Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking ΔΔG Predictors on Membrane Proteins: Data Provenance, Held-Out Evaluation and Lipid Composition

Anusree Mulanthala Raj

sree.anu.mraj@gmail.com

Independent Researcher https://orcid.org/0009-0000-1713-5700

Research Article

Keywords: membrane proteins, protein stability, ΔΔG prediction, benchmarking, data provenance, beta-barrel, BamA, held-out evaluation, lipid composition

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11876747/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Benchmarking ΔΔG Predictors on Membrane Proteins: Data Provenance, Held-Out Evaluation and Lipid Composition

Anusree Mulanthala Raj
Independent Researcher, Kerala, India
Correspondence: sree.anu.mraj@gmail.com
ORCID: 0009-0000-1713-5700

Abstract

Predicting ΔΔG for membrane proteins is challenging with currently available computational methods because most of these predictors have been developed and tested against soluble proteins. This study traced back 379 records from the MutTMP Stab-pred dataset to verify the values reported against the original data and conditions of the experiments. Of these records, 271 were associated with 12 integral membrane proteins. After checking the sources, 25 records did not meet the inclusion criteria, which leaves 246 records from nine proteins. Of these, 142 depended on modelling assumptions reported in the original studies, mostly a single denaturant m value used for all mutants of a protein. DDGun, ESM-2 and ThermoMPNN were compared using the remaining 246 entries. For each of these three predictors, Spearman correlation coefficients of 0.49, 0.43 and 0.43 were obtained, respectively, and no statistically significant differences were found between the methods. Higher correlations were obtained for β-barrel proteins than for α-helical proteins, although the effects of protein fold and protein identity could not be separated. Pearson's correlation decreased from 0.80 under row-wise cross-validation to 0.60 when holding out entire proteins, when checked in a reconstructed MutTMB Stab-pred model. Similarly, when holding out individual structure files, excluding generated reverse mutations from the scoring decreased the correlation from 0.66 to 0.45. In the case of BamA, published measurements across two lipid environments indicated that 81 of the 88 alanine variants were less destabilised in PE-containing than in PG-containing membranes; however, none of the methods captured this effect because none considers lipid composition. Hence, these results show that how the experimental data are chosen and how predictors are tested can strongly change the reported performance for membrane proteins.

Introduction

Membrane proteins comprise a sizeable portion of encoded proteins. One early study estimated that 20–30% of all open reading frames in genomic sequences encode α-helical membrane proteins [1]. Since many of these membrane proteins are vital in cellular processes and important in drug target discovery [2,3], understanding their stability and function remains crucial. When a mutation alters the stability, it can promote misfolding of the protein and disrupt cellular processes [4]. However, predicting the effect of amino acid substitutions on membrane protein stability continues to be challenging. Traditionally, the methods used to predict changes in folding free energy (ΔΔG) have been developed and tested on soluble proteins since the data on membrane proteins remain limited [5,6]. Hence, their performance may not translate equally well to membrane proteins. In 2016, Kroncke et al. assessed the performance of ten such methods against experimental membrane-protein stability measurements [5]. The study found that all methods had correlation coefficients below 0.4.

Since the study by Kroncke et al. [5], many newer methods using different techniques have been developed to predict changes in protein stability. These include DDGun, RaSP, ThermoMPNN and ESM-2. DDGun is an untrained statistical method that uses evolutionary sequence-based information to predict ΔΔG without using experimental stability data to train the algorithm [7]. RaSP enlists a neural network to predict stability changes using structural representations and ΔΔG values calculated with Rosetta for training [8,9]. ThermoMPNN uses a structure-based neural network trained on stability measurements obtained experimentally to predict stability changes that occur upon mutation [10,11]. ESM-2 is a protein language model trained on amino acid sequences, which can be used to score mutations based on the probabilities assigned to wild-type and mutant amino acids without using any input of experimental values [12,13]. These different methods offer a variety of ways to assess the effect of mutation on a protein. Although these newer methods have broadened the range of approaches that can be used for stability prediction in proteins, their performance in predicting the changes to stability for membrane proteins has not been thoroughly established. All four were published after the comparison by Kroncke et al. [5].

Integral membrane proteins are embedded in the membrane, and hence their stability can be affected by the membrane environment. In experimental data, consideration should be given to the environment of these proteins, since the variation in the lipid or detergent environment in which membrane-protein stability is measured during experiments can influence the stability

changes associated with a mutation [6,14,15]. BamA illustrates this, as George et al. measured the effects of 88 alanine substitutions at lipid-facing residues of its β-barrel in vesicles composed of 95% phosphatidylcholine and 5% of either phosphatidylethanolamine (PE) or phosphatidylglycerol (PG) [15]. The results showed that the measured stability changes differed between the two lipid environments of PE and PG. Experimental data have also shown that the lipid composition alters the folding free energy of OmpA, through changes in bilayer forces such as hydrophobic mismatch and curvature stress [16]. The MPTherm database has assembled experimental thermodynamic measurements from the literature together with their experimental condition information, sequence, and structure [17]. Data from this database have been used to develop several membrane-protein stability models [18,19,20].

Membrane-specific methods which can predict changes in membrane protein stability upon mutations have also been developed. One such predictor is mCSM-membrane, which was trained on experimental mutation data to predict the effects of substitutions on transmembrane protein stability [21]. More recently, Reddy et al. developed MutTMP Stab-pred, introducing separate models for α-helical and β-barrel membrane proteins (MutTMH Stab-pred and MutTMB Stab-pred) using 379 experimentally measured mutations from the MPTherm database [22]. They reported that the hold-out correlations for the α-helical and β-barrel models were 0.70 and 0.92, respectively, while the mean absolute errors (MAEs) in leave-one-protein-out cross-validation were 1.36 and 1.31 kcal/mol, respectively. Their data and evaluation code are publicly available, which allowed the effect of different evaluation procedures on the reported results to be examined.

A single pooled correlation can be misleading when certain proteins are overrepresented in a dataset. Also, multiple PDB files may be associated with a given protein, while data from the experiments reported across publications may vary in both the thermodynamic measurement annotations and the approach taken to compute ΔΔG [23]. These factors can impact how comparable experimental measurements are and also affect how evaluations are interpreted based on held-out proteins or structures. The augmented MutTMP evaluation datasets also include synthetic reverse mutations, which are derived from experimental measurements [22]; although this is common practice, it can affect reported correlations [21,24]. This study uses the 379 published mutation records from MutTMP Stab-pred [22], re-examines the source data and protein assignments and compares DDGun, ESM-2, and ThermoMPNN using direct mutation measurements and assesses their performance by fold and protein rather than relying just on a single pooled estimate. Since the RaSP modified software workflow did not reproduce the

predictions against those archived by the RaSP authors [8] within a prespecified tolerance for the one protein compared, RaSP was evaluated separately as a secondary analysis. Additionally, the study also investigated how the different definitions of held-out data impact the evaluation of the model which was reconstructed from MutTMB Stab-pred.

These predictors do not include lipid composition as an input, and hence the same mutation receives a single prediction for a given sequence or structure input rather than separate predictions in two different lipid environments. George et al. showed that the differences in lipid environment can affect experimentally measured ΔΔG values [15]. Thus, to examine how predictor performance differs between experimental lipid environments, predictions provided by these three predictors were assessed against experimentally measured changes in BamA stability in PC:PE and PC:PG vesicles. The membrane-specific predictors mCSM-membrane and MutTMB Stab-pred were also assessed against experimental BamA stability data. These analyses assess the performance of the predictors, taking into account variations in dataset construction, evaluation procedures and experimental conditions. In addition to comparing newer predictors, this study explores the experimental sources and assumptions of stability measurements, including cases where the original source could not be identified. The study investigates how choices of holdout and scoring of reverse mutations affect performance estimates in a reconstructed model and how the predictions compare for the same BamA mutations measured in two different lipid environments.

Methods

Dataset

The files alpha.csv and beta.csv were downloaded from the MutTMP Stab-pred website [22] on 7 October 2026. Together, they contained 379 single-site substitutions with a measured ΔΔG, and according to the authors [22], these values were obtained from the MPTherm database [17]. The files contained 254 records from α-helical membrane proteins and 125 records from β-barrel membrane proteins. ΔΔG is defined as the unfolding free energy of the mutant minus the unfolding free energy of the wild type, as per MPTherm [17], so a negative value is destabilising (Supplementary Methods).

Cohorts

Of the 379 downloaded entries, 343 were retained, which were associated with structures classified as integral or conditionally inserting membrane proteins. Of these, 273 entries

associated with membrane-spanning mutations were selected. A further final reduction to 271 entries was made by retaining mutations from 12 integral membrane proteins. These 271 entries comprise the LEGACY cohort.

Each LEGACY record was then traced to its original source to compare the dataset value with the published measurement (see Supplementary Table S1). A primary publication was identified for 218 records [25-47], while 49 records could only be traced to the compilation by Kroncke et al. [5], including 12 of the OmpA records whose values also appear with an opposite sign convention in the study of Hong et al. [48], and the original source of the remaining four records could not be identified. Based on the information obtained from the source, records were classified into one of four classes.

1) INELIGIBLE_GENOTYPE (11 records): Records in this class represent measurements made on constructs that differ from the substitutions listed in the dataset.

2) INELIGIBLE_QUANTITY (14 records): Records in this class represent values which do not correspond to the target change in unfolding free energy of an integral membrane protein.

3) CONDITIONAL (142 records): Records in this class consist of values which depend on documented modelling assumptions. Of the 142 records in this class, 140 were calculated using the same value of denaturant m for each variant rather than fitting separately for each mutant.

4) ELIGIBLE (104 records): The remaining records were classified as ELIGIBLE. These are records where no qualifications were identified.

The 246 records classified as ELIGIBLE or CONDITIONAL together are referred to as the PRIMARY cohort. Of these 246 records, 138 are from α-helical proteins, and 108 from β-barrel proteins. Excluding the 4 records for which no original source was identified gives the PRIMARY_SV cohort of 242 records. Results are reported for the PRIMARY cohort. The results for PRIMARY_SV and LEGACY are used for comparison in sensitivity analyses.

Predictors

DDGun [7,49] was run in sequence mode using its standalone release (commit bcf112c), with sequence profiles generated using HHblits [50,51] against the uniclust30_2018_08 database [52]. ESM-2 [12] scores were calculated using the masked marginal method [13] with the model esm2_t33_650M_UR50D. For each mutation, a score was calculated by subtracting the log-probability of the wild-type amino acid from that of the mutant amino acid, and these ESM-2

scores were used only for rank statistics. ThermoMPNN [10] was run using commit 370f76ec with the default checkpoint and ProteinMPNN backbone weight version v_48_020 [53].

RaSP [8] was run using the authors' Colab notebook with modifications to the software environment (see Supplementary Methods). To test whether the archived predictions could be reproduced in this modified environment, PMP-22 was scored using the same AlphaFold model used by the RaSP authors. Only one of the ten scores agreed with the authors' previous predictions to within the tolerance of 0.05, which was set beforehand (Supplementary Table S11). RaSP was therefore moved from the primary comparison to a secondary analysis presented in the Supplementary information. This decision was taken after the four-predictor results were known, and the original four-predictor analysis is therefore reported as a sensitivity analysis. Both RaSP and ThermoMPNN used the structures associated with the dataset records, either PDB structures [54] or, in the case of PMP-22, an AlphaFold model [55,56]. Two LEGACY records have no such structure files, leaving 269 scored records. Because RaSP and ThermoMPNN prediction numbers assign positive values to destabilising mutations, their predictions were multiplied by −1 to fit with the experimental ΔΔG convention. Supplementary Table S10 provides further information on model versions, settings and sign conventions.

Statistics

Spearman's ρ was used to determine the correlation between predicted and measured ΔΔG values. It was calculated for the pooled dataset as well as separately for each structural fold. Ninety-five percent confidence intervals were estimated using 10,000 bootstrap replicates [57]. In each replicate, proteins were sampled with replacement, thus keeping all measurements from each sampled protein together. Comparisons of the predictors were also performed using the bootstrap method. Comparisons of β-barrel and α-helical proteins were also included in the analysis.

Bonferroni-adjusted confidence intervals were calculated for each of the three pairwise comparisons among DDGun [7,49], ESM-2 [12] and ThermoMPNN [10], as well as the three fold contrasts of β-barrel vs α-helix predictions, one for each predictor. These comparison groups were defined after the decision to include RaSP as a secondary analysis. Therefore, Bonferroni-adjusted intervals for the original four-predictor analysis, comprising six pairwise predictor comparisons and four between-fold comparisons, are also reported. These analyses used the nine proteins in the PRIMARY cohort, with confidence intervals calculated using protein-level

bootstrap resampling. Proteins were resampled within each fold for the comparison of β-barrel and α-helical proteins.

Held-out unit

The MutTMB Stab-pred model was refitted using its published features and model settings [22], while varying the definition of the held-out groups. The resulting performance estimates describe the reconstructed evaluation and are not attributed to the authors' archived fitted model. The input was the development table from the MutTMB Stab-pred reproducibility package [22], containing 214 rows: 107 measured mutations and their 107 generated reverse mutations. This represented five structure files, four UniProt proteins and three homology clusters. A gradient boosting regressor [58] was employed to build the model using the six features and model settings specified in the model package.

Five different holdout approaches were used to evaluate the model. In each of these analyses, either a single data row, a direct–reverse mutation pair, a structure file, a UniProt protein or a homology cluster was held out. The model was then refitted using the remaining records, and predictions were made for those held-out records. The held-out predictions from each of these analyses were pooled separately and evaluated using Pearson's r and MAE. Performance was evaluated across all 214 rows and then separately for the 107 experimentally measured mutations, using the metrics as in the original evaluation.

BamA

Experimental folding free-energy changes for 88 alanine variants of *Escherichia coli* BamA in PC:PE and PC:PG vesicles were obtained from Tables S8 and S9 of George et al. [15]. ΔΔG_F is calculated as ΔG_F of the wild type minus ΔG_F of the mutant, so a negative value indicates a destabilising mutation as per George et al. [15].

All 88 mutations were created on a BamA background devoid of cysteine residues and containing substitutions C690S and C700S. For structure-based predictions, the AlphaFold model of BamA (UniProt P0A940) [55,56] was modified by removing the signal peptide and introducing both serine substitutions. For predictions based on amino acid sequence, the complete cysteine-free BamA protein including the signal peptide was used.

The 88 mutations were analysed using DDGun [7,49], ESM-2 [12] and ThermoMPNN [10]. Membrane-specific predictions were obtained using mCSM-membrane [21] and MutTMB Stab-

pred [22]. mCSM-membrane [21] was run through its web server, while MutTMB Stab-pred [22] was run using its distributed standalone model. Predictions obtained using RaSP are included in the Supplementary Information. Since MutTMB Stab-pred was developed using the β-barrel subset of the 379-record dataset examined in this study, it was evaluated here only on BamA. mCSM-membrane [21] was developed using the compilation of Kroncke et al. [5], which overlaps with the 379-record dataset, so it was also evaluated only on BamA.

Each method produced one prediction per variant, which was correlated with the PC:PG values, the PC:PE values and the difference between them (PC:PE minus PC:PG). Spearman's rank correlation coefficients (ρ) were calculated across all variants. 95% confidence intervals were computed using 10,000 bootstrap replicates in which variants were resampled. The predictions for each variant are provided in Supplementary Table S8.

Results

Cohorts

Of the 379 entries present in the MutTMP Stab-pred files [22], 271, 246 and 242 were assigned to the LEGACY, PRIMARY and PRIMARY_SV cohorts, respectively (Figure 1, Supplementary Tables S1 and S2). PRIMARY consisted of 138 records from four α-helical proteins and 108 records from five β-barrel proteins (Supplementary Table S3). There was an uneven distribution of records across these nine proteins, with bacteriorhodopsin contributing 66 records, PagP contributing 61, and GlpG contributing 51, totalling 178 of the 246 records (72.4%). The remaining six proteins each contributed between 2 and 29 records.

Performance of three main predictors

In the PRIMARY cohort, Spearman's ρ between predicted and measured ΔΔG for DDGun [7,49], ESM-2 [12] and ThermoMPNN [10] was 0.49 (0.30 to 0.61), 0.43 (0.15 to 0.58) and 0.43 (0.31 to 0.59), respectively (Figure 3A, Supplementary Table S4). The MAE was calculated as 1.13 kcal/mol (0.97 to 1.32) for ThermoMPNN and 1.27 kcal/mol (1.08 to 1.44) for DDGun. No MAE was reported for ESM-2 because its scores are dimensionless and are not expressed in kcal/mol [13].

The three predictors performed similarly. The biggest difference in ρ was between DDGun [7,49] and ESM-2 (+0.062), and its 95% interval was from −0.098 to +0.296. For all three pairs, the

interval included zero, with or without Bonferroni adjustment (Supplementary Table S6A). So none of the three predictors performed statistically significantly better than the others.

β-barrel and α-helical proteins

For each of the three predictors, the ρ value was higher for β-barrel proteins compared to α-helical proteins, and the values were 0.61 and 0.31 for DDGun, 0.59 and 0.31 for ESM-2, and 0.57 and 0.33 for ThermoMPNN, respectively (Figure 3A). For DDGun, the difference in ρ (β-barrel minus α-helical) was +0.300, with intervals of +0.110 to +0.664 before Bonferroni correction and +0.067 to +0.691 after correction. For ThermoMPNN, the difference in ρ (β-barrel minus α-helical) was +0.240, with intervals of +0.007 to +0.601 before Bonferroni correction and −0.179 to +0.717 after Bonferroni correction. The difference for ESM-2 in ρ was +0.276, and both intervals included zero (uncorrected −0.033 to +0.726) (Supplementary Table S6B). Given that the PRIMARY cohort only contained four α-helical and five β-barrel proteins, the effect of structural fold could not be separated from protein identity. Therefore, these analyses were considered exploratory.

Individual proteins

Only bacteriorhodopsin, PagP and GlpG had at least 30 records; therefore, Spearman correlations were calculated for these proteins (Figure 3B, Supplementary Table S5). The Spearman correlations for DDGun were 0.51 for bacteriorhodopsin, 0.59 for PagP and 0.13 for GlpG. Values for ESM-2 were 0.44, 0.63 and 0.45 for bacteriorhodopsin, PagP, GlpG respectively and ThermoMPNN values were 0.38, 0.42 and 0.26, for bacteriorhodopsin, PagP, GlpG respectively. In the case of GlpG, DDGun had the lowest correlation among the three predictors, although its pooled point estimate in PRIMARY was the highest.

Leaving out one protein at a time altered the pooled Spearman correlation at most 0.07, 0.10 and 0.06 for DDGun, ESM-2 and ThermoMPNN respectively (Supplementary Table S4). For ESM-2, the largest changes occurred when PagP was removed, causing ρ to be lower by 0.10, while for DDGun, removing PagP also lowered ρ by 0.06. Removing GlpG showed the largest change for DDGun, which increased ρ by 0.07.

Sensitivity analyses

In PRIMARY_SV, ρ was 0.50 (0.31 to 0.62), 0.42 (0.14 to 0.57) and 0.42 (0.29 to 0.59) for DDGun, ESM-2 and ThermoMPNN, respectively (Supplementary Table S4, Supplementary

Figure S1). As with PRIMARY, the confidence intervals for all three pairwise predictor comparisons included zero (Supplementary Table S6A). The two datasets (PRIMARY and PRIMARY_SV) differ by just four records. As a result, the similar results across both of these cohorts were foreseen and thus cannot be treated as independent evidence.

For DDGun, in PRIMARY_SV, the intervals for the fold contrast did not include zero, either before Bonferroni adjustment (+0.117 to +0.664) or after adjustment (+0.071 to +0.691). For ESM-2, both intervals for the fold contrast included zero. For ThermoMPNN, the unadjusted interval included zero, with its lower limit close to zero (−0.003 to +0.600) and hence this contrast is therefore regarded as borderline (Supplementary Table S6B).

Of the 269 records assessed in LEGACY, ρ was 0.41, 0.39 and 0.41 for DDGun, ESM-2, and ThermoMPNN, respectively (Supplementary Table S4). These correlations were lower than those in the PRIMARY analysis. In LEGACY, which includes two other α-helical proteins, the interval for DDGun's fold contrast included zero (+0.29, −0.17 to +0.73; Supplementary Table S6B). As in the case of PRIMARY, the intervals for all three pairwise comparisons among the three predictors (DDGun, ESM-2 and ThermoMPNN) included zero (Supplementary Table S6A).

RaSP was run in a modified software environment in Colab notebook. In the recalculated scores only one of ten agreed with the authors' archived predictions within the prespecified tolerance for PMP-22, which is the one protein compared. Thus, it was included only as a secondary analysis, and ρ was 0.28 (0.21 to 0.38) in PRIMARY for RaSP (Supplementary Table S4). Including RaSP and applying the original adjustment for six pairwise predictor comparisons and four between-fold comparisons did not change the conclusions for the other three predictors, DDGun, ESM-2 and ThermoMPNN (Supplementary Table S6, Supplementary Figures S2 and S3).

Testing on held-out data in the MutTMB Stab-pred model

The performance of the refitted MutTMB Stab-pred model depended on which data were held out (Figure 2, Supplementary Table S7). When all 214 data rows were evaluated, Pearson's r was 0.80 when a single data row was held out and was also 0.80 when a direct–reverse mutation pair was held out. Pearson's r decreased to 0.66 when a structure file was held out and further to 0.60 when either a UniProt protein or a homology cluster was held out. Mean absolute error increased from 1.12 kcal/mol when a single data row was held out to 1.48, 1.64

and 1.67 kcal/mol when a structure file, UniProt protein and homology cluster were held out, respectively.

When predictions were evaluated only for the 107 experimentally measured mutations, Pearson's r decreased to 0.68 and 0.67 when single data rows and direct–reverse mutation pairs were held out, respectively. Pearson's r was further reduced to 0.45, 0.43 and 0.44 when a structure file, UniProt protein and homology cluster were held out, respectively. Thus, when a structure file was held out, Pearson's r was 0.66 for all 214 records and 0.45 when only the 107 experimentally measured mutations were evaluated.

BamA in two lipid environments

In the case of PC:PE vesicles, Spearman correlation coefficients between predicted and measured ΔΔG for the 88 BamA variants [15] were 0.60 (0.45 to 0.72), 0.57 (0.40 to 0.70) and 0.51 (0.32 to 0.66) for ThermoMPNN, DDGun, and MutTMB Stab-pred, respectively (Figure 4B, Supplementary Table S9). The correlation for mCSM-membrane [21] was 0.15 (−0.06 to 0.35), while that for ESM-2 was 0.10 (−0.12 to 0.31).

In the case of PC:PG vesicles, the correlation coefficients were 0.49 (0.31 to 0.63), 0.48 (0.29 to 0.64) and 0.46 (0.27 to 0.61) for DDGun, ThermoMPNN and MutTMB Stab-pred, respectively. The ρ for mCSM-membrane [21] was 0.22 (0.004 to 0.43), while that for ESM-2 was 0.19 (−0.03 to 0.38).

The folding free energy of wild-type BamA was −13.04 ± 0.34 kcal/mol in PC:PE vesicles and −13.12 ± 0.61 kcal/mol in PC:PG vesicles [15]. So the wild-type value was only 0.08 kcal/mol higher in PC:PE than in PC:PG. Reanalysis of the published BamA measurements [15] showed that 81 of 88 variants showed a positive difference in ΔΔG_F between the two lipid environments (PC:PE minus PC:PG) with a mean difference of +0.82 kcal/mol (Figure 4A) [15]. The correlation of each method's prediction with the measured PC:PE minus PC:PG difference ranged from mCSM-membrane showing lowest correlation with the measured value of −0.15 to ThermoMPNN showing 0.08. Confidence intervals for all predictions included zero (Figure 4B, Supplementary Table S9). Since none of the methods takes lipid composition into consideration, each variant got a single prediction, which was compared to PC:PE values, PC:PG values and their difference. RaSP results for BamA, from the modified execution, are given in Supplementary Table S9 and Supplementary Figure S4.

Discussion

Experimental measurements of the stability of membrane proteins are limited [5,6]. As a consequence, the same experimental measurements from the same databases and source studies are reused across different predictors, and benchmarking studies [5,17-22]. This study uses the MutTMP Stab-pred dataset [22], which was developed from MPTherm [17] and compiles experimental stability data from different laboratories. Of the records evaluated, 271 mutations in 12 proteins were found to meet the initial structural criteria. The PRIMARY cohort consisted of 246 records from nine proteins, three of which accounted for 178 (72%). The number of proteins and studies from which these measurements are derived is very limited and thus, it is crucial to evaluate the effect of experimental and benchmarking assumptions on the conclusions drawn from them. Any new predictors or models developed using these datasets will also be influenced by their experimental sources and the types of proteins represented, while evaluation choices will dictate how well they appear to perform. The results of this study extend beyond the data set analysed here. They illustrate three points that should be taken into account when benchmarking membrane protein stability predictors. First, what do these experimental values represent? Second, how is predictive performance evaluated? Third, what are the conditions under which protein stability was measured?

Although DDGun [7,49], ESM-2 [12,13] and ThermoMPNN [10,11] employ different methods, their pooled Spearman correlations with respect to the experimental values of ΔΔG were in the range of 0.43 to 0.49 (Figure 3). No statistically significant differences were observed among the three predictors DDGun [7,49], ESM-2 [12,13] and ThermoMPNN. In contrast to DDGun and ThermoMPNN, ESM-2 provides dimensionless values. Thus, Spearman's rank correlation was used to evaluate the performance of ESM-2, and mean absolute error in kcal/mol was not calculated. The limited number of proteins in these datasets limits the comparison of predictors. The largest pairwise difference was only +0.062, while the confidence interval for this value was between −0.098 to +0.296. Across individual proteins, pooled correlations also hid any difference in performance. For example, in the case of GlpG, DDGun had the lowest correlation even though it had the highest pooled correlation. Overall, the three predictors performed better when predicting β-barrels than when predicting α-helical membrane proteins. Only for DDGun was this difference statistically significant after application of the Bonferroni correction. PagP accounted for 61 of the 108 β-barrel records, and removing it from the data set decreased the pooled Spearman correlation of ESM-2 by 0.10. With four α-helical and five β-barrel proteins in PRIMARY, the fold cannot be dissected from the protein identity. In LEGACY, which includes two

additional α-helical proteins, the interval for DDGun also includes zero and therefore also remains a hypothesis to be tested on more proteins.

Not all recorded ΔΔG represent the same quantity. Tracing these records to the sources identified 11 records with experimental constructs that did not align with named substitutions, and 14 records that did not represent the unfolding free energy of integral membrane proteins. A further 142 records were dependent on modelling assumptions that were documented. Of these, 140 records assumed a constant denaturant m value for all variants of the protein, rather than assigning an individual m value for each variant. Further, 49 records could be traced back only to the compilation of Kroncke et al. [5,48]. Since the conditional records comprise more than half of PRIMARY, the use of a common denaturant m value for all variants of a protein should be interpreted with caution. The experimental measurements also varied in the denaturants, membrane mimetics and analysis methods used. Kroncke et al. [5] also pointed out that an agreement between the stabilities of a single membrane protein measured using different methods had yet to be extensively tested. Correlations in LEGACY were generally lower than in PRIMARY. Additionally, LEGACY contained different proteins than those in PRIMARY; thus, decreased correlations in LEGACY may not be attributable to either the inclusion criteria or the record quality of the datasets. The record-level annotations make these decisions transparent and also allow others to apply different rules for testing. Hence, the main question when evaluating these predicted values is whether the records are comparable for the question being asked rather than whether the original measurements or their compilation itself are valid by themselves.

The performance of the models is also dependent upon the method of evaluation. Previously, Reddy et al. [22] had already performed protein- and homology-cluster-level evaluation on their models. Here, reconstructed the MutTMB Stab-pred model was used to study the effect of holdout sets and scoring sets, and the development subset consisted of 107 measured and 107 reverse mutations. For the 214 rows, the Pearson correlation was 0.80 when a row or direct–reverse pair was held out. When the entire structure file was held out the correlation lowered to 0.66 and when a UniProt protein or a homology cluster held out, it was 0.60 (Figure 2). Holding out a structure file instead of a UniProt protein, where four proteins are represented by five files, raised r from 0.60 to 0.66. Scoring generated reverse mutations with the measured ones raised the values even further. At the structure-file level, r was 0.66 and 0.45 on all 214 rows and 107 measured mutations respectively, but the MAE changed only from 1.48 to 1.49 kcal/mol. Reverse mutations are used in training datasets to enforce antisymmetry [24]. In these cases,

the ΔΔG value is simply assigned the negative value of the experimentally measured ΔΔG value. When evaluating the performance of the model, reverse mutations in the test set increase the spread of the ΔΔG values and raise Pearson's correlation coefficient [21,24]. In the present study the effect was much greater on Pearson's correlation coefficient than mean absolute error (MAE). Because of this, the real concern is how predictive performance is evaluated, and not the use of reverse mutations itself during training. A similar pattern was observed in the archived predictions included with the package of MutTMB Stab-pred, although these predictions could not be reproduced exactly using the distributed code (Supplementary Methods). For any model trained on augmented data or on several structures per protein, performance on measured mutations with proteins or homology clusters held out is the most reliable way to understand how it will perform on a new protein.

Measurements of protein stability are also dependent on the lipid environment of the membrane. Studies have shown that protein stability can be changed by the lipid composition [14,16]. In the case of BamA measurements [15] in two different lipid environments, the difference in mutation effects had a positive point estimate for 81 of 88 variants, with a mean of +0.82 kcal/mol. The wild-type folding free energies differed by only +0.08 kcal/mol. ThermoMPNN, DDGun and MutTMB Stab-pred agreed moderately with mutation effects in each environment (ρ 0.46 to 0.60), while ESM-2 and mCSM-membrane [21] had lower agreement. None of these five predictors tracked the difference between the two environments, and all five intervals included zero. This is expected because none of these predictors takes lipid compositions as input, so each variant receives the exact same score under both conditions. This makes it impossible to directly predict differences between them. Thus, these rankings within these membranes for each of the predictors do not prove that they capture lipid-dependent effects. These results point to the need for an experimental dataset that records membrane composition and predictors that can use this information and take membrane-specific information as input.

This study has several limitations. PRIMARY consists of 246 records across nine proteins, with three of them providing most of the measurements. Consequently, the protein cluster bootstrap intervals reflect variability among these nine proteins. Therefore, they cannot be used to generalise the results to membrane protein as a whole. Relying on a single published source also means that other datasets might differ in experimental conditions and protein composition included in them. Classifications of ineligibility records and source tracing were based on a single researcher's manual review of the publications, without any other independent verification. However, the reasons for these decisions are given in Supplementary Table S1.

More than half of the records in PRIMARY were conditional records, mostly because of shared values of denaturant m, and the results thus remain dependent on these assumptions. RaSP [8] was run in a modified environment. Only one prediction agreed with the archived predictions from the authors within the prespecified tolerance for the ten PMP-22 substitutions that were examined. As a result, the RaSP results here are therefore not a direct test of the published version. Membrane-specific methods were compared for only one set of experimental data from BamA, alanine substitutions and two lipid compositions.

Overall, these results suggest that the reported performance of a membrane-protein stability predictor is dependent on what is considered an experimental measurement, how the evaluations are performed, and what membrane conditions are included. These questions are not just specific to the present dataset, but also need to be evaluated when constructing future datasets for predictors. In the future, any benchmark including membrane proteins should describe in more detail the assay and assumptions used in the experimental methods, and predictor performance should be evaluated using experimentally measured mutations and protein-level holdouts. Benchmarks should also include datasets that have some description of lipid composition or conditions that were used in measuring them in experiments. With this context provided in detail for each value, results from different studies can be compared, and it would make it easier to see where membrane-aware predictors still fall short.

Data and code availability

Data and code supporting the results from this study are available at https://github.com/sreeanumraj-mulanthala/membrane-protein-stability-predictors. The repository includes the annotation table of MutTMP Stab-pred records (Supplementary Table S1), predictions of 271 LEGACY records, prepared structure for BamA and the analysis scripts, and predictions for variants of BamA. Measurements of BamA are provided in Supplementary Tables S8 and S9 of George et al. [15]. The measured ΔΔG values for MutTMP Stab-pred can be found on their website [22].

Competing interests

The author has no competing interests.

Funding

No funding was received for this work.

Author contributions

A.M.R. designed the study, performed the analyses, created the figures and wrote the text for this manuscript.

Acknowledgement

The developers of MutTMP Stab-pred are thanked for providing the complete dataset with a complete reproducibility package, which made the held-out analysis and BamA predictions possible for this study. The developers of MutTMP Stab-pred, MPTherm, DDGun, ESM-2, ThermoMPNN, RaSP and mCSM-membrane are thanked for openly sharing their code and servers.

Disclosure of assistance

Claude Opus 5.5 (Anthropic, accessed October and November 2026) was used to assist with code development, debugging and refinement of analysis scripts. This code was checked by the author and was run on her own computer. During the preparation of the manuscript, Grammarly was used to improve clarity of language. The author has checked the manuscript after the edit and takes full responsibility for the contents.

References

1. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Science. 1998 Apr;7(4):1029–38. doi:10.1002/pro.5560070420

2. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006 Dec;5(12):993–6. doi:10.1038/nrd2199

3. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017 Jan;16(1):19–34. doi:10.1038/nrd.2016.230

4. Marinko JT, Huang H, Penn WD, Capra JA, Schlebach JP, Sanders CR. Folding and Misfolding of Human Membrane Proteins in Health and Disease: From Single Molecules to Cellular Proteostasis. Chemical Reviews. 2019 May 8;119(9):5537–606. doi:10.1021/acs.chemrev.8b00532

5. Kroncke BM, Duran AM, Mendenhall JL, Meiler J, Blume JD, Sanders CR. Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. Biochemistry. 2016 Sep 13;55(36):5002–9. doi:10.1021/acs.biochem.6b00537

6. Fleming KG. Energetics of Membrane Protein Folding. Annu Rev Biophys. 2014 May 6;43(1):233–55. doi:10.1146/annurev-biophys-051013-022926

7. Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinformatics. 2019 Jul;20(S14):335. doi:10.1186/s12859-019-2923-1

8. Blaabjerg LM, Kassem MM, Good LL, Jonsson N, Cagiada M, Johansson KE, et al. Rapid protein stability prediction using deep learning representations. eLife. 2023 May 15;12:e82593. doi:10.7554/eLife.82593

9. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput. 2016 Dec 13;12(12):6201–12. doi:10.1021/acs.jctc.6b00819

10. Dieckhaus H, Brocidiacono M, Randolph NZ, Kuhlman B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. Proc Natl Acad Sci USA. 2024 Feb 6;121(6):e2314853121. doi:10.1073/pnas.2314853121

11. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, et al. Mega-scale experimental analysis of protein folding stability in biology and design. Nature. 2023 Aug 10;620(7973):434–44. doi:10.1038/s41586-023-06328-6

12. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123–30. doi:10.1126/science.ade2574

13. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. 2021. doi:10.1101/2021.07.09.450648

14. Sanders MR, Findlay HE, Booth PJ. Lipid bilayer composition modulates the unfolding free energy of a knotted α-helical membrane protein. Proc Natl Acad Sci USA. 2018 Feb 20;115(8): E1799–E1808. doi:10.1073/pnas.1714668115

15. George A, Raj AM, Patil AG, Kumari V, Mahalakshmi R. Lipid-regulated assembly mechanisms and functional energetics of the essential bacterial chaperone BamA. Chem Sci. 2026;17(6):3103–21. doi:10.1039/D5SC07027A

16. Hong H, Tamm LK. Elastic coupling of integral membrane protein stability to lipid bilayer forces. Proc Natl Acad Sci USA. 2004 Mar 23;101(12):4065–70. doi:10.1073/pnas.0400358101

17. Kulandaisamy A, Sakthivel R, Gromiha MM. MPTherm: database for membrane protein thermodynamics for understanding folding and stability. Briefings in Bioinformatics. 2021 Mar 22;22(2):2119–25. doi:10.1093/bib/bbaa064

18. Kulandaisamy A, Zaucha J, Frishman D, Gromiha MM. MPTherm-pred: Analysis and Prediction of Thermal Stability Changes upon Mutations in Transmembrane Proteins. Journal of Molecular Biology. 2021 May;433(11):166646. doi:10.1016/j.jmb.2020.09.005

19. Reddy PR, Kulandaisamy A, Gromiha MM. TMH Stab-pred: Predicting the stability of α-helical membrane proteins using sequence and structural features. Methods. 2023 Oct;218:118–24. doi:10.1016/j.ymeth.2023.08.005

20. Reddy PR, Kulandaisamy A, Gromiha MM. TMB Stab-pred: Predicting the stability of transmembrane β-barrel proteins using their sequence and structural signatures. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2025 Jul;1873(4):141070. doi:10.1016/j.bbapap.2025.141070

21. Pires DEV, Rodrigues CHM, Ascher DB. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. Nucleic Acids Research. 2020 Jul 2;48(W1):W147–53. doi:10.1093/nar/gkaa416

22. Reddy PR, Kulandaisamy A, Gromiha MM. MutTMP Stab-pred: Predicting Thermodynamic Stability Changes Upon Mutations in α-helical and β-barrel Membrane Proteins. Journal of Molecular Biology. 2026 Oct;170052. doi:10.1016/j.jmb.2026.170052

23. Pancotti C, Benevenuta S, Birolo G, Alberini V, Repetto V, Sanavia T, et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Briefings in Bioinformatics. 2022 Mar 10;23(2):bbab555. doi:10.1093/bib/bbab555

24. Pucci F, Bernaerts KV, Kwasigroch JM, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics. 2018 Nov 1;34(21):3659–65. doi:10.1093/bioinformatics/bty348

25. Paslawski W, Lillelund OK, Kristensen JV, Schafer NP, Baker RP, Urban S, et al. Cooperative folding of a polytopic α-helical membrane protein involves a compact N-terminal nucleus and nonnative loops. Proc Natl Acad Sci USA. 2015 Jun 30;112(26):7978–83. doi:10.1073/pnas.1424751112

26. Moon CP, Fleming KG. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. Proc Natl Acad Sci USA. 2011 Jun 21;108(25):10174–7. doi:10.1073/pnas.1103979108

27. Iyer BR, Mahalakshmi R. Hydrophobic Characteristic Is Energetically Preferred for Cysteine in a Model Membrane Protein. Biophysical Journal. 2019 Jul;117(1):25–35. doi:10.1016/j.bpj.2019.05.024

28. Marx DC, Fleming KG. Influence of Protein Scaffold on Side-Chain Transfer Free Energies. Biophysical Journal. 2017 Aug;113(3):597–604. doi:10.1016/j.bpj.2017.06.032

29. Huysmans GHM, Baldwin SA, Brockwell DJ, Radford SE. The transition state for folding of an outer membrane protein. Proc Natl Acad Sci USA. 2010 Mar 2;107(9):4099–104. doi:10.1073/pnas.0911904107

30. Fleming KG, Engelman DM. Specificity in transmembrane helix–helix interactions can define a hierarchy of stability for sequence variants. Proc Natl Acad Sci USA. 2001 Dec 4;98(25):14340–4. doi:10.1073/pnas.251367498

31. Curnow P, Di Bartolo ND, Moreton KM, Ajoje OO, Saggese NP, Booth PJ. Stable folding core in the folding transition state of an α-helical integral membrane protein. Proc Natl Acad Sci USA. 2011 Aug 23;108(34):14133–8. doi:10.1073/pnas.1012594108

32. Otzen DE. Mapping the folding pathway of the transmembrane protein DsbB by protein engineering. Protein Engineering Design and Selection. 2011 Jan 1;24(1–2):139–49. doi:10.1093/protein/gzq079

33. Schlebach JP, Narayan M, Alford C, Mittendorf KF, Carter BD, Li J, et al. Conformational Stability and Pathogenic Misfolding of the Integral Membrane Protein PMP22. Journal of the American Chemical Society. 2015 Jul 15;137(27):8758–68. doi:10.1021/jacs.5b03743

34. Harris NJ, Findlay HE, Simms J, Liu X, Booth PJ. Relative Domain Folding and Stability of a Membrane Transport Protein. Journal of Molecular Biology. 2014 Apr;426(8):1812–25. doi:10.1016/j.jmb.2014.01.012

35. Joh NH, Oberai A, Yang D, Whitelegge JP, Bowie JU. Similar Energetic Contributions of Packing in the Core of Membrane and Water-Soluble Proteins. J Am Chem Soc. 2009 Aug 12;131(31):10846–7. doi:10.1021/ja904711k

36. Gupta A, Mahalakshmi R. Reversible folding energetics of Yersinia Ail barrel reveals a hyperfluorescent intermediate. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2020 Feb;1862(2):183097. doi:10.1016/j.bbamem.2019.183097

37. Sanchez KM, Gable JE, Schlamadinger DE, Kim JE. Effects of Tryptophan Microenvironment, Soluble Domain, and Vesicle Size on the Thermodynamics of Membrane Protein Folding: Lessons from the Transmembrane Protein OmpA. Biochemistry. 2008 Dec 2;47(48):12844–52. doi:10.1021/bi800860k

38. Yohannan S, Faham S, Yang D, Grosfeld D, Chamberlain AK, Bowie JU. A C_α−H···O Hydrogen Bond in a Membrane Protein Is Not Stabilizing. J Am Chem Soc. 2004 Mar 3;126(8):2284–5. doi:10.1021/ja0317574

39. Baker RP, Urban S. Architectural and thermodynamic principles underlying intramembrane protease function. Nat Chem Biol. 2012 Sep;8(9):759–68. doi:10.1038/nchembio.1021

40. Chaturvedi D, Mahalakshmi R. Position-Specific contribution of interface tryptophans on membrane protein energetics. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2018 Feb;1860(2):451–7. doi:10.1016/j.bbamem.2017.11.003

41. Figueira-Mansur J, Aguilera EA, Stoque RM, Ventura GT, Mohana-Borges R. Mutations in the dimer interfaces of the dengue virus capsid protein affect structural stability and impair RNA-capsid interaction. Sci Rep. 2019 Feb 26;9(1):2829. doi:10.1038/s41598-019-39185-3

42. Joh NH, Min A, Faham S, Whitelegge JP, Yang D, Woods VL, et al. Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. Nature. 2008 Jun;453(7199):1266–70. doi:10.1038/nature06977

43. Yohannan S, Yang D, Faham S, Boulting G, Whitelegge J, Bowie JU. Proline Substitutions are not Easily Accommodated in a Membrane Protein. Journal of Molecular Biology. 2004 Jul;341(1):1–6. doi:10.1016/j.jmb.2004.06.025

44. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA. 2004 Jan 27;101(4):959–63. doi:10.1073/pnas.0306077101

45. Curnow P, Booth PJ. The transition state for integral membrane protein folding. Proc Natl Acad Sci USA. 2009 Jan 20;106(3):773–8. doi:10.1073/pnas.0806953106

46. Sanchez KM, Kang G, Wu B, Kim JE. Tryptophan-Lipid Interactions in Membrane Protein Folding Probed by Ultraviolet Resonance Raman and Fluorescence Spectroscopy. Biophysical Journal. 2011 May;100(9):2121–30. doi:10.1016/j.bpj.2011.03.018

47. Schlebach JP, Woodall NB, Bowie JU, Park C. Bacteriorhodopsin Folds through a Poorly Organized Transition State. Journal of the American Chemical Society. 2014 Nov 26;136(47):16574–81. doi:10.1021/ja508359n

48. Hong H, Park S, Flores Jiménez RH, Rinehart D, Tamm LK. Role of Aromatic Side Chains in the Folding and Thermodynamic Stability of Integral Membrane Proteins. J Am Chem Soc. 2007 Jul 1;129(26):8320–7. doi:10.1021/ja068849o

49. Montanucci L, Capriotti E, Birolo G, Benevenuta S, Pancotti C, Lal D, et al. DDGun: an untrained predictor of protein stability changes upon amino acid variants. Nucleic Acids Research. 2022 Jul 5;50(W1):W222–7. doi:10.1093/nar/gkac325

50. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012 Feb;9(2):173–5. doi:10.1038/nmeth.1818

51. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019 Dec;20(1):473. doi:10.1186/s12859-019-3019-7

52. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017 Jan 4;45(D1):D170–6. doi:10.1093/nar/gkw1081

53. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science. 2022 Oct 7;378(6615):49–56. doi:10.1126/science.add2187

54. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 Jan 1;28(1):235–42. doi:10.1093/nar/28.1.235

55. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9. doi:10.1038/s41586-021-03819-2

56. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2022 Jan 7;50(D1):D439–44. doi:10.1093/nar/gkab1061

57. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Statist. 1979 Jan 1;7(1):1–26. doi:10.1214/aos/1176344552

58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.

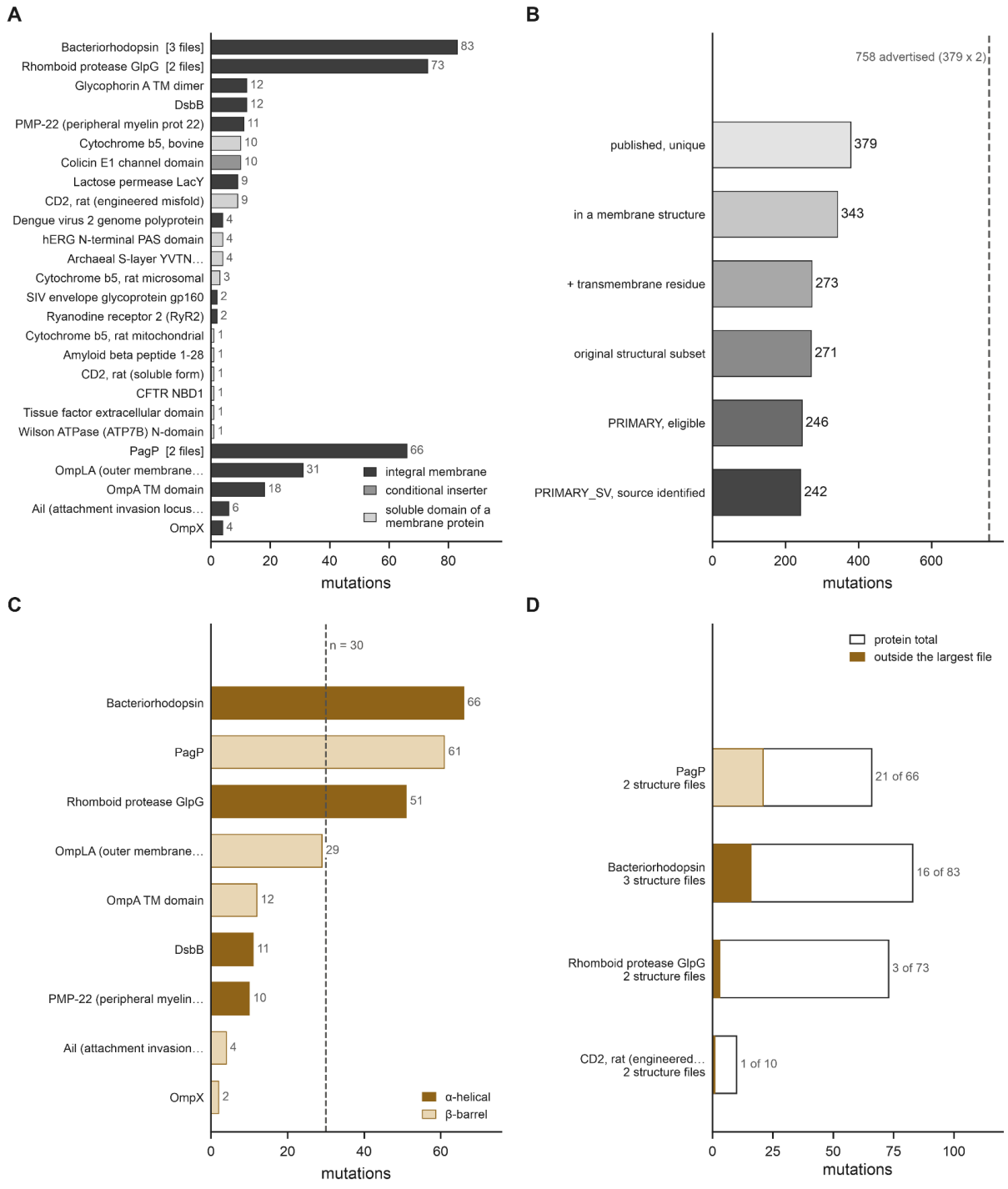


Figure 1. The MutTMP Stab-pred records and how the cohorts were selected from them.

(A) Number of mutations per protein for the 379 published unique records after the 30 structure records were resolved to 25 proteins. Bars are shaded to represent protein association with the membrane, including integral membrane proteins, conditionally inserted proteins, and soluble domains of membrane

proteins. The two entries of CD2 (an engineered misfolded and soluble form) represent the same protein (UniProt P08921). Thus, the 26 bars in the figure represent 25 different proteins. Long protein names are shortened here and given in full in Supplementary Table S1. (B) How many records are left after each step: 379 published unique records; 343 on the structure of an integral or conditionally inserting membrane protein; 273 where the mutated residue is marked as membrane-spanning; 271 in integral membrane proteins (LEGACY, labelled "original structural subset"); 246 in PRIMARY (ELIGIBLE and CONDITIONAL records, nine proteins); and 242 in PRIMARY_SV (records with an identified source). The dashed line marks the 758 entries listed on the download page. Each mutation is counted as both direct and reverse, and thus listed twice. The line is for reference and was not used in selection. (C) Mutations per protein in PRIMARY, shaded by fold (four α-helical and five β-barrel proteins). Bacteriorhodopsin, PagP and GlpG contribute 178 of the 246 records (72%) and are the only proteins to reach 30 records (dashed line), the threshold for the per-protein correlations in Figure 3B. (D) Proteins which are listed in more than one structure across the 379 records are represented. Open bars show all mutations of the protein; filled bars show those outside its largest structure file. These proteins account for 232 of the 379 records (61%). If the largest structures of the proteins are excluded, 41 of the mutations remain in the protein's other files.

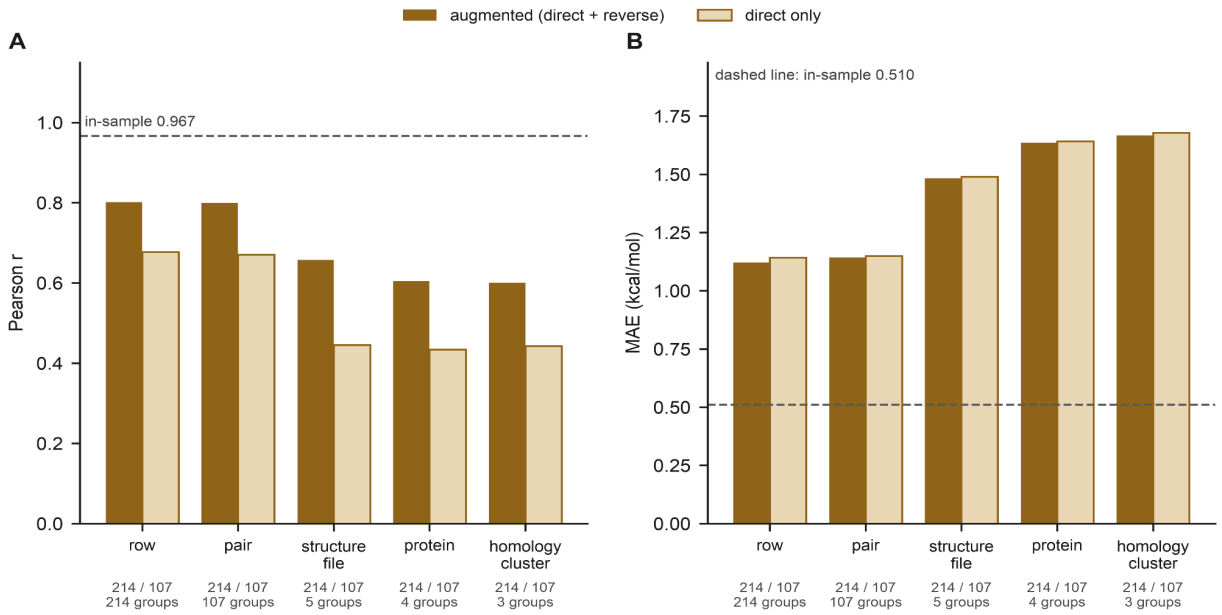


Figure 2. The estimated performance of the refitted MutTMB Stab-pred model depends on what is held out and what is scored.

The gradient boosting model was refitted according to the published feature set and model settings to the 214-row development table of the MutTMB Stab-pred reproducibility package [22]. The development set consisted of 107 mutations and their reverse mutations derived from five protein structure files, four entries in UniProt and three homology clusters. Predictions were made for one row at a time, with one direct-reverse pair, one structure file, one UniProt entry or one homology cluster held out. Pearson correlation (r) and mean absolute error (MAE) between the predicted and measured ΔΔG are shown in (A) and (B), respectively. Dark bars represent estimates for the 214 rows (measured and reverse mutations). Light bars represent estimates for the 107 measured mutations. The numbers below the bars represent the rows scored and the number of groups held out. Dashed lines show correlation (r = 0.967) and mean absolute error (0.510 kcal/mol) for the 214 rows, which is the performance of the model when it was fitted and evaluated on the same rows. These are for reference and should not be interpreted as held-out estimates. Holding out a PDB entry partitions the data in the same way as holding out a structure file, and is not shown. No confidence intervals are shown, and differences in r < 0.01 are not interpreted (Supplementary Methods). The estimates are for the refitted model and should not be attributed to the authors' previously fitted models (Supplementary Methods).

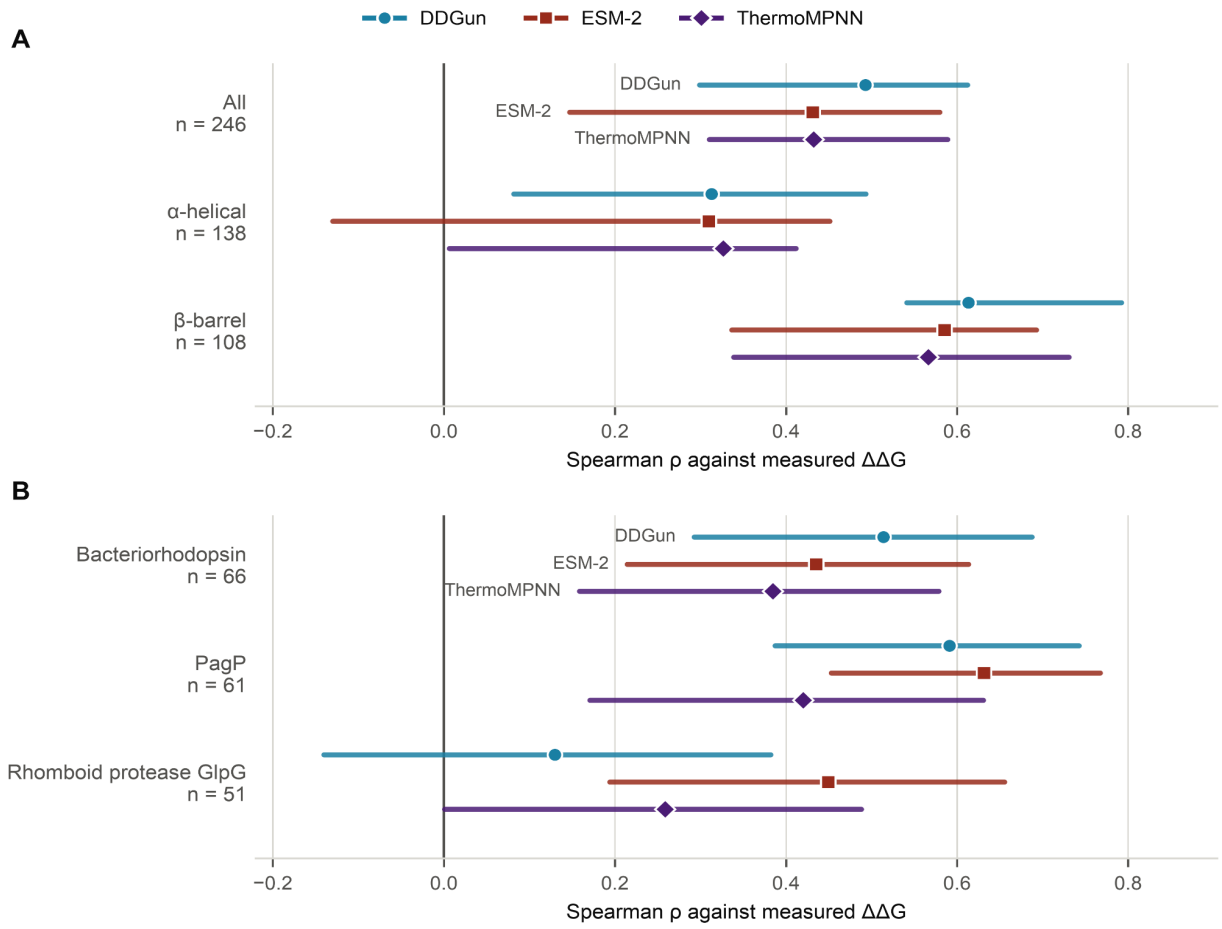


Figure 3. Comparison of DDGun, ESM-2 and ThermoMPNN with Measured ΔΔG in the PRIMARY Cohort.

(A) Spearman's ρ for predicted and measured ΔΔG for 246 records. Spearman's ρ for α-helical (138) and β-barrel (108) records are shown separately. Point estimates are represented by points, and bars are 95% percentile intervals from 10,000 bootstrap replicates in which proteins were resampled with replacement, and the records of each protein were kept together. For the α-helical and β-barrel records, only records for proteins of that fold were resampled. The intervals are not adjusted for multiple comparisons. See Supplementary Table S6 for adjusted intervals for the differences between predictors and between folds. PRIMARY consists of nine proteins (4 α-helical and 5 β-barrel), accounting for the width of the intervals. (B) Spearman's ρ within each protein with at least 30 records: bacteriorhodopsin (n = 66), PagP (n = 61) and GlpG (n = 51). Bars show 95% confidence intervals from record resampling within the proteins. Substitutions at the same site are correlated; for example, there are 19 records for PagP sharing the same site, confidence intervals are narrower than a resampling of sites would give and are not comparable with those in A. Predictors are shown as circles (DDGun), squares (ESM-2) and

diamonds (ThermoMPNN). The same order is used for each group. The analysis for PRIMARY_SV is shown in Supplementary Figure S1 and with RaSP included in Supplementary Figures S2 and S3.

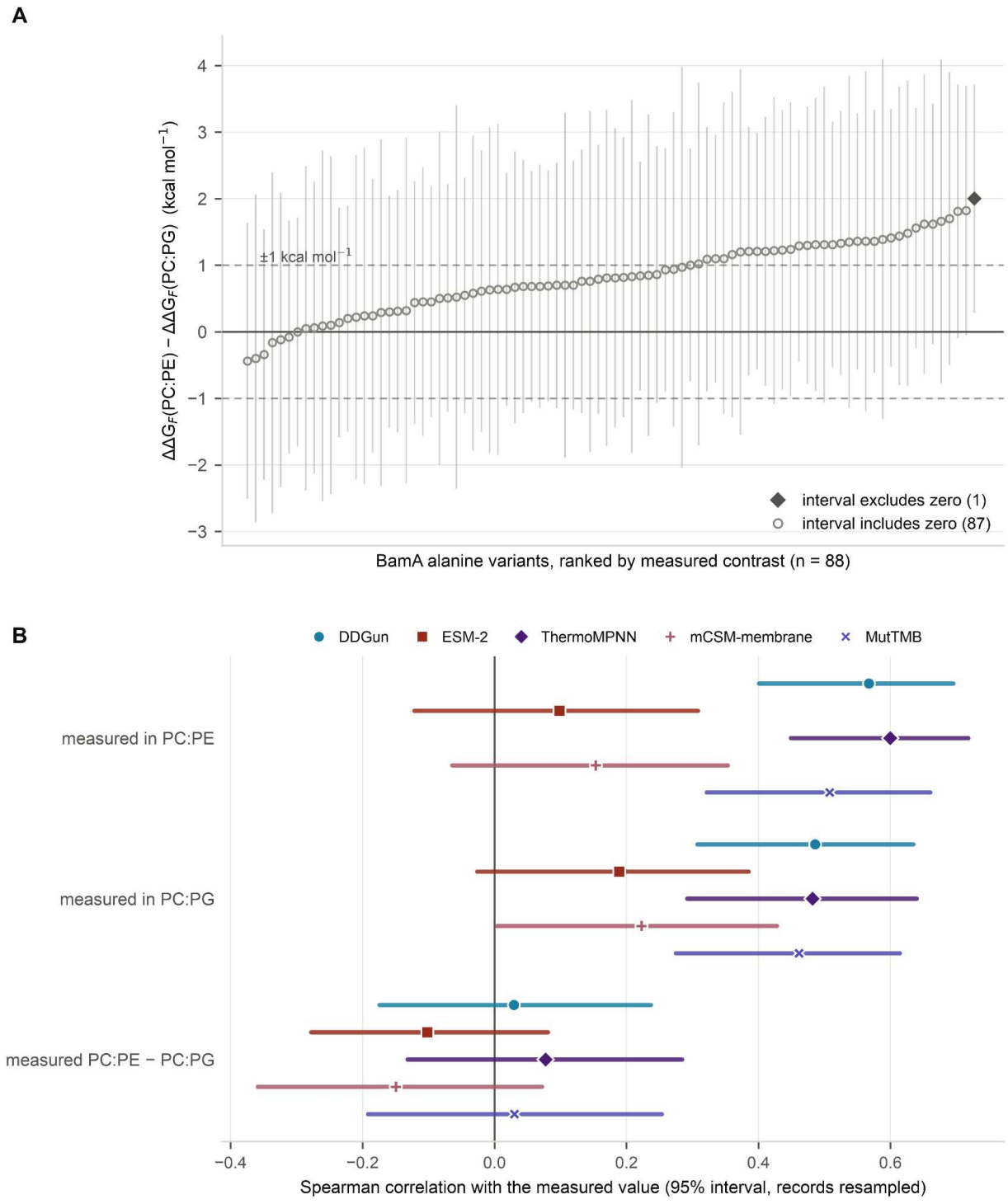


Figure 4. Differences between lipid environments and agreement between five methods for BamA.

(A) Difference in folding free energy change, ΔΔG_F (PC:PE)- ΔΔG_F (PC:PG) for the 88 alanine-substituted lipid-facing residues in the transmembrane domain of BamA, ranked by that difference. PC:PE and PC:PG are large unilamellar vesicles of 95% DMPC with 5% DMPE or 5% DMPG (20 mM borate, pH 10, lipid-to-protein ratio of 1042:1, 25 °C) and these values are from George et al. [15]. Bars show the difference plus or minus the two published uncertainties combined. Each is scaled by a Student t value at Welch-Satterthwaite degrees of freedom, giving nominal 95% coverage. They are pointwise and not adjusted for the 88 variants. Intervals including zero are shown in open circles, and those excluding zero are shown by filled diamonds. No interval extends entirely beyond ±1 kcal/mol (represented by dashed lines, n = 0). The difference was positive for 81 of the 88 variants with an average value of +0.82 kcal/mol [15]. In comparison, the difference in the wild-type free energy of folding was 0.08 kcal/mol. No interval was assigned to the mean of the 88 variants, because they all share the same wild-type measurements. The number of variants for which the interval does not include zero is dependent on how the published uncertainties are combined (supplementary methods). (B) Spearman's ρ between each of the five methods and the measured ΔΔG_F for PC:PE, PC:PG and PC:PE - PC:PG. The three main predictors are DDGun, ESM-2 and ThermoMPNN, and mCSM-membrane [21] and MutTMB Stab-pred [22] are membrane-specific predictors. Neither mCSM-membrane [21] nor MutTMB Stab-pred [22] were trained on BamA. Each method makes one prediction per variant, and these predictions are used in all three analyses. Bars represent 95% confidence intervals from 10,000 bootstrap replicates in which variants were resampled. Given that BamA is a single protein, these confidence intervals reflect uncertainty for BamA only and are therefore narrower compared to the protein-clustered intervals shown in Figure 3. Differences are taken at the published precision of 0.01 kcal/mol, so equal differences are ranked as ties. Predictions for RaSP are shown in Supplementary Figure S4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- SupplementaryTableS1.csv
- SupplementaryTableS8.csv