Selection mode	Availability in strategies	Description
Amino acid frequency	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal frequency in the multiple sequence alignment
Mutational landscape	FUNC, FLEX	suggests amino acid residues fulfilling the criterion of minimal probability of preservation of protein function
Sequence consensus	CONS	suggests amino acid residues fulfilling the criteria of at least one of approaches implemented in sequence consensus strategy: (i) majority approach or (ii) frequency ratio approach
Correlated positions	CORREL	suggests amino acid residues fulfilling the criterion of minimal frequency of co-occurrence with some other specific residue from coupled position
Manual	ALL	manual selection of amino acid residues

FUNC – Analysis of functional hot spots; FLEX – Analysis of stability hot spots / structural flexibility approach; CONS – Analysis of stability hot spots / sequence consensus approach; CORREL – Analysis of correlated hot spots

Supplementary Data 1: Design and implementation of predictor of deleteriousness of amino acid substitutions (RAPHYD)

RAPHYD (**RA**pid **PHY**logenetic predictor of **D**eleteriousness) is a newly developed predictor of the impact of amino acid substitutions on protein function (Figure 1). It is inspired by MAPP algorithm (1) which, at each position, quantifies the variation of seven physico-chemical properties selected by expert judgment and calculates the deviation of all possible amino acid replacements from this variation. The lower the deviation, the higher is the probability that a given amino acid substitution will be tolerated at a given position (1). More specifically, RAPHYD employs a multiple-sequence alignment and phylogenetic tree with known branch lengths to estimate the deleteriousness of substitutions. To obtain these required inputs, HotSpot Wizard pipeline is utilized which involves the usage of ClustalΩ (2) and FastTree (3). Based on the topology and branch lengths of the tree, the weights are calculated for each sequence to obtain the phylogenetic correlation among all sequences in the alignment. This is performed by Felsenstein's algorithm (4) that calculates the weighted average of the "best weights" obtained by rooting the tree at the midpoint of each branch. Subsequently, the sequence weights are multiplied by the relative frequency of each amino acid occurring at the analyzed position to obtain "alignment summary" matrix. This summary is interpreted using a matrix of selected physico-chemical properties. Such an interpretation is expressed by multiplication of alignment summary matrix with matrices of physico-chemical properties. Finally, a constraint violation is measured for each position of the sequence. Dissimilarity scale between original and substituting amino acid in combination with the conservation rate (obtained in previous steps) yields the probability that substituting amino acid is neutral.

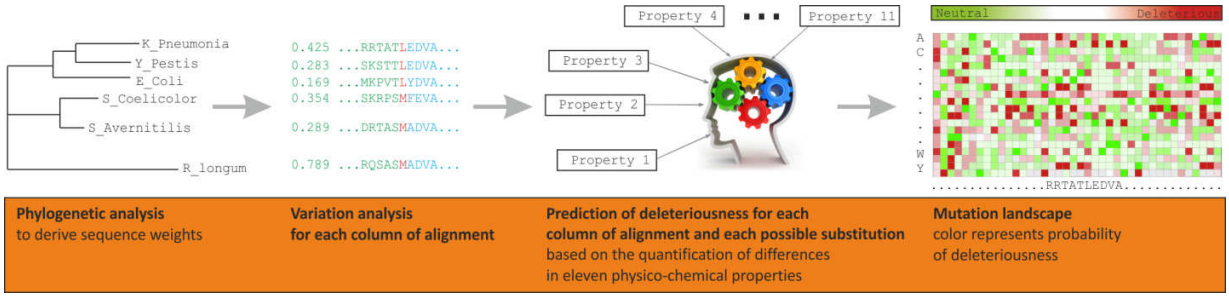

Figure 1. Workflow diagram of RAPHYD.

While the computation core of RAPHYD follows the methodology of MAPP, it differs in the set of utilized physico-chemical and biophysical properties. RAPHYD uses properties chosen from AAindex database (5) by supervised feature selection technique, combining forward selection and backward elimination in several consecutive rounds. This selection was performed on a training dataset collected from nine systematic mutagenesis studies, which contained 5,282 mutations in seven different proteins (Table 1). These mutagenesis studies represent a source of massively mutated proteins and provide reliable data measured under the same laboratory conditions, which made them effective for the following feature selection. These mutations were divided into neutral and deleterious subsets according to Yampolski et al (6). Using this training dataset, the forward selection process was used to select 25 properties with the highest accuracies. This upper bound on a number of properties was set as an acceptable trade-off

between the time demands and the size of a property space for the following backward elimination. With the utilization of the backward elimination, this subspace of properties was re-analyzed by successive reduction to five properties. The forward selection was then processed again starting from the subset of properties, for which the backward elimination reported the highest accuracy. Such rounds combining the forward selection and backward elimination were repeated five times, which led to the final set of eleven parameters (Table 2).

Table 1. Composition of the training dataset.

Protein	Source species	PDB ID	Number of mutations		Reference
			deleterious	neutral	
Interleukin-3	*Homo Sapiens*	1JLI	369	384	(7)
Alpha amylase	*Geobacillus stearothermophilus*	1HVX	442	2,205	(8)
Barnase	*Escherichia coli*	1B2X	33	585	(9)
Reverse transcriptase	HIV	1REV	257	108	(10)
Protease	HIV	3PHV	227	107	(11)
Nuclease	*Staphylococcus aureus*	1EY0	160	109	(12–14)
Protein V	Phage f1	1GVP	200	96	(15)

Table 2. Overview of eleven physico-chemical and biophysical parameters utilized by RAPHYD.

AAindex name	AAindex description	Reference
ROBB760109	Information measure for N-terminal turn	(16)
ROBB760105	Information measure for extended	(16)
OOBM850103	Optimized transfer energy parameter	(17)
MUNV940102	Free energy in alpha-helical region	(18)
OOBM850104	Optimized average non-bonded energy per atom	(17)
RADA880101	Transfer free energy from chx to wat	(19)
WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O	(20)
NADH010101	Hydropathy scale based on self-information values in the 2-state model (ASA=5%)	(21)
CORJ870103	PRIFT index	(22)
MITS020101	Amphiphilicity index	(23)
ENGD860101	Hydrophobicity index	(24)

In order to correctly assess the performance of developed predictor (25), its performance was evaluated using two independent testing datasets. For this purpose, a collection of another four systematic mutagenesis studies including 7,496 mutations was used to construct the testing dataset 1 (Table 3), again employing the same rules for division of mutations as described for training dataset. In contrast to these massively mutated proteins, the testing dataset 2 consisted of a diverse set of 1,100 proteins with only a handful mutations per protein (2,864 mutations in total) extracted from Protein Mutant Database (26). The records with annotations [=], i.e., no change of activity, were considered as neutral, while the records with any other annotations were considered as deleterious. All mutations with conflicting annotations, e.g., [=] and [++] at the same time, were excluded. To ensure the correctness of comparison between the RAPHYD tool and six other established tools, i.e., MAPP (1), PhD-SNP (26),

PolyPhen-1 (27), PolyPhen-2 (28), SIFT (29) and SNAP (30), the mutations present in the training dataset of any of these tools were also excluded from both testing datasets.

Table 1. Composition of the testing dataset 1.

Protein	Source species	PDB ID	Number of mutations		Reference
			deleterious	neutral	
Lactose repressor	*Escherichia coli*	1LBI	879	1,449	(32)
Serine protease	*Bacillus subtilis*	1ST3	1,589	2,331	(33)
Lysozyme	Phage T4	1LYD	329	578	(34)
β-Lactamase	*Pseudomonas aeruginosa*	1DDK	310	31	(35)

On both employed datasets, the RAPHYD tool provided more accurate predictions than the second best tool, which was moreover different for each dataset, i.e., SNAP for testing dataset 1 and PolyPhen-1 for testing dataset 2 (Table 4). Coupled with nearly instantaneous calculations, when the input alignment and tree are provided by HotSpot Wizard workflow, the observed prediction performance makes the RAPHYD tool well suitable for the task of prioritizing amino acid substitutions within the process of library design.

Table 4. Performance of RAPHYD with the testing datasets.

Performance metrics*	Dataset	PhD-SNP	PolyPhen-1	PolyPhen-2	SIFT	SNAP	MAPP	RAPHYD
Accuracy	Testing dataset 1	0.640	0.711	0.697	0.658	0.741	0.703	0.751
	Testing dataset 2	0.632	0.658	0.635	0.652	0.629	0.637	0.660
	Average	**0.636**	**0.684**	**0.666**	**0.655**	**0.685**	**0.670**	**0.706**
Matthews correlation coefficient	Testing dataset 1	0.284	0.419	0.417	0.354	0.478	0.417	0.504
	Testing dataset 2	0.257	0.304	0.297	0.330	0.249	0.288	0.307
	Average	**0.270**	**0.361**	**0.357**	**0.342**	**0.363**	**0.353**	**0.406**
Area under the receiver operating characteristics curve	Testing dataset 1	0.716	0.747	0.708	0.723	0.804	0.825	0.814
	Testing dataset 2	0.673	0.681	0.611	0.678	0.663	0.658	0.707
	Average	**0.694**	**0.714**	**0.659**	**0.701**	**0.734**	**0.741**	**0.761**
Sensitivity	Testing dataset 1	0.786	0.803	0.910	0.923	0.723	0.878	0.697
	Testing dataset 2	0.690	0.676	0.843	0.838	0.629	0.810	0.583
	Average	**0.738**	**0.739**	**0.877**	**0.880**	**0.676**	**0.844**	**0.640**
Specificity	Testing dataset 1	0.493	0.620	0.485	0.393	0.759	0.529	0.805
	Testing dataset 2	0.574	0.639	0.427	0.466	0.630	0.463	0.738
	Average	**0.534**	**0.630**	**0.456**	**0.429**	**0.694**	**0.496**	**0.771**
Percent of evaluated mutations	Testing dataset 1	100	100	100	100	100	99	100
	Testing dataset 2	100	98	99	78	95	94	100
	Average	**100**	**99**	**99**	**89**	**97**	**96**	**100**

* all metrics are calculated with normalized numbers

References

1. Stone,E.A. and Sidow,A. (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, **15**, 978–986.

2. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

3. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One*, **5**, e9490.

4. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

5. Kawashima,S., Ogata,H. and Kanehisa,M. (1999) AAindex: Amino Acid Index Database. *Nucleic Acids Res.*, **27**, 368–369.

6. Yampolsky,L.Y. and Stoltzfus,A. (2005) The exchangeability of amino acids in proteins. *Genetics*, **170**, 1459–1472.

7. Olins,P.O., Bauer,S.C., Braford-Goldberg,S., Sterbenz,K., Polazzi,J.O., Caparon,M.H., Klein,B.K., Easton,A.M., Paik,K. and Klover,J.A. (1995) Saturation mutagenesis of human interleukin-3. *J. Biol. Chem.*, **270**, 23754–23760.

8. Cuevas,W.A., Estell,D.A., Hadi,S.H., Lee,S.-K., Ramer,S.W., Shaw,A., Toppozada,A.R. and Weyler,W. (2011) Geobacillus stearothermophilus alpha-amylase (amys) variants with improved properties.

9. Axe,D.D., Foster,N.W. and Fersht,A.R. (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry (Mosc.)*, **37**, 7157–7166.

10. Wrobel,J.A., Chao,S.F., Conrad,M.J., Merker,J.D., Swanstrom,R., Pielak,G.J. and Hutchison,C.A. (1998) A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 638–645.

11. Loeb,D.D., Swanstrom,R., Everitt,L., Manchester,M., Stamper,S.E. and Hutchison,C.A. (1989) Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.

12. Green,S.M., Meeker,A.K. and Shortle,D. (1992) Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. *Biochemistry (Mosc.)*, **31**, 5717–5728.

13. Meeker,A.K., Garcia-Moreno,B. and Shortle,D. (1996) Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry (Mosc.)*, **35**, 6443–6449.

14. Shortle,D., Stites,W.E. and Meeker,A.K. (1990) Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry (Mosc.)*, **29**, 8033–8041.

15. Zabin,H.B., Horvath,M.P. and Terwilliger,T.C. (1991) Approaches to predicting effects of single amino acid substitutions on the function of a protein. *Biochemistry (Mosc.)*, **30**, 6230–6240.

16. Robson,B. and Suzuki,E. (1976) Conformational properties of amino acid residues in globular proteins. *J. Mol. Biol.*, **107**, 327–356.

17. Kubota,M., Ooi,Y., Obatake,M., Kubota,Y. and Om,T. (1985) Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteins. *Bull Inst Chem Res Kyoto Univ*, **63**, 82–94.

18. Muñoz,V. and Serrano,L. (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins*, **20**, 301–311.

19. Radzicka,A. and Wolfenden,R. (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry (Mosc.)*, **27**, 1664–1670.

20. Wilce,M.C.J., Aguilar,M.-I. and Hearn,M.T.W. (1995) Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Anal. Chem.*, **67**, 1210–1219.

21. Naderi-Manesh,H., Sadeghi,M., Arab,S. and Moosavi Movahedi,A.A. (2001) Prediction of protein surface accessibility with information theory. *Proteins*, **42**, 452–459.

22. Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**, 659–685.

23. Mitaku,S., Hirokawa,T. and Tsuji,T. (2002) Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*, **18**, 608–616.

24. Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.

25. Smialowski,P., Frishman,D. and Kramer,S. (2010) Pitfalls of supervised feature selection. *Bioinformatics*, **26**, 440–443.

26. Kawabata,T., Ota,M. and Nishikawa,K. (1999) The Protein Mutant Database. *Nucleic Acids Res.*, **27**, 355–357.

27. Capriotti,E., Calabrese,R. and Casadio,R. (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, **22**, 2729–2734.

28. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

29. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

30. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

31. Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.

32. Markiewicz,P., Kleina,L.G., Cruz,C., Ehret,S. and Miller,J.H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as 'spacers' which do not require a specific sequence. *J. Mol. Biol.*, **240**, 421–433.

33. Aehle,W., Cascao-Pereira,L.G., Estell,D.A., Goedegebuur,F., Jr,J.T.K., Poulose,A.J. and Schmidt,B.F. (2010) Compositions and methods comprising serine protease variants.

34. Rennell,D., Bouvier,S.E., Hardy,L.W. and Poteete,A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, **222**, 67–88.

35. Materon,I.C. and Palzkill,T. (2001) Identification of residues critical for metallo-beta-lactamase function by codon randomization and selection. *Protein Sci. Publ. Protein Soc.*, **10**, 2556–2565.

Supplementary Data 2: Effectiveness of the mutagenesis targeting the highly mutable positions

The strategy for detection of functional hot spots relies on HotSpot Wizard's prediction of safe mutations outside the positions essential for protein structure or function. The effectivity of eliminating truly conserved positions was assessed by comparing HotSpot Wizard results with the data from systematic mutagenesis studies (Figure 1). The mutations were classified according to their viability into three categories: deleterious, slightly deleterious, and neutral/positive (1). Table 1 shows the comparison between the proportion of deleterious mutations at positions predicted to be highly mutable by HotSpot Wizard (mutability grades 6-9) and the proportion of deleterious mutations within the entire protein. The obtained results suggest that the mutagenesis targeting the highly mutable positions is nearly four times less likely to lead to the loss of function. Figure 2 shows the trend of increasing ratio of deleterious mutations with decreasing mutability.

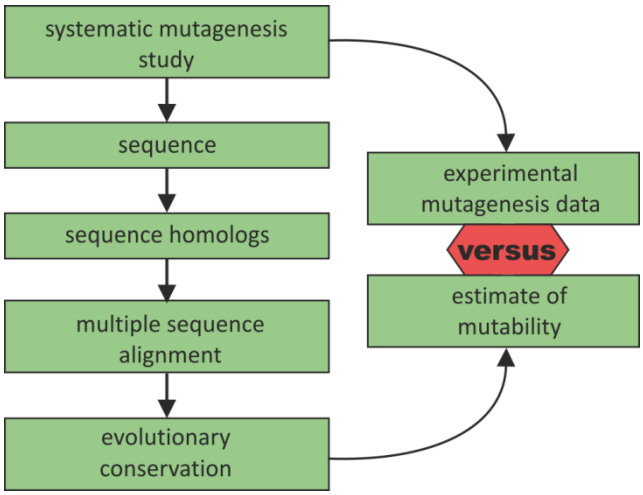

Figure 1. Design of validation experiments.

Table 1. Summary of employed systematic mutagenesis studies.

Protein	Source species	PDB ID	# mutations	Proportion of deleterious mutations		Reference
				region of hot spots	entire protein	
Serine protease	*Bacillus subtilis*	1ST3	4,199	11.3 %	39.5 %	(2)
Alpha amylase	*Geobacillus stearothermophilus*	1HVX	2,666	9.3 %	16.7 %	(3)
Lysozyme	Phage T4	1LYD	1,918	0.6 %	9.1 %	(4)
Lactose repressor	*Escherichia coli*	1LBI	4,038	6.0 %	25.8%	(5)
Barnase	*Escherichia coli*	1B2X	676	0.4 %	4.9 %	(6)

Protease	HIV	3PHV	336	35.2 %	47.3 %	(7)
Reverse transcriptase	HIV	1REV	366	8.1 %	31.1 %	(8)
Weighted average			**14,199**	**7.8 %**	**26.4 %**	

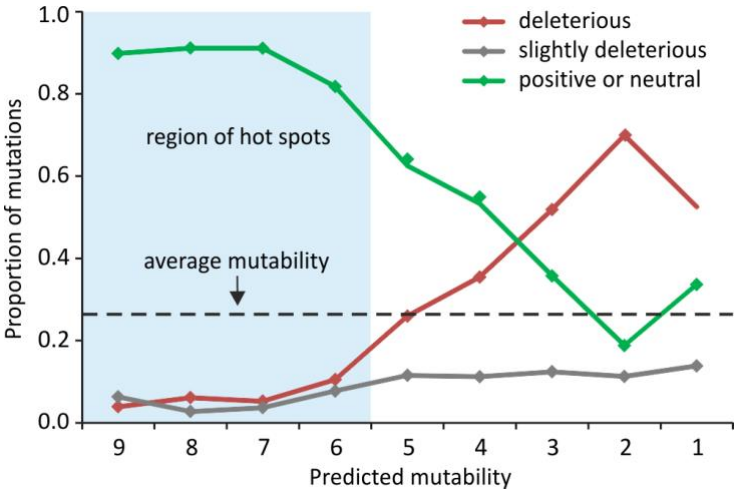

Figure 2. Proportion of deleterious, slightly deleterious and positive/neutral mutations as a function of predicted mutability degree at corresponding protein positions.

References

1. Yampolsky,L.Y. and Stoltzfus,A. (2005) The exchangeability of amino acids in proteins. *Genetics*, **170**, 1459–1472.

2. Aehle,W., Cascao-Pereira,L.G., Estell,D.A., Goedegebuur,F., Jr,J.T.K., Poulose,A.J. and Schmidt,B.F. (2010) Compositions and methods comprising serine protease variants. *Patent EP2647692 A3*.

3. Cuevas,W.A., Estell,D.A., Hadi,S.H., Lee,S.-K., Ramer,S.W., Shaw,A., Toppozada,A.R. and Weyler,W. (2011) *Geobacillus stearothermophilus* alpha-amylase (amys) variants with improved properties. *EP2623591 A3*.

4. Rennell,D., Bouvier,S.E., Hardy,L.W. and Poteete,A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, **222**, 67–88.

5. Markiewicz,P., Kleina,L.G., Cruz,C., Ehret,S. and Miller,J.H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as 'spacers' which do not require a specific sequence. *J. Mol. Biol.*, **240**, 421–433.

6. Axe,D.D., Foster,N.W. and Fersht,A.R. (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry*, **37**, 7157–7166.

7. Loeb,D.D., Swanstrom,R., Everitt,L., Manchester,M., Stamper,S.E. and Hutchison,C.A. (1989) Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.

8. Wrobel,J.A., Chao,S.F., Conrad,M.J., Merker,J.D., Swanstrom,R., Pielak,G.J. and Hutchison,C.A. (1998) A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 638–645.