Supplemental Information

Supplemental Figures

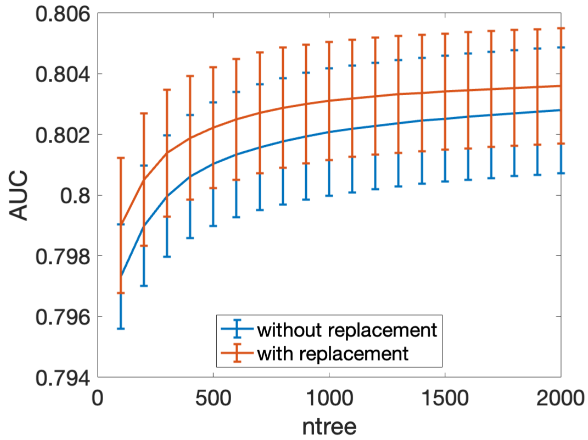

Figure S1. Random forest number of trees optimization, based on cross validation. AUC is shown vs. number of trees. Error bars represent the standard error. The number of trees selected (500) improved performance by 0.05% over the previous value.

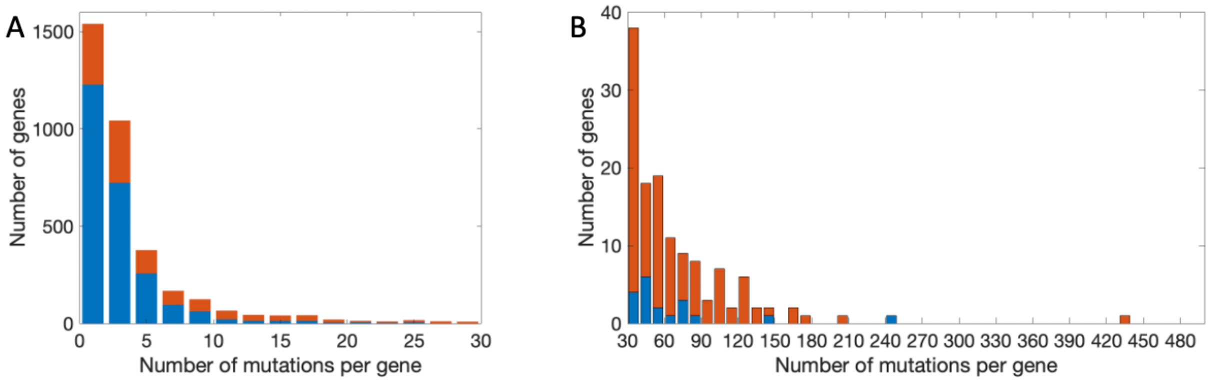

Figure S2. Distributions of numbers of mutations per gene, for benign (blue, lower) and pathogenic (orange, upper) mutations. A) Leftmost portion of distribution. B) Rightmost portion of distribution.

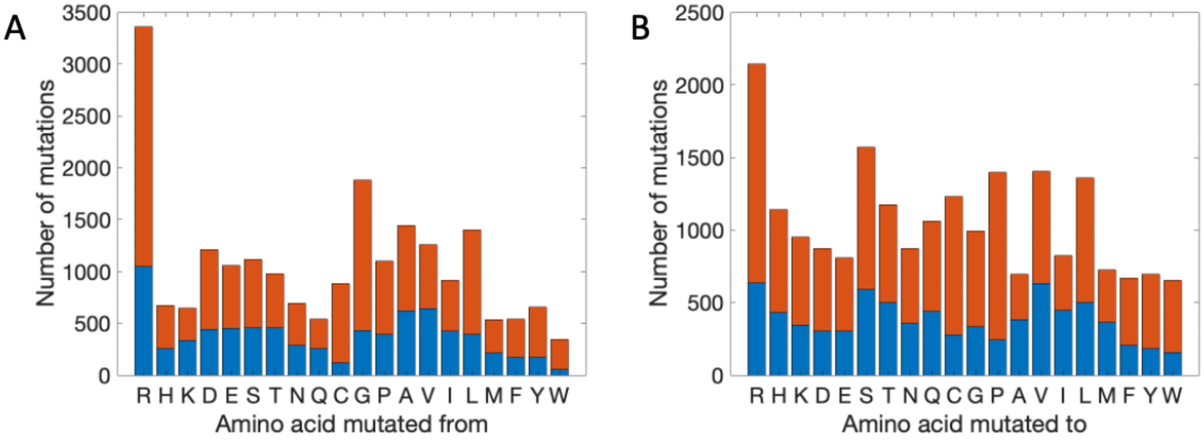

Figure S3. Number of mutations for each residue type, for benign (blue, lower) and pathogenic (orange, upper) mutations. A) Amino acid mutated from (WT residue type). B) Amino acid mutated to (mutant residue type).

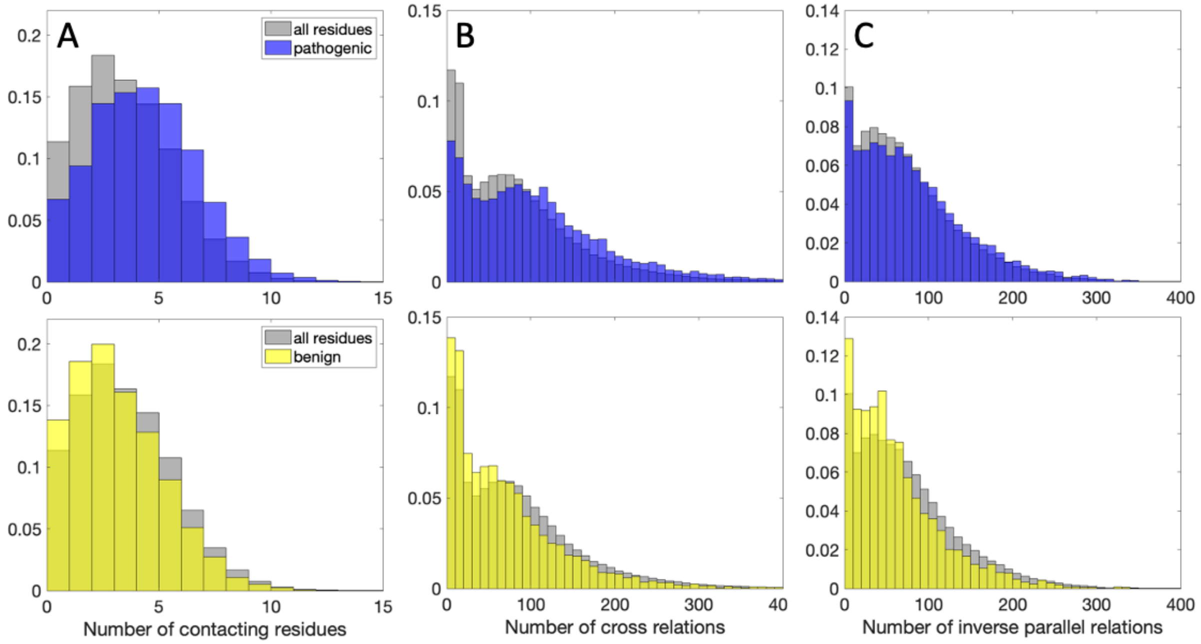

Figure S4. Histograms of topological information for mutations from the ADDRESS database. Blue (foreground, top) indicates pathogenic mutations, yellow (foreground, bottom) indicates benign mutations, and gray (background, both panels) indicates all residues within proteins in the database. A) Number of residues in contact with the mutated residue (6 or more atom-atom contacts within 5 Angstroms). B) Number of contacts in cross relation with a contact formed with the mutated residue. C) Number of contacts in inverse parallel relation with a contact formed with the mutated residue. Additional features are shown in Figure S1.

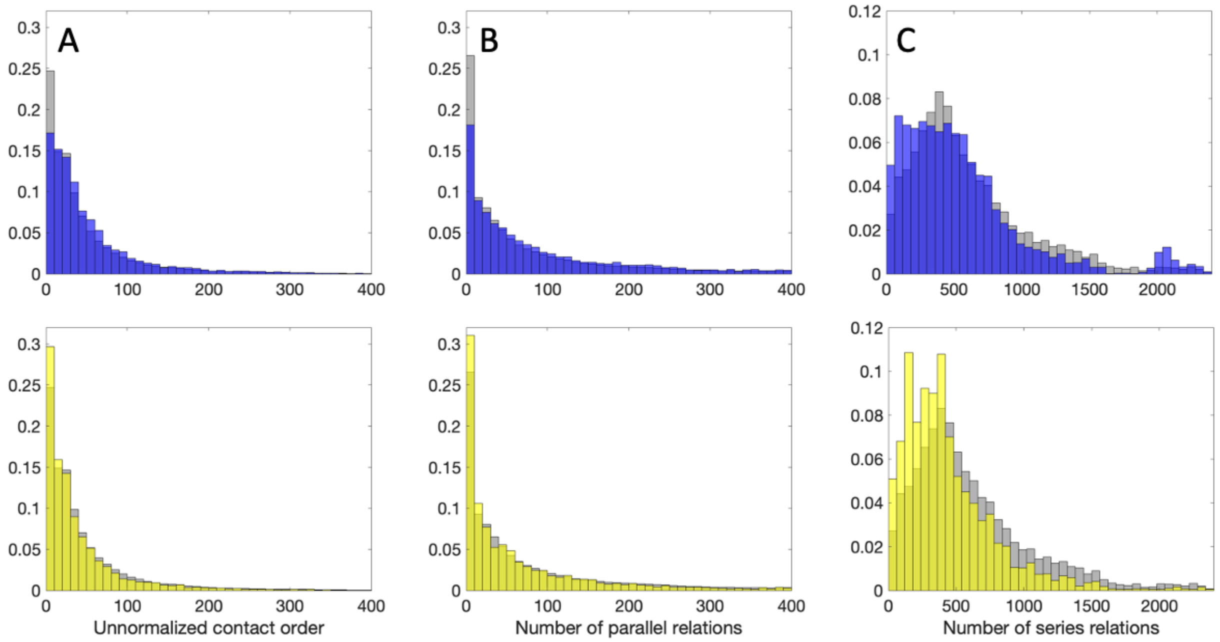

Figure S5. Histograms of topological information for mutations. Colors are as in Figure S2. A) Local contact order, defined as the mean interchain distance to contacting residues. B) Number of contacts in parallel relation with a contact formed with the mutated residue. C) Number of contacts in series relation with a contact formed with the mutated residue.

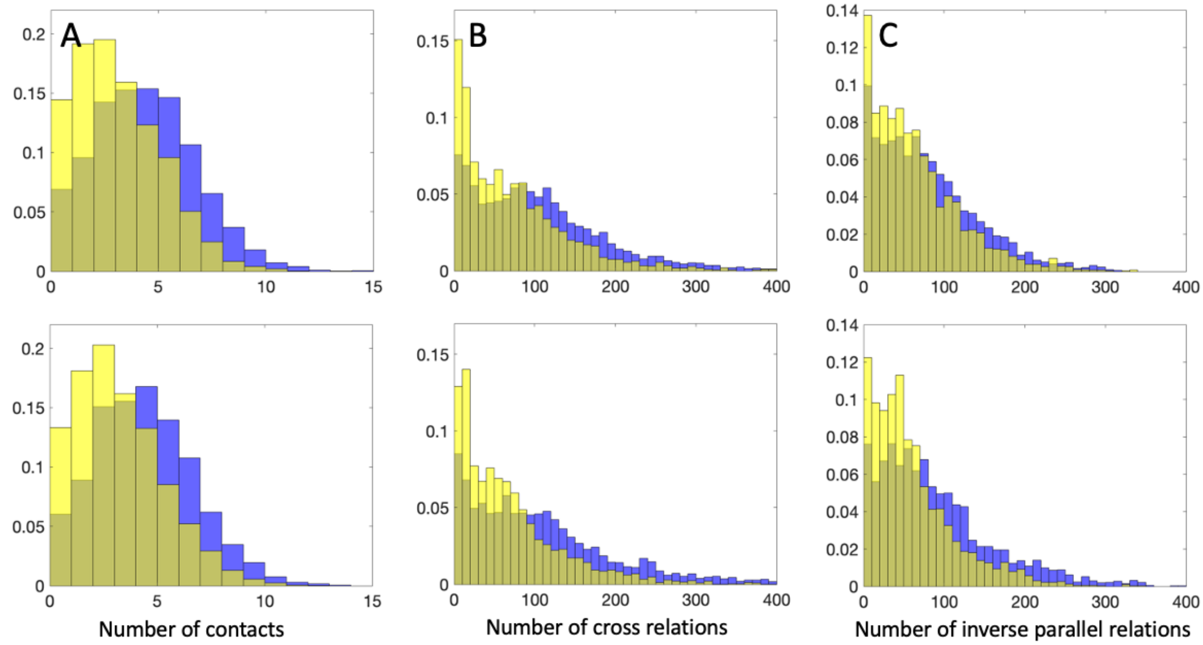

Figure S6. Topological measures for proteins in essential and non-essential genes. Top: essential genes. Bottom: non-essential genes. A) Number of residues in contact with the mutated residue. B) Number of contacts in cross relation with a contact formed with the mutated residue. C) Number of contacts in inverse parallel relation with a contact formed with the mutated residue.

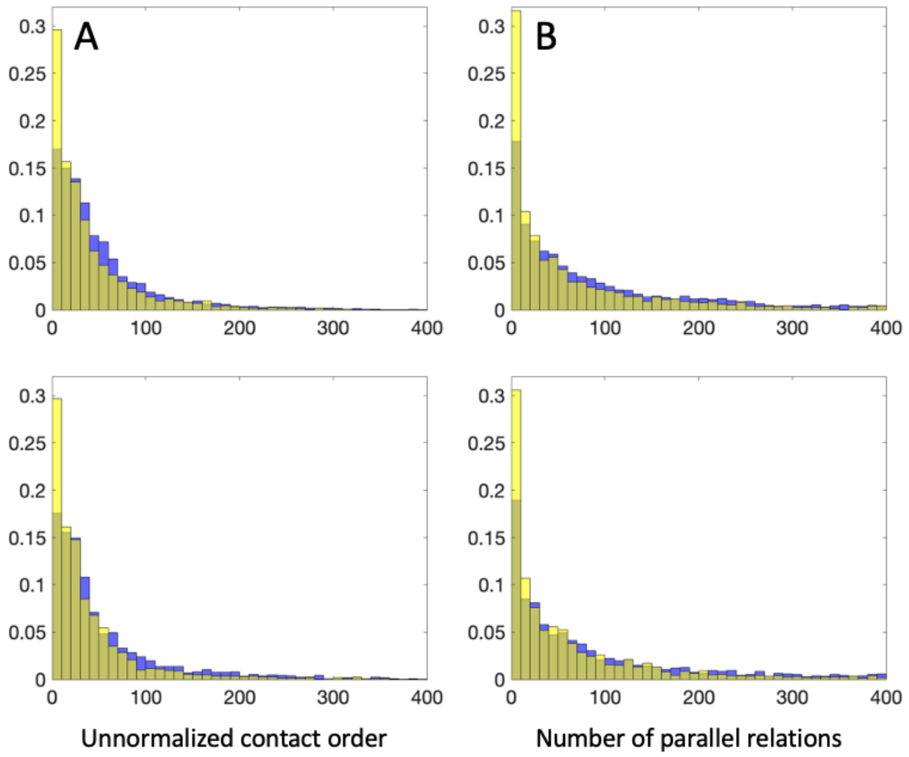

Figure S7. Topological measures for proteins in essential and non-essential genes. Top: essential genes. Bottom: non-essential genes. A) Local contact order. B) Number of contacts in parallel relation with a contact formed with the mutated residue.

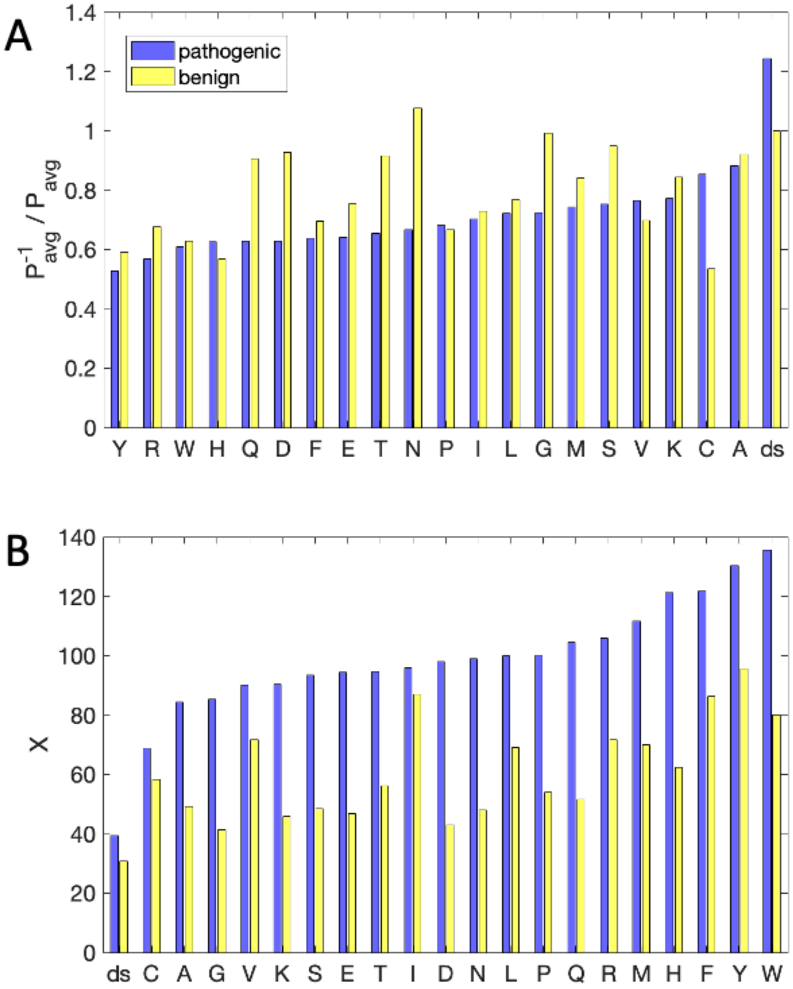

Figure S8. Relations information by residue type. ds: disulfide. Residue types are ordered according to values for pathogenic mutations. A) Mean number of inverse parallel relations divided by mean number of parallel relations. B) Mean number of cross relations.

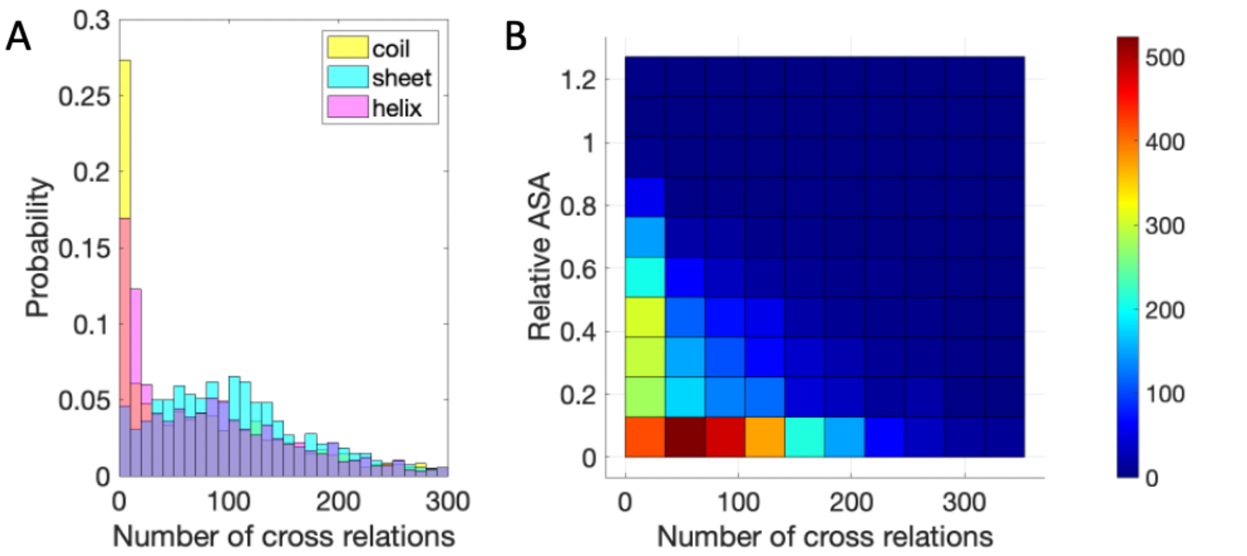

Figure S9. Comparison of cross relation to structural features from MISCAST database. A) Histogram of number of cross relations for residues exhibiting coil, sheet, and helix secondary structures. T-test p-value for helix to coil: 1x10^{-7}, helix to sheet: 0.03, coil to sheet: 9x10^{-12}. B) Two-dimensional histogram showing relative accessible surface area vs. number of cross relations. r = -0.35, p = 6x10^{-144}.

Supplemental Tables

Table S1. Numbers of genes and mutations with specific GO annotations

	GO_term_ID	GO_term_def	NGenes	NMutations
1	GO:0043167	ion binding	1494	13084
2	GO:0019899	enzyme binding	599	4940
3	GO:0003677	DNA binding	380	2530
4	GO:0016301	kinase activity	319	2183
5	GO:0003723	RNA binding	302	1356
6	GO:0016491	oxidoreductase activity	272	3252
7	GO:0030234	enzyme regulator activity	257	1542
8	GO:0008092	cytoskeletal protein binding	213	1583
9	GO:0008289	lipid binding	210	1934
10	GO:0008233	peptidase activity	174	1359
11	GO:0008134	transcription factor binding	163	1190
12	GO:0003700	DNA-binding transcription factor activity	162	1142
13	GO:0005198	structural molecule activity	108	898
14	GO:0022857	transmembrane transporter activity	107	929
15	GO:0016887	ATPase activity	97	1043
16	GO:0004518	nuclease activity	75	336

17	GO:0016791	phosphatase activity	73	345
18	GO:0016829	lyase activity	73	663
19	GO:0008168	methyltransferase activity	65	358
20	GO:0003924	GTPase activity	58	499
21	GO:0042393	histone binding	54	178
22	GO:0016874	ligase activity	54	465
23	GO:0016757	transferase activity transferring glycosyl groups	52	500
24	GO:0016853	isomerase activity	52	606
25	GO:0030674	protein binding bridging	51	438
26	GO:0016810	hydrolase activity acting on carbon-nitrogen (but not peptide) bonds	51	279
27	GO:0016746	transferase activity transferring acyl groups	50	292
28	GO:0016798	hydrolase activity acting on glycosyl bonds	44	1121
29	GO:0016779	nucleotidyltransferase activity	43	351
30	GO:0004386	helicase activity	38	201
31	GO:0003729	mRNA binding	35	200
32	GO:0032182	ubiquitin-like protein binding	29	144
33	GO:0051082	unfolded protein binding	28	95
34	GO:0016765	transferase activity transferring alkyl or aryl (other than methyl) groups	25	272
35	GO:0008135	translation factor activity RNA binding	10	15
36	GO:0019843	rRNA binding	7	31
37	GO:0003735	structural constituent of ribosome	3	10
38	GO:0030555	RNA modification guide activity	0	0
39	GO:0030533	triplet codon-amino acid adaptor activity	0	0

Table S2. Significance of difference in means of topological relations (T-test), for pathogenic vs. benign mutations. Relations, contact order, and long-range order are not normalized by chain length.

	p-value	**p-value, lower distance cutoff**
Number of contacts	2x10^{-298}	2x10^{-257}
Cross relations	8x10^{-195}	1x10^{-163}
Parallel relations	2x10^{-66}	2x10^{-54}
Inverse parallel relations	3x10^{-99}	4x10^{-88}
Series relations	4x10^{-58}	6x10^{-52}
Local contact order	9x10^{-43}	5x10^{-35}
Long range order	5x10^{-259}	1x10^{-226}

Table S3. Random Forests performance for pathogenicity prediction, for a range of ntree values

num features	ntree	features					AUC
2	50	FoldX	essential				0.689
2	200	X	essential				0.692
2	500	FoldX	essential				0.693
2	1000	X	essential				0.694
3	50	aa1	ncontacts	essential			0.738
3	200	aa1	ncontacts	essential			0.741
3	500	aa1	X	essential			0.741
3	1000	aa1	ncontacts	essential			0.742
4	50	aa1	X	P-1	essential		0.769
4	200	aa1	X	position	essential		0.769
4	500	aa1	X	position	essential		0.770
4	1000	aa1	X	position	essential		0.769
5	50	aa1	ncontacts	P-1	position	essential	0.787
5	200	aa1	ncontacts	P-1	position	essential	0.792
5	500	aa1	X	P-1	position	essential	0.794
5	1000	aa1	ncontacts	P-1	positiion	essential	0.793
14	50	all					0.798
14	200	all					0.807
14	500	all					0.807
14	1000	all					0.808

Table S4. Feature importance for Random Forest model

	0	1	Mean Decrease in Accuracy	Mean Decrease in Gini
essential	94.18	86.57	105.11	708.30
P^{-1}	19.52	48.80	64.56	642.51
aa1	92.93	0.62	61.82	941.78
X	22.36	35.96	61.66	699.51
FoldX	44.92	23.21	51.46	851.69
CS	18.01	24.96	46.78	427.11
CO	9.62	31.60	46.54	517.49
EvoEF	44.19	-0.93	45.80	734.39
aa2	78.49	-18.73	44.82	976.39
CP	14.61	26.50	43.10	391.61
CP^{-1}	13.58	29.90	41.59	405.54

P	9.78	28.61	39.54	493.03
ncontacts	24.45	16.30	36.83	292.46
LRO	11.85	22.35	31.41	229.85
position	10.54	28.02	28.04	642.13

Data are sorted by mean decrease in accuracy.
Essential: whether or not the gene is essential.
P^{-1}: number of inverse parallel relations
aa1: amino acid mutated from
X: number of cross relations.
FoldX: FoldX ΔΔG
CS: number of concerted parallel relations
CO: unnormalized local contact order
EvoEF: EvoEF ΔΔG
aa2: amino acid mutated from
CP^{-1}: number of concerted inverse parallel relations
CP: number of concerted parallel relations
P: number of parallel relations
ncontacts: number of contacts formed by the mutated residue and other residues
LRO: unnormalied local long range order
position: the position of the residue along the chain divided by chain length