Supplementary

Supplementary Methods

Sign Convention of Experimental Values

Experimental ΔΔG values indicate the unfolding free energy of the mutant minus that of the wild type, as defined in MPTherm [17]. Negative values therefore represent destabilisation. However, the MutTMP Stab-pred supplementary text [22] and website interpret negative values as stabilisation. In this study, MPTherm's convention was followed and compared to the original publications. Curnow and Booth [45] also state that negative values indicate destabilisation. The twelve OmpA values attributed to Hong et al. [48], who use the reverse order of subtraction, correspond to the dataset values once their sign is reversed. Therefore, no sign change was made to the downloaded experimental values.

Cohorts

36 soluble domain records were removed first from the 379 records. Of the 14 records classified as INELIGIBLE_QUANTITY, 12 represented association free energies.

PRIMARY_SV contained 242 records, of which 136 were α-helices, and 106 were β-barrels. Four records, 1THQA_L57Q and 1THQA_S91K from PagP, and 3B45A_S181A and 3B45A_A265V from GlpG, had no identified original sources.

Initially, PRIMARY_SV contained 223 records. The source of 19 records of substitutions at V111 of PagP, given in Marx and Fleming [28], was identified later. These values, however, represent folding free energies with a single denaturant m value for all the variants, and therefore were classified as CONDITIONAL. But after the sources were identified, they were moved to PRIMARY_SV; hence the total increased to 242 records. Changes made did not affect records in PRIMARY or LEGACY.

Predictors

DDGun

DDGun followed the same sign convention as the experimental data; hence, no sign change was required. Local predictions were checked using the same sequences and mutations in 145 records across three structures (1PY6, 2XOV and 3GP6) using the DDGun web server [49]. The

local and server predictions agreed to within 0.1 kcal/mol, the resolution of DDGun's output, for 139 of the 145 records. Identical predictions were obtained for 89 records. The Spearman correlation for the two sets of predictions was 0.995. Local and server predictions for the mutation bacteriorhodopsin L13A differed by 1.7 kcal/mol, with the local prediction being −2.3 kcal/mol and the server prediction being −0.6 kcal/mol. Using the agreement rules that were established before the comparison, the two sets did not align. No reason could be established for this discrepancy, and thus local predictions were kept as is. Excluding the mutation in L13A or using the server value in place of the local value, changed DDGun's ρ by at most 0.004 but did not affect any of the conclusions drawn for DDGun. However, excluding this mutation moved the lower bound of ThermoMPNN's unadjusted fold contrast in PRIMARY_SV from −0.003 to +0.002. This small change is the reason the unadjusted contrast is described as borderline.

ThermoMPNN and RaSP values

To check the extracted predictions, forty-two records were randomly selected (seed 20261114) from the analysis tables of 269 scored records, three records were selected from each of the fourteen structures, except two structures with only two records each (1QJ8 and 3B45), which were taken completely and five records were selected for bacteriorhodopsin structure 1C3W.

RaSP and ThermoMPNN values were extracted from the raw prediction files for each of these records using a different method. In this method, residue positions were mapped from the protein structure files as opposed to SEQRES. All 42 values were in agreement for both predictors. RaSP predictions agreed to the sixth decimal place stored, and ThermoMPNN predictions were exact. Sign conversions were also correct. The same was also done for the 88 BamA variants, and the values were in agreement.

RaSP Environment

The authors' Colab notebook states that support for the notebook has been discontinued and suggests that the user installs it locally. The local environment documented as Python 3.6, PyTorch 1.2.0 and OpenMM 7.3.1, could not be installed on the computer available to the author. Therefore, the authors' notebook was used to run RaSP, and the other required packages were manually installed. All runs used an NVIDIA Tesla T4 GPU with OpenMM 8.6.1 and PDBFixer 1.12.0. The corpus predictions used Python 3.13.15, PyTorch 2.11.0 and numpy 2.1.3, the BamA predictions Python 3.13.16, PyTorch 2.11.0 and numpy 2.4.6; and the PMP-22 comparison Python 3.13.16, PyTorch 2.14.1 and numpy 2.4.6. Updated and modified versions

of the notebook have been published and used in other studies [59-61]. All 30 model files were identical across runs. Predictions for a test protein (1CTF) were identical on each of the three occasions.

Comparison of the authors' archive with RaSP

PMP-22 was re-scored using the same AlphaFold model as that used in the authors' archived human proteome predictions [8]. The sequence, residue numbering, score definition and amino-acid frequency table were the same in both. Of the 10 PMP-22 substitutions in the benchmark, only one was in agreement with the archived prediction, within the tolerance of 0.05 (Supplementary Table S11). Across all 3,200 variants of PMP-22, the two sets showed a Pearson correlation of 0.93 and least-squares slope of 1.007. The median absolute difference was 0.29. 317 variants agreed within 0.05. The discrepancies occurred throughout the length of the protein and were not limited to particular residues. These were also not concentrated in regions with low model confidence. The reasons for the differences could not be determined.

Statistics

Proteins were identified based on their UniProt accessions for bootstrap analysis. Bonferroni correction was performed separately on two sets of comparisons. This comprises three comparisons between DDGun, ESM-2 and ThermoMPNN, and three fold contrasts (β-barrel minus α-helical) predictions. In the case of the initial analysis with four predictors, the correction encompassed six pairwise comparisons and four fold contrasts. Confidence intervals for individual proteins (Figure 3B) were calculated by resampling records within each protein.

Held-out analysis

The refitted model was run on an Apple arm64 Mac using the MutTMB Stab-pred reproducibility package environment (Python 3.13.5, scikit-learn 1.8.0, NumPy 2.3.5, and pandas 2.2.3). Other environments gave very similar results. Differences in Pearson's r were observed only to the third decimal place (differences in r below 0.01 were not interpreted). Spearman’s ρ was calculated, as shown in Supplementary Table S7, which is used to test sensitivity.

Reconstruction of held-out evaluation and archived predictions

The package contains archived predictions from leave-one-protein-out (LOPO) and leave-one-homology-cluster-out (LOHCO) analyses. Running the code in the package with the software

versions specified by the authors did not reproduce those predictions. For LOPO, all 214 records were matched using row identifiers, and the measured values were the same. None of the predictions agreed within 10^{-6}. The mean absolute difference between the two sets of predictions was 0.7486, and the maximum difference was 3.8775. The correlation between the two sets of predictions was 0.8845. For LOHCO, corresponding values were 0.9970, 4.8255 and 0.8862, respectively. The reasons for these differences could not be determined; the results shown in Figure 2 were derived from the refitted model.

Separate evaluations of archived predictions were also performed using the direction label of each record to identify mutations that were measured experimentally. For LOPO, Pearson's correlation was 0.74 and 0.57 for the 214 records and 107 measured mutations, respectively. The mean absolute error for both groups was 1.31 kcal/mol. For LOHCO, the correlations were 0.71 and 0.55 for the 214 records and the 107 measured mutations, respectively; the mean absolute error was 1.52 kcal/mol for both cases.

The archived predictions also follow the same trend shown in Figure 2, with lower correlations when only experimentally measured mutations were evaluated. The LOPO results across all 214 records are in agreement with those values reported for MutTMB Stab-pred (r = 0.74, MAE = 1.31 kcal/mol). However, the published LOHCO values were r = 0.72 and MAE = 1.35 kcal/mol [22], which differed from those calculated using the archived predictions.

All measured and predicted values for the 214 records were present in each of the archived files. All records were evaluated. The 107 experimentally measured mutations were identified by their direct mutation labels, with each paired with a reverse mutation carrying the negative of the measured value. Predictions from three held-out homology clusters averaged to r = 0.58 and a mean absolute error of 1.33 kcal/mol. MAE values and the correlation coefficient obtained differed from the published LOHCO values. The reason for the discrepancy could not be determined.

BamA

For the structure-based methods, the AlphaFold model of BamA (UniProt P0A940, model version 6) was used. The signal peptide from residues 1 to 20 was removed, which will leave residues 21 to 810 with their original numbering and coordinates. C690S and C700S were introduced by replacing each cysteine Sγ atom with a serine Oγ atom, at 1.417 Å along the same Cβ–Sγ direction, without energy minimisation or side-chain repacking. Every other atom

remained unchanged. Positions 690 and 700 are at least 19 Å (Cα–Cα) from every one of the 88 scanned residues. The pLDDT of the 88 scanned residues had a median of 98.4, with 83 at or above 90 and 2 below 70 (median 93.8 over all 790 residues). DDGun and ESM-2 input was the full 810-residue precursor sequence, which carries C690S and C700S. Thus, residues 21 to 810 were identical to the structure, and the signal peptide was also present. The effect of the signal peptide on these two predictions was not tested.

The published uncertainties were standard deviations obtained from five independent datasets for the wild type and two datasets for each variant [15]. The wild-type contribution to each variant's published ΔΔG_F uncertainty was estimated by subtracting the squared mutant ΔG_F uncertainty from the squared ΔΔG_F uncertainty and taking the square root.

For each variant, the difference between the two lipid environments was calculated using three approaches:

1. Using the published standard deviation for each of the two lipids, the uncertainty was determined using a Student’s t distribution with Welch-Satterthwaite approximation [62,63]. This approach is shown in Figure 4A.
2. Standard errors were calculated by dividing the wild-type and mutant contributions by the square roots of the respective dataset numbers (5 and 2). Standard errors were then combined to determine an interval using the normal quantile of 1.96.
3. The same standard errors were used, but the interval was calculated using a Student's t quantile instead of the normal quantile.

The three methods produced 1, 37 and 17 variants with zero excluded intervals, respectively. Intervals were entirely above 1 kcal/mol for 0, 2 and 1 variants, respectively. An additional calculation using the published standard deviations with the normal quantile gave corresponding counts of 5 and 0. Since these counts depended on how uncertainties were combined and intervals determined, they were not interpreted any further.

Of the 88 BamA variants, 34 variants showed differences of more than 1 kcal/mol between the two lipid environments. All 88 variants, however, share the same wild-type values, and hence intervals for individual variants are correlated. Consequently, each of the intervals describes one variant only, and no joint statement is made. An interval for the mean difference was also not reported.

Supplementary Tables

Supplementary Table S1. Annotation of the records. Available as a separate file (Supplementary_Table_S1.csv) containing 379 records in the MutTMP Stab-pred files. Each record in the file provides identifiers containing the following information: protein identifier (protein, UniProt accession, structure, substitution), fold, topology and class of the membrane, last selection step (see Supplementary Table S2), and which cohort the record belongs to. Measurements can be found in the MutTMP Stab-pred files [22] and can be associated with each record using the identifier Mutation_ID. In the case of the 271 LEGACY records, the table also provides the name of the source document, the extent of the attribution to the primary publication or compilation, if any, the nature of the relationship of the value to the source value and data set value, and the eligibility class of the record. A record marked as source_identified means that it has an identified source for its value, and not that the value was reproduced from that source.

Supplementary Table S2. Selection steps and definitions of the cohort. Records remaining after each selection step applied to the 379 unique records published in the MutTMP Stab-pred files, broken down by α-helices and β-barrels, and records eliminated at each step.

Step	Cohort	Rule	Records	α-helical	β-barrel	Removed	Removed by
1	published unique	every unique record in the MutTMP Stab-pred alpha.csv and beta.csv releases	379	254	125	0	
2	in a membrane structure	membrane class of the deposited structure is integral or conditional	343	218	125	36	36 soluble domain
3	+ transmembrane residue	and the residue is annotated as membrane-spanning (M) in the source three-state topology	273	160	113	70	70 not membrane-spanning
4	LEGACY	and the membrane class is integral; the original structural subset	271	158	113	2	2 conditional

5	PRIMARY	and the eligibility class is ELIGIBLE or CONDITIONAL	246	138	108	25	11 INELIGIBLE GENOTYPE; 14 INELIGIBLE QUANTITY
6	PRIMARY_SV	and a source was identified for the value	242	136	106	4	2 PagP; 2 Rhomboid protease GlpG

Supplementary Table S3. Records per protein in each cohort. LEGACY includes two records without accompanying structure files (Dengue virus 2 genome polyprotein), which were not assessed; thus, 269 LEGACY records were assessed.

Protein	UniProt	Fold	LEGACY	PRIMARY	PRIMARY_SV
Bacteriorhodopsin	P02945	α-helical	66	66	66
PagP	P37001	β-barrel	61	61	59
Rhomboid protease GlpG	P09391	α-helical	51	51	49
OmpLA (outer membrane phospholipase A)	P0A921	β-barrel	29	29	29
OmpA TM domain	P0A910	β-barrel	17	12	12
Glycophorin A TM dimer	P02724	α-helical	12	0	0
DsbB	P0A6M2	α-helical	11	11	11
PMP-22 (peripheral myelin prot 22)	Q01453	α-helical	10	10	10
Lactose permease LacY	P02920	α-helical	6	0	0
Ail (attachment invasion locus protein)	Q8D0Z7	β-barrel	4	4	4

Dengue virus 2 genome polyprotein	P29991	α-helical	2	0	0
OmpX	P0A917	β-barrel	2	2	2
Total			271	246	242

Supplementary Table S4. Performance of predictors in LEGACY, PRIMARY and PRIMARY_SV. Agreement of each predictor with measured ΔΔG in the three cohorts. Intervals are 95% percentile intervals from 10,000 bootstrap replicates where proteins were sampled with replacement and records of each protein kept together. For fold-specific values, proteins were sampled within that fold. MAE is not applicable (n.a.) for ESM-2 as it is unitless. The last column shows the largest change in pooled ρ when that particular protein was removed. RaSP was run in a modified execution environment and is reported as a secondary analysis (Supplementary Methods).

Cohort	Predictor	n	Spearman ρ (95% interval)	ρ, α-helical (95% interval)	ρ, β-barrel (95% interval)	Pearson r (95% interval)	MAE, kcal/mol (95% interval)	Largest change in ρ on leaving out one protein
PRIMARY	DDGun	246	0.49 (0.30 to 0.61)	0.31 (0.08 to 0.49)	0.61 (0.54 to 0.79)	0.51 (0.23 to 0.63)	1.27 (1.08 to 1.44)	+0.07 (GlpG)
PRIMARY	ESM-2	246	0.43 (0.15 to 0.58)	0.31 (−0.13 to 0.45)	0.59 (0.34 to 0.69)	n.a.	n.a.	−0.10 (PagP)
PRIMARY	ThermoMPNN	246	0.43 (0.31 to 0.59)	0.33 (0.01 to 0.41)	0.57 (0.34 to 0.73)	0.43 (0.26 to 0.60)	1.13 (0.97 to 1.32)	−0.06 (OmpLA)
PRIMARY	RaSP (secondary)	246	0.28 (0.21 to 0.38)	0.31 (0.07 to 0.40)	0.39 (0.30 to 0.56)	0.17 (0.05 to 0.33)	1.60 (1.42 to 1.78)	+0.02 (GlpG)
PRIMARY_SV	DDGun	242	0.50 (0.31 to 0.62)	0.32 (0.08 to 0.49)	0.62 (0.55 to 0.79)	0.54 (0.27 to 0.63)	1.24 (1.08 to 1.39)	+0.07 (GlpG)

PRIMARY_SV	ESM-2	242	0.42 (0.14 to 0.57)	0.30 (−0.13 to 0.44)	0.58 (0.34 to 0.69)	n.a.	n.a.	−0.10 (PagP)
PRIMARY_SV	ThermoMPNN	242	0.42 (0.29 to 0.59)	0.32 (0.01 to 0.41)	0.56 (0.34 to 0.73)	0.42 (0.25 to 0.60)	1.11 (0.94 to 1.29)	−0.06 (OmpLA)
PRIMARY_SV	RaSP (secondary)	242	0.27 (0.18 to 0.37)	0.29 (0.07 to 0.40)	0.37 (0.28 to 0.56)	0.16 (0.03 to 0.31)	1.58 (1.40 to 1.78)	+0.03 (GlpG)
LEGACY	DDGun	269	0.41 (0.14 to 0.57)	0.23 (−0.11 to 0.46)	0.52 (0.10 to 0.78)	0.40 (0.06 to 0.58)	1.41 (1.17 to 1.77)	−0.07 (PagP)
LEGACY	ESM-2	269	0.39 (0.09 to 0.55)	0.29 (−0.21 to 0.42)	0.49 (0.06 to 0.69)	n.a.	n.a.	−0.11 (PagP)
LEGACY	ThermoMPNN	269	0.41 (0.26 to 0.55)	0.32 (−0.15 to 0.39)	0.49 (−0.05 to 0.73)	0.37 (0.14 to 0.55)	1.18 (0.98 to 1.42)	−0.07 (OmpLA)
LEGACY	RaSP (secondary)	269	0.27 (0.17 to 0.35)	0.27 (−0.19 to 0.37)	0.36 (0.17 to 0.46)	0.15 (0.004 to 0.29)	1.61 (1.44 to 1.80)	−0.02 (PagP)

Supplementary Table S5. Spearman correlation for each protein with at least 30 records. Intervals are 95% confidence intervals from resampling records within the protein. Since substitutions at the same site are correlated, these intervals are narrower than resampling of sites would provide and are not comparable with those in Supplementary Table S4.

Cohort	Protein	n	DDGun	ESM-2	ThermoMPNN	RaSP (secondary)
PRIMARY	Bacteriorhodopsin	66	0.51 (0.29 to 0.69)	0.44 (0.21 to 0.61)	0.38 (0.16 to 0.58)	0.40 (0.19 to 0.58)
PRIMARY	PagP	61	0.59 (0.39 to 0.74)	0.63 (0.45 to 0.77)	0.42 (0.17 to 0.63)	0.29 (0.02 to 0.52)
PRIMARY	Rhomboid protease GlpG	51	0.13 (−0.14 to 0.38)	0.45 (0.19 to 0.66)	0.26 (0.001 to 0.49)	0.20 (−0.08 to 0.45)

PRIMARY_SV	Bacteriorhodopsin	66	0.51 (0.29 to 0.69)	0.44 (0.21 to 0.61)	0.38 (0.16 to 0.58)	0.40 (0.19 to 0.58)
PRIMARY_SV	PagP	59	0.60 (0.40 to 0.75)	0.63 (0.44 to 0.76)	0.39 (0.14 to 0.60)	0.26 (−0.01 to 0.49)
PRIMARY_SV	Rhomboid protease GlpG	49	0.13 (−0.14 to 0.38)	0.44 (0.18 to 0.64)	0.22 (−0.04 to 0.45)	0.14 (−0.15 to 0.40)
LEGACY	Bacteriorhodopsin	66	0.51 (0.30 to 0.69)	0.44 (0.22 to 0.62)	0.38 (0.16 to 0.58)	0.40 (0.19 to 0.58)
LEGACY	PagP	61	0.59 (0.39 to 0.74)	0.63 (0.45 to 0.77)	0.42 (0.17 to 0.63)	0.29 (0.03 to 0.52)
LEGACY	Rhomboid protease GlpG	51	0.13 (−0.14 to 0.38)	0.45 (0.19 to 0.66)	0.26 (0.0001 to 0.49)	0.20 (−0.08 to 0.45)

Supplementary Table S6A. Pairwise predictor contrasts. Pairwise differences in Spearman's ρ for the first and second predictors, computed within each bootstrap replicate for the same records, with proteins which are resampled. Bonferroni adjustment is shown for each family of comparisons. Bonferroni correction has been applied within each family of comparisons defined after the RaSP decision (three comparisons) and before (six comparisons). LEGACY is shown for the original family only as a sensitivity analysis.

Cohort	Family	Comparison	Difference in ρ	95% interval	Bonferroni-adjusted interval	Comparisons in family
PRIMARY	three predictors (revised)	DDGun minus ESM-2	+0.062	−0.098 to +0.296	−0.148 to +0.381	3
PRIMARY	three predictors (revised)	DDGun minus ThermoMPNN	+0.061	−0.103 to +0.199	−0.140 to +0.219	3
PRIMARY	three predictors (revised)	ESM-2 minus ThermoMPNN	−0.001	−0.277 to +0.157	−0.370 to +0.178	3
PRIMARY	four predictors (original)	DDGun minus ESM-2	+0.062	−0.098 to +0.296	−0.179 to +0.428	6
PRIMARY	four predictors (original)	DDGun minus RaSP	+0.209	+0.026 to +0.296	−0.042 to +0.321	6
PRIMARY	four predictors (original)	DDGun minus ThermoMPNN	+0.061	−0.103 to +0.199	−0.158 to +0.228	6

PRIMARY	four predictors (original)	ESM-2 minus RaSP	+0.147	−0.103 to +0.266	−0.250 to +0.292	6
PRIMARY	four predictors (original)	ESM-2 minus ThermoMPNN	−0.001	−0.277 to +0.157	−0.400 to +0.190	6
PRIMARY	four predictors (original)	RaSP minus ThermoMPNN	−0.148	−0.277 to −0.027	−0.331 to −0.006	6
PRIMARY_SV	three predictors (revised)	DDGun minus ESM-2	+0.079	−0.078 to +0.303	−0.129 to +0.383	3
PRIMARY_SV	three predictors (revised)	DDGun minus ThermoMPNN	+0.080	−0.087 to +0.221	−0.123 to +0.245	3
PRIMARY_SV	three predictors (revised)	ESM-2 minus ThermoMPNN	+0.002	−0.277 to +0.170	−0.370 to +0.195	3
PRIMARY_SV	four predictors (original)	DDGun minus ESM-2	+0.079	−0.078 to +0.303	−0.162 to +0.429	6
PRIMARY_SV	four predictors (original)	DDGun minus RaSP	+0.236	+0.054 to +0.316	−0.006 to +0.345	6
PRIMARY_SV	four predictors (original)	DDGun minus ThermoMPNN	+0.080	−0.087 to +0.221	−0.144 to +0.255	6
PRIMARY_SV	four predictors (original)	ESM-2 minus RaSP	+0.157	−0.100 to +0.287	−0.250 to +0.318	6
PRIMARY_SV	four predictors (original)	ESM-2 minus ThermoMPNN	+0.002	−0.277 to +0.170	−0.400 to +0.208	6
PRIMARY_SV	four predictors (original)	RaSP minus ThermoMPNN	−0.155	−0.286 to −0.030	−0.334 to −0.009	6
LEGACY (sensitivity)	four predictors (original)	DDGun minus ESM-2	+0.018	−0.175 to +0.249	−0.266 to +0.361	6
LEGACY (sensitivity)	four predictors (original)	DDGun minus RaSP	+0.135	−0.080 to +0.261	−0.172 to +0.300	6
LEGACY (sensitivity)	four predictors (original)	DDGun minus ThermoMPNN	−0.002	−0.209 to +0.149	−0.294 to +0.197	6
LEGACY (sensitivity)	four predictors (original)	ESM-2 minus RaSP	+0.117	−0.124 to +0.241	−0.246 to +0.271	6

LEGACY (sensitivity)	four predictors (original)	ESM-2 minus ThermoMPNN	−0.019	−0.306 to +0.139	−0.396 to +0.171	6
LEGACY (sensitivity)	four predictors (original)	RaSP minus ThermoMPNN	−0.137	−0.285 to −0.027	−0.359 to −0.002	6

Supplementary Table S6B. Fold contrasts for β-barrel minus α-helical folds. Differences in Spearman's ρ for β-barrel and α-helical folds (β-barrel minus α-helical) with the proteins which were resampled for each fold type. Bonferroni adjustment is shown for the three comparisons within the three-predictor family and for the four comparisons within the original family. Folds cannot be dissociated from protein identities in these datasets, and therefore these comparisons are exploratory.

Cohort	Family	Predictor	ρ, α-helical	ρ, β-barrel	Difference (β minus α)	95% interval	Bonferroni-adjusted interval	Contrasts in family	Proteins (α / β)
PRIMARY	three predictors (revised)	DDGun	0.31	0.61	+0.300	+0.110 to +0.664	+0.067 to +0.691	3	4 / 5
PRIMARY	three predictors (revised)	ESM-2	0.31	0.59	+0.276	−0.033 to +0.726	−0.161 to +0.772	3	4 / 5
PRIMARY	three predictors (revised)	ThermoMPNN	0.33	0.57	+0.240	+0.007 to +0.601	−0.179 to +0.717	3	4 / 5
PRIMARY	four predictors (original)	DDGun	0.31	0.61	+0.300	+0.110 to +0.664	+0.032 to +0.697	4	4 / 5
PRIMARY	four predictors (original)	ESM-2	0.31	0.59	+0.276	−0.033 to +0.726	−0.202 to +0.779	4	4 / 5
PRIMARY	four predictors (original)	RaSP	0.31	0.39	+0.081	−0.053 to +0.365	−0.096 to +0.428	4	4 / 5
PRIMARY	four predictors (original)	ThermoMPNN	0.33	0.57	+0.240	+0.007 to +0.601	−0.224 to +0.736	4	4 / 5

PRIMARY_SV	three predictors (revised)	DDGun	0.32	0.62	+0.296	+0.117 to +0.664	+0.071 to +0.691	3	4 / 5
PRIMARY_SV	three predictors (revised)	ESM-2	0.30	0.58	+0.280	−0.023 to +0.719	−0.154 to +0.767	3	4 / 5
PRIMARY_SV	three predictors (revised)	ThermoMPNN	0.32	0.56	+0.241	−0.003 to +0.600	−0.169 to +0.713	3	4 / 5
PRIMARY_SV	four predictors (original)	DDGun	0.32	0.62	+0.296	+0.117 to +0.664	+0.029 to +0.697	4	4 / 5
PRIMARY_SV	four predictors (original)	ESM-2	0.30	0.58	+0.280	−0.023 to +0.719	−0.199 to +0.774	4	4 / 5
PRIMARY_SV	four predictors (original)	RaSP	0.29	0.37	+0.082	−0.068 to +0.372	−0.110 to +0.462	4	4 / 5
PRIMARY_SV	four predictors (original)	ThermoMPNN	0.32	0.56	+0.241	−0.003 to +0.600	−0.211 to +0.736	4	4 / 5
LEGACY (sensitivity)	four predictors (original)	DDGun	0.23	0.52	+0.287	−0.172 to +0.731	−0.408 to +0.829	4	6 / 5
LEGACY (sensitivity)	four predictors (original)	ESM-2	0.29	0.49	+0.195	−0.287 to +0.753	−0.443 to +0.869	4	6 / 5
LEGACY (sensitivity)	four predictors (original)	RaSP	0.27	0.36	+0.091	−0.131 to +0.546	−0.259 to +0.622	4	6 / 5
LEGACY (sensitivity)	four predictors (original)	ThermoMPNN	0.32	0.49	+0.176	−0.339 to +0.689	−0.549 to +0.835	4	6 / 5

Supplementary Table S7. Held-out unit and scoring basis for the refitted MutTMB Stab-pred model. Performance of the refitted MutTMB Stab-pred model by held-out unit, scored against 214 rows (107 measured and 107 generated reverse mutations) and on 107 measured mutations only. Holding out one PDB entry partitions the data in the same way holding out a structure file would. Values shown in-sample are for reference only and are not estimates of the held-out values. MAE in kcal/mol.

Held-out unit	Held-out groups	r, all 214 rows	MAE, all 214 rows	ρ, all 214 rows	r, 107 measured	MAE, 107 measured	ρ, 107 measured
none (in-sample, reference)	1	0.97	0.51	0.96	0.95	0.52	0.93
one row	214	0.80	1.12	0.84	0.68	1.14	0.65
one direct-reverse pair	107	0.80	1.14	0.84	0.67	1.15	0.64
one structure file	5	0.66	1.48	0.71	0.45	1.49	0.45
one PDB entry	5	0.66	1.48	0.71	0.45	1.49	0.45
one UniProt protein	4	0.60	1.64	0.64	0.43	1.64	0.47
one homology cluster	3	0.60	1.67	0.63	0.44	1.68	0.48

Supplementary Table S8. Predictions for 88 BamA alanine variants. Available as a separate file (Supplementary_Table_S8.csv). One row entry for each alanine variant predicted in the cysteine-free background (C690S/C700S) using the precursor numbering. Units for DDGun_ddG, RaSP_ddG, ThermoMPNN_ddG, mCSM-membrane_ddG and MutTMB_ddG are kcal/mol, and negative values represent destabilising substitutions. ESM-2_score is a log-likelihood ratio of the mutation compared to the wild type, and positive values indicate the mutation is more probable under the model. Values have not been rounded and represent the outputs of the respective sign change methods as described in Supplementary Table S10. Due to a lack of input for the lipid composition, each method makes a single prediction for both PC:PE and PC:PG. Measured values are PC:PE and PC:PG differences not reported and are given in Tables S8 and S9 of George et al. [15].

Supplementary Table S9. Agreement of six methods with measured values of BamA. Spearman's ρ and 95% confidence intervals based on 10,000 bootstrap replicates in which variants were resampled. Also included is Pearson's correlation and mean absolute error (MAE) in kcal/mol without intervals. Each method predicts one value per variant and is compared with PC:PE value, PC:PG values and difference between them. RaSP results are from the modified execution environment.

Method	Target	n	Spearman ρ (95% interval)	Pearson r	MAE, kcal/mol
DDGun	PC:PE	88	0.57 (0.40 to 0.70)	0.60	0.75
DDGun	PC:PG	88	0.49 (0.31 to 0.63)	0.51	0.67
DDGun	PC:PE minus PC:PG	88	0.03 (−0.17 to 0.24)	0.05	n.a.
ESM-2	PC:PE	88	0.10 (−0.12 to 0.31)	n.a.	n.a.
ESM-2	PC:PG	88	0.19 (−0.03 to 0.38)	n.a.	n.a.
ESM-2	PC:PE minus PC:PG	88	−0.10 (−0.28 to 0.08)	n.a.	n.a.
ThermoMPNN	PC:PE	88	0.60 (0.45 to 0.72)	0.58	1.14
ThermoMPNN	PC:PG	88	0.48 (0.29 to 0.64)	0.45	0.73
ThermoMPNN	PC:PE minus PC:PG	88	0.08 (−0.13 to 0.28)	0.11	n.a.
mCSM-membrane	PC:PE	88	0.15 (−0.06 to 0.35)	0.12	0.79
mCSM-membrane	PC:PG	88	0.22 (0.004 to 0.43)	0.17	0.68
mCSM-membrane	PC:PE minus PC:PG	88	−0.15 (−0.36 to 0.07)	−0.09	n.a.
MutTMB Stab-pred	PC:PE	88	0.51 (0.32 to 0.66)	0.53	0.86
MutTMB Stab-pred	PC:PG	88	0.46 (0.27 to 0.61)	0.46	0.99

MutTMB Stab-pred	PC:PE minus PC:PG	88	0.03 (−0.19 to 0.25)	0.02	n.a.
RaSP (secondary)	PC:PE	88	0.51 (0.34 to 0.64)	0.46	1.12
RaSP (secondary)	PC:PG	88	0.49 (0.32 to 0.62)	0.44	0.78
RaSP (secondary)	PC:PE minus PC:PG	88	−0.04 (−0.26 to 0.18)	−0.04	n.a.

Supplementary Table S10. Versions, inputs, settings, sign conventions and basis for inclusion of the six methods. The details of version, input, setting and sign conversion for the six methods used. Under the standard experimental convention, negative values indicate destabilisation. ThermoMPNN and RaSP outputs were multiplied by −1 to match this sign convention.

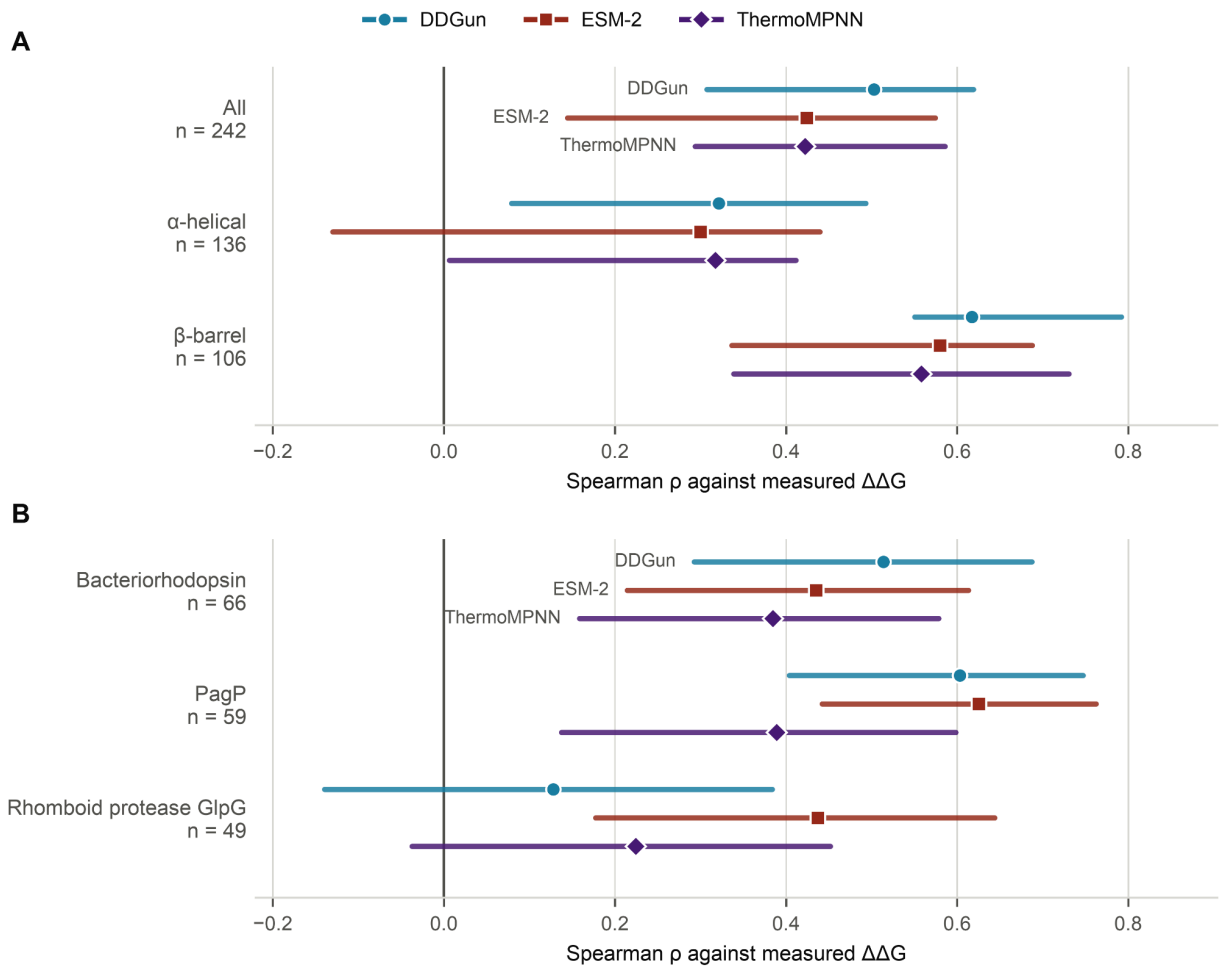
Method	Version	Input	Settings	Sign convention	Role
DDGun	Commit bcf112c; HHblits 3.3.0; uniclust30_2018_08	SEQRES sequence; full-length Cys-less BamA	Sequence mode, default settings	Unchanged	Primary
ESM-2	esm2_t33_650M_UR50D; CPU, float32	Same as DDGun	Masked marginal; log P(mutant) − log P(WT)	Unchanged; dimensionless	Primary
ThermoMPNN	Commit 370f76ec; default checkpoint; v_48_020 weights	Dataset structures; AlphaFold for PMP-22 and Cys-less BamA (21–810)	Default saturation scan	Reversed	Primary
RaSP	Authors' Colab notebook; 30 model files	Same as ThermoMPNN	Default settings; score_ml	Reversed	Secondary

mCSM-membrane	Web server	Prepared BamA structure	Batch mode, defaults	Unchanged	BamA only
MutTMB Stab-pred	Distributed model; scikit-learn 1.0.2	Prepared BamA structure	Individual substitutions	Unchanged	BamA only

Supplementary Table S11. RaSP scores for the ten PMP-22 substitutions in the benchmark, from the modified environment and from the authors' archive. Raw RaSP score (score_ml) on the AlphaFold model of PMP-22, prior to the sign change. Positive values indicate a destabilised substitution. One of ten substitutions agreed within a tolerance of 0.05, which was set before the predictions.

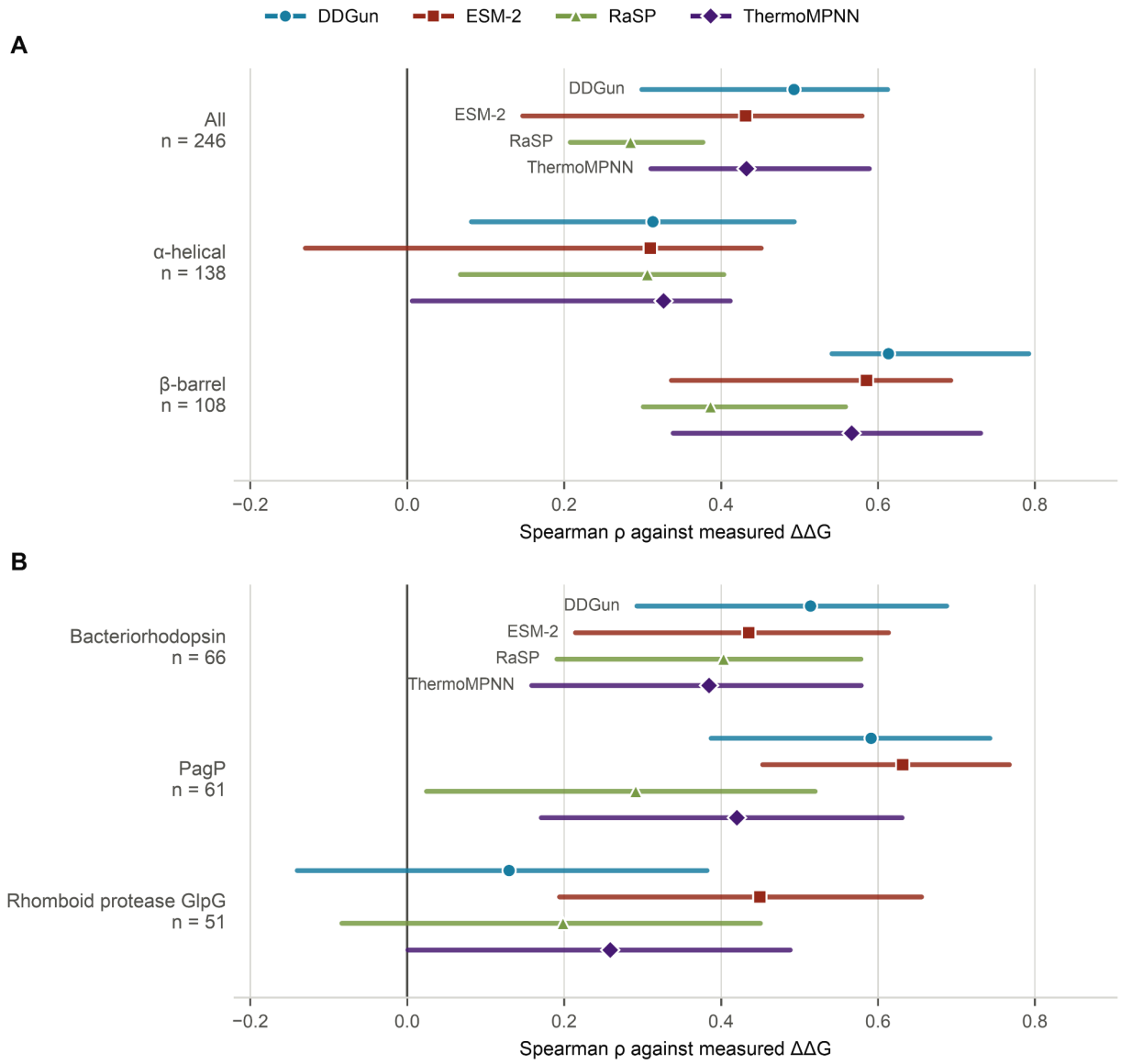
Substitution	The prediction obtained	Archive	Absolute difference
L16P	9.263	9.242	0.021
G150D	6.666	6.956	0.290
M69K	3.375	4.370	0.995
S22F	3.702	5.725	2.023
H12Q	1.055	1.301	0.246
G107V	1.287	2.359	1.073
A67T	0.353	0.414	0.061
I137V	0.224	0.322	0.098
T118M	−0.311	0.256	0.567
S76I	−0.460	1.014	1.474

Supplementary Figure



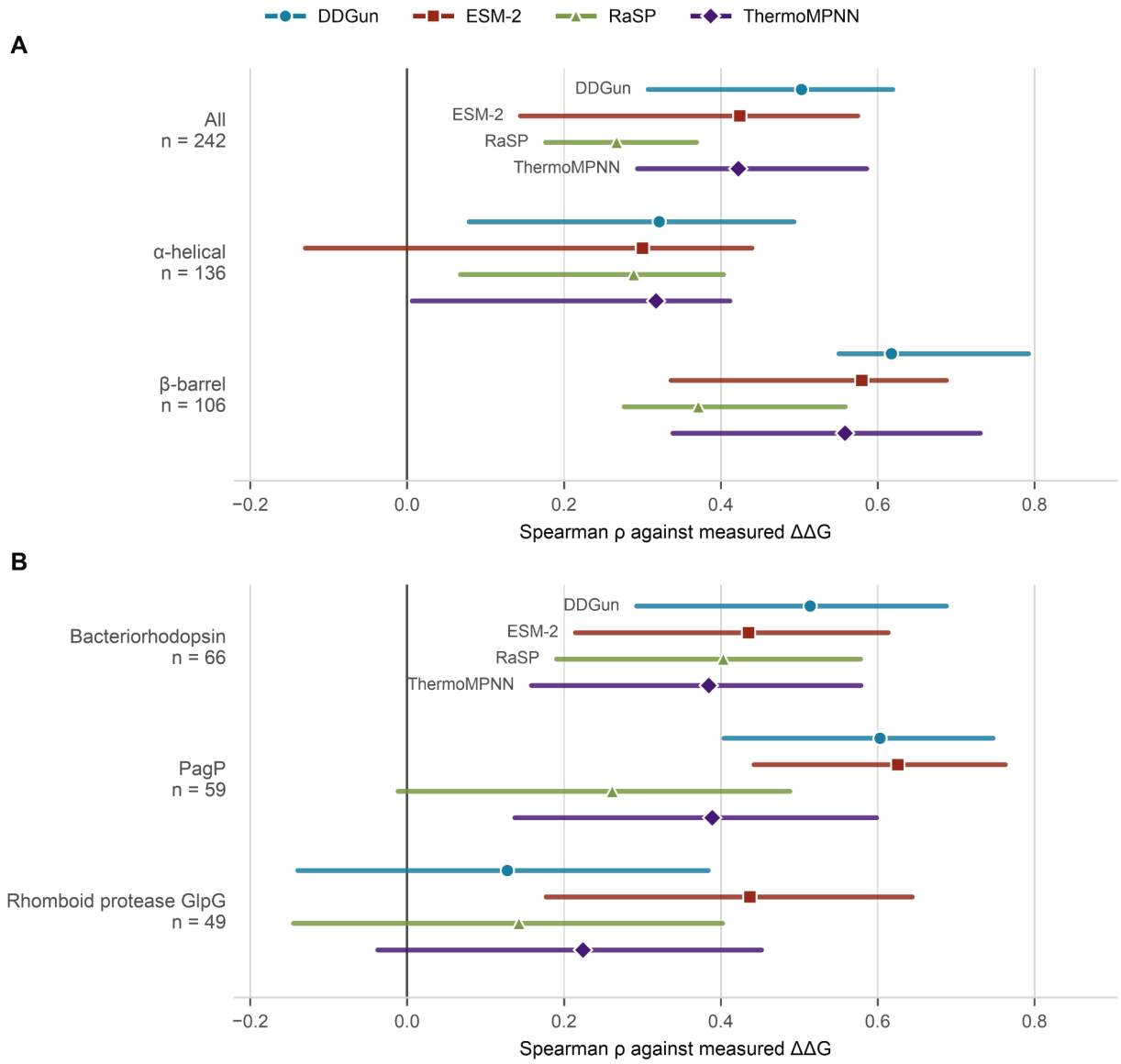
Supplementary Figure S1. Agreement between DDGun, ESM-2 and ThermoMPNN and measured ΔΔG in PRIMARY_SV.

As in Figure 3, for the 242 records with identified sources (136 α-helical, 106 β-barrel), from nine proteins (four α-helical and five β-barrel). (A) Spearman's ρ for all records and for each fold. Bars are 95% confidence intervals from 10,000 bootstrap replicates where proteins were resampled with replacement, and records within each protein were kept together; for each fold, only the proteins for that fold were resampled. (B) Spearman's ρ for each protein with at least 30 records, bacteriorhodopsin (n = 66), PagP (n = 59) and GlpG (n = 49). Circles, DDGun; squares, ESM-2; diamonds, ThermoMPNN. Bars are 95% confidence intervals from resampling records within the protein. They are narrower than a resampling of sites would give and are not comparable with those in A.



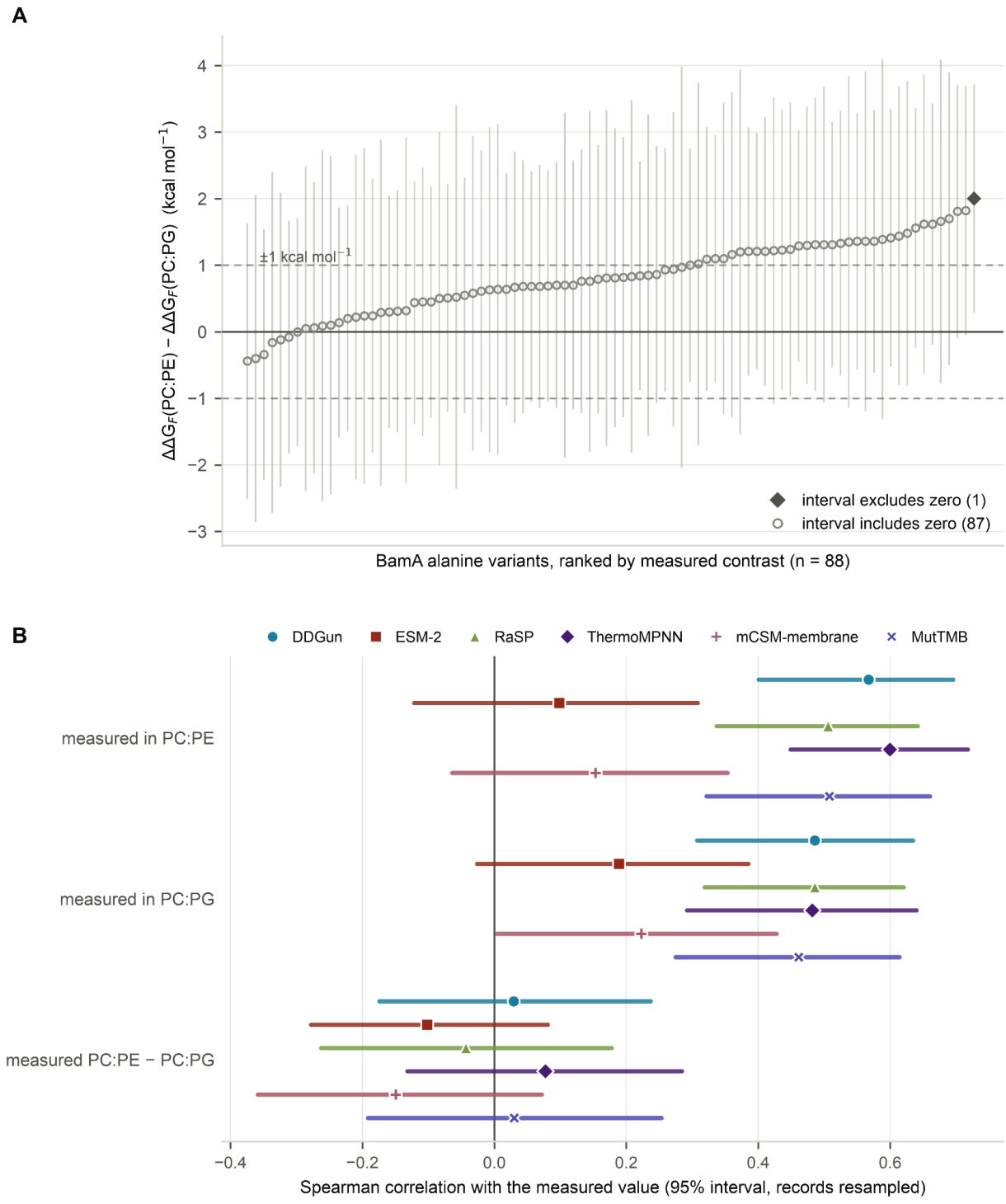
Supplementary Figure S2. Performance of the four originally defined predictors with measured ΔΔG in PRIMARY.

As in Figure 3, except showing the added RaSP prediction (246 records: 138 α-helical and 108 β-barrel). RaSP values are from the modified execution environment and are reported as a secondary analysis (see Supplementary Methods, RaSP environment). (A) Spearman's ρ for all records and each fold, with protein resampled 95% confidence intervals as in Figure 3A. (B) Spearman's ρ for each of the three proteins (bacteriorhodopsin, n = 66; PagP, n = 61; and GlpG, n = 51), with record-resampled 95% confidence intervals as in Figure 3B. Circles, DDGun; squares, ESM-2; triangles, RaSP; diamonds, ThermoMPNN. Contrasts for the four predictors of each family are shown in Supplementary Tables S6A and S6B.



Supplementary Figure S3. Performance of the four originally specified predictors with measured ΔΔG in PRIMARY_SV.

As in Supplementary Figure S2 for the 242 records of PRIMARY_SV (136 α-helical and 106 β-barrel). In B bacteriorhodopsin (n = 66), PagP (n = 59) and GlpG (n = 49). Circles, DDGun; squares, ESM-2; triangles, RaSP; diamonds, ThermoMPNN.



Supplementary Figure S4. Representations of BamA in the two different lipid environments, with RaSP included.

(A) As in Figure 4A: the difference ΔΔG_F (PC:PE) – ΔΔG_F (PC:PG) for the 88 alanine variants, in order of that difference, using the data of George et al. [15]. Bars show the difference plus or minus the

combined published uncertainty multiplied by a Student's t quantile at Welch-Satterthwaite degrees of freedom, giving nominal 95% coverage. They are pointwise and not adjusted for the 88 variants. Open circles indicate that the interval includes zero; filled diamonds indicate that the interval excludes zero. Values in open circles are from 87 variants including zero, filled diamonds are from 1 variant excluding zero. No interval lies entirely beyond ±1 kcal/mol (dashed lines; n = 0). (B) As in Figure 4B, for six methods: Spearman's ρ between each method and the measured ΔΔG_F for BamA in PC:PE, PC:PG and PC:PE - PC:PG. Values for RaSP come from the modified execution environment, as described in Supplementary Methods, Predictors. Bars show 95% confidence intervals based on 10,000 bootstrap replicates in which variants were resampled. Differences are taken at the published precision of 0.01 kcal/mol, so equal differences are ranked as ties. Circles, DDGun; squares, ESM-2; triangles, RaSP; diamonds, ThermoMPNN; plus signs, mCSM-membrane; crosses, MutTMB Stab-pred (MutTMB in the key).

Additional Reference for Supplementary

59. Gerasimavicius L, Livesey BJ, Marsh JA. Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. Protein Science. 2023 Jul;32(7):e4688. doi:10.1002/pro.4688

60. Diaz DJ, Gong C, Ouyang-Zhang J, Loy JM, Wells J, Yang D, et al. Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. Nat Commun. 2024 Jul 23;15(1):6170. doi:10.1038/s41467-024-49780-2

61. Dawson JE, Smith IN, Tushar AM, Eng C. Elucidating PTEN conformational dynamics and phosphatase regulation via integrative modeling and mutation prediction. Structure. 2025 Sep;33(9):1533-1546.e10. doi:10.1016/j.str.2025.06.002

62. Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin. 1946 Dec;2(6):110-114. doi:10.2307/3002019

63. Welch BL. The generalization of 'Student's' problem when several different population variances are involved. Biometrika. 1947;34(1–2):28–35. doi:10.1093/biomet/34.1-2.28