Quantitative Analysis of Multiplex H-bonds: Supporting Information

Esther S. Brielle*,† and Isaiah T. Arkin*,‡

†*The Alexander Grass Center for Bioengineering, Benin School of Computer Science and Engineering. The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem, 9190400, Israel.*

‡*The Alexander Silberman Institute of Life Sciences. Department of Biological Chemistry. The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem, 9190400, Israel.*

E-mail: esther.brielle@mail.huji.ac.il; arkin@huji.ac.il

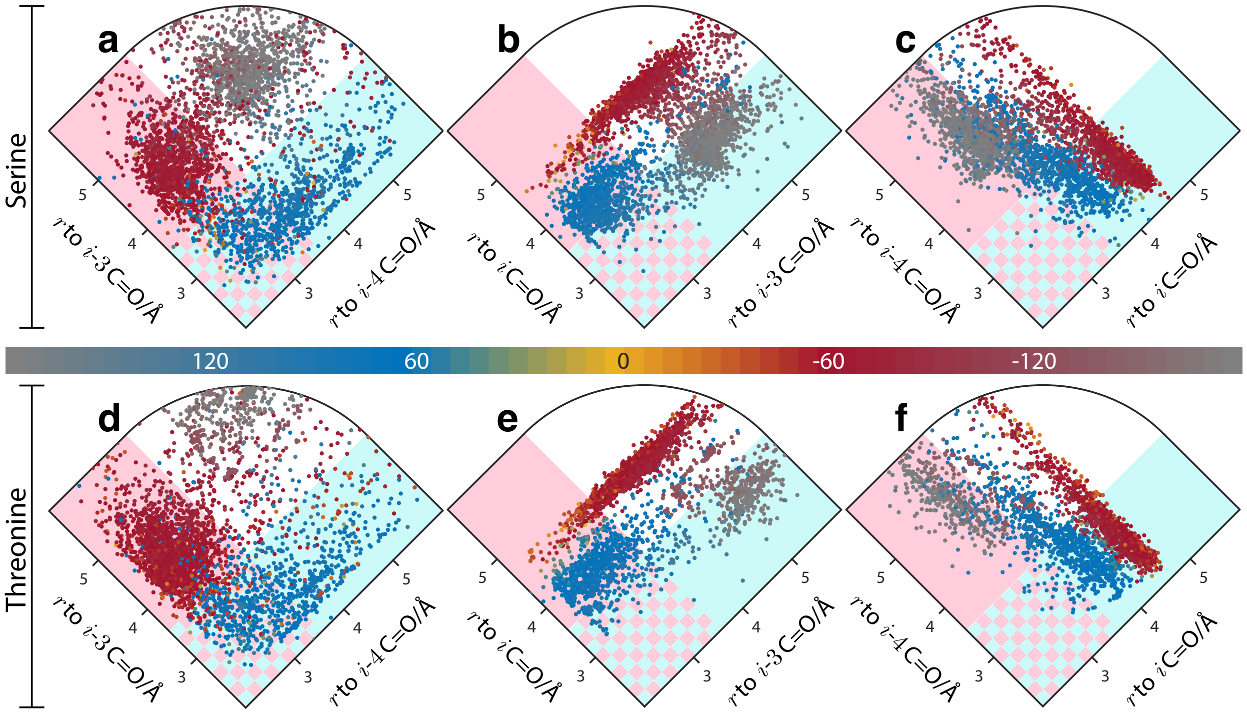

Figure S1: Analysis of H-bonding by serine (top row) and threonine (bottom row) hydroxyl O's to carbonyl O's located at the i, $i-3$ or $i-4$ positions, as a function of side chain rotamer. Each point is colored according to the color scale based on that residues' $\mathcal{X}_1$ dihedral angle. The cyan and pink shaded regions indicate residues whose Oγ is close enough (within 3.5 Å) to H-bond to the carbonyl group specified along the axis. The residues are from a dataset of non-redundant transmembrane helices.[1–3]

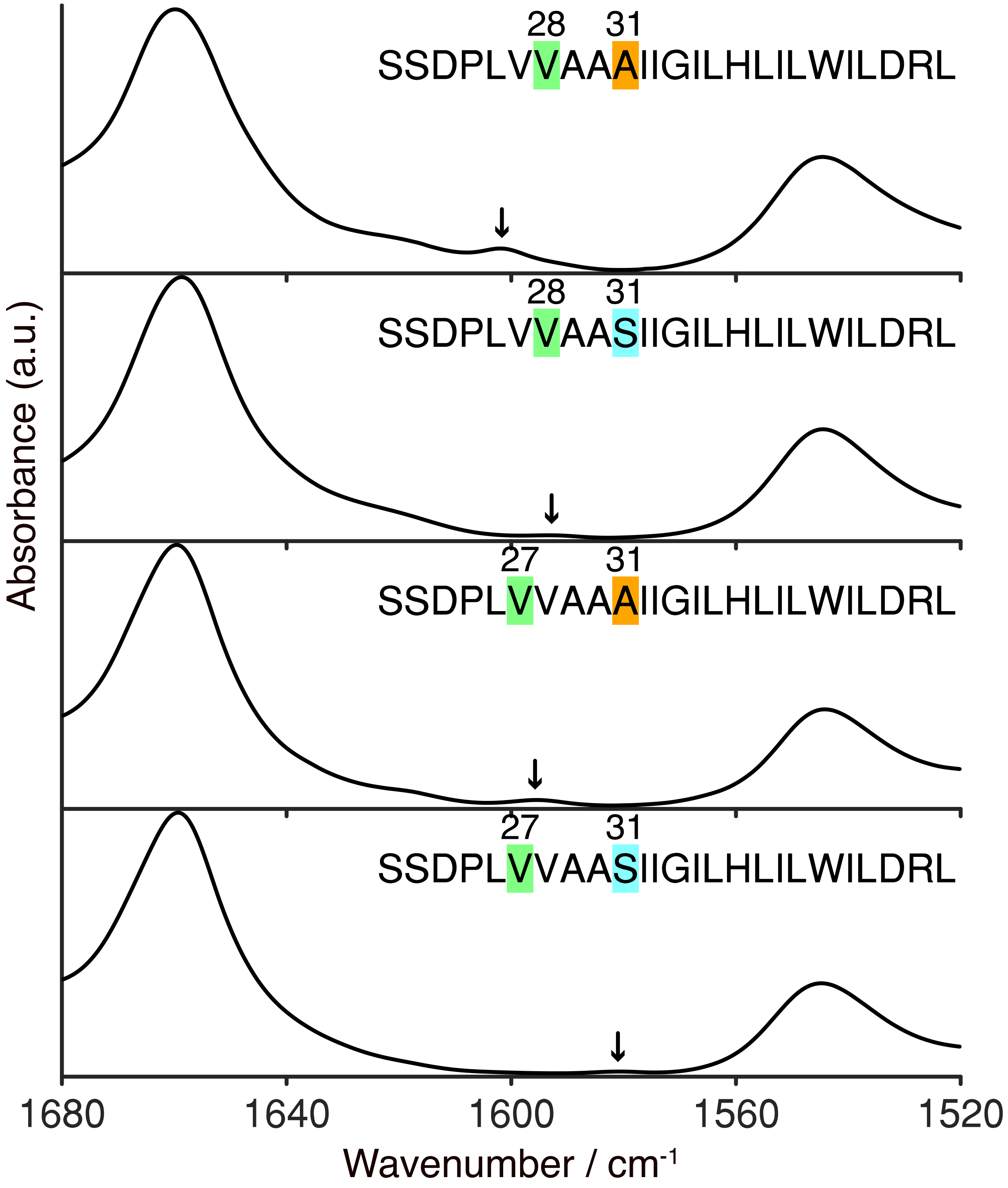

Figure S2: Infrared spectra for all M2 peptide labeling schemes showing the amide I and amide II peak locations (dashed verticle lines). The spectra are normalized according to the amide I band. The top two graphs have V28 ($i - 3$) labeled, and the bottom two graphs have V27 ($i - 4$) labeled, as indicated. Arrows indicate isotope-edited peak locations.

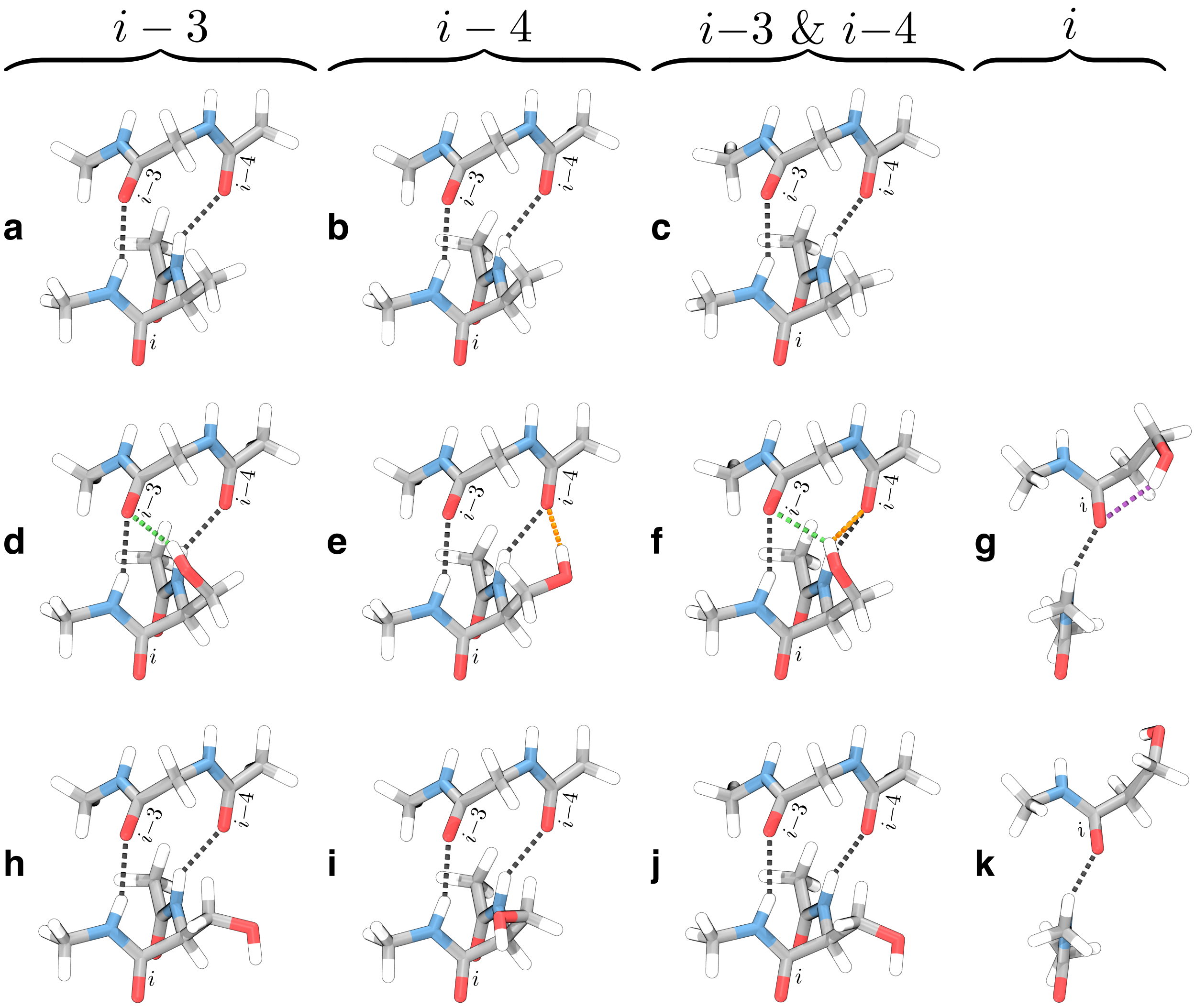

Figure S3: The serine multiplex H-bonding and alanine canonical H-bonding system mimetics with accurate coordinates used for DFT calculations. This figure represents the same structures as those in Fig. 5 along with the same numbering a-k. While Figure 5 portrays the structures schematically, this figure depicts them with their accurate geometry. The locations of the i, $i-3$, and $i-4$ carbonyls are indicated. Canonical H-bonds are depicted in black, while the bonds between the hydroxyl groups to the i, $i-3$, and $i-4$ carbonyls are colored in purple, green and orange, respectively.

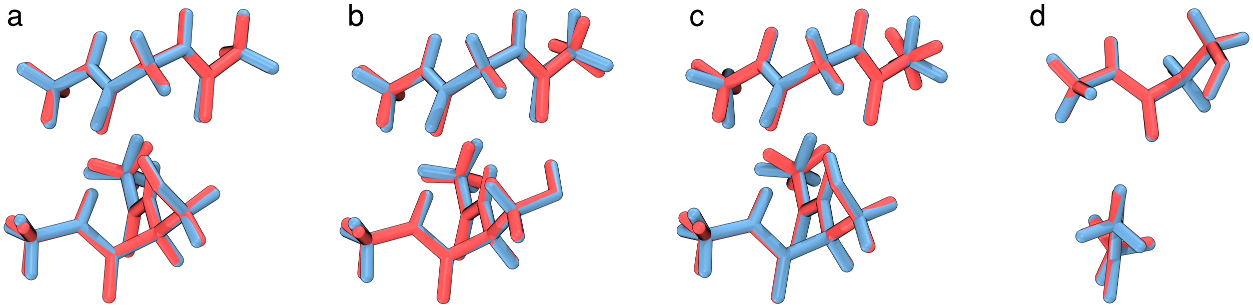

Figure S4: DFT optimization of the mimetics of the serine side chain in a multiplex H-bond, with the before (red) and after (blue) optimization structures overlayed. (a), (b), (c), and (d) are the pre- and post-optimization structures of Fig. S3 d, e, f, and g, respectively. All hydrogen atoms (except amine hydrogens) as well as backbone carbonyl groups were allowed to optimize. All other heavy atoms were restrained. The RMSDs are (a) 0.15 Å for $i-3$, (b) 0.19 Å for $i-4$, (c) 0.35 Å for $i-3$ and $i-4$, and (d) 0.068 Å for i.

Table S1: Prevalence of amino acids in transmembrane helices of membrane proteins in the TOPDB[4,5] and in the PDBTM[1–3] databases. Bitopic and polytopic prevalence values are calculated from TOPDB. Bitopic proteins are those that traverse the membrane once (single-pass), while polytopic are those that traverse the membrane more than once (multi-pass). Non-helical prevalence values are calculated from PDBTM. The prevalence in full proteomes is according to a range of values found in literature.[6–8] Hydrophobicity ($\Delta G_{\mathrm{Water}\rightarrow\mathrm{Oil}}$) is according to the GES scale.[9] Residues that we consider to be capable of H-bonding are Cys, Thr, Ser, Tyr, His, Gln, Asn, Glu, Lys, Asp, and Arg.

Amino acid	Hydrophobicity	Bitopic TM proteins	Polytopic TM proteins	All helical TM proteins	Non-helical TM proteins	Full proteomes
Phe	3.7 kcal/mol	7.19%	9.21%	7.7-9.0%	8.5%	3.6-4.0%
Met	3.4 kcal/mol	2.61%	3.70%	3.5-3.6%	3.5%	2.3-2.4%
Ile	3.1 kcal/mol	13.02%	11.33%	11.5-11.6%	8.7%	5.3-6.7%
Leu	2.8 kcal/mol	23.24%	17.58%	17.8-18.1%	13.3%	8.9-10.2%
Val	2.6 kcal/mol	15.39%	10.87%	11.3-12.1%	6.7%	6.6-8.2%
Cys	2 kcal/mol	2.83%	2.22%	1.5-2.3%	0.6%	0.8-1.9%
Trp	1.9 kcal/mol	1.80%	2.30%	2.3%	3.4%	1.0-1.4%
Ala	1.6 kcal/mol	11.53%	10.34%	10.2-10.5%	14.7%	7.8-8.8%
Thr	1.2 kcal/mol	4.21%	5.39%	5.3-5.4%	6.5%	4.9-5.9%
Gly	1 kcal/mol	7.87%	8.08%	7.5-8.1%	5.2%	7.2-7.4%
Ser	0.6 kcal/mol	4.03%	6.01%	5.2-5.8%	6.1%	4.7-6.8%
Pro	-0.2 kcal/mol	1.32%	2.46%	2.2-2.4%	6.0%	4.4-5.2%
Tyr	-0.7 kcal/mol	2.63%	3.59%	3.2-3.5%	2.0%	3.0-3.3%
His	-3 kcal/mol	0.38%	0.77%	0.7-0.9%	0.9%	1.9-2.3%
Gln	-4.1 kcal/mol	0.38%	1.22%	1.1-1.4%	3.1%	3.2-4.2%
Asn	-4.8 kcal/mol	0.35%	1.81%	1.7-1.9%	2.1%	3.4-4.3%
Glu	-8.2 kcal/mol	0.19%	0.80%	0.7-1.4%	3.1%	6.3-8.6%
Lys	-8.8 kcal/mol	0.47%	0.71%	0.7-1.6%	1.4%	5.6-7.8%
Asp	-9.2 kcal/mol	0.22%	0.75%	0.7-1.1%	1.6%	5.3-5.4%
Arg	-12.3 kcal/mol	0.35%	0.87%	0.8-1.6%	2.1%	5.1-6.2%

Table S2: H-bonding configuration of serine and threonine residues in a dataset of non-redundant α-helical membrane proteins.[1–3] A permissive cutoff distance of 3.5 Å between the hydroxyl Oγ and the appropriate acceptor was used for classification.

Partner	Serine	Threonine
Total serine/threonine	4057 (100.0%)	4199 (100.0%)
Serines/threonines H-bonding	3509 (86.5%)	3789 (90.2%)
Waters	210 (5.2%)	122 (2.9%)
Ion	17 (0.4%)	11 (0.3%)
Ligand	36 (0.9%)	31 (0.7%)
C=O at $i-0$	1251 (30.8%)	640 (15.2%)
C=O at $i-3$	1034 (25.5%)	936 (22.3%)
C=O at $i-4$	1646 (40.6%)	2595 (61.8%)
Inter-helical backbone C=O	86 (2.1%)	38 (0.9%)
Inter-helical side chain C=O	20 (0.5%)	7 (0.2%)
Inter-helical side chain OH	50 (1.2%)	52 (1.2%)
Inter-helical side chain N	8 (0.2%)	3 (0.1%)
Inter-helical side chain SH	2 (0.0%)	2 (0.0%)
Long distance backbone C=O (>±10Å)	283 (7.0%)	161 (3.8%)
side chain OH	225 (5.5%)	238 (5.7%)
side chain C=O	181 (4.5%)	130 (3.1%)
side chain SH	20 (0.5%)	8 (0.2%)
side chain N	70 (1.7%)	88 (2.1%)
Other	44 (1.1%)	109 (2.6%)

References

(1) Kozma, Dániel and Simon, István and Tusnády, Gábor E, PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res* **2013**, *41*, D524–9.

(2) Tusnády, Gábor E and Dosztányi, Zsuzsanna and Simon, István, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res* **2005**, *33*, D275–8.

(3) Tusnády, Gábor E and Dosztányi, Zsuzsanna and Simon, István, Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* **2004**, *20*, 2964–72.

(4) Dobson, L.; Langó, T.; Reményi, I.; Tusnády, G. E. Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res* **2015**, *43*, D283–9.

(5) Tusnády, G. E.; Kalmár, L.; Simon, I. TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res* **2008**, *36*, D234–9.

(6) Doolittle, R. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G., Ed.; Plenum Press, New York, 1989; Chapter Redundancies in protein sequences, pp 599–623.

(7) Brooks, D. J.; Fresco, J. R.; Lesk, A. M.; Singh, M. Evolution of Amino Acid Frequencies in Proteins Over Deep Time: Inferred Order of Introduction of Amino Acids into the Genetic Code. *Molecular Biology and Evolution* **2002**, *19*, 1645–1655.

(8) Hormoz, S. Amino acid composition of proteins reduces deleterious impact of mutations. *Scientific reports* **2013**, *3*, 2919.

(9) Engelman, D. M.; Steitz, T. A.; Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **1986**, *15*, 321–53.