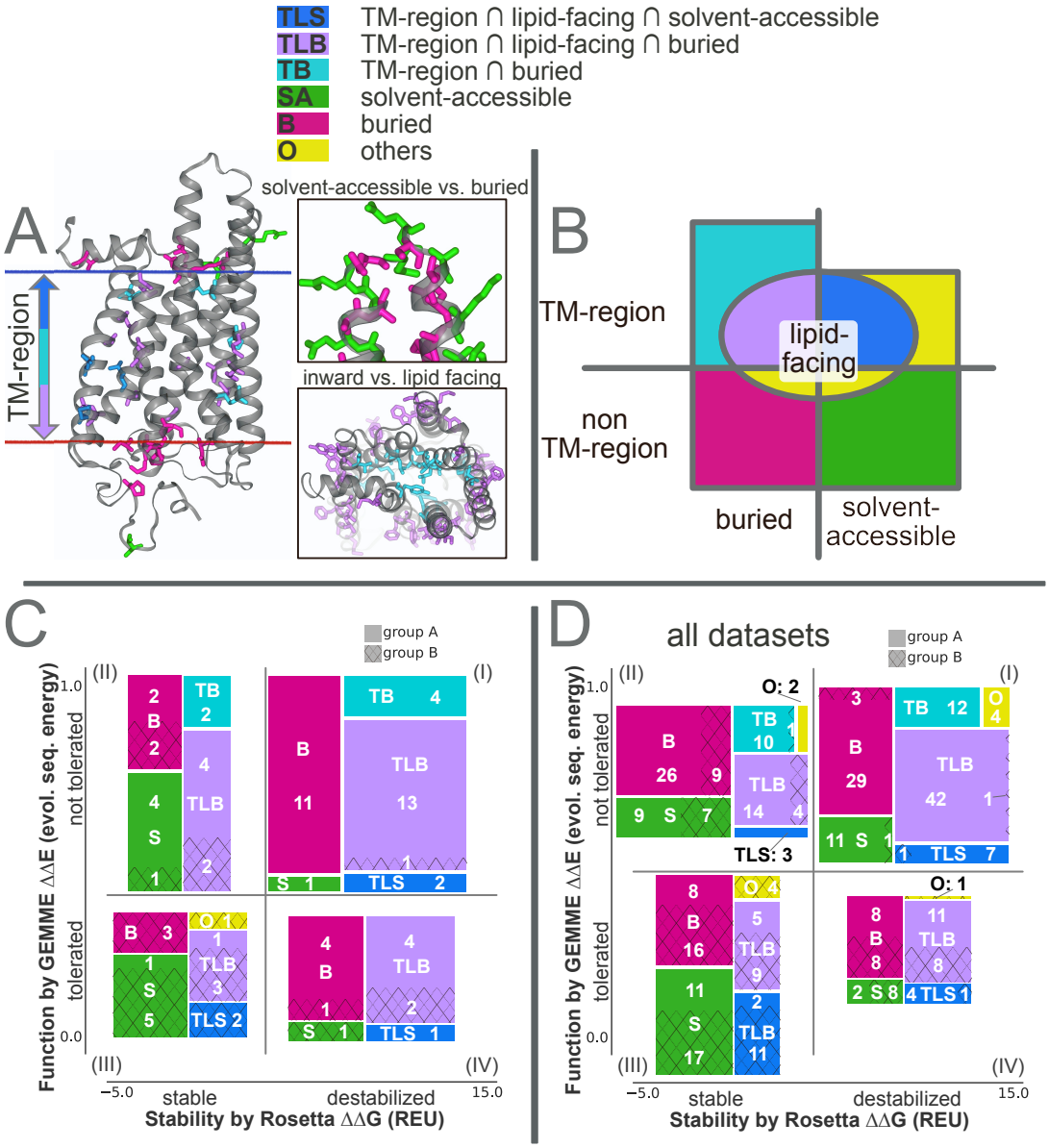


Figure S4: Protein regions and their overlaps and analysis of GPCR variant classification performance and structural differences of variant effects. (A) [left] Illustration of the different residue categories used in this work on OPSD, namely, whether they are inside (dark blue, turquoise, and violet) or outside of the transmembrane (TM) region (green and pink), and their orientation towards the membrane (lipid-facing: blue and violet), and whether they are solvent-accessible or buried. The structure on the left shows disease-associated variants, while the two inserts on the right illustrate solvent-accessible vs. buried, and inwards vs. lipid-facing, more generally and are not restricted to a disease relationship. Variants labeled with *other* are rare combinations, e.g., residues within the TM region that are solvent accessible but do not face the membrane, and some at the intersection between the membrane and the solvent (like lipid-facing but not placed within the TM region; see note in Limitations section). (B) Overview about the protein region distribution. (C) Similar as Fig.3C, variant counts in the four quadrants, separated by their position in the proteins, are shown for *group A* (pathogenic, full) and *group B* (benign and/or non-rare, hashed) variants. (D) Variant counts as seen in (C) are shown summed over all 15 proteins.

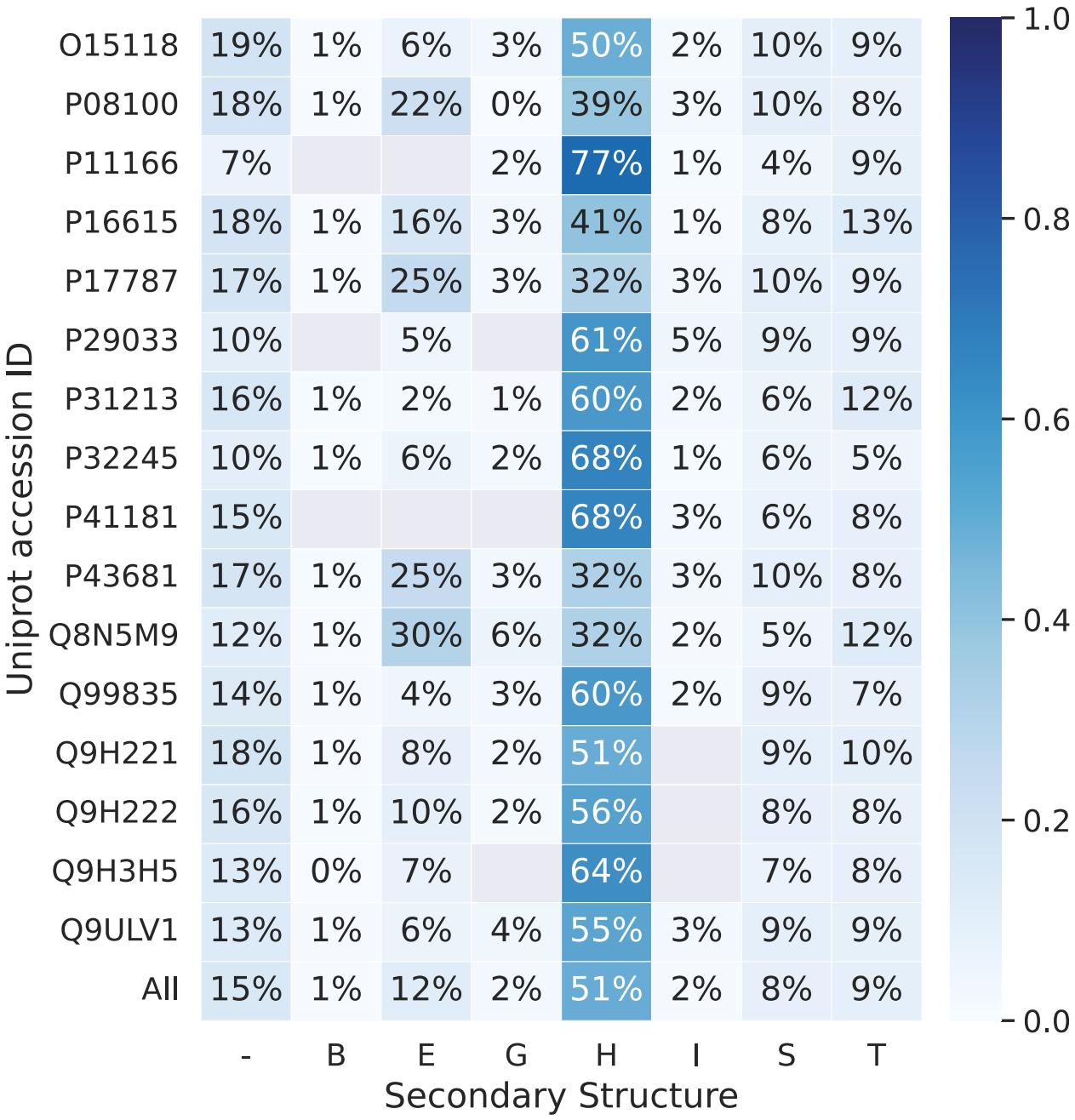


Figure S5: Secondary structure of target dataset calculated using DSSP (3). Abbreviation stand for H = α-helix; B = residue in isolated β-bridge; E = extended strand, participates in β ladder; G = 3-helix (3_{10} helix); I = 5 helix (π-helix); T = hydrogen bonded turn; S = bend; - = unstructured

Supplementary Tables

Table S1: Experimental ΔΔG datasets (all from E. coli) used for benchmarking. Multiple variant counts indicate different pH, labeling tags or temperatures.

uniprotID	protein	#variants	reference
P09391	GLPG	6, 3	(4)
P09391	GLPG	69	(5)
P09391	GLPG	142	(6)
P09391	GLPG	20, 20, 8, 8	(7)
P09391	GLPG	2	(8)
P0A910	OMPA	19, 20	(9)
P0A910	OMPA	15	(10)
P0A921	PA1	36	(11)
P0A921	PA1	6, 6, 6	(12)
P0A921	PA1	49	(13)
P37001	PAGP	20	(14)
P37001	PAGP	19, 19	(15)

Table S2: Number of variants annotated by ClinVar or gnomAD for those proteins with at least one experimentally resolved structure per cellular compartment

	all	Extracellular	Cytoplasmic	Transmembrane	other
total	281,220	112,643	61,112	29,944	77,521
benign	2,089	957	495	169	4,029
benign ⋂ gnomAD	1,951	883	467	155	446
pathogenic	4,030	1,430	1,118	807	3,822
patho. ⋂ gnomAD	1,046	326	320	208	192
VUS (ClinVar)	18,882	7,592	5,962	1,974	3,354
VUS ⋂ gnomAD	9,685	3,777	3,190	908	1,810
only gnomAD	256,219	102,664	53,537	26,994	73,024

Table S3: Variant counts, AUC and variant counts within the quadrants for each protein class. Cutoffs are taken from the complete dataset.

protein class info			after filtering		AUC		Q I		Q II		Q III		Q IV	
class	# proteins	subselection	group A	group B	ΔΔG	ΔΔE	group A	group B	group A	group B	group A	group B	group A	group B
Cell Junction	1	all	33	16	0.42	0.84	8	1	12	0	8	11	5	4
		TM region	24	8	0.47	0.78	6	1	9	0	4	6	5	1
Enzyme	2	all	24	10	0.62	0.83	12	0	8	3	2	4	2	3
		TM region	6	33	1	0.83	4	0	0	1	0	1	2	1
GPCR	4	all	54	24	0.79	0.83	31	1	12	5	2	14	9	4
		TM region	31	11	0.81	0.87	19	1	6	2	1	5	5	3
Ion channel	3	all	11	12	0.72	0.81	6	1	3	3	2	8	0	20
		TM region	4	3	0.42	0.75	1	0	2	1	1	2	0	0
Transporter	5	all	98	42	0.63	0.82	48	3	29	9	12	20	9	10
		TM region	44	10	0.71	0.97	27	0	11	1	2	6	4	3

Additional Supplementary Tables

Additional Supplementary Table 1:

Information of variants for each X-ray PDB with at least 1 benign and 1 pathogenic variant: 2022_05_05-count_hMP_anno_splitPDB_Xray_publish.xlsx

The extended supplemental table is a collection of variant information (ClinVar and gnomAD per cellular compartment) per protein (each in a separate worksheet tab) per PDB-ID and chain. This is only generated for proteins, where at least one benign and one pathogenic variant is located in an experimentally resolved region. Further the StrucSel score (see Methods and Materials) is given, including further information about the PDB

Additional Supplementary Table 2:

Variant counts per protein: 2022_11_11-count_hMP_anno_nonsyndel_PDB_publish.xlsx

The extended supplemental table is a collection of all displayed data and additional information on the data shown in the main manuscript and the supplement, including worksheet tabs:

- *2022_05_05-count_hMP_Clinvar_gnomad_PDB_nonsyndel_df*: contains a statistic of variant counts and PDB ids for each human membrane proteins that was experimentally resolved (data fetched by 2022-05-05)
- *Variant_annotation*: variant count from ClinVar and gnomAD for all human membrane proteins separated by cellular compartments
- *Variant_annotation_hP*: variant and protein count from ClinVar and gnomAD for all human proteins
- *Variant_annotation_PDB*: variant count from ClinVar and gnomAD for all human membrane proteins that are located in experimentally resolved protein regions, separated by cellular compartments
- *Category_variant_annotation*: variant count from ClinVar and gnomAD for all human membrane proteins separated by their membrane protein category and further subdivided into the variants located in the membrane bilayer; protein counts per category are also added.
- *Category_variant_annotation_PDB*: variant count from ClinVar and gnomAD for all human membrane proteins that are located in experimentally resolved protein regions, separated by their membrane protein category and further subdivided into the variants located in the membrane bilayer; protein counts per category are also added.
- *exp_ddg_benchmark*: data used for Rosetta stability benchmark (Supplementary Material table S1).
- *X-ray_set*: X-ray protein information table (equal to table 1
- *X-ray_set_app*: extended X-ray protein information table including variant counts after each sequential filtering steps and GEMME/MSA statistics
- *X-ray_set_app_AUC*: further extended X-ray protein information table (from worksheet tab *X-ray_set_app*) including additionally the AUC calculations (error via bootstrapping) for each filtered set of variants and the sequential filtered remaining data.
- *classes*: AUC and quadrant variant counts for each protein class in total and in the TM region.

Supporting References

REFERENCES

1. Koehler Leman, J., S. Lyskov, S. M. Lewis, J. Adolf-Bryfogle, R. F. Alford, K. Barlow, Z. Ben-Aharon, D. Farrell, J. Fell, W. A. Hansen, A. Harmalkar, J. Jeliazkov, G. Kuenze, J. D. Krys, A. Ljubetič, A. L. Loshbaugh, J. Maguire, R. Moretti, V. K. Mulligan, M. L. Nance, P. T. Nguyen, S. Ó Conchúir, S. S. Roy Burman, R. Samanta, S. T. Smith, F. Teets, J. K. S. Tiemann, A. Watkins, H. Woods, B. J. Yachnin, C. D. Bahl, C. Bailey-Kellogg, D. Baker, R. Das, F. DiMaio, S. D. Khare, T. Kortemme, J. W. Labonte, K. Lindorff-Larsen, J. Meiler, W. Schief, O. Schueler-Furman, J. B. Siegel, A. Stein, V. Yarov-Yarovoy, B. Kuhlman, A. Leaver-Fay, D. Gront, J. J. Gray, and R. Bonneau, 2021. Ensuring scientific reproducibility in bio-macromolecular modeling via extensive, automated benchmarks. *Nature Communications* 12:6947. https://www.nature.com/articles/s41467-021-27222-7.

2. Alford, R. F., P. J. Fleming, K. G. Fleming, and J. J. Gray, 2021. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. *Biophysical Journal* 120:4635. https://www.sciencedirect.com/science/article/pii/S0006349521007530.

3. Touw, W. G., C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, 2015. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research* 43:D364–368.

4. Gaffney, K. A., and H. Hong, 2018. The rhomboid protease GlpG has weak interaction energies in its active site hydrogen bond network. *Journal of General Physiology* 151:282–291. https://doi.org/10.1085/jgp.201812047.

5. Paslawski, W., O. K. Lillelund, J. V. Kristensen, N. P. Schafer, R. P. Baker, S. Urban, and D. E. Otzen, 2015. Cooperative folding of a polytopic α-helical membrane protein involves a compact N-terminal nucleus and nonnative loops. *Proceedings of the National Academy of Sciences* 112:7978–7983. https://www.pnas.org/content/112/26/7978.

6. Baker, R. P., and S. Urban, 2012. Architectural and thermodynamic principles underlying intramembrane protease function. *Nature Chemical Biology* 8:759–768. https://www.nature.com/articles/nchembio.1021.

7. Guo, R., K. Gaffney, Z. Yang, M. Kim, S. Sungsuwan, X. Huang, W. L. Hubbell, and H. Hong, 2016. Steric trapping reveals a cooperativity network in the intramembrane protease GlpG. *Nature Chemical Biology* 12:353–360. https://www.nature.com/articles/nchembio.2048.

8. Min, D., R. E. Jefferson, J. U. Bowie, and T.-Y. Yoon, 2015. Mapping the energy landscape for second-stage folding of a single membrane protein. *Nature Chemical Biology* 11:981–987. https://www.nature.com/articles/nchembio.1939.

9. Hong, H., S. Park, R. H. Flores Jiménez, D. Rinehart, and L. K. Tamm, 2007. Role of Aromatic Side Chains in the Folding and Thermodynamic Stability of Integral Membrane Proteins. *Journal of the American Chemical Society* 129:8320–8327. https://doi.org/10.1021/ja068849o.

10. Hong, H., G. Szabo, and L. K. Tamm, 2006. Electrostatic couplings in OmpA ion-channel gating suggest a mechanism for pore opening. *Nature Chemical Biology* 2:627–635. https://www.nature.com/articles/nchembio827.

11. Moon, C. P., and K. G. Fleming, 2011. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proceedings of the National Academy of Sciences* 108:10174–10177. https://www.pnas.org/content/108/25/10174.

12. Stanley, A. M., and K. G. Fleming, 2007. The Role of a Hydrogen Bonding Network in the Transmembrane β-Barrel OMPLA. *Journal of Molecular Biology* 370:912–924. https://www.sciencedirect.com/science/article/pii/S0022283607006213.

13. McDonald, S. K., and K. G. Fleming, 2016. Aromatic Side Chain Water-to-Lipid Transfer Free Energies Show a Depth Dependence across the Membrane Normal. *Journal of the American Chemical Society* 138:7946–7950. https://doi.org/10.1021/jacs.6b03460.

14. Marx, D. C., and K. G. Fleming, 2017. Influence of Protein Scaffold on Side-Chain Transfer Free Energies. *Biophysical Journal* 113:597–604. https://www.cell.com/biophysj/abstract/S0006-3495(17)30682-3.

15. Huysmans, G. H. M., S. A. Baldwin, D. J. Brockwell, and S. E. Radford, 2010. The transition state for folding of an outer membrane protein. *Proceedings of the National Academy of Sciences* 107:4099–4104. https://www.pnas.org/content/107/9/4099.