Supporting Information

Computational Design of Stable and Soluble Biocatalysts

Milos Musil[1,2,3,#], Hannes Konegger[1,3,#], Jiri Hon[1,2,3,#], David Bednar[1,3], Jiri Damborsky[1,3,*]

[1] Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic;

[2] IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 61266 Brno, Czech Republic;

[3] International Clinical Research Center, St. Anne's University Hospital, Pekarska 53, 656 91 Brno, Czech Republic

# These authors contributed equally
* The author for correspondence: jiri@chemi.muni.cz

Table S1. Datasets for prediction of protein stability.

Dataset	Stabilizing/Neutral	Destabilizing	Proteins	Source
S238[1]	45	193	25	ProTherm (Feb 2013)
S1676[1]	453	1,223	70	ProTherm (Feb 2013)
S2648[2]	568	2,080	131	ProTherm
S350[2]	90	260	67	ProTherm
S2155[3]	NA	NA	79	ProTherm (Dec 2004)
S3366[4]	836	2,530	NA	Prethermut
S1480[5]	464	1,016	NA	NA
S1859[6]	NA	NA	64	NA
S1210[7]	NA	NA	NA	NA
S595[8]	NA	NA	NA	NA
S918[9]	NA	NA	27	NA
S3421[10]	NA	NA	150	NA
S1615[11]	462	1,153	42	ProTherm
S388[11]	44	340	17	ProTherm
S1573[12]	315	1,258	93	ProTherm
S1925[13]	NA	NA	55	NA
S3463[14]	NA	NA	NA	NA
S1948[15]	NA	NA	NA	NA
S1765[16]	NA	NA	NA	NA
S1538[17]	NA	NA	NA	NA
S1603[17]	NA	NA	NA	NA
S1626[4]	461	1165	93	ProTherm (in part)
S2399[18]	NA	NA	113	ProTherm
Trudeau[19]	34	231	1	Experimental

NA – information was not available in the article

Table S2. Software tools for prediction of protein stability.

Method	Basis of prediction	Availability	Input	Output	Mutations	Dataset	Validation
Machine learning							
EASE-MM[1]	SVM	web	sequence	ddG	single	S1676	10-fold crossvalidation
MuStab[2]	SVM	web (unavailable)	sequence	binary + confidence	single	S1480	5-fold crossvalidation
ProMaya[3]	Random forest	web	structure	ddG	single	S2648, S2155	5 and 10-fold crossvalidation
mCSM[4]	Graph based	web	structure	ddG	single	S2648, S350, S1925	5 and 10-fold crossvalidation
ELASPIC[5]	SVM + HMM	web	structure	ddG	single/multiple	S3463	20-fold crossvalidation
MuPro[6]	SVM	web	seq/struct	ddG	single	S1615	20-fold crossvalidation
I-Mutant2.0[7]	SVM	web	seq/struct	ddG	single	S1948	crossvalidation
TopologyNet[8]	Deep learning	web	structure	ddG	single	S2648, S350	5-fold crossvalidation
PROTS-RF[9]	Random forest	SA	structure	ddG	single/multiple	S2155	5-fold crossvalidation
MAESTRO[10]	ANNs + SVM + multiple linear regression + statistical potentials	SA/web	structure	ddG + confidence	single/multiple, disulfide bridges	S2648, S350, S1925, S1765	5/10/20-fold crossvalidation and performance test
Iptree-stab[11]	Decision tree	web (unavailable)	partial sequence	binary	Single	S1859	4/10/20-fold crossvalidation
INPS-MD[12]	Support Vector Regression	web	sequence	ddG	Single	S2648	10-fold crossvalidation
iStable[13]	SVM	web	structure	ddG	Single	S2648, S1948	5-fold crossvalidation
Prethermut[14]	SVM + RF	SA	structure	ddG	single/multiple	S3366	10-fold crossvalidation
Force field calculations							
PopMusic[15]	SEEF	web	structure	ddG	single	S2648	5-fold crossvalidation
FoldX[16]	SEEF	SA	structure	ddG	single	NA	NA
CUPSAT[17]	Atom potentials and torsion angles	web	structure	ddG	single	S1538, S1603	3/4/5-fold crossvalidation
Rosetta[18]	PEEF	SA	structure	ddG	single/multiple	S1210	20-fold crossvalidation
ERIS[19]	PEEF	SA	structure	ddG	single	S595	crossvalidation
CC/PBSA[20]	PEEF	SA	structure	ddG	single	NA	5-fold crossvalidation
DMutant[21]	Amino acid potentials and torsion angles	SA	structure	ddG	single	S918	independent
SDM[22]	SEEF	web	structure	ddG	single	S2648, S350	independent
HotMusic[23]	SEEF	web	structure	dTm	single	S1626	5-fold crossvalidation
STRUM[24]	SEEF	SA/web	structure	ddG	single	S3421	5-fold crossvalidation

AUTO-MUTE[25]	SEEF/ML	SA	structure	binary/ddG	single	NA	NA
Phylogenetic analysis							
HotStopWizard[26]	CA	web	seq/struct	hotspots	single/multiple	NA	NA
FastML[27]	ML	web	MSA + tree	seq	multiple	Protein sequence databases such as UniProt	NA
RaXML[28]	ML	SA/web	MSA	phylogeny	multiple		NA
MLGO[29]	ML	web	MSA + tree	seq + phylogeny	multiple		NA
Ancestors[30]	ML	web (unavailable)	MSA + tree	seq + PP	multiple		NA
PARANA[31]	MP	SA	MSA + tree	biological networks	multiple		NA
HandAlign[32]	BA	SA	MSA + tree	seq + PP + phylogeny	multiple		NA
TreeTime[33]	BA	SA	MSA + tree	seq + PP + phylogeny	multiple		NA
PAML[34]	ML	SA	MSA + tree	seq + PP + phylogeny	multiple		NA
PhyloBot[35]	ML	web	MSA	seq + PP + phylogeny	multiple		NA
MaxAlike[36]	ML	web	MSA + tree	seq + PP + seq logo	multiple		NA
Hybrid methods							
FireProt[37]	Evolution + energy	web	structure	mutations + ddG	multiple	S1573	performance test
PROSS[38]	Evolution + energy	web	structure	mutations	multiple	Trudeau	NA
FRESCO[39]	Evolution + energy	SA	structure	mutations	multiple	experimental	Experimental
Other methods							
pStab[40]	Equilibrium thermodynamics fitting on Wako–Saito–Muñoz–Eaton model	web	structure	unfolding curves	charged residues	NA	NA
Encom[41]	Normal mode analysis	web (unavailable)	structure	ddG	single		
Neemo[42]	Residue interaction networks	web	structure	ddG	single	S2399	independent

SA – Stand alone; CA – Conservation analysis; ML – Maximum likelihood; PEEF – Physical force-field; SEEF – Statistical force-field; MP – Maximum parsimony; BA – Bayessian; NA – Information not available in the article; PP = Posterior Probabilities; Characteristics of datasets is provided in Table S1; Method – hyperlinks refer to the web pages of the method

Table S3. Datasets for prediction of protein solubility.

Name	Description	Contents	AV	Advantages	Disadvantages	Value	Method	PS
Protein sequences								
eSOL[20,21]	Solubility of entire ensemble *E.coli* proteins individually synthesized by PURE system	4,132 proteins	Y	highly consistent dataset, solubility value in %, effect of chaperones	only *E.coli* proteins, in vitro system, low number of negative samples (26 cytosolic proteins), especially after chaperones added	0-100 %	Ratio of supernatant and non-centrifuged protein fraction	Y
TargetTrack[22]	Data from Protein Structure Initiative project. Previously known as PepcDB or TargetDB.	297,404 proteins, 961,548 trials	Y	the largest source of experimental data, description of experimental protocols used	low-quality trial annotations, especially of unsuccessful trials, solubility might be either over- or underestimated depending on extraction method, unreliable annotation of expression system, strict database pre-processing can significantly reduce database size	No explicit value, binary solubility has to be deduced from trial status	Mixed	N
NESG[23]	Subset of TargetTrack. Results from high-throughput platform developed by North East Structural Genomics Consortium.	9,644 proteins	Y*	consistent data from uniform protein production pipeline of the NESG	created between 2001 and 2008 in the first PSI project phase - the high throughput pipeline might not reflect current advances in experimental methods	Integer score from 0 to 5	Yield in supernatant after low-speed centrifugation	Y
HGPD[24,25]	Data from genome-scale experiment to assess the overexpression and the solubility of human full-length cDNA in an *in vivo E. coli* expression system and a wheat germ cell-free expression system	5,100 proteins expressed in *E.coli*, 2,932 proteins expressed in wheat germ cell-free system, 289 proteins expressed in *Brevibacillus*	N	consistent expression and solubility data from uniform pipeline, DNA-level information	only human cDNA	Binary	Detection of specific activities of the 14 C-Leu and 35 S-Met radioisotopes. Binary solubility based on ratio of signal intensity of soluble fraction and signal intensity of whole sample	Y
Periscope[26]	Solubility of proteins expressed in periplasm of *E. coli*.	98 proteins	Y	unique data on expression in *E. coli* periplasm.	very small dataset	Three state: low, medium, high	Literature search	N
AMYPdb[27]	Online database dedicated to amyloid precursor families and to their amino acid sequence signatures.	12,069 proteins, 6,454 patterns	Y	amyloid sequence patterns derived from known amyloid families	not actively maintained and enriched, result of database mining	Binary	Literature search, keyword mining in UniProtKB, extraction of PROSITE motifs	N

Protein fragments								
AmylHex & AmylFrag[28]	A data set of six-residue peptides including positive and negative examples of fibril formation	158 hexapeptides, 45 amyloidogenic protein fragments	Y	one of the first sets of fibril-forming fragments	strong overrepresentation (51%) of point mutations of the amyloidogenic hexapeptide STVIIE	Binary	Literature search	N
WALTZ-DB[29]	Experimentally verified amyloidogenic hexapeptides	1089 peptides	Y	many samples experimentaly validatated by authors	only 244 amyloidogenic peptides	Binary	Fourier Transform Infrared Spectroscopy, Proteostat Dye Binding, Transmission Electron Microscopy, FoldX Modelling of Structural Zipper class	Y
AmyLoad[30]	Amyloidogenic and non amyloidogenic protein fragments, experimentally or computationally characterized.	1481 protein fragments	Y	aggregated from various datasets, additional manual curation and references	only 444 amyloidogenic fragments	Binary	Data selected from WALTZ-DB, AmylHex, AmylFrag and validation datasets of AGGRESCAN and TANGO, detailed information obtained by manual inspection of over 90 publications	N
HPA[31]	Data from high-throughput screening of human protein fragments used for antibody screening (Protein Epitope Signature Tags - PrESTs). Part of Human Protein Atlas project.	16,082 protein fragments ranging from 20 to 150 amino acids	Y	consistent high-throughput expression and solubility data, DNA-level information	only human protein fragments, fragmentation prevents folding into globular protein	Integer score from 0 to 5.	Protein concentration after separating protein precipitate using centrifugation	Y
CPAD[32]	Amyloid peptides and aggregation rates upon mutations. Amyloid peptides with known structure. Verified aggregation prone regions.	1,681 peptides 2,356 agg. rate changes upon mutation, 76 agg. prone regions (APR)	Y	unique resource for validating mutation effect on protein aggregation	no clear database structure, not easily downloadable	Binary amyloidogenicity, continuous aggregation rate	Literature search, other data taken from GAP dataset, WALTZ-DB, PDB	N
Protein variants								
OptSolMut[33]	Mixed single-point and multi-point protein variants.	137 variants of 19 proteins.	Y	multi-point mutations, nearly balanced amount of positive and negative samples	small dataset	Binary	Literature search	N
CamSol[34]	Mixed single-point and multi-point protein variants.	56 variants of 19 proteins.	Y	multi-point mutations	very small dataset, only three mutation decreasing solubility	Three levels, '-', neutral, '+'	Literature search	N
PON-Sol[35]	Single-point protein variants	443 variants of 71 proteins	Y	unique resource for validating mutation effect on protein solubility	small dataset, 222 mutations with no effect, only 85 increasing solubility and 136 decreasing solubility	Five levels: '--', '-', neutral, '+', '++'	Literature search	N

AV – Availability; PS – Primary source; *Available only at request; Name – hyperlinks refer to the web pages of the dataset

Table S4. Software tools for prediction of protein solubility.

Method	Approach	Type[a]	Availability[b]	Input	Output	Dataset source	Dataset size	Validation[c]
Protein sequence solubility								
Revised Wilkinson-Harrison[36,37]	Discriminant analysis	ML	Equation	Sequence	Propensity	own experiments	81 proteins	no independent test set, ACC 88 %
SOLpro[38]	Two-layer SVM	ML	SA - Linux, web	Sequence	Propensity	TargetTrack, SwissProt, PDB	17,408 proteins	10-fold crossvalidation, MCC 0.487, ACC 60%, MCC 0.20 on newer test set[39]
PROSO II[40]	Logistic regression, Parzen window	ML, SS	web	Sequence	Propensity	TargetTrack, PDB	82,299 proteins	10-fold cross-validation, MCC 0.421, ACC 64%, MCC 0.34 on newer test set[39]
ESPRESSO[24]	SVM	ML, SP	web	Sequence, expression system	Propensity, binary decision, mutations increasing solubility	HGPD	5,100 proteins (*E. coli* expression system) 2,932 proteins (wheat germ cell-free expression system) 289 (*Brevibacillus* expression system)	MCC 0.42 for property-based solubility in E.coli
ccSOL omics[41,42]	SVM	ML	web	Sequence	Propensity, profile	TargetTrack	36,990 proteins	10-fold cross-validation, ACC 78%
Periscope[26]	SVM	ML	web	Sequence	Propensity	literature	98 proteins expressed in periplasm of *E. coli*	independent test set of 15 proteins ACC 78%, PC 0.77
Protein-Sol[43]	Linear regression	ML	web	Sequence	Propensity	eSOL	2,395 proteins	no independent solubility test set, ACC 90% on train set
DeepSol[44]	CNN	ML	SA - Python	Sequence	Propensity	PROSO II unfiltered set	69,420 proteins	ACC 77%, MCC 0.55
SoluProt[under review]	Random forests	ML	SA - Python	Sequence	Propensity	TargetTrack	10,912 proteins	ACC 58% on independent balanced test set of 3,788 proteins from NESG dataset
Solubility profile								
Zyggregator[45,46]	Linear regression	ML	web	Sequence, pH	Profile	literature	79 variants of 15 proteins	leave-one-out cross-validation, PC 0.91, validated on several case studies
AGGRESCAN[47,48]	Custom regression	ML	web	Sequence	Profile	own experiments	20 AB42 variants at position 19	validated on various protein sets from literature
TANGO[49]	Custom regression and statistical potentials	ML	web, SA - Linux, Windows, Mac OS	Sequence, pH, temperature, ionic strength, concentration, N-, C-term protection	Profile	literature	179 fragments of 21 proteins and 71 peptides from human disease-related proteins	MCC 0.70 on 71 experimentally measured peptides
BETASCAN[50]	Pairwise probabilistic analysis	ML	web, SA - Perl	Sequence	Profile	PDB	not published	validated on 120 protein fragments from TANGO dataset, ACC 80%

ZipperDB[51]	Threading	FF	web	Sequence	Profile	own experiments	16 hexapeptide zipper crystal structures	experimental validation on 12 hexapeptides, ACC 100%
WALTZ[52]	PSSM	ML	web	Sequence, pH	Segments	own experiments, AmylHex	278 hexapeptides	cross-validation ACC 60-80%
FoldAmyloid[53]	Custom regression and statistical potentials	ML	web	Sequence	Profile	PDB[54]	3,769 protein structures	validated on dataset derived from TANGO and AmylHex (407 peptides), ACC 75%
PASTA 2.0[55]	Custom regression and statistical potentials	ML	web	Sequence	Profile	TANGO, httNT[56], AmylHex, PDB[54], AmylPred2	424 peptides and 33 amyloidogenic proteins	leave-one-out cross-validation, AUC 0.85
ArchCandy[57]	Amino acid pairing	SP	SA - Java	Sequence	Segments	literature, DisProt[58]	73 proteins	no independent test set ACC 95%
AmylPred2[59]	Majority	MP	web	Sequence	Segments	literature	33 amyloidogenic proteins	no independent test set as complete dataset was used to optimize consensus threshold MCC 0.22
MetAmyl[60]	Logistic [24]regression	MP	web	Sequence	Profile	WALTZ	278 hexapeptides	leave-one-out cross-validation on AmylPred2 dataset, MCC 0.23
Effect of mutations on solubility								
OptSolMut[33]	Linear programming	ML	SA - Binary	Structure	Propensity	literature	137 variants of 19 proteins	10-fold cross-validation, ACC 76%, MCC 0.55
CamSol[34]	Custom regression	ML, SC	web	Sequence or structure	Profile, mutations increasing solubility	literature	56 variants of 19 proteins	no independent test set 7 mutations verified experimentally with PC 0.98
AGGRESCAN3D[61]	Custom regression	ML, SC	web	Structure	Profile	AGGRESCAN	20 AB42 variants at position 19	Validated on 129 variants of 29 proteins from literature, ACC 94%
SolubiS[62,63]	Statistical and physical potentials (empirical force field)	FF	web, SA - YASARA plugin	Structure	Profile, ddG of mutations to selected gatekeepers	none	none	experimental validation on two proteins
PON-Sol[35]	Random forests	ML	web	Sequence	Propensity, mutation effect	literature	443 variants of 71 proteins	5-fold cross-validation, ACC 43% on blind test set (three-state prediction)
SODA[64]	Custom regression	ML, SC	web	Sequence or structure	Mutation landscape	PON-Sol	201 mutations	5-fold cross-validation, ACC 59-67%, ACC 100% on CamSol dataset

[a]SC – spatial corrections; SP – sequence patterns; ML – machine learning; MP – meta predictor; SS – sequence similarity; [b]SA – stand-alone application; [c]ACC – accuracy; PC – Pearson correlation; MCC – Mathew's correlation coefficient; AUC – area under the ROC curve; Method – hyperlinks refer to the web pages of the method;

Table S5. Comparison of the existing tools using S350 dataset.

Method	PCC	RMSE
PopMuSiC 2.0[2]	0.67	1.16
PEAT-SA[65]	0.50	1.92
AUTO-MUTE[66]	0.46	1.42
CUPSAT[17]	0.37	1.46
DMutant[9]	0.48	1.38
Eris[8]	0.35	1.49
I-Mutant 2.0[15]	0.29	1.50
I-Mutant 3.0[67]	0.53	1.35
MuPro[68]	0.41	1.43
Neemo[18]	0.67	1.16
Pro-Maya[3]	0.79	0.96
Prethermut[69]	0.72	1.12
SDM[70]	0.52	1.80
mCSM[13]	0.73	1.08
INPS[71]	0.68	1.26
STRUM[10]	0.79	0.98
TopologyNet 1.0[72]	0.74	1.07
TopologyNet 2.0[72]	0.81	0.94
MAESTRO[16]	0.70	1.13
SDM2[70]	0.61	1.29
iStable[73]	0.68	1.39
Rosetta[7]	0.69	0.72

PCC – Pearson Correlation Coefficient; RMSE – Root Mean Square Error

References

(1) Folkman, L.; Stantic, B.; Sattar, A.; Zhou, Y. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *J. Mol. Biol.* **2016**, *428*, 1394–1405.

(2) Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; Rooman, M. PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality. *BMC Bioinformatics* **2011**, *12*, 151.

(3) Wainreb, G.; Wolf, L.; Ashkenazy, H.; Dehouck, Y.; Ben-Tal, N. Protein Stability: A Single Recorded Mutation Aids in Predicting the Effects of Other Mutations in the Same Amino Acid Site. *Bioinforma. Oxf. Engl.* **2011**, *27*, 3286–3292.

(4) Pucci, F.; Bourgeas, R.; Rooman, M. Predicting Protein Thermal Stability Changes upon Point Mutations Using Statistical Potentials: Introducing HoTMuSiC. *Sci. Rep.* **2016**, *6*, 23257.

(5) Teng, S.; Srivastava, A. K.; Wang, L. Sequence Feature-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions. *BMC Genomics* **2010**, *11 Suppl 2*, S5.

(6) Huang, L.-T.; Gromiha, M. M.; Ho, S.-Y. IPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations. *Bioinforma. Oxf. Engl.* **2007**, *23*, 1292–1293.

(7) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. *Proteins* **2011**, *79*, 830–838.

(8) Yin, S.; Ding, F.; Dokholyan, N. V. Eris: An Automated Estimator of Protein Stability. *Nat. Methods* **2007**, *4*, 466–467.

(9) Hoppe, C.; Schomburg, D. Prediction of Protein Thermostability with a Direction- and Distance-Dependent Knowledge-Based Potential. *Protein Sci. Publ. Protein Soc.* **2005**, *14*, 2682–2692.

(10) Quan, L.; Lv, Q.; Zhang, Y. STRUM: Structure-Based Prediction of Protein Stability Changes upon Single-Point Mutation. *Bioinforma. Oxf. Engl.* **2016**, *32*, 2936–2946.

(11) Capriotti, E.; Fariselli, P.; Casadio, R. A Neural-Network-Based Method for Predicting Protein Stability Changes upon Single Point Mutations. *Bioinforma. Oxf. Engl.* **2004**, *20 Suppl 1*, i63-68.

(12) Musil, M.; Stourac, J.; Bendl, J.; Brezovsky, J.; Prokop, Z.; Zendulka, J.; Martinek, T.; Bednar, D.; Damborsky, J. FireProt: Web Server for Automated Design of Thermostable Proteins. *Nucleic Acids Res.* **2017**, *45*, W393–W399.

(13) Pires, D. E. V.; Ascher, D. B.; Blundell, T. L. MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinforma. Oxf. Engl.* **2014**, *30*, 335–342.

(14) Witvliet, D. K.; Strokach, A.; Giraldo-Forero, A. F.; Teyra, J.; Colak, R.; Kim, P. M. ELASPIC Web-Server: Proteome-Wide Structure-Based Prediction of Mutation Effects on Protein Stability and Binding Affinity. *Bioinforma. Oxf. Engl.* **2016**, *32*, 1589–1591.

(15) Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* **2005**, *33*, W306-310.

(16) Laimer, J.; Hofer, H.; Fritz, M.; Wegenkittl, S.; Lackner, P. MAESTRO--Multi Agent Stability Prediction upon Point Mutations. *BMC Bioinformatics* **2015**, *16*, 116.

(17) Parthiban, V.; Gromiha, M. M.; Schomburg, D. CUPSAT: Prediction of Protein Stability upon Point Mutations. *Nucleic Acids Res.* **2006**, *34*, W239-242.

(18) Giollo, M.; Martin, A. J. M.; Walsh, I.; Ferrari, C.; Tosatto, S. C. E. NeEMO: A Method Using Residue Interaction Networks to Improve Prediction of Protein Stability upon Mutation. *BMC Genomics* **2014**, *15 Suppl 4*, S7.

(19) Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R. L.; Aharoni, A.; Silman, I.; Sussman, J. L.; Tawfik, D. S.; Fleishman, S. J. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016**, *63*, 337–346.

(20) Niwa, T.; Ying, B.-W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal Protein Solubility Distribution Revealed by an Aggregation Analysis of the Entire Ensemble of Escherichia Coli Proteins. *Proc. Natl. Acad. Sci.* **2009**, *106*, 4201–4206.
(21) Niwa, T.; Kanamori, T.; Ueda, T.; Taguchi, H. Global Analysis of Chaperone Effects Using a Reconstituted Cell-Free Translation System. *Proc. Natl. Acad. Sci.* **2012**, *109*, 8937–8942.
(22) Helen M. Berman, M. J. G., Andrei Kouranov, David I. Micallef, John Westbrook; Protein Structure Initiative network of investigators. Protein Structure Initiative - TargetTrack 2000-2017 - All Data Files, 2017. https://doi.org/10.5281/zenodo.821654.
(23) Price, W. N.; Handelman, S. K.; Everett, J. K.; Tong, S. N.; Bracic, A.; Luff, J. D.; Naumov, V.; Acton, T.; Manor, P.; Xiao, R.; Rost, B.; Montelione, G. T.; Hunt, J. F. Large-Scale Experimental Studies Show Unexpected Amino Acid Effects on Protein Expression and Solubility in Vivo in E. Coli. *Microb. Inform. Exp.* **2011**, *1*, 6.
(24) Hirose, S.; Noguchi, T. ESPRESSO: A System for Estimating Protein Expression and Solubility in Protein Expression Systems. *PROTEOMICS* **2013**, *13*, 1444–1456.
(25) Hirose, S.; Kawamura, Y.; Yokota, K.; Kuroita, T.; Natsume, T.; Komiya, K.; Tsutsumi, T.; Suwa, Y.; Isogai, T.; Goshima, N.; Noguchi, T. Statistical Analysis of Features Associated with Protein Expression/Solubility in an in Vivo Escherichia Coli Expression System and a Wheat Germ Cell-Free Expression System. *J. Biochem. (Tokyo)* **2011**, *150*, 73–81.
(26) Chang, C. C. H.; Li, C.; Webb, G. I.; Tey, B.; Song, J.; Ramanan, R. N. Periscope: Quantitative Prediction of Soluble Protein Expression in the Periplasm of *Escherichia Coli*. *Sci. Rep.* **2016**, *6*, 21844.
(27) Pawlicki, S.; Le Béchec, A.; Delamarche, C. AMYPdb: A Database Dedicated to Amyloid Precursor Proteins. *BMC Bioinformatics* **2008**, *9*, 273.
(28) Thompson, M. J.; Sievers, S. A.; Karanicolas, J.; Ivanova, M. I.; Baker, D.; Eisenberg, D. The 3D Profile Method for Identifying Fibril-Forming Segments of Proteins. *Proc. Natl. Acad. Sci.* **2006**, *103*, 4074–4078.
(29) Beerten, J.; Van Durme, J.; Gallardo, R.; Capriotti, E.; Serpell, L.; Rousseau, F.; Schymkowitz, J. WALTZ-DB: A Benchmark Database of Amyloidogenic Hexapeptides. *Bioinformatics* **2015**, *31*, 1698–1700.
(30) Wozniak, P. P.; Kotulska, M. AmyLoad: Website Dedicated to Amyloidogenic Protein Fragments. *Bioinformatics* **2015**, *31*, 3395–3397.
(31) Sastry, A.; Monk, J.; Tegel, H.; Uhlen, M.; Palsson, B. O.; Rockberg, J.; Brunk, E. Machine Learning in Computational Biology to Accelerate High-Throughput Protein Expression. *Bioinformatics* **2017**, *33*, 2487–2495.
(32) Thangakani, A. M.; Nagarajan, R.; Kumar, S.; Sakthivel, R.; Velmurugan, D.; Gromiha, M. M. CPAD, Curated Protein Aggregation Database: A Repository of Manually Curated Experimental Data on Protein and Peptide Aggregation. *PLOS ONE* **2016**, *11*, e0152949.
(33) Tian, Y.; Deutsch, C.; Krishnamoorthy, B. Scoring Function to Predict Solubility Mutagenesis. *Algorithms Mol. Biol.* **2010**, *5*, 33.
(34) Sormanni, P.; Aprile, F. A.; Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **2015**, *427*, 478–490.
(35) Yang, Y.; Niroula, A.; Shen, B.; Vihinen, M. PON-Sol: Prediction of Effects of Amino Acid Substitutions on Protein Solubility. *Bioinformatics* **2016**, *32*, 2032–2034.
(36) Wilkinson, D. L.; Harrison, R. G. Predicting the Solubility of Recombinant Proteins in Escherichia Coli. *Biotechnol. Nat. Publ. Co.* **1991**, *9*, 443–448.
(37) Davis, G. D.; Elisee, C.; Newham, D. M.; Harrison, R. G. New Fusion Protein Systems Designed to Give Soluble Expression InEscherichia Coli. *Biotechnol. Bioeng.* **1999**, *65*, 382–388.
(38) Magnan, C. N.; Randall, A.; Baldi, P. SOLpro: Accurate Sequence-Based Prediction of Protein Solubility. *Bioinformatics* **2009**, *25*, 2200–2207.
(39) Chang, C. C. H.; Song, J.; Tey, B. T.; Ramanan, R. N. Bioinformatics Approaches for Improved Recombinant Protein Production in Escherichia Coli: Protein Solubility Prediction. *Brief. Bioinform.* **2014**, *15*, 953–962.

(40) Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; Frishman, D. PROSO II - a New Method for Protein Solubility Prediction: PROSO II. *FEBS J.* **2012**, *279*, 2192–2200.
(41) Agostini, F.; Vendruscolo, M.; Tartaglia, G. G. Sequence-Based Prediction of Protein Solubility. *J. Mol. Biol.* **2012**, *421*, 237–241.
(42) Agostini, F.; Cirillo, D.; Livi, C. M.; Delli Ponti, R.; Tartaglia, G. G. CcSOL Omics: A Webserver for Solubility Prediction of Endogenous and Heterologous Expression in Escherichia Coli. *Bioinformatics* **2014**, *30*, 2975–2977.
(43) Hebditch, M.; Carballo-Amador, M. A.; Charonis, S.; Curtis, R.; Warwicker, J. Protein–Sol: A Web Tool for Predicting Protein Solubility from Sequence. *Bioinformatics* **2017**, *33*, 3098–3100.
(44) Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G.-Y.; Bensmail, H.; Mall, R. DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction. *Bioinformatics* **2018**, *34*, 2605–2613.
(45) DuBay, K. F.; Pawar, A. P.; Chiti, F.; Zurdo, J.; Dobson, C. M.; Vendruscolo, M. Prediction of the Absolute Aggregation Rates of Amyloidogenic Polypeptide Chains. *J. Mol. Biol.* **2004**, *341*, 1317–1326.
(46) Tartaglia, G. G.; Pawar, A. P.; Campioni, S.; Dobson, C. M.; Chiti, F.; Vendruscolo, M. Prediction of Aggregation-Prone Regions in Structured Proteins. *J. Mol. Biol.* **2008**, *380*, 425–436.
(47) de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Ventura, S. Mutagenesis of the Central Hydrophobic Cluster in Abeta42 Alzheimer's Peptide. Side-Chain Properties Correlate with Aggregation Propensities. *FEBS J.* **2006**, *273*, 658–668.
(48) Conchillo-Solé, O.; de Groot, N. S.; Avilés, F. X.; Vendrell, J.; Daura, X.; Ventura, S. AGGRESCAN: A Server for the Prediction and Evaluation of "Hot Spots" of Aggregation in Polypeptides. *BMC Bioinformatics* **2007**, *8*, 65.
(49) Fernandez-Escamilla, A.-M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nat. Biotechnol.* **2004**, *22*, 1302–1306.
(50) Bryan, A. W.; Menke, M.; Cowen, L. J.; Lindquist, S. L.; Berger, B. BETASCAN: Probable β-Amyloids Identified by Pairwise Probabilistic Analysis. *PLoS Comput. Biol.* **2009**, *5*, e1000333.
(51) Goldschmidt, L.; Teng, P. K.; Riek, R.; Eisenberg, D. Identifying the Amylome, Proteins Capable of Forming Amyloid-like Fibrils. *Proc. Natl. Acad. Sci.* **2010**, *107*, 3487–3492.
(52) Maurer-Stroh, S.; Debulpaep, M.; Kuemmerer, N.; de la Paz, M. L.; Martins, I. C.; Reumers, J.; Morris, K. L.; Copland, A.; Serpell, L.; Serrano, L.; Schymkowitz, J. W. H.; Rousseau, F. Exploring the Sequence Determinants of Amyloid Structure Using Position-Specific Scoring Matrices. *Nat. Methods* **2010**, *7*, 237–242.
(53) Garbuzynskiy, S. O.; Lobanov, M. Y.; Galzitskaya, O. V. FoldAmyloid: A Method of Prediction of Amyloidogenic Regions from Protein Sequence. *Bioinformatics* **2010**, *26*, 326–332.
(54) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Lobanov, M. Y. Prediction of Amyloidogenic and Disordered Regions in Protein Chains. *PLoS Comput. Biol.* **2006**, *2*, e177.
(55) Walsh, I.; Seno, F.; Tosatto, S. C. E.; Trovato, A. PASTA 2.0: An Improved Server for Protein Aggregation Prediction. *Nucleic Acids Res.* **2014**, *42*, W301–W307.
(56) Roland, B. P.; Kodali, R.; Mishra, R.; Wetzel, R. A Serendipitous Survey of Prediction Algorithms for Amyloidogenicity: Survey of Prediction Algorithms for Amyloidogenicity. *Biopolymers* **2013**, *100*, 780–789.
(57) Ahmed, A. B.; Znassi, N.; Château, M.-T.; Kajava, A. V. A Structure-Based Approach to Predict Predisposition to Amyloidosis. *Alzheimers Dement.* **2015**, *11*, 681–690.
(58) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z.; Elofsson, A.; Gasparini, A.; Hatos, A.; Kajava, A. V.; Kalmar, L.; Leonardi, E.; Lazar, T.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Meszaros, A.; Minervini, G.; Murvai, N.; Pujols, J.; Roche, D. B.; Salladini, E.; Schad, E.; Schramm, A.; Szabo, B.; Tantos, A.; Tonello, F.; Tsirigos, K. D.; Veljković, N.; Ventura, S.; Vranken, W.; Warholm, P.;

Uversky, V. N.; Dunker, A. K.; Longhi, S.; Tompa, P.; Tosatto, S. C. E. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45*, D219–D227.
(59) Tsolis, A. C.; Papandreou, N. C.; Iconomidou, V. A.; Hamodrakas, S. J. A Consensus Method for the Prediction of 'Aggregation-Prone' Peptides in Globular Proteins. *PLoS ONE* **2013**, *8*, e54175.
(60) Emily, M.; Talvas, A.; Delamarche, C. MetAmyl: A METa-Predictor for AMYLoid Proteins. *PLoS ONE* **2013**, *8*, e79722.
(61) Zambrano, R.; Jamroz, M.; Szczasiuk, A.; Pujols, J.; Kmiecik, S.; Ventura, S. AGGRESCAN3D (A3D): Server for Prediction of Aggregation Properties of Protein Structures. *Nucleic Acids Res.* **2015**, *43*, W306–W313.
(62) De Baets, G.; Van Durme, J.; van der Kant, R.; Schymkowitz, J.; Rousseau, F. Solubis: Optimize Your Protein: Fig. 1. *Bioinformatics* **2015**, *31*, 2580–2582.
(63) Van Durme, J.; De Baets, G.; Van Der Kant, R.; Ramakers, M.; Ganesan, A.; Wilkinson, H.; Gallardo, R.; Rousseau, F.; Schymkowitz, J. Solubis: A Webserver to Reduce Protein Aggregation through Mutation. *Protein Eng. Des. Sel.* **2016**, *29*, 285–289.
(64) Paladin, L.; Piovesan, D.; Tosatto, S. C. E. SODA: Prediction of Protein Solubility from Disorder and Aggregation Propensity. *Nucleic Acids Res.* **2017**, *45*, W236–W240.
(65) Johnston, M. A.; Søndergaard, C. R.; Nielsen, J. E. Integrated Prediction of the Effect of Mutations on Multiple Protein Characteristics. *Proteins* **2011**, *79*, 165–178.
(66) Masso, M.; Vaisman, I. I. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. *Adv. Bioinforma.* **2014**, *2014*, 278385.
(67) Capriotti, E.; Fariselli, P.; Rossi, I.; Casadio, R. A Three-State Prediction of Single Point Mutations on Protein Stability Changes. *BMC Bioinformatics* **2008**, *9*, S6.
(68) Cheng, J.; Randall, A.; Baldi, P. Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines. *Proteins* **2006**, *62*, 1125–1132.
(69) Tian, J.; Wu, N.; Chu, X.; Fan, Y. Predicting Changes in Protein Thermostability Brought about by Single- or Multi-Site Mutations. *BMC Bioinformatics* **2010**, *11*, 370.
(70) Pandurangan, A. P.; Ochoa-Montaño, B.; Ascher, D. B.; Blundell, T. L. SDM: A Server for Predicting Effects of Mutations on Protein Stability. *Nucleic Acids Res.* **2017**, *45*, W229–W235.
(71) Savojardo, C.; Fariselli, P.; Martelli, P. L.; Casadio, R. INPS-MD: A Web Server to Predict Stability of Protein Variants from Sequence and Structure. *Bioinforma. Oxf. Engl.* **2016**, *32*, 2542–2544.
(72) Cang, Z.; Wei, G.-W. TopologyNet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLoS Comput. Biol.* **2017**, *13*, e1005690.
(73) Chen, C.-W.; Lin, J.; Chu, Y.-W. IStable: Off-the-Shelf Predictor Integration for Predicting Protein Stability Changes. *BMC Bioinformatics* **2013**, *14 Suppl 2*, S5.