Support Information

DeepDDG: Predicting the Stability Change of Protein Point Mutations using Neural Networks

Huali Cao,[†] Jingxue Wang[†], Liping He[†], Yifei Qi[*,†,‡], and John. Z. Zhang[*,†,‡,§]

[†]*Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, 200062, China*

[‡]*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China*

[§]*Department of Chemistry, New York University, New York, New York 10003, USA*

[*]Correspondence:
yfqi@chem.ecnu.edu.cn
John.Zhang@nyu.edu

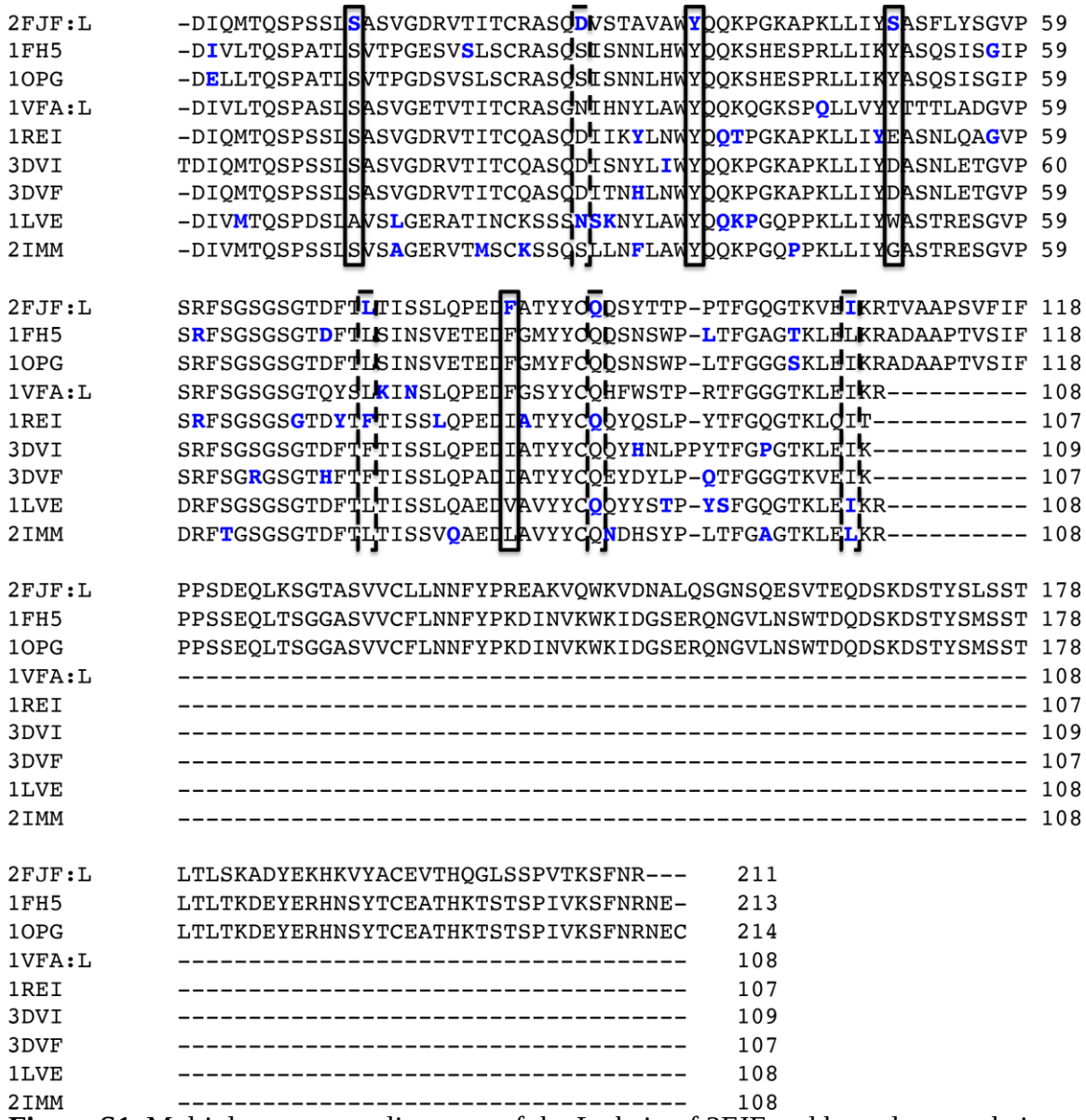

Figure S1. Multiple sequence alignment of the L chain of 2FJF and homologous chains within the training set. Blue letters indicate mutated sites. Solid and dashed boxes indicate mutations in 2FJF. Only mutations in solid boxes were included in comparison of ΔTm and predicted ΔΔG values.

Table S1. Highest sequence identity for each test protein upon comparison to the training proteins.*

PDB ID of test proteins	PDB ID of training protein with highest sequence identity	Sequence identity
2JOF	2RPN	0.197
1LVM	1XZO	0.215
3FFN	1KFW	0.222
4HE7	1SHG	0.222
5VP3	3BDC	0.227
1BA3	1AMQ	0.228
1JL9	3FIL	0.228
1HCQ	1LZ1	0.229
2NTE	5T43	0.231
1E0L	1BF4	0.235
1FT8	1BLC	0.236
1DXX	1LS4	0.237
2Q98	3HHR	0.237
1NMV	2MMX	0.240
1E0W	1ANK	0.240
1G3P	2C9Q	0.241
1F8I	3MBP	0.241
4BUQ	1RN1	0.242
1FC1	2A01	0.243
1GLU	1HME	0.244
2H3F	1LS4	0.245
1GUA	1FKJ	0.245
2JUC	1A43	0.245
3D2C	1AM7	0.246
1J8I	1TTG	0.247
1IV7	1CDC	0.247
1O6X	1IGV	0.248
3C2I	1H7M	0.248
2N7Z	1TEN	0.248
2PR5	1AKK	0.249
2ARF	1BD8	0.249
1JLV	1ANK	0.249
1PRG	2DRI	0.249
1L6H	1DIV	0.250
1BNL	1NFI	0.250
2CLR	1POH	0.250
1A0F	3BDC	0.250

*Sequence identity was calculated using the *pairwise2* module in Biopython.

Table S2. Features used by the neural network for the target and neighboring residues.

Category	Property	Normalization	#bits	Target residue	Neighbor residue
Backbone dihedral	*sin* and *cos* of backbone phi, psi, and omega	N	6	Y	Y
Residue solvent accessible surface area	Residue solvent accessible surface area	Standard values from Naccess	1	Y	Y
Secondary structure	One-hot code of secondary structure	N	3	Y	Y
Num. hydrogen bond	#Backbone-backbone hydrogen bond with the target residue	N	1	N	Y
	#Backbone-sidechain hydrogen bond with the target residue	N	1	N	Y
	#sidechain-sidechain hydrogen bond with the target residue	N	1	N	Y
	#sidechain-backbone hydrogen bond with the target residue	N	1	N	Y
Distance and orientation of the neighbor residue	Ca-Ca distance to the target residue	20 Å	1	N	Y
	Ca-Ca unit vector to the target residue	N	3	N	Y
	Ca-C unit vector of the neighbor residue	N	3	N	Y
	Ca-N unit vector of the neighbor residue	N	3	N	Y
PSSM	PSSM score of the wide type and mutant residue	N	2	Y	N
PFS	PFS between the wildtype and mutant target residue and neighbor residue	N	2	Y	N
Protein design probability	Probability of the wildtype residue at the target residue from protein design network	N	1	Y	N
	Probability of the mutant residue at the	N	1	Y	N

	target residue from protein design network				
Amino acid type	5-bit code of residue type of the wildtype target residue	N	5	Y	N
	5-bit code of residue type of the mutant target residue	N	5	Y	N
	5-bit code of residue type of the neighbor residue	N	5	N	Y

Table S3. Correlation, slope and intercept obtained during linear fitting of the experimental and calculated ΔΔG values for the test set.

Method	R	Slope	Intercept
MUpro1.1	0.190	0.36	-0.25
DynaMut	0.393	0.49	-0.47
EASE-MM	0.402	0.52	-0.12
iStable	0.427	0.66	-0.18
PopMusic	0.443	0.55	-0.08
STRUM	0.447	0.85	0.06
I-Mutant3.0	0.453	0.76	0.12
mCSM	0.467	0.67	0.03
SDM	0.483	0.46	-0.45
DUET	0.515	0.62	-0.09
DeepDDG	0.557	0.67	0.02
iDeepDDG	0.563	0.75	0.05

Table S4. Highest sequence identity of each protein in the ΔT_m set to the training proteins*

PDB ID of test proteins	PDB ID of training protein with highest global sequence identity	Sequence identity
1AQH	1KFW	0.23
1H8V	1OPG	0.24
1OSI	4Q0M	0.25
1XAS	1ANK	0.24
2FJF chain H	1VFA	0.27
2FJF chain L	1FH5	0.60
GK	2A01	0.24

*Sequence identity was calculated using the *pairwise2* module in Biopython.

Table S5. Pearson's correlation between predicted ΔΔG values and experimental stabilities on stabilizing and destabilizing mutations from the PTEN and TPMT datasets.

Method	PTEN		TPMT	
	Stabilizing (513)*	Destabilizing (2455)*	Stabilizing (794)*	Destabilizing (2570)*
DeepDDG	0.025	0.565	0.013	0.553
iDeepDDG	0.008	0.523	0.007	0.526
PopMusic	0.048	0.534	0.039	0.408
DUET	-0.003	0.426	-0.007	0.364
EASE_MM	0.002	0.411	0.014	0.345
mCSM	-0.004	0.397	-0.025	0.349
STRUM	0.019	0.397	0.165	0.414
SDM	0.012	0.411	0.028	0.319
I-mutant	-0.021	0.388	-0.054	0.354
iStable	0.051	0.322	0.028	0.357
MUpro	-0.015	0.247	0.036	0.239

*Numbers in parentheses are total number of mutations.