

Supplementary Material for:

**A Universal Hydrophobicity Scale for Multi-Span Membrane Proteins**

**Supplementary Table (I):**

**Composition of the membrane protein database used for the derivation of the UHS**

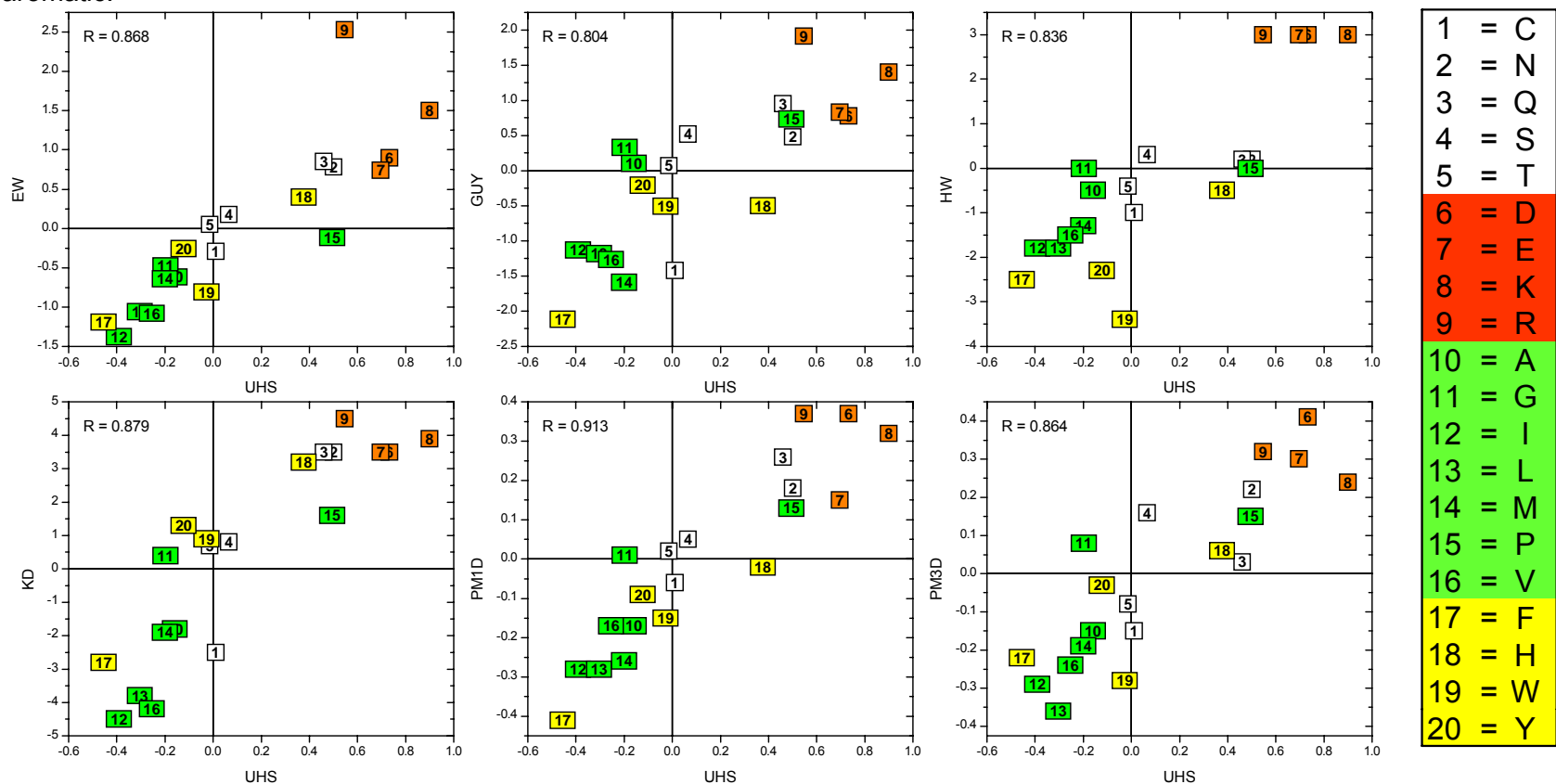
The MP database used for the derivation of the UHS was constructed by culling a complete list of multi-span MPs from the PDBTM with the PISCES server. The resulting MP database consists of 60 MPs and was divided into five parts for cross-validation. Each part contained approximately the same number of  $\alpha$ -helices and  $\beta$ -sheets. Since the MPs in the database had very different sizes the number of proteins in the different datasets vary. The table shows the datasets that were used for cross-validation, where the columns represent the number of the dataset, the number of proteins in the dataset, the PDB code of the proteins, the number of  $\alpha$ -helices and  $\beta$ -strands and the number of proteins in the database that were purely  $\alpha$ -helical,  $\beta$ -barrels and which contained both secondary structure elements (from left to right). For cross-validation the free energies were derived five times for four of the datasets and tested on the remaining one.

<i>dataset</i>	<i>#proteins</i>	<i>PDB code of proteins</i>	<i>#<math>\alpha</math>-helices</i>	<i>#<math>\beta</math>-strand</i>	<i><math>\alpha</math></i>	<i><math>\beta</math></i>	<i><math>\alpha+\beta</math></i>
1	7	1I78, 1KMO, 1QFG, 1R3J, 1V54, 2BL2, 7AHL	196	209	2	1	4
2	11	1PPJ, 1S3E, 1U7G, 1XRD, 1YMG, 1ZLL, 2BG9, 2CFQ, 2ERV, 2FGQ, 2MPR	196	213	4	1	6
3	9	1C17, 1M0K, 1OKC, 1QJP, 1UUN, 1WAZ, 1YC9, 1YCE, 1YEW	192	208	4	1	4
4	16	1EK9, 1EQ8, 1HXX, 1K24, 1KPL, 1P49, 1QD6, 1QJ8, 1T16, 1THQ, 1UYN, 1WP1, 1XME, 2A65, 2F2B, 2FBW	196	217	4	1	11
5	17	1AFO, 1BA4, 1BZK, 1FDM, 1KQF, 1NKZ, 1NQE, 1P4T, 1RWT, 1RZH, 1U19, 1WPG, 1XKW, 1Y4Z, 1ZZA, 2F1V, 2POR	197	209	7	1	9

## Supplementary Figure (1):

### Correlation plots of the UHS with other scales

Plots showing the correlation of the hydrophobicity values in kcal/mol between the UHS and the scales from EW, Guy, HW, KD, PM1D and PM3D. The correlation coefficients are shown in the upper left corner of the plots. The amino acids are **numbered according to the scheme on the right** and colored according to their class: white = polar, red = charged, green = apolar, yellow = aromatic.



## Supplementary Figure (2):

### Prediction of trans-membrane spans using a window for averaging

The figure shows the sliding-window approach for averaging the free energies for the prediction of trans-membrane spans from a protein sequence. The free energy is calculated as an average of the free energies of the amino acids located in the window where the middle residue has the highest weight. The result of the free energy is assigned to the central residue of the window.



**Supplementary Table (II):**

**Over-prediction of amino acids in the soluble region as being in the membrane**

To assess the over-prediction of amino acids in solution as being in the trans-membrane region the scales were tested on a dataset of non-redundant soluble proteins. The set was created by culling the PDB with the PISCES server as described in the Methods section. The set consisted of 2569 proteins with 3538 chains and 526,422 residues. The agreements are given in %.

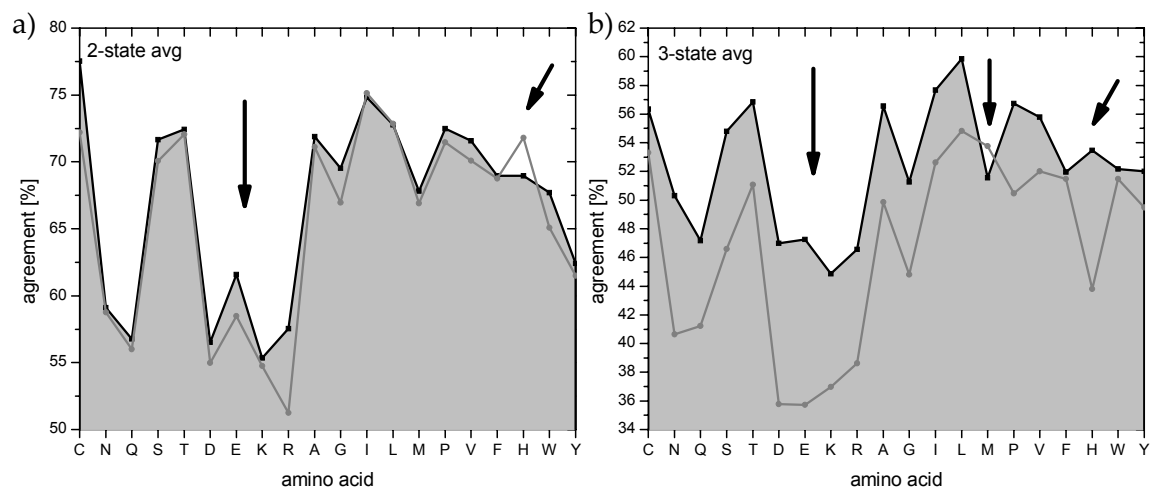
	predicted SOL	predicted TM
HWvH	100	0
WW	95.6	4.4
GES	86.3	13.7
UHS	85.7	14.3
Janin	74.5	25.2
KD	63.2	36.8
PM3D	53.2	46.7
Guy	51.7	48.3
PM1D	50.2	49.7
HW	49.4	50.5
EW	44.3	55.6

### Supplementary Figure (3):

#### Performance of the UHS as seen for the individual amino acid averages

The figure shows agreements between the predicted and actual locations for the individual amino acids. Figure (a) shows the performance of the UHS (black) and the GES (gray) in two-state scenario (TM and SOL) where the averages of the diagonal matrix elements (compare Table (IV)) are plotted against the amino acids. Figure (b) shows the performance of the UHS (black) and the WW (gray) in the three-state scenario with the averages of the diagonal matrix elements (compare Table (V)). For both scenarios a window length of 15 residues was used for averaging. The details are given in the Results and Discussion section:

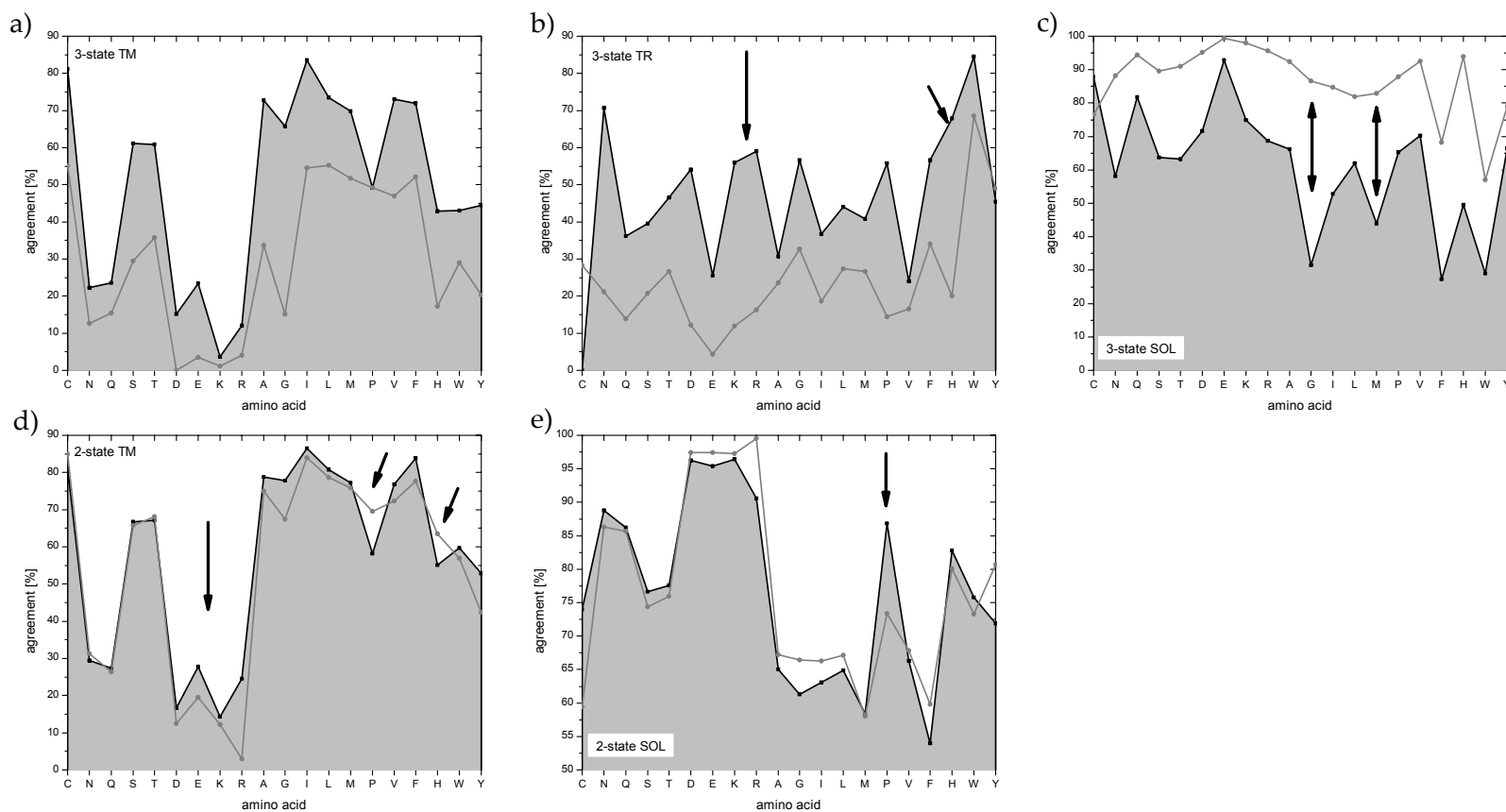
"Comparing the GES scale with the UHS, the average agreements have increased most for Arg (51% to 58%), Cys (72% to 78%), and Glu (58% to 62%). Note that the average agreement in the UHS is lower than in the GES scale only for His (72% to 69%). This indicates a slightly better representation of polar residues in the present UHS."



## Supplementary Figure (4):

### Performance of the UHS as seen for the individual amino acids in the different regions

The figure shows the individual amino acid agreements in the three-state (a-c) and two-state (d & e) scenario at a 15 residue window length for the UHS (black line) and the WW (gray in the upper panel) or the GES (gray in the lower panel). (a) 3-state TM agreement; (b) 3-state TR agreement; (c) 3-state SOL agreement; (d) 2-state TM agreement; (e) 2-state SOL agreement. **It can be seen that in the three-state scenario "the polar residues Arg, Asn, Asp, Glu, Gln, His, Lys, and Ser are predicted in a more balanced manner in the UHS than in the WW scale. When comparing the overall prediction accuracies, all amino acids either display an improvement or at least a similar accuracy for the UHS. Highest changes are observed for Asp and Glu (from 36% to 47%), Asn (from 41% to 50%), and His (from 44% to 53%)." (see Results and Discussion).**



## Supplementary Information:

### ***The UHS is largely independent of the protein fold***

We systematically excluded folds when deriving the UHS to address the question whether or not our scale is biased towards protein folds represented in the PDB. The following five folds were excluded one by one: aquaporins, outer membrane proteins, porins, bacteriorhodopsin, and the potassium channel (see Supplementary Table (III)).

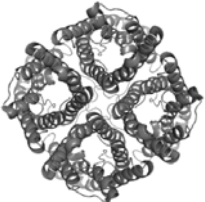



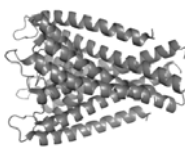
The hydrophobicity values that were derived without these different folds deviate on average 0.6 standard deviations from the UHS with a maximal deviation of three standard deviations for Glu in class 3. The largest deviations occur for classes 2 and 3. These changes are small in actual numbers given the range of hydrophobicity values. This indicates that the hydrophobicity value derived here is mostly an amino acid centered property largely independent of the fold of the protein. Further, the five resulting "leave-one-fold-out" UHS scales were used to predict TM and SOL regions within the "left-out" folds. The results of this experiment are summarized in Supplementary Table (IV).

The performance of these "leave-one-fold-out" UHS scales agrees on average to within 2.4% accuracy compared to the performance of the UHS scale. The largest deviations are 3% (SOL) for class 1 (1.51% for the average), 3.2% (SOL) for class 2 (1.92% for the average), 8.8% (TM) for class 3 (2.95% for the average), 1.4% (SOL) for class 4 (0.71% for the average), and 3.7% (SOL) for class 5 (1.84% for the average). This supports our argument that the UHS scale is largely fold independent.

### Supplementary Table (III):

#### The UHS is largely independent of the protein fold (continued)

Classes of proteins that were excluded from the derivation to assess the performance of the scale on novel folds. The UHS was derived when the following folds were excluded one by one and then tested on the excluded folds.

class #	1	2	3	4	5
class	aquaporins	outer membrane proteins	porins	bacterio-rhodopsin	potassium channel
#proteins	2	3	3	1	1
PDB ID	1YMG 2F2B	1EK9 1WP1 1YC9	1HXX 2FGQ 2MPR	1M0K	1R3J
#AAs	1912	3885	3273	666	412
fold					



**Supplementary Table (IV):**

**The UHS is largely independent of the protein fold (continued)**

The table shows the performance of the UHS for folds that have not been used for the derivation of the scale. #1 to #5 are the class numbers from Supplementary Table (III).

		<i>PDB</i>			<i>PDB</i>		
		<i>TM</i>	<i>SOL</i>	<i>avg</i>	<i>TM</i>	<i>SOL</i>	<i>avg</i>
<i>pred</i>	<i>TM</i>	#1	85.0	60.4	#2	47.0	16.3
	<i>SOL</i>		14.9	38.9		53.0	83.5
				61.95			65.26
<i>pred</i>	<i>TM</i>	#3	18.1	9.2	#4	92.4	54.7
	<i>SOL</i>		81.5	90.7		7.6	43.3
				54.42			67.84
<i>pred</i>	<i>TM</i>	#5	92.9	46.3			
	<i>SOL</i>		7.1	50.8			
				71.84			

### Supplementary Table (V):

#### The performance of the MHS in the two-state scenario

The prediction quality of the MHS was assessed by cross-validation and by testing the scale on the bacterial part of the MP database (*bact* in this table). The agreements of the MHS from cross-validation are very high for SOL (89.0%) and somewhat lower for the TM region (77.2%). The average agreement is therefore 83.1% which is the highest agreement of a hydrophobicity scale in this paper. When the MHS is tested on a bacterial dataset, the agreement in SOL decreases to 51.0%, leaving an average agreement of 67.74%. These results are somewhat expected considering that the database used for the MHS only consists of  $\alpha$ -helical proteins that are easier to predict than  $\beta$ -barrels (see below). In contrast, the bacterial database includes  $\beta$ -barrel proteins explaining the lower agreement on this set.

		<i>PDB</i>		
		<i>TM</i>	<i>SOL</i>	<i>avg</i>
<i>pred</i>	<i>TM</i>	<i>MHS</i>	77.2	10.9
	<i>SOL</i>		22.8	89.0
				83.08
<i>pred</i>	<i>TM</i>	<i>bact</i>	51.0	15.5
	<i>SOL</i>		49.0	84.5
				67.74