Supplementary Data

EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models

Lukas Folkman, Bela Stantic, Abdul Sattar, Yaoqi Zhou*

Supplementary Figures

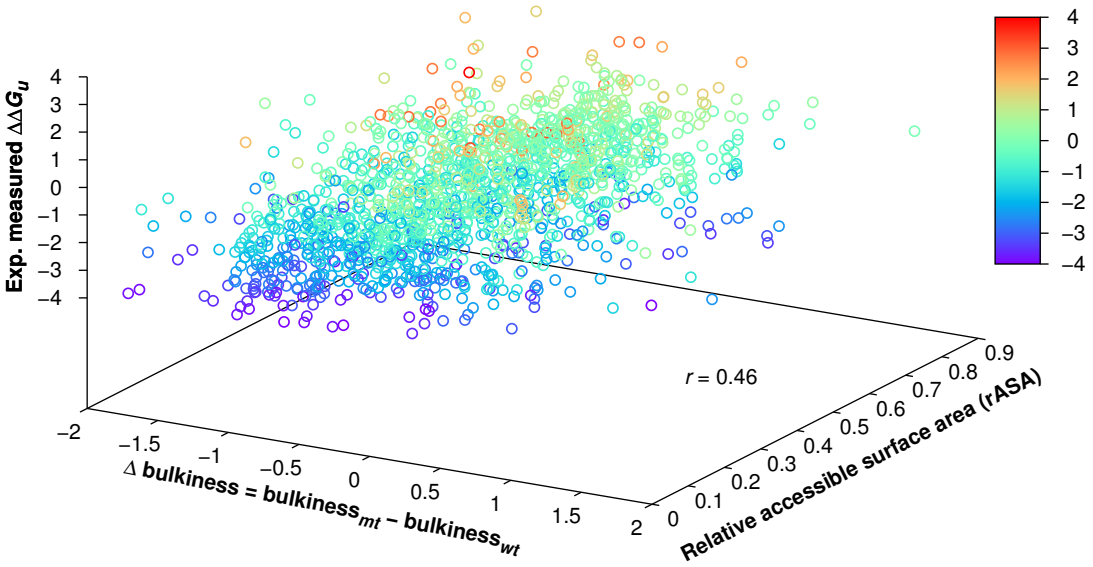

Figure S1: Experimentally measured stability changes ($\Delta\Delta G_u$) as a function of the amino acid parameter Δ bulkiness and predicted structural property relative accessible surface area (rASA) for the S1676 dataset. Δ bulkiness denotes the difference of the bulkiness of the mutant (bulkiness$_{mt}$) and wild-type (bulkiness$_{wt}$) amino acids. $\Delta\Delta G_u$ predicted based on Δ bulkiness and rASA with a *linear* support vector machine (SVM) model yielded a Pearson correlation coefficient (r) of 0.46. The figure shows that the introduction of a bulkier (relative to wild-type) amino acid in the protein core (low rASA) has a tendency to destabilise the protein structure.

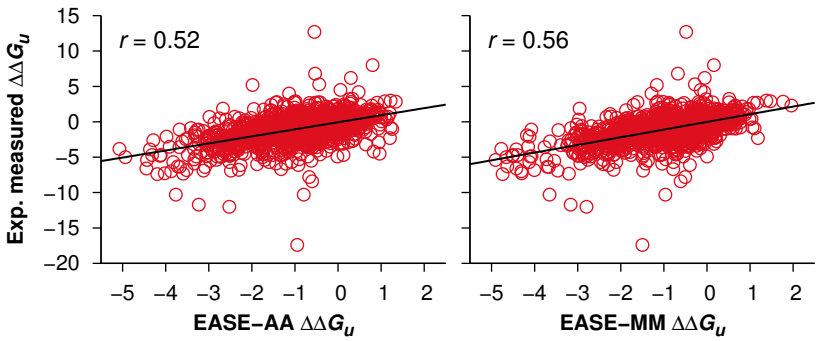

Figure S2: Experimentally measured stability changes ($\Delta\Delta G_u$) as a function of $\Delta\Delta G_u$ predicted with EASE-AA and EASE-MM for the S1676 dataset. The black lines are the linear regression fits.

*Corresponding author

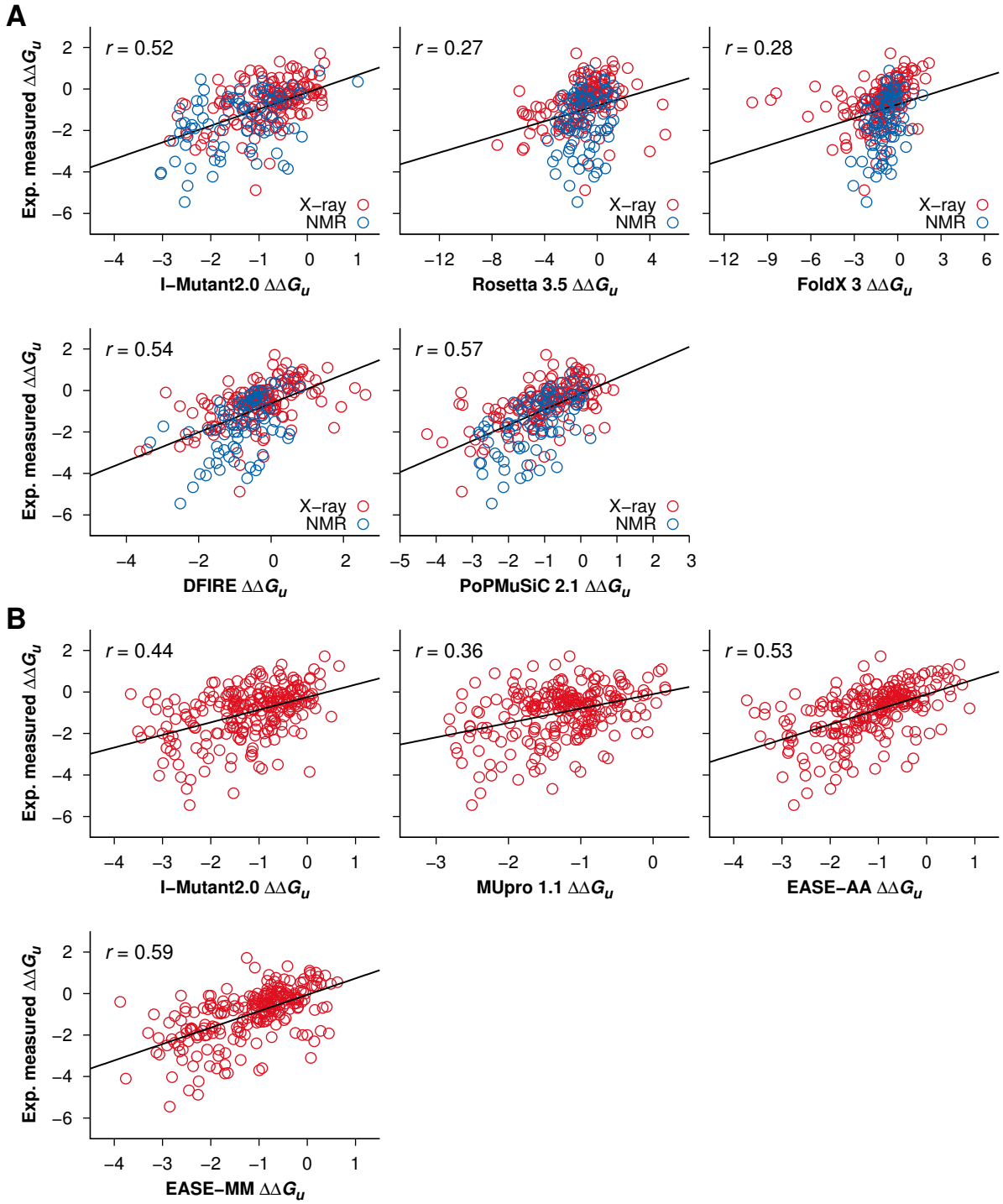

Figure S3: Experimentally measured stability changes ($\Delta\Delta G_u$) from the S236 dataset as a function of $\Delta\Delta G_u$ predicted with the five *structure-based* methods **(A)** and four *sequence-based* methods **(B)** including EASE-MM. Four predictions which caused atomic clashes during structure optimisation with Rosetta ($E_{rep} > 7$) were removed from the Rosetta plot. For the structure-based methods **(A)**, X-ray denotes predictions for proteins with high-resolution (≤ 3 Å) crystal structures (157 mutations), and NMR denotes predictions for protein structures determined with nuclear magnetic resonance (79 mutations). The black lines are the linear regression fits.

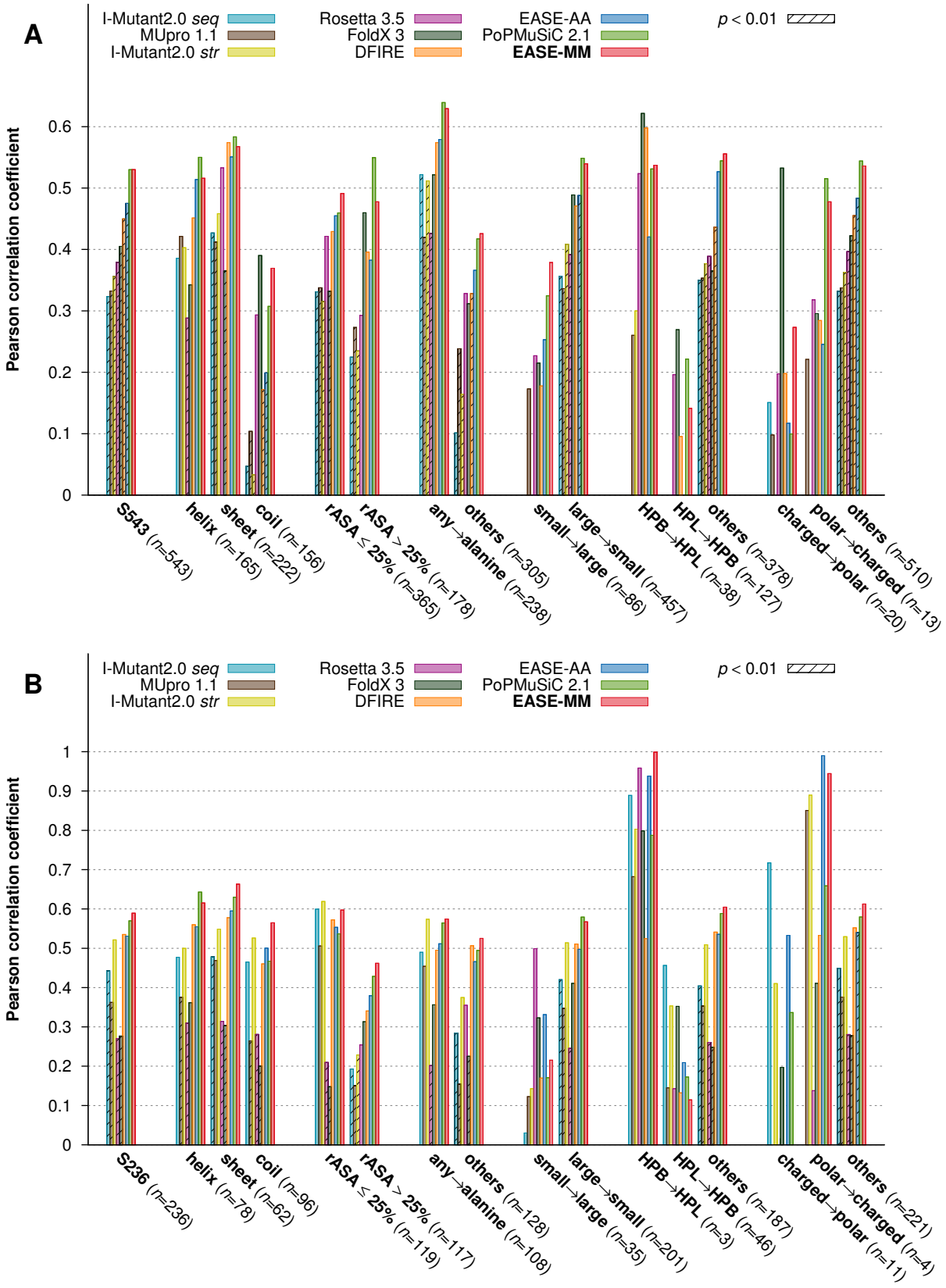

Figure S4: Pearson correlation coefficient (r) as the performance of EASE-MM and the eight compared methods for different types of mutations from the S543 **(A)** and S236 **(B)** datasets. The striped bars show results which are statistically different from EASE-MM (Williams' test, $p < 0.01$). Some methods yielded a negative correlation, which is shown here as a missing bar. The secondary structure elements (helix, sheet, coil) and relative accessible surface area (rASA) of the mutation site were calculated with DSSP [1]. We also divided mutations based on the type of the wild-type and mutant amino acids (denoted as 'wild-type→mutant'). Small and large amino acids were defined based on the non-hydrogen atom counts. Amino acids were grouped based on their side-chains as hydrophobic (HPB): A, V, I, L, M, F, Y, W; polar: S, T, N, Q; charged: D, E, K, R, H; and hydrophilic (HPL): polar + charged.

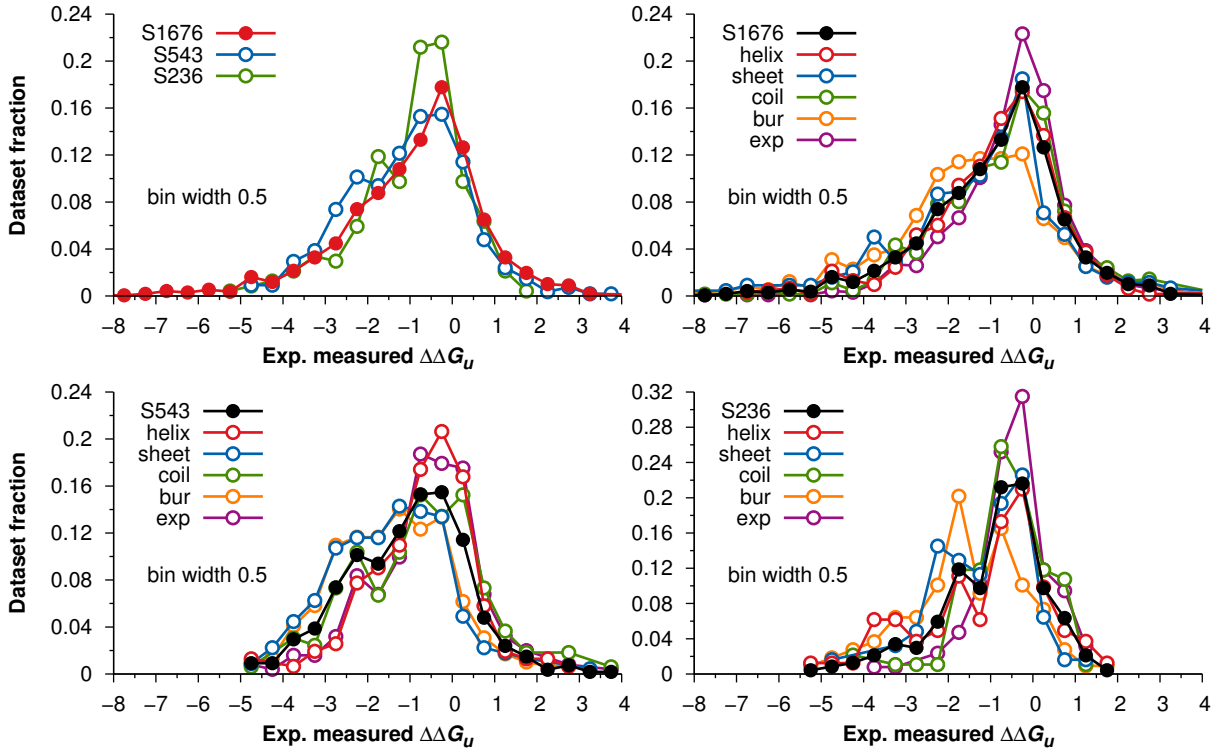

Figure S5: The distributions of the experimentally measured stability changes ($\Delta\Delta G_u$) for the three different datasets and for the five data partitions of each dataset. The five data partitions were created based on SPIDER [2] predictions.

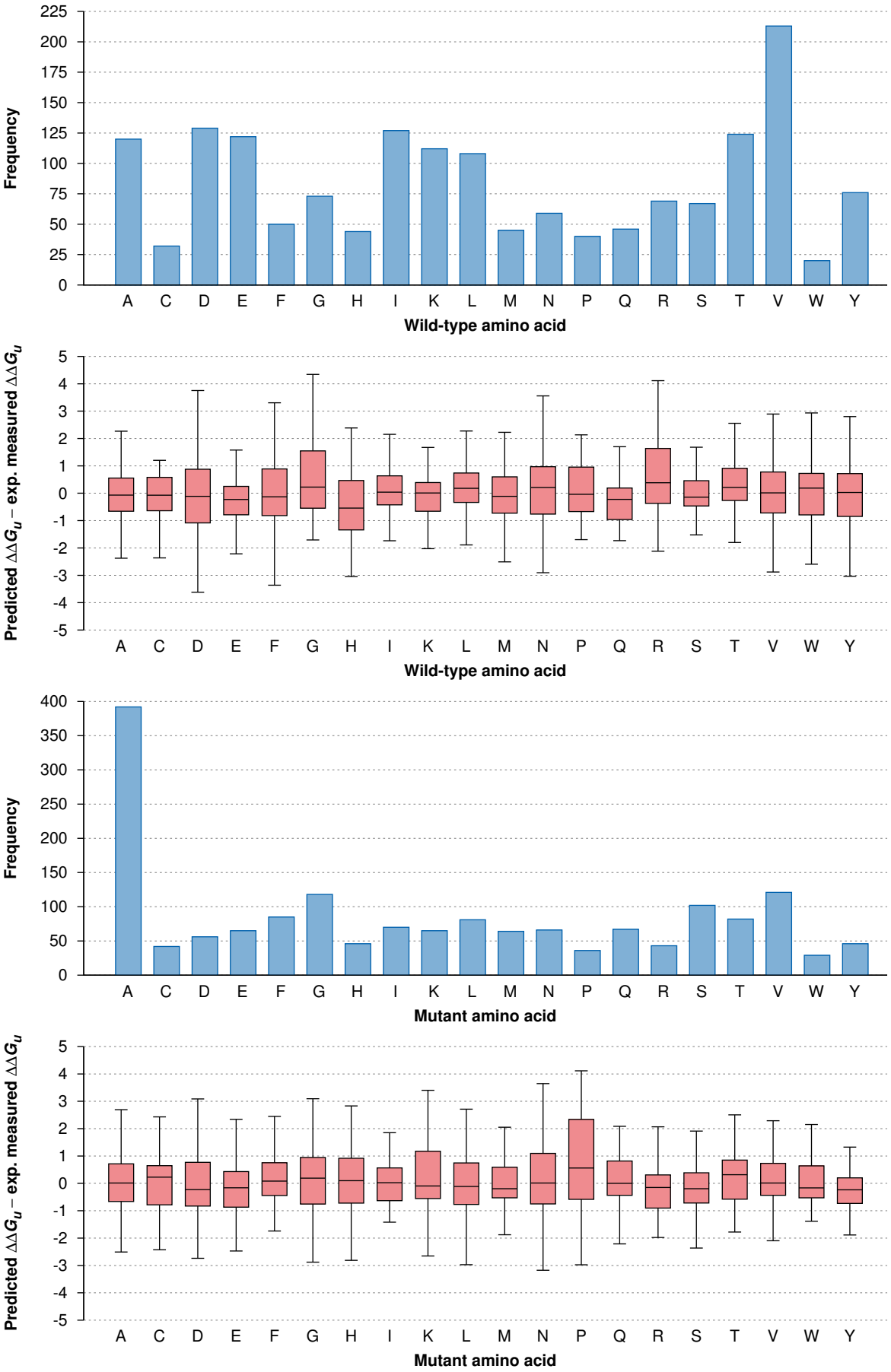

Figure S6: Frequencies and EASE-MM's prediction errors for different wild-type and mutant amino acid types from the S1676 dataset.

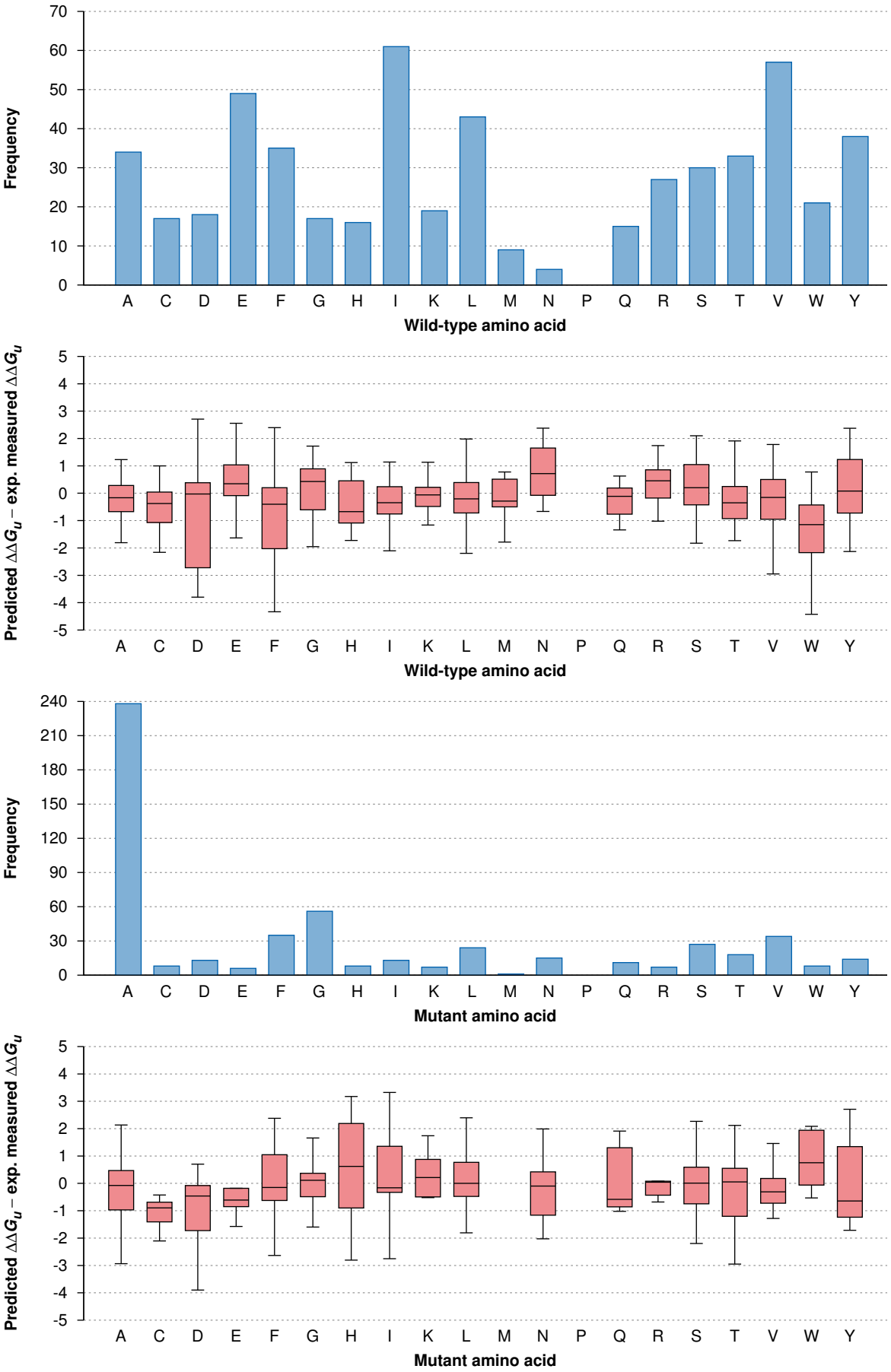

Figure S7: Frequencies and EASE-MM's prediction errors for different wild-type and mutant amino acid types from the S543 dataset.

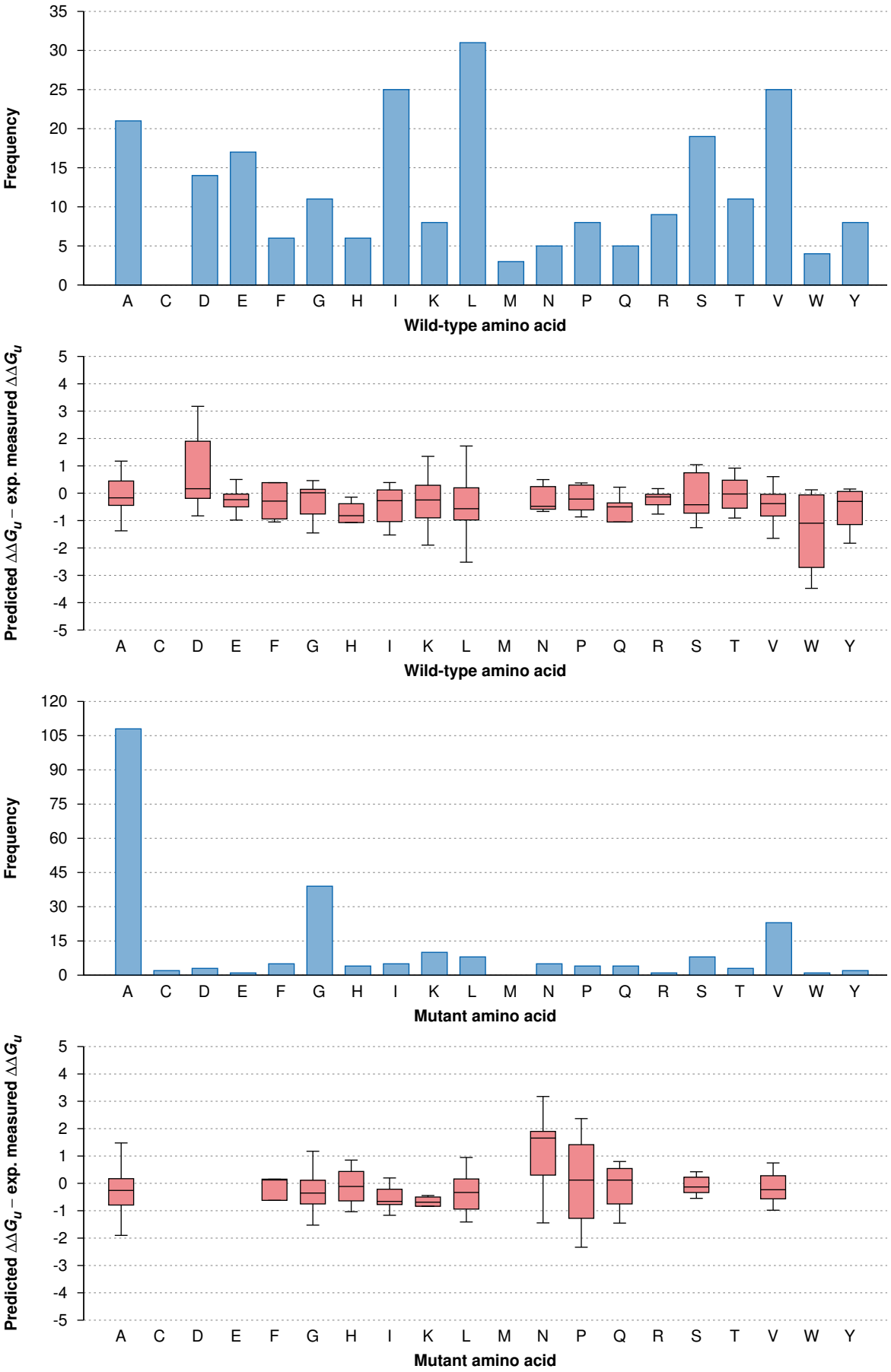

Figure S8: Frequencies and EASE-MM's prediction errors for different wild-type and mutant amino acid types from the S236 dataset.

Supplementary Tables

Table S1: Individual predictive features ranked by their correlation with experimentally measured stability changes ($\Delta\Delta G_u$) on the S1676 dataset

Feature name	r^a	p^a	**Definition**[b]
Δ bulkiness	0.348	6.7×10^{-49}	*amino acid parameter*, Gromiha *et al.* [3], Δ bulkiness = bulkiness$_{mt}$ − bulkiness$_{wt}$, Table S5
Δ hydrophobicity	0.339	3.1×10^{-46}	*amino acid parameter*, Meiler *et al.* [4], Δ hydrophobicity = hydrophobicity$_{mt}$ − hydrophobicity$_{wt}$, Table S5
Δ steric parameter	0.328	1.9×10^{-43}	*amino acid parameter*, Meiler *et al.* [4], Δ steric parameter = steric parameter$_{mt}$ − steric parameter$_{wt}$, Table S5
Δ sheet tendency	0.309	1.7×10^{-38}	*amino acid parameter*, Meiler *et al.* [4], Δ sheet tendency = sheet tendency$_{mt}$ − sheet tendency$_{wt}$, Table S5
Δ polarisability	0.279	2.6×10^{-31}	*amino acid parameter*, Meiler *et al.* [4], Δ polarisability = polarisability$_{mt}$ − polarisability$_{wt}$, Table S5
Δ PSSM	0.271	1.2×10^{-29}	*evolutionary feature*, PSSM was generated with PSI-BLAST [5]; Δ PSSM = PSSM$_{mt}$ − PSSM$_{wt}$, PSSM$_{wt}$ and PSSM$_{mt}$ are the probabilities of the wild-type and mutant amino acids at the mutation site, respectively
rASA	0.268	6.0×10^{-29}	*predicted structural property*, relative accessible surface area of the mutated residue was predicted with SPIDER [2]
Δ volume	0.265	2.0×10^{-28}	*amino acid parameter*, Meiler *et al.* [4], Δ volume = volume$_{mt}$ − volume$_{wt}$, Table S5
Δ flexibility	−0.202	6.5×10^{-17}	*amino acid parameter*, Vihinen *et al.* [6], Δ flexibility = flexibility$_{mt}$ − flexibility$_{wt}$, Table S5
PSSM$_{wt}$	−0.179	1.8×10^{-13}	*evolutionary feature*, PSSM was generated with PSI-BLAST [5]; PSSM$_{wt}$ is the probability of the wild-type amino acid at the mutation site
sheet probability	−0.132	5.3×10^{-8}	*predicted structural property*, probability that the mutation site is located in a sheet was predicted with SPIDER [2]
Δ ionisation[c]	−0.131	6.7×10^{-8}	*amino acid parameter*, Gromiha *et al.* [3], Δ ionisation = ionisation$_{mt}$ − ionisation$_{wt}$, Table S5
property entropy	−0.122	4.9×10^{-7}	*evolutionary feature*, overall conservation of the mutation site expressed as property entropy with respect to six amino acid 'property' groups [7]; the property entropy was calculated from a multiple sequence alignment of the 30 most similar sequences ranked by *e*-value with PSI-BLAST [5] (see Materials and Methods)
Δ compressibility	−0.091	1.8×10^{-4}	*amino acid parameter*, Gromiha *et al.* [3], Δ compressibility = compressibility$_{mt}$ − compressibility$_{wt}$, Table S5
coil probability	0.084	5.9×10^{-4}	*predicted structural property*, probability that the mutation site is located in a coil was predicted with SPIDER [2]
Δ isoelectric point	0.067	6.0×10^{-3}	*amino acid parameter*, Meiler *et al.* [4], Δ isoelectric point = isoelectric point$_{mt}$ − isoelectric point$_{wt}$, Table S5
Δ helix tendency	0.054	0.026	*amino acid parameter*, Meiler *et al.* [4], Δ helix tendency = helix tendency$_{mt}$ − helix tendency$_{wt}$, Table S5
helix probability	0.040	0.100	*predicted structural property*, probability that the mutation site is located in a helix was predicted with SPIDER [2]
disorder probability	0.022	0.361	*predicted structural property*, probability that the mutation site is in a disordered region of the protein was predicted with SPINE-D [8]

[a] r, Pearson correlation coefficient; p, probability that r is different from 0 due to random chance.
[b] *wt* and *mt* refer to the wild-type and mutant amino acids, respectively.
[c] equilibrium constant with reference to the ionisation property of COOH group

Table S2: Predictive features selected with the sequential forward floating selection algorithm for the five models of EASE-MM, ranked by their contributions to the respective models

		r decrease upon removing[b]		
Model	**Feature**[a]	**Relative**	**Absolute**	r **(single feature)**[c]
helix	rASA[d]	23.7%	0.117	0.295
	Δ helix tendency	9.2%	0.046	0.095
	Δ volume	4.4%	0.022	0.278
	Δ bulkiness	3.8%	0.019	0.321
	Δ compressibility	3.4%	0.017	0.160
	Δ isoelectric point	2.2%	0.011	0.193
	helix probability	0.3%	0.002	0.186
	coil probability	0.0%	0.000	0.182
	features combined			0.495
sheet	Δ PSSM[e]	5.3%	0.033	0.314
	Δ volume	4.7%	0.029	0.443
	Δ hydrophobicity	4.6%	0.029	0.449
	Δ compressibility	3.7%	0.023	0.109
	Δ helix tendency	1.7%	0.011	0.075
	sheet probability	0.9%	0.005	0.119
	coil probability	0.5%	0.003	0.002
	Δ steric parameter	0.2%	0.001	0.503
	disorder probability	0.2%	0.001	0.091
	Δ bulkiness	0.1%	0.000	0.533
	features combined			0.618
coil	Δ hydrophobicity	19.5%	0.087	0.233
	Δ flexibility	5.7%	0.026	0.227
	rASA[d]	3.4%	0.015	0.212
	Δ polarisability	2.2%	0.010	0.045
	Δ PSSM[e]	1.5%	0.007	0.143
	sheet probability	1.1%	0.005	0.129
	$PSSM_{wt}$ [e]	0.9%	0.004	0.063
	coil probability	0.5%	0.002	0.219
	Δ volume	0.3%	0.001	0.092
	features combined			0.449
buried	Δ isoelectric point	6.0%	0.037	0.089
	Δ bulkiness	5.6%	0.034	0.514
	Δ PSSM[e]	4.4%	0.027	0.274
	rASA[d]	2.7%	0.016	0.135
	Δ polarisability	1.7%	0.010	0.434
	Δ volume	1.5%	0.009	0.428
	Δ flexibility	1.0%	0.006	0.262
	Δ sheet tendency	0.8%	0.005	0.410
	features combined			0.612
exposed	Δ volume	19.2%	0.071	0.076
	helix probability	15.9%	0.059	0.008
	rASA[d]	6.8%	0.025	0.107
	Δ hydrophobicity	6.5%	0.024	0.141
	sheet probability	6.3%	0.023	0.075
	Δ helix tendency	4.4%	0.016	0.004
	Δ flexibility	2.2%	0.008	0.015
	$PSSM_{wt}$ [e]	1.3%	0.005	0.097
	features combined			0.370

[a] Δ, the change between the mutant and wild-type amino acids
[b] Decrease in Pearson correlation coefficient (r) for the given data partition (*e.g.*, helix) upon removing the given feature from the given model (*e.g.*, *helix*)
[c] Pearson correlation coefficient (r) of a single feature for the given data partition (*e.g.*, helix)
[d] rASA, relative accessible surface area
[e] Δ PSSM = $PSSM_{mt} - PSSM_{wt}$; $PSSM_{wt}$, PSSM probability of the wild-type amino acids; $PSSM_{mt}$, PSSM probability of the mutant amino acids; PSSM, position-specific scoring matrix

Table S3: Comparison of the prediction performance when swapping the five different models of EASE-MM and their corresponding data partitions on the S1676 dataset

	S1676 data partition									
	helix		**sheet**		**coil**		**buried**		**exposed**	
Model	r^a	p^b	r^a	p^b	r^a	p^b	r^a	p^b	r^a	p^b
helix	**0.50**	—	0.55	2.5×10^{-3}	0.37	9.3×10^{-3}	—	—	—	—
sheet	0.38	1.0×10^{-4}	**0.62**	—	0.38	7.7×10^{-3}	—	—	—	—
coil	0.40	4.2×10^{-4}	0.53	1.8×10^{-4}	**0.45**	—	—	—	—	—
buried	—	—	—	—	—	—	**0.61**	—	0.21	1.1×10^{-6}
exposed	—	—	—	—	—	—	0.51	5.7×10^{-7}	**0.37**	

[a] r, Pearson correlation coefficient; correlation coefficients of the 'matching' models (*i.e.*, the *helix* model for the helix data partition) are highlighted in bold.
[b] p, probability that the correlation coefficients (r) of the given model and the 'matching' model (*i.e.*, the *helix* model for the helix data partition) are different due to random chance (Williams' test for comparing correlation coefficients).

Table S4: Comparison of the prediction performance of EASE-MM when the structural properties are predicted from the sequence with SPIDER, calculated from the structure with DSSP, or drawn randomly.

Method	**Dataset**	**SS[a] and ASA[a]**	r^a	p^a	**RMSE[a]**
EASE-MM	S543	SPIDER[b]	0.53	—	1.22
		DSSP[c]	0.53	0.973	1.24
		random[d]	0.36	1.1×10^{-7}	1.36
	S236	SPIDER[b]	0.59	—	1.03
		DSSP[c]	0.57	0.446	1.06
		random[d]	0.31	2.2×10^{-3}	1.27

[a] SS, secondary structure; ASA, accessible surface area; r, Pearson correlation coefficient; p, probability that the correlation coefficients (r) of the given method and that of EASE-MM based on SPIDER are different due to random chance (Williams' test for comparing correlation coefficients); RMSE, root mean square error.
[b] SS and ASA were predicted from the protein *sequence* using SPIDER [2].
[c] SS and ASA were calculated from the protein *structure* using DSSP [1].
[d] The tests were repeated ten times, each time with *randomly drawn* SS and ASA; results were averaged.

Table S5: Scaled values of the 11 amino acid parameters which were implemented as candidate predictive features

AA[a]	**H[b]**	**V[b]**	**P[b]**	**IP[b]**	**HT[b]**	**ST[b]**	**GSI[b]**	**F_0[b]**	**F_1[b]**	**F_2[b]**	**C[b]**	**B[b]**	**EC[b]**
Ala	−0.171	−0.677	−0.680	−0.170	0.900	−0.476	−0.350	−0.044	−0.234	−0.269	0.587	−0.099	0.829
Asp	−0.767	−0.281	−0.417	−0.900	−0.155	−0.635	−0.213	−0.103	0.900	0.014	−0.475	−0.082	0.247
Cys	0.508	−0.359	−0.329	−0.114	−0.652	0.476	−0.140	−0.642	−0.773	−0.035	−0.433	0.094	−0.388
Glu	−0.696	−0.058	−0.241	−0.868	0.900	−0.582	−0.230	0.347	0.480	0.021	−0.900	0.105	0.565
Phe	0.646	0.412	0.373	−0.272	0.155	0.318	0.363	−0.863	−0.504	−0.113	−0.673	0.721	0.035
Gly	−0.342	−0.900	−0.900	−0.179	−0.900	−0.900	−0.900	0.701	0.527	−0.050	0.378	−0.900	0.829
His	−0.271	0.138	0.110	0.195	−0.031	−0.106	0.384	−0.480	−0.186	−0.255	−0.297	0.115	−0.088
Ile	0.652	−0.009	−0.066	−0.186	0.155	0.688	0.900	−0.332	−0.662	−0.411	−0.288	0.879	−0.900
Lys	−0.889	0.163	0.066	0.727	0.279	−0.265	−0.088	0.339	0.844	0.900	−0.375	0.317	0.547
Leu	0.596	−0.009	−0.066	−0.186	0.714	−0.053	0.213	−0.590	−0.115	−0.064	−0.288	0.879	0.865
Met	0.337	0.087	0.066	−0.262	0.652	−0.001	0.110	−0.738	−0.900	−0.893	−0.205	0.370	0.724
Asn	−0.674	−0.243	−0.329	−0.075	−0.403	−0.529	−0.213	0.516	0.242	0.000	−0.166	0.031	0.265
Pro	0.055	−0.294	−0.900	−0.010	−0.900	0.106	0.247	0.059	0.868	0.014	0.900	0.487	0.212
Gln	−0.464	−0.020	−0.110	−0.276	0.528	−0.371	−0.230	0.870	0.416	−0.319	−0.403	0.192	0.529
Arg	−0.900	0.466	0.373	0.900	0.528	−0.371	0.105	−0.066	0.416	−0.206	0.430	0.175	−0.106
Ser	−0.364	−0.544	−0.637	−0.265	−0.466	−0.212	−0.337	0.900	0.575	−0.050	−0.024	−0.300	0.600
Thr	−0.199	−0.321	−0.417	−0.288	−0.403	0.212	0.402	0.192	0.599	0.028	−0.212	0.323	0.406
Val	0.331	−0.232	−0.285	−0.191	−0.031	0.900	0.677	−0.480	−0.385	−0.120	−0.127	0.896	0.794
Trp	0.900	0.900	0.900	−0.209	0.279	0.529	0.479	−0.900	−0.464	−0.900	−0.074	0.900	0.900
Tyr	0.188	0.541	0.417	−0.274	−0.155	0.476	0.363	−0.634	−0.361	−0.659	−0.738	0.546	0.582

[a] AA denotes an amino acid in the standard three-letter code.
[b] H, hydrophobicity; V, volume; P, polarisability; IP, isoelectric point; HT, helix tendency; ST, sheet tendency; GSI, graph shape index (steric parameter); F_0, flexibility with no rigid neighbours; F_1, flexibility with one rigid neighbour; F_2, flexibility with two rigid neighbours; C, compressibility; B, bulkiness; and EC, equilibrium constant with reference to the ionisation property of COOH group.

References

[1] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.

[2] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific Reports 5 (2015) 11476.

[3] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, Journal of Protein Chemistry 18 (5) (1999) 565–578.

[4] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, Molecular modeling annual 7 (9) (2001) 360–369.

[5] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research 25 (17) (1997) 3389.

[6] M. Vihinen, E. Torkkila, P. Riikonen, Accuracy of protein flexibility predictions, Proteins: Structure, Function, and Bioinformatics 19 (2) (1994) 141–149.

[7] L. A. Mirny, E. I. Shakhnovich, Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, Journal of Molecular Biology 291 (1) (1999) 177–196.

[8] T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, Y. Zhou, SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method, Journal of Biomolecular Structure and Dynamics 29 (4) (2012) 799–813.