Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity

Minghui Li, Marharyta Petukh, Emil Alexov and Anna R. Panchenko

Supplementary Materials

Supplemental Data

Table S1. Ten single and eight multiple mutants with unrealistic VMD initial models where mutated residues had large steric clashes with adjacent residues that could not be fixed by minimization procedure (see Methods).

Protein	Mutation
1CHO_EFG_I	AI12R
1PPF_E_I	AI15R
1CHO_EFG_I	TI14W
1PPF_E_I	TI17F
1PPF_E_I	TI17Y
1PPF_E_I	TI17W
1R0R_E_I	TI12W
3SGB_E_I	TI11W
1PPF_E_I	GI32R
3SGB_E_I	NI30L
1CHO_EFG_I	AI12R_LI15R
1CHO_EFG_I	AI12R_TI14K_LI15R
1CHO_EFG_I	AI12R_TI14P_LI15R
1CHO_EFG_I	PI11S_AI12R_TI14P_LI15R
1CHO_EFG_I	KI10R_AI12R_TI14K_LI15R
1CHO_EFG_I	KI10R_PI11F_AI12R_TI14K_LI15R
1CHO_EFG_I	AI12R_TI14K_LI15R_EI16S
1PPF_E_I	AI15R_LI18R

Protein: The PDB entry for the complex, followed by chain identifiers of two subunits separated by the underscore. **Mutation**: residue number of mutation in the 'cleaned' pdb files (renumbering of residue in pdb file starting from one). The first character is amino acid of the original residue, the second character is chain identifier, the third to penultimate characters indicate the residue number, and the last character indicates the mutant amino acid. If multiple mutations are present, they are separated by '_'.

Table S2. Correlation between predicted and experimental values of *ΔΔG* for different simulation protocols. All calculations were performed with *Pred1* energy function. R - Pearson correlation coefficient between experimental and predicted *ΔΔG* values, R^{CV} - five-fold cross-validated correlation coefficient and RMSE, root-mean squared error, are shown for the case of training/testing on single mutations of NM set.

Simulation method	**Water model**	**Flexibility**	**Epsilon**	**CONC**	**R(R^{CV})**	**RMSE (kcal mol^{-1})**
Minimization	Explicit water	Flexible backbone	1	0.0	0.62(0.59)	1.24
			2	0.0	0.63(0.61)	1.22
				0.05	0.62(0.60)	1.23
				0.1	0.61(0.60)	1.24
			4	0.0	0.60(0.58)	1.26
		Restrained backbone	2	0.0	0.62(0.61)	1.23
	Implicit water	Flexible backbone	2	0.0	0.50(0.48)	1.36
MD simulation	Explicit water	Flexible backbone	2	0.0	0.35(0.26)	1.48

Epsilon: dielectric constant. CONC: ion concentration (Mol L^{-1}, M).

Table S3. The optimal fitting coefficients and standardized coefficients from multiple linear regression performed on SKEMPI set. γ is in kcal mol^{-1} nm^{-2} and δ is in kcal mol^{-1}.

Training SKEMPI	Equation	Parameters	Energy term	Regression coefficients(p-value, standard deviation)	Standardized regression coefficients
Single mutations	Pred1	α	$\Delta\Delta E_{vdw}$	0.226 (2e-16, 0.013)	0.344
		β	$\Delta\Delta G_{solv}$	0.130 (2e-16, 0.007)	0.399
		γ	ΔSA_{mut}	0.045 (2e-16, 0.000)	0.169
		δ		1.678 (2e-16, 0.114)	
Single mutations	Pred2	α	$\Delta\Delta E_{vdw}$	0.122 (3.83e-16, 0.015)	0.186
		β	$\Delta\Delta G_{solv}$	0.101 (2e-16, 0.007)	0.308
		γ	ΔSA_{mut}	0.043 (2e-16, 0.000)	0.161
		ε	$\Delta\Delta G_{BM}$	0.446 (2e-16, 0.044)	0.222
		λ	$\Delta\Delta G_{FD}$	0.168 (4.33e-12, 0.024)	0.148
		δ		1.326 (2e-16, 0.113)	
Multiple mutations	Pred1	α	$\Delta\Delta E_{vdw}$	0.098 (1.10e-11, 0.014)	0.245
		β	$\Delta\Delta G_{solv}$	0.151 (2e-16, 0.011)	0.483
		γ	ΔSA_{mut}	0.038 (0.018, 0.000)	0.084
		δ		1.978 (1.02e-07, 0.367)	

Table S4. Accuracy of prediction for different types of amino acid substitutions categorized by their charge. Negatively charged amino acids (D, E), neutral amino acids (A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V), and positively charged amino acids (R, K). R is calculated for SKEMPI single mutation set using *Pred2* energy function. Only statistically significant correlation coefficients are shown (p-value < 0.01).

	Mutant		
	Negative	Neutral	Positive
Wild-type	R/# mutations	R/# mutations	R/# mutations
Negative	-	0.33/232	-
Neutral	0.72/86	0.58/1042	0.48/89
Positive	0.81/33	0.67/300	-

Table S5. Accuracy of prediction for different types of amino acid substitutions categorized by the side chain volume. Small (A, G, S), medium (N, D, C, Q, E, H, I, L, K, M, P, T, V), and large (R, F, W, Y) amino acids. R is calculated for SKEMPI single mutations set using *Pred2* energy function. Only statistically significant correlation coefficients are shown (p-value < 0.01).

	Mutant		
	Small	**Medium**	**Large**
Wild-type	R/# mutations	R/# mutations	R/# mutations
Small	0.52/97	0.51/123	0.67/39
Medium	0.61/590	0.58/450	0.34/130
Large	0.63/210	0.64/142	0.58/63

Table S6. Residue-residue pairs that have hydrogen bonds and salt bridges formed in the final minimized structure for wild type (WT-MM), 500-step minimized structure for mutant (Mutant-MM), average structure obtained using 1ns of MD simulations for wild type (WT-MD) and mutant (Mutant-MD). Red color highlights hydrogen bonds and salt bridges formed by mutated residue of L15E and R85A. Residue contacts include those formed between main chain atoms and side chain atoms.

Name	WT-MM	WT-MD	Mutant-MM	Mutant-MD
1CHO_I_L15E Between two partners Partner1: chain F and G Partner2: chain I	F_D35 I_R18 F_F41 I_Y17 F_Y146 I_N33 G_G193 I_L15 G_S214 I_L15 G_G216 I_C13 (2) *F_D35 I_R18*	F_F41 I_Y17 F_Y146 I_N33 G_G193 I_L15 G_S214 I_L15 G_G216 I_C13	F_D35 I_R18 F_F41 I_Y17 F_Y146 I_N33 G_G193 I_E15 G_S214 I_E15 G_G216 I_C13 (2) *F_D35 I_R18*	F_F41 I_Y17 F_C58 I_R18 G_S190 I_E15 G_G193 I_E15 G_S195 I_E15 G_S214 I_E15 G_G216 I_C13 G_S217 I_E15 (3) G_S217 I_C13
1IAR_A_R85A Between two partners Partner1: chain A Partner2: chain B	A_T6 B_S70 A_E9 B_S70 A_E9 B_Y127 A_E9 B_Y13 A_E9 B_Y183 A_K12 B_H131 A_Q78 B_D125 A_R81 B_D67 (2) A_R81 B_D125 A_R85 B_D67 (3) A_R85 B_D125 A_R88 B_D72 (2) A_R88 B_D67 (2) A_N89 B_A71 *A_R81 B_D125 (2)* *A_R81 B_D67* *A_R85 B_D125 (2)* *A_R85 B_D67* *A_R88 B_D67 (2)* *A_R88 B_D72 (2)*	A_E9 B_S70 A_E9 B_Y13 A_E9 B_Y183 A_T13 B_Y127 A_R81 B_D67 (3) A_R85 B_D67 (2) A_R88 B_D72 (2) A_R88 B_D67 A_N89 B_A71 *A_R81 B_D67 (2)* *A_R85 B_D125 (2)* *A_R85 B_D67* *A_R88 B_D72*	A_T6 B_S70 A_E9 B_S70 A_E9 B_Y127 A_E9 B_Y13 A_E9 B_Y183 A_K12 B_H131 A_Q78 B_D125 A_R81 B_D67 (2) A_R81 B_D125 A_R88 B_D72 (2) A_R88 B_D67 (2) A_N89 B_A71 *A_R81 B_D125 (2)* *A_R81 B_D67* *A_R88 B_D67 (2)* *A_R88 B_D72 (2)*	A_E9 B_S70 A_E9 B_Y13 A_E9 B_Y183 A_T13 B_Y127 A_R81 B_D67 (2) A_R81 B_V68 A_R88 B_D72 (2) A_R88 B_D67 A_N89 B_A71 *A_R81 B_D67 (2)* *A_R88 B_D67 (2)* *A_R88 B_D72 (2)*

Salt bridges are shown in italic and others correspond to hydrogen bonds. The number in a bracket is the number of bonds formed within each residue-residue pair. Hydrogen bonds are defined using the following criteria: first, the maximum distance between acceptor (N/O/S atoms) and hydrogen is 2.5 Å; second, the minimum angle of donor-hydrogen-acceptor is 90°.[1] Salt bridges between two charged residues are defined using the following criteria: a maximal distance of 4 Å between two charged atoms (N/O).[2]

Figure S1. Schema of the simulation protocols

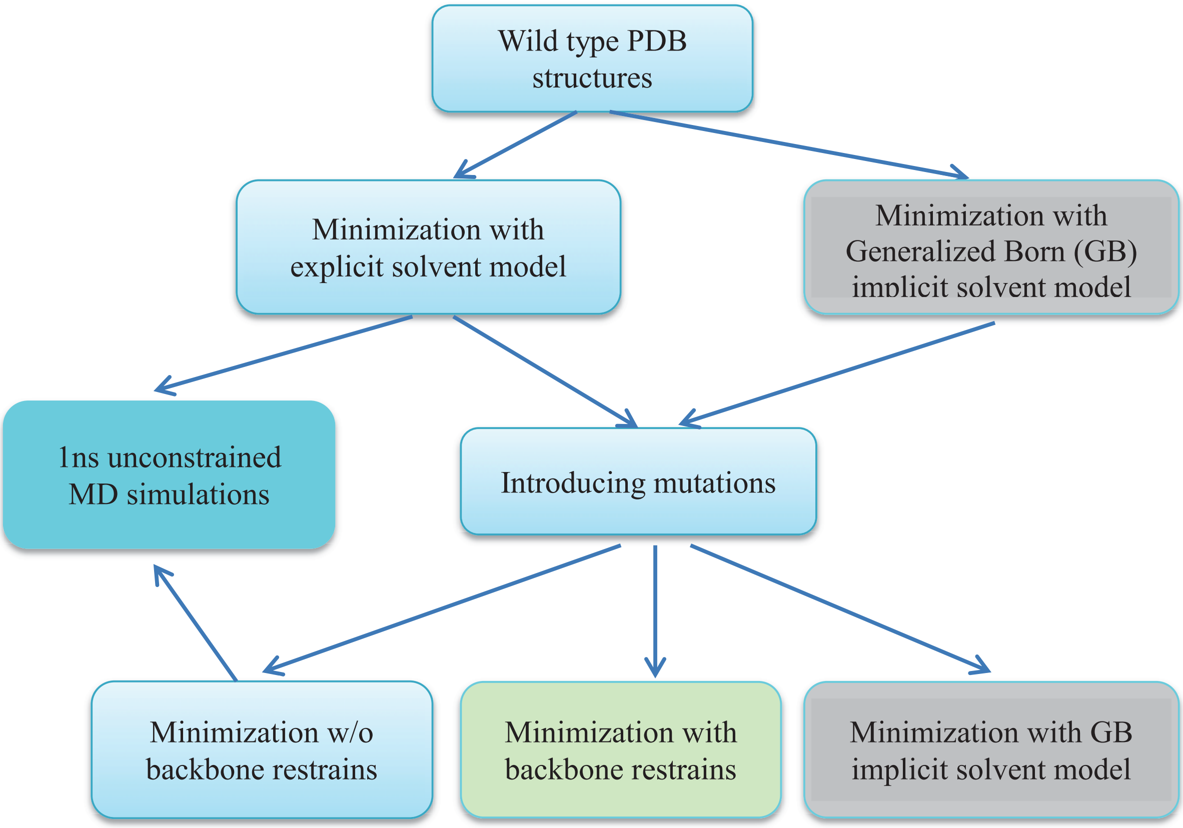

Wild type PDB structures
Minimization with explicit solvent model
Minimization with Generalized Born (GB) implicit solvent model
1ns unconstrained MD simulations
Introducing mutations
Minimization w/o backbone restrains
Minimization with backbone restrains
Minimization with GB implicit solvent model

Figure S2. The distribution of the system size for protein-protein complexes. The number of atoms includes all atoms in the solvated system (number of atoms in proteins, number of atoms in solvent and added ions).

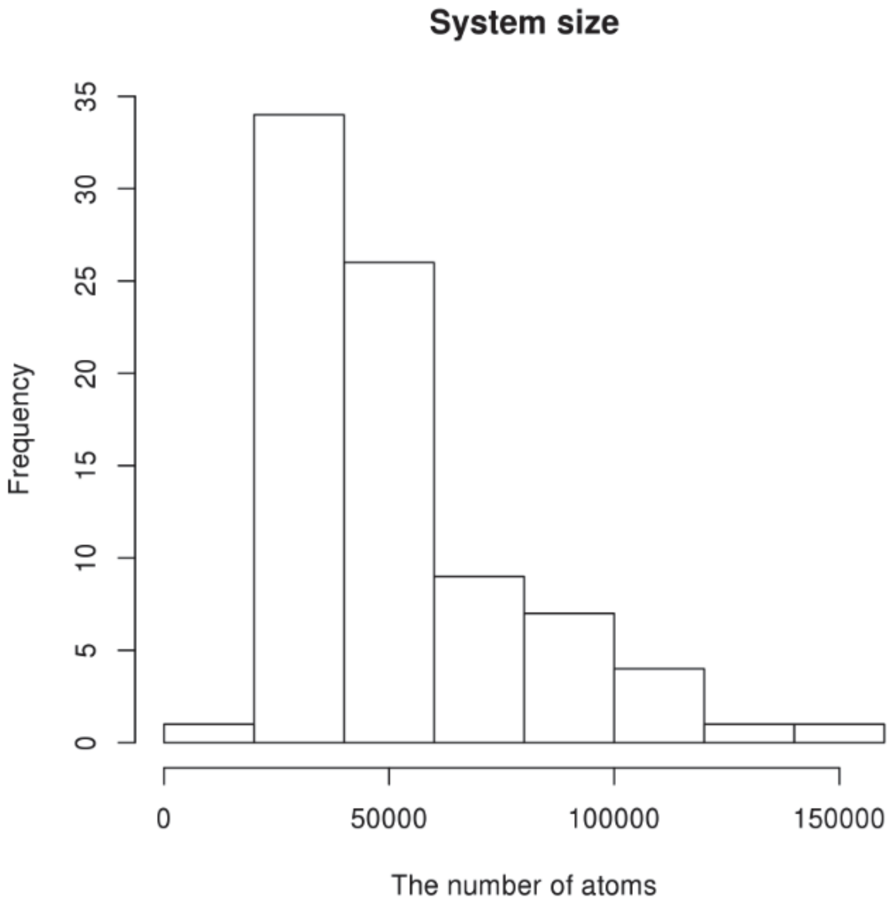


Figure S3. Dependence of correlation coefficient between experimental $\Delta\Delta G_{exp}$ and predicted $\Delta\Delta G_{pred1}$ on the number of minimization steps and number of frames in MD simulations. Training and fitting was done on NM single mutations set (A); SKEMPI single mutations set (B); NM multiple mutations set (C); SKEMPI multiple mutations set (D); on NM single mutations set with MD simulation performed for mutant only (minimized structure is used for wild type) (E); on NM single mutations set with MD simulation performed for both mutant and wild type (F).

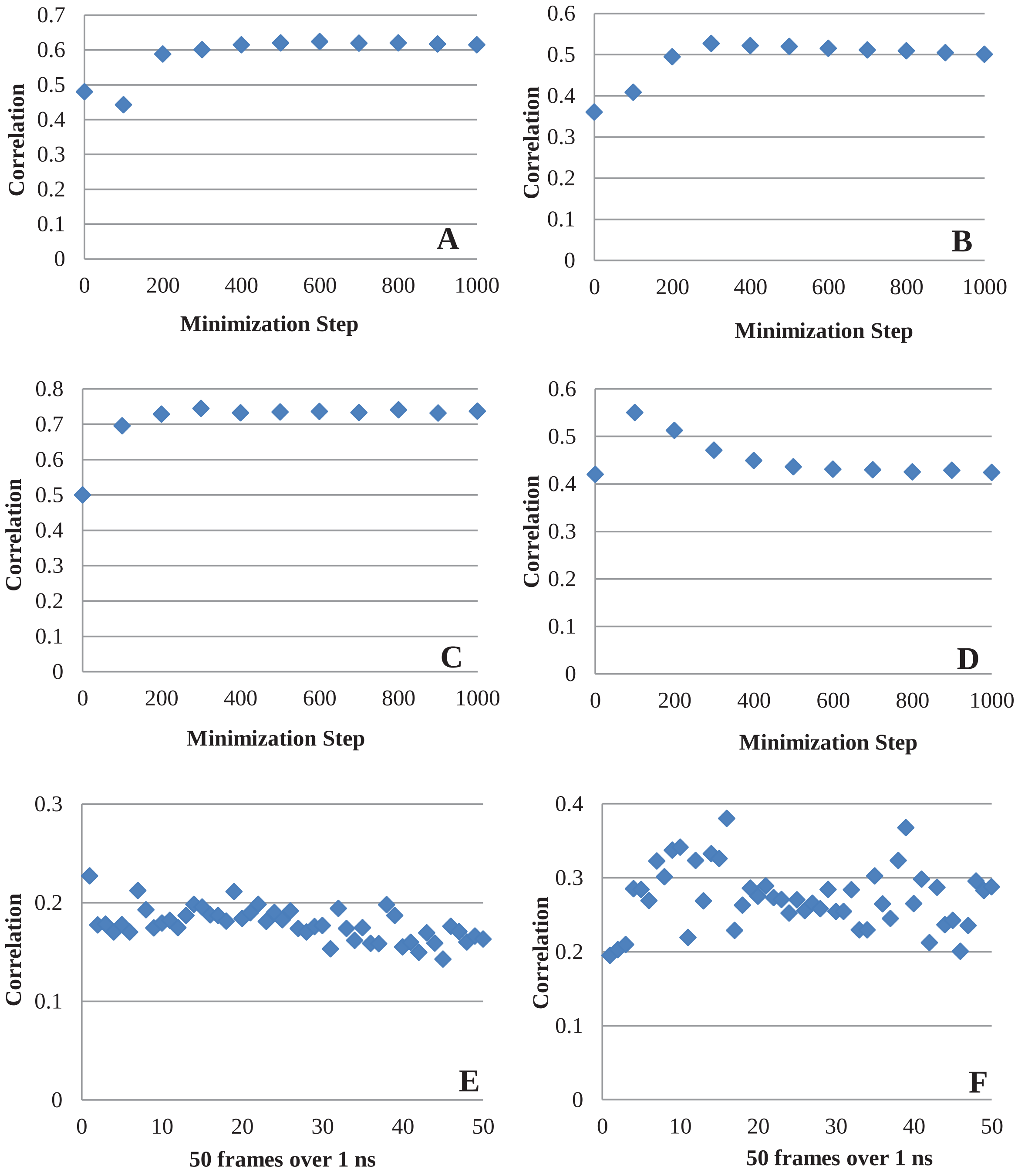


Figure S4. **The correlation between experimental and predicted *Pred1 ΔΔG* values for each protein complex for 500 and 10,000 minimization steps for single mutants from NM set.** R = 0.63 at 500 step and R = 0.59 at 10,000 step for all single mutations of NM set.

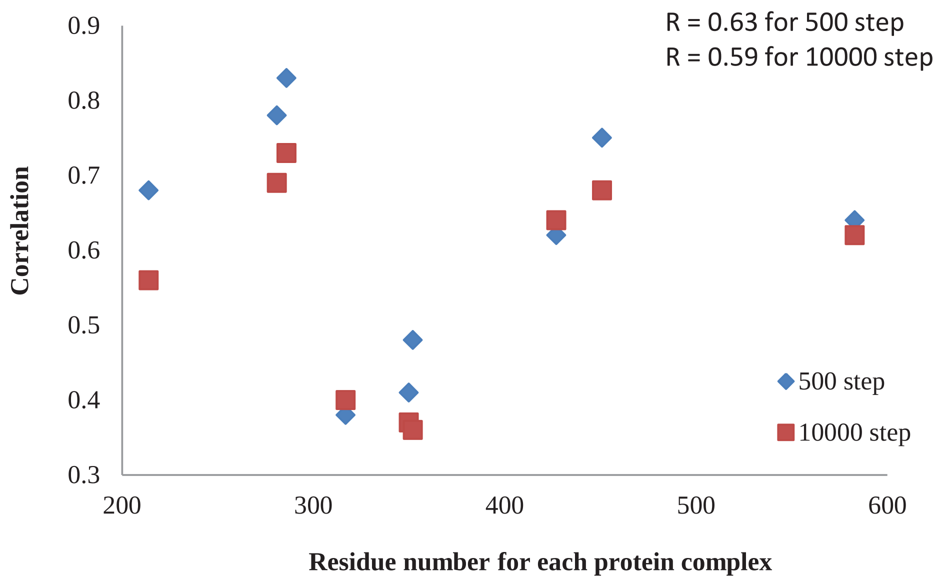


Figure S5. Distribution of the Root mean square deviation (RMSD, Å) of backbone atoms for 242 single mutants from NM set. 500 frames are extracted from every mutant's MD trajectory. Overall results from 121000 structures are shown in the figure.

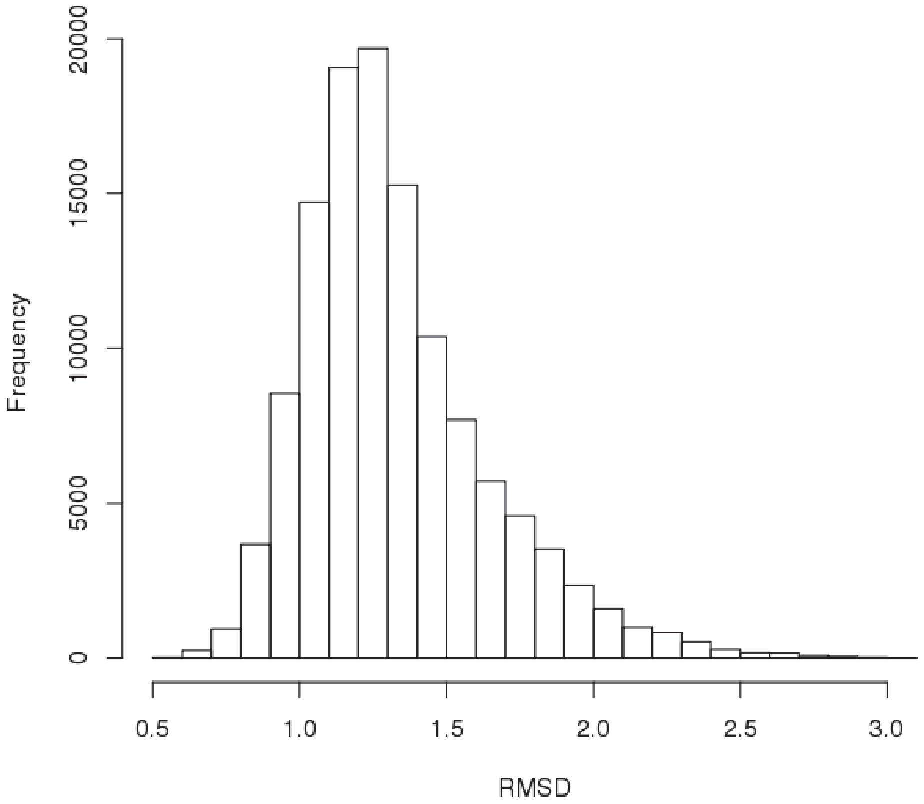


Figure S6. Average local heavy atom RMSD values between the minimized mutant structure and the initial non-minimized mutant models for different types of amino acid substitutions categorized by charge, side chain volume. RMSD is calculated for the mutated residues and residues within 4Å from the mutant site. "Small/Large" refers to small amino acids substituted into large amino acid. Substitutions from and to Proline are provided as separate bars because of specific properties of this residue.

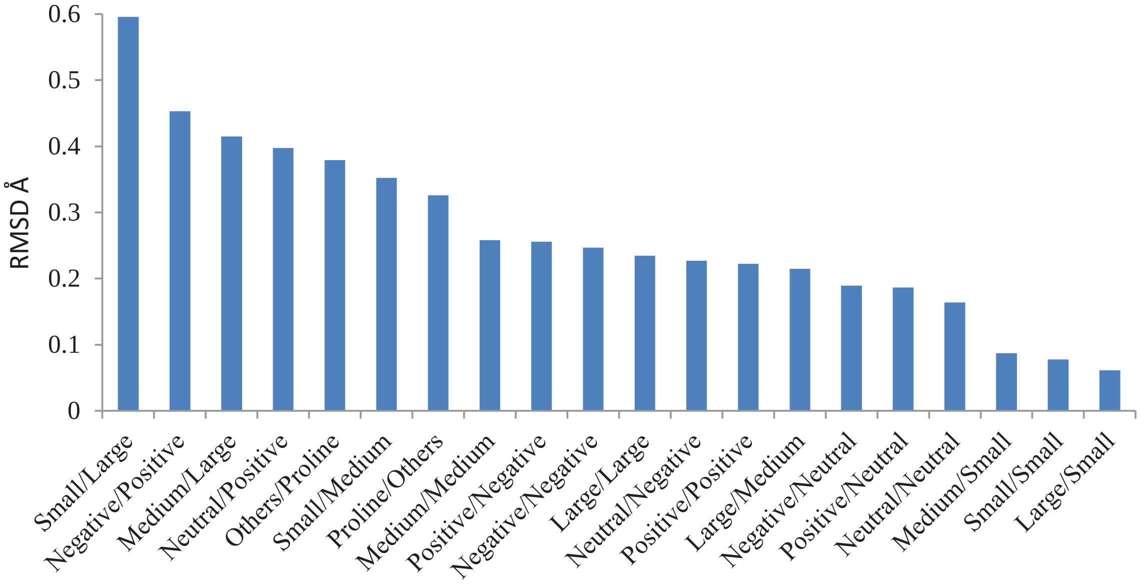
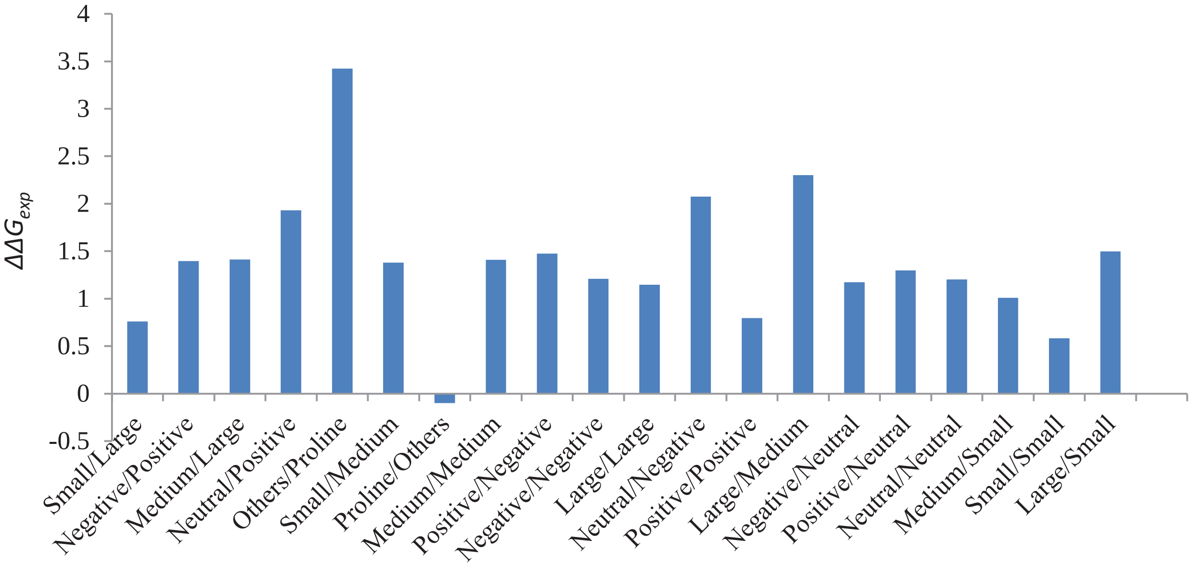


Figure S7. Average values of experimental *ΔΔG*$_{exp}$ for different types of amino acid substitutions categorized by charge, side chain volume. Substitutions from and to Proline are provided as separate bars because of specific properties of this residue.



Supplemental procedures

Definitions of models Pred3 and Pred4

$$\Delta\Delta G_{pred3} = \alpha\Delta E_{vdw\_mut} + \beta\Delta E_{coul\_mut} + \gamma\Delta G_{solv\_mut} + \delta\Delta G_{vac\_mut} + \varepsilon\Delta SA_{mut} + \epsilon\Delta E_{vdw\_wt} + \zeta\Delta E_{coul\_wt} + \kappa\Delta G_{solv\_wt} + \lambda\Delta G_{vac\_wt} + \mu\Delta SA_{wt} + \omega$$

$$\Delta\Delta G_{pred4} = \alpha\Delta E_{vdw\_mut} + \beta\Delta E_{coul\_mut} + \gamma\Delta G_{solv\_mut} + \delta\Delta G_{vac\_mut} + \varepsilon\Delta SA_{mut} + \epsilon\Delta E_{vdw\_wt} + \zeta\Delta E_{coul\_wt} + \kappa\Delta G_{solv\_wt} + \lambda\Delta G_{vac\_wt} + \mu\Delta SA_{wt} + \nu\Delta\Delta G_{BM} + \xi\Delta\Delta G_{FD} + \omega$$

ΔE_{vdw} - Van der Waals interaction between proteins, calculated as a difference between energies of complex and each monomer, equation (2)

ΔE_{coul} - Coulomb electrostatic interaction, calculated as a difference between energies of complex and each monomer, equation (2)

ΔG_{solv} - Polar solvation energy of solute in water obtained from Poisson-Boltzmann equation

ΔG_{vac} - Polar solvation energy of solute in vacuum obtained from Poisson-Boltzmann equation

ΔSA - Interface area of complex

$\Delta\Delta G_{BM}$: Changes of binding energy between mutant and wild type obtained by BeAtMuSiC;

$\Delta\Delta G_{FD}$: Changes of binding energy between mutant and wild type obtained by FoldX.

Definitions of regions for different locations of mutations for Figure 3. COR: ΔrASA > 0 & rASAm > 25% & rASAc < 25%; RIM: ΔrASA > 0 & rASAc > 25%; SUP: ΔrASA > 0 & rASAm < 25%; INT: rASAc < 25% & ΔrASA = 0; SUR: rASAc > 25% & ΔrASA = 0. ΔrASA = rASAm – rASAc; rASAm = relative ASA in monomer; rASAc = relative ASA in complex.[3]

Definition of standardized regression coefficients: Each variable can be standardized by subtracting its mean and dividing by the standard deviation. Standardization of coefficients is usually done to answer the question, which of the independent variables has a greater effect on the dependent variable in a multiple regression analysis.

Supplemental References

(1) Li, M.; Zheng, W.: All-Atom Structural Investigation of Kinesin–Microtubule Complex Constrained by High-Quality Cryo-Electron-Microscopy Maps. *Biochemistry (Mosc.)* **2012**, *51*, 5022-5032.

(2) Li, M.; Zheng, W.: All-Atom Molecular Dynamics Simulations of Actin–Myosin Interactions: A Comparative Study of Cardiac α Myosin, β Myosin, and Fast Skeletal Muscle Myosin. *Biochemistry (Mosc.)* **2013**, *52*, 8393-8405.

(3) Levy, E. D.: A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J. Mol. Biol.* **2010**, *403*, 660-670.