Computational and Structural Biotechnology Journal 20 (2022) 6339–6347

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Data article

SoluProtMut^{DB}: A manually curated database of protein solubility changes upon mutations

Jan Velecký [a], Marie Hamsikova [a,b], Jan Stourac [a,b], Milos Musil [a,c], Jiri Damborsky [a,b], David Bednar [a,b,*], Stanislav Mazurenko [a,b,*]

[a] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlarska 2, Brno 61137, Czech Republic
[b] International Clinical Research Center, St. Anne's University Hospital Brno, Pekarska 53, Brno 65691, Czech Republic
[c] Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Bozetechova 2, Brno 61200, Czech Republic

ARTICLE INFO

Article history:
Received 24 August 2022
Received in revised form 4 November 2022
Accepted 4 November 2022
Available online 9 November 2022

Keywords:
Mutational database
Protein engineering
Soluble expression
Protein yield
Machine learning
Protein aggregation

ABSTRACT

Protein solubility is an attractive engineering target primarily due to its relation to yields in protein production and manufacturing. Moreover, better knowledge of the mutational effects on protein solubility could connect several serious human diseases with protein aggregation. However, we have limited understanding of the protein structural determinants of solubility, and the available data have mostly been scattered in the literature. Here, we present SoluProtMut^{DB} – the first database containing data on protein solubility changes upon mutations. Our database accommodates 33 000 measurements of 17 000 protein variants in 103 different proteins. The database can serve as an essential source of information for the researchers designing improved protein variants or those developing machine learning tools to predict the effects of mutations on solubility. The database comprises all the previously published solubility datasets and thousands of new data points from recent publications, including deep mutational scanning experiments. Moreover, it features many available experimental conditions known to affect protein solubility. The datasets have been manually curated with substantial corrections, improving suitability for machine learning applications. The database is available at loschmidt.chemi.muni.cz/soluprotmutdb.

1. Introduction

Protein mutational databases accumulate results from experiments examining how mutations introduced to a protein affect a selected property. Several such databases have arisen recently, including FireProt^{DB} [1] for the protein stability data for single-point mutants, the MPTherm [2] database for membrane protein thermodynamics, or D3DistalMutation [3] for enzyme activity. However, there has not been any mutational solubility database yet despite solubility being a basic characteristic of any globular protein. Moreover, high solubility is essential for high-dosing protein therapeutics or for efficient protein production [4,5]. The lowered solubility of a body protein due to a mutation may also cause a disease [6]. And neither too low nor too high solubility is required for successful structure determination of a protein in the crystalline form.

Prediction of solubility change upon mutation is thus an important problem. Several predictors for this task were developed, usually using mutational solubility data sets for training collected independently from the literature [7–10]. While these attempts showed great promise, the training datasets were rather limited in the number of entries and their annotations. These limitations provide a possible explanation as to why recent studies comparing the predictors revealed significant room for improvement, as the latest predictors did not exceed the correct prediction ratio of 70% [10,11].

The data available in the solubility datasets come mostly from small-scale experiments. These often search for a solubilizing mutation to a particular protein in order to enhance its insufficient solubility. A small-scale experiment measures only a small number of mutants and only one direction of solubility change is often observed among all of them. Another drawback is that these experiments may be incomparable due to the different conditions under which they were conducted. Most typically, a variant of an electrophoresis assay

* Corresponding authors at: Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlarska 2, Brno 61137, Czech Republic.

E-mail addresses: 222755@mail.muni.cz (D. Bednar), mazurenko@mail.muni.cz (S. Mazurenko).

https://doi.org/10.1016/j.csbj.2022.11.009
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

and protein staining is used to assess protein solubility through mass separation, e.g., the SDS–PAGE assay. Other, less frequent methods include Western blotting, where the soluble fraction of protein of interest is separated and marked via antigen binding.

In contrast, high-throughput experiments provide many results from a single run. Apart from the clear advantage of obtaining a large amount of data at once, they allow a more precise comparison thanks to the elimination of setup differences. High-throughput methods typically measure solubility indirectly through another property, e.g., fluorescence, which can be achieved in an automated manner more easily. For instance, in recent studies by Whitehead's group [12,13], fluorescence-activated cell sorting (FACS) was used to select solubilizing mutations out of almost all possible single-point variants. While such a strategy is usually applied to one protein at a time, it has the potential to provide the sufficient data abundance for modern data-hungry machine learning (ML) methods [14].

Here we present a database incorporating solubility data from several sources (Fig. 1): (i) curated data from OptSolMut [7], CamSol [8], A3D [9] and PON-Sol [10] datasets, (ii) recently conducted deep mutational scanning (DMS) of solubility at Whitehead's research group [12,13], (iii) our own literature search for solubility experiments, and (iv) data from high-throughput experiments currently conducted in our laboratories.

The database goes beyond the basic reporting of introduced mutations and their effects on protein solubility. We performed an extensive manual curation of each entry based on the original publications. We also keep track of the experimental setup wherever possible as it has a major influence on the experimental outcome [17]. This setup has two main components: expression-related conditions (how the protein was produced) and assay-related conditions (how the solubility was measured). For instance, the expression conditions include host cells, the temperature, and induction times used. Assays differ mainly in the physical property used to measure solubility change. Finally, the data are annotated with dataset memberships, links to UniProt [15] and its annotations, and HotSpot Wizard [16] features per sequence or structure as depicted in Fig. 1.

While the database will serve as a valuable source of insights for protein engineers, structural biologists, or biochemists, we have made our database convenient for the broad ML and data science communities as well, e.g., to facilitate using the deposited data in the development and testing of predictive models. All the aforementioned experimental conditions and annotations are utilizable as features. We also performed a systematization of reported changes and created a flexible Export Wizard. The systematization deals with the verbally-assessed changes – these are discrete and inexact values with no scale specified by the authors. Export Wizard allows exporting the filtered data and converting the values to the desired classes to be used in a target model.

With the advent of high-throughput screening methods, we may see a flood of mutational solubility data published, and SoluProtMut^{DB} should serve as a central depository for this type of data. A centralized and regularly updated depository for mutational solubility data will facilitate the *in silico* engineering of protein solubility, which is critical in biopharmacy, biotechnology, or structural biology. The depository will also be useful for data scientists, ML engineers, protein engineers and medical doctors.

2. Materials and methods

2.1. Data from small-scale experiments

The cornerstones of SoluProtMut^{DB} are four mutational solubility datasets, published between 2010 and 2017, which we merged together: OptSolMut [7], CamSol [8], A3D [9], and PON-Sol [10]. Every datapoint in each of these datasets represents a mutated variant of a particular protein, where the protein is specified either by its sequence or Protein Data Bank ID (PDB ID) and labeled according to the effect on the protein solubility. While none of the datasets is fully contained in another, they do overlap significantly. Therefore, we ensured that each datapoint is contained in the final database only once and assigned to all the datasets it appears in. We also added new data from the updated PON-Sol dataset [11]. Furthermore, as all these datasets only comprise the solubility data from publications before 2017, we conducted a data search in more recent literature and added new results.

We carried out manual validation and curation of the datasets against the source publications as the data are not in a machine-readable format in most of the source publications. We found and resolved a substantial number of discrepancies of the following types by correction or removal of the affected datapoints: reports of changes in properties with no clear relation to solubility; measurements which are not present in the source publication; wrong values; wrong positions or residues of substitutions.

During the manual processing of the publications, we additionally extracted the data that do not appear in the published datasets. These include reported experimental conditions, such as measurement assay, host organism and strain, temperature, pH, and concentration method used; originally reported numerical changes in solubility; and even more than hundred instances of measured protein variants that were left unnoticed by the authors of the datasets. We also distinguish the types of solubility the continuous values referred to: the soluble fraction, soluble concentration, or total concentration.

During the validation, we assigned a UniProt accession number (UniProt AC) of an original variant to every datapoint and renumbered the mutated positions with respect to the UniProt sequence. This was necessary as the proteins in the datasets are only assigned with PDB IDs or protein/gene names, which are, however, less reliable, stable, or not unique in comparison to UniProt ACs in the long term. In the case of PDBs, one structure can refer to several proteins, and a single protein typically has multiple relevant PDBs with new and refined structures of proteins appearing over time.

2.2. Deep mutational scanning data

The eminent source is the data collected at Whitehead's research group – the first use of DMS for solubility screening. The group measured the soluble expression of the levoglucosan kinase, TEM-1 β-lactamase, and pyrrolidine ketide synthase variants in *E. coli* or yeast assays [12,13]. Their DMS approach consisted of three steps. The first step was comprehensive saturation mutagenesis across the entire protein, which yielded a cell library of all possible single-point mutants. The second step was the selection of cells with soluble protein. And the third step was deep sequencing – measuring the frequencies of the variants before and after the selection procedure of the second step by sampling and sequencing them. The authors explored two selection procedures: Tat-export and FACS. In the former, soluble protein provided antibiotic resistance and was required for cell survival. In the latter, the fluorescence change upon binding with a fluorescence-enabled antibody or a green-fluorescence-protein (GFP) tag was exploited as the proxy to protein solubility, and then the cells with higher solubility were sorted out using FACS. The enrichment ratio for each variant was calculated based on the number of reads before and after the selection, normalized, and reported as the score for the effect of the mutations on protein solubility.

To make these continuous scores comparable with the discrete values reported in the other literature, we binned them into 5 levels according to the threshold of 0.15, suggested by the authors (that is $+10\%$ on a linear scale) to label enhancing mutations, and a threshold of $+50\%$ for significantly enhancing mutations. Symmet-

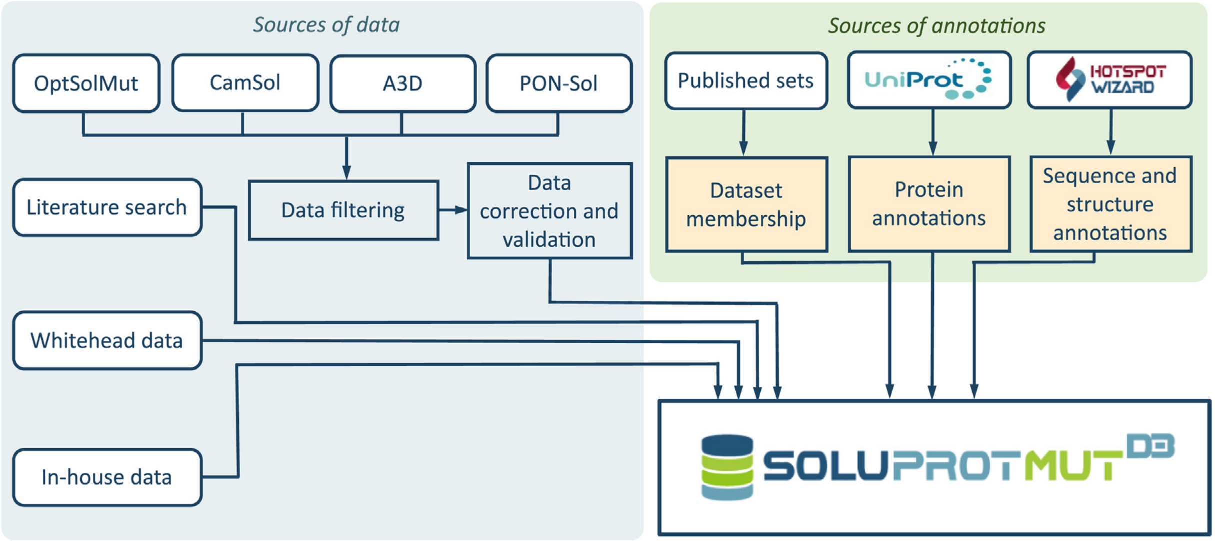


Fig. 1. The data sources of SoluProtMut^{DB} and their processing. The primary sources are the merged data from the earlier published datasets of protein-solubility predictors and the high-throughput data from Whitehead's group [12,13]. The datasets have been manually checked with the original publications and corrected accordingly. Apart from these, we conducted an extensive literature search and deposited more recently published data and the data collected in our laboratories. The information about a dataset membership and UniProt [15] and HotSpot Wizard 3.0 [16] annotations were added to the entries.

rically, we used -10% and $-33.\overline{3}\%$ to label slightly and significantly deteriorating mutations, respectively. The remaining datapoints were binned into the neutral class. During this process, we also omitted the scores of nonsense mutations and those having statistically insignificant enrichment values due to the low number of reads.

2.3. In-house data

In addition to the published literature, the database contains the data from medium-throughput experiments on haloalkane dehalogenase, recently conducted by our research group [18].[1] Our assay, validated by comparison with SDS–PAGE on multiple proteins, measures solubility through fluorescence activity introduced by the split-GFP approach. The mutant library was created with error-prone PCR, and randomly selected mutants were measured and sequenced. Measuring was conducted in replicates, and the mutants with statistically insignificant results were discarded. This resulted in 22 datapoints available in the database.

2.4. Systematization of values

By analyzing the literature, we identified five patterns appearing in solubility experiments for a mutation effect assessment. We systematized these patterns into reporting systems as per Table 1 to make the reported changes comparable even when they come from different publications and are described in different terms. These differences are partially due to the use of various assays as their precision varies, and sometimes the effect was not quantifiable. In other cases, incomplete information was published. For example, in experiments aiming to solubilize a particular protein, only verbal assessment is often reported for mutants not improving solubility.

We distinguish the orientation (positive, negative, or neutral) of an effect and, whenever applicable, also its significance (slight or significant). Altogether, up to five discrete values are defined: *significantly/slightly deteriorating*, *neutral*, and *slightly/significantly enhancing*. This system suggests different resolutions in different experiments, e.g., a value from the 5-value system should be more precise than from the 3-value system. Hence, if one mutation is enhancing in the 3-value system and another is slightly enhancing in the 5-value system, we can assume the former to be at least as enhancing as the latter, and possibly substantially more.

Table 1
The comparison table between reported solubility changes in various reporting systems. The considered reporting systems (columns) consist of 2 to 5 possible values of measured effects on solubility (rows), spanning from –– (significantly deteriorating) over neutral (N) to ++ (significantly enhancing). For example, a substantial deteriorating change in solubility could be reported as simply deteriorating in the 2- or 3-value systems or non-enhancing in the unipolar system.

real change	reported change					real change
	unipolar	2-value	3-value	4-value	5-value	
++	enhancing			significantly enhancing		++
+				slightly enhancing		+
N	non-enhancing	deteriorating	neutral		neutral	neutral
–				slightly deteriorating		–
––				significantly deteriorating		––

2.5. Annotations

In addition to the data extracted from the literature, we annotated proteins on sequence and structure levels. As all the sequences were mapped to UniProt through their accession numbers, we extracted protein names, species of origin, InterPro families, and Enzyme Commission numbers from there. We also manually linked proteins with their structures in PDB. We prioritized the X-ray crystallographic structures with the highest resolution, without ligands or mutations. The assigned structures were then used as an input to HotSpot Wizard (HSW) [19] to obtain additional sequence and structural features.

HSW sequence features come from multiple sequence alignment of homologous sequences. HSW obtains these sequences by a BLAST search [20] against the UniRef90 database [21] and clusters them using the UCLUST algorithm [22] with a 90% sequence identity. Sorted by the coverage of the BLAST query, the top 200 cluster-representing sequences are selected and subsequently aligned using Clustal Omega [23]. The resulting alignment is then employed (i) to estimate the conservation score for each position

[1] https://loschmidt.chemi.muni.cz/soluprotmutdb/protein/103.

using the Jensen-Shannon divergence [24], (ii) to identify correlated positions using the consensus prediction of several tools integrated with HSW, and (iii) to identify potential back-to-consensus mutations, i.e., the positions in the multiple sequence alignment where an amino acid in the query sequence differs from the majority of amino acids at conserved positions.

Apart from sequence features, the following structural features are included: (i) the protein secondary structure calculated by DSSP [25], (ii) the accessible surface area calculated with the Shrake and Rupley algorithm [26], (iii) average B-factors for protein residues [27], (iv) protein pockets identified by the fpocket tool [28], and (v) protein tunnels and their bottlenecks calculated by CAVER [29]. Only the tunnels connected with catalytic pockets are stored in the database. The structural features are mapped back onto UniProt sequences using the SIFTS database [30].

2.6. Database structure

Measurement results of differential solubility experiments are at the core of our database. Each result is linked to a protein variant defined by a particular protein and a set of substitutions in its sequence. The effect of any protein variant on solubility contains a difference in the measured property compared to the original protein variant, both measured under the same experimental setup. This setup includes the host cell, assay, or temperature used and is linked to the corresponding results. The corresponding protein is identified by UniProt AC, and the mutated positions are based on the UniProt indexing. Each result has its alphanumerical *accession code*, which is meant to be stable, searchable, and therefore citable. In addition, each result may be linked to one or more published datasets.

3. Results

The basic statistics summarizing the content of the database are given in Table 2. The total number of datapoints consists of (i) merged 764 (610 unique) datapoints from the previously published datasets, (ii) Whitehead's DMS data – accounting for 32 081 of the datapoints, (iii) 279 new measurements from the literature and (iv) 22 measurements carried out in-house.

The data reveal that a random mutation likely has a desolubilizing effect, as shown in the mutational effect distribution in Fig. 2. Only 18% of mutants increase solubility and just one third of them significantly. This is confirmed when the distribution is plotted per protein (Fig. 3). The three most frequent proteins from small-scale experiments, on the other hand, display a strong distribution bias compared to the DMS data and the 'Other' category alike. The exact ratio is protein-dependent.

While the database size is several orders of magnitude larger than the sizes of the prior datasets, the results from the high-throughput experiments from Whitehead's group dominate the deposited data. The exhaustiveness of Whitehead's data provides the database with great variability in mutated positions and in combinations of substituted and target amino-acid pairs (Fig. 4) but is limited to only three proteins. The protein variability of the database is provided by the rest of the data - Fig. 3 contrasts the entry counts for these three proteins with the remaining ones.

We kept the FAIR principles (Findable, Accessible, Interoperable, Reusable) [31] in mind during the database development. In addition to making the data accessible and searchable online (see the section 3.1) and exportable in a machine-readable format (see the section 3.2), we also assigned a unique *accession code* (SPMDB AC) to each entry of a measurement result. The accession code is an identifier that is stable in time and can be used for searching or linking. Our database crosslinks SPMDB AC with UniProt, PDB, and InterPro databases.

Table 2
Current statistics of the database. The most recent numbers are available at loschmidt.chemi.muni.cz/soluprotmutdb as the database is regularly updated with new data.

Datapoints	32 992
Mutant variants	17 392
of which multi-point	157
Publications	110
Proteins	103

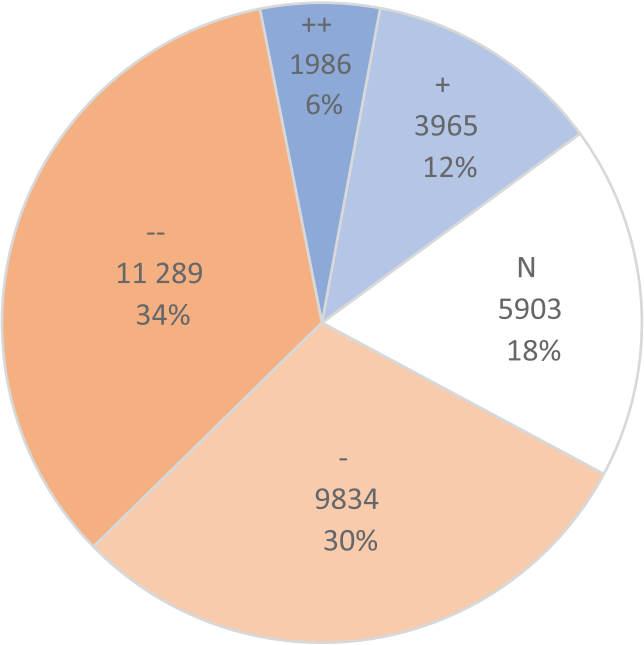


Fig. 2. The distribution of protein variants in the database by their mutational effects on solubility. The distribution is divided into 5 levels: neutral (N), slightly/significantly desolubilizing (–/– –) and solubilizing (**+/++**). Notably, two thirds of the mutants show a deteriorating effect.

3.1. Interface

SoluProtMut^{DB} has a user-friendly web interface enabling its users to browse, search, and export the data. The 'Show all' option in the navigation bar leads to the result table listing all the entries available in the database (Fig. 5). To filter these entries, the search at the top of the page can be used in two ways: (i) a full-text search by protein names, UniProt accession codes, PDB identifiers, InterPro entries, EC numbers, publications, dataset names, organisms, host cells, or SPMDB AC; or (ii) an advanced search capable of combining several queries on database fields (Fig. 6). The displayed data in the search results can be exported using Export Wizard by clicking the 'Export' button (see the section 3.2).

Protein and variant pages can be accessed from the result table by clicking on a protein name or mutation, respectively. A variant page shows all measurements for the particular protein variant. A protein page shows basic information about the protein, such as UniProt AC, species, EC number, assigned InterPro families, or the table containing experimental data for this protein. In addition, interactive ProtVista tracks [32] visualize the following sequence features: the secondary structure, catalytic sites, natural variants, amino-acid charges, catalytic pockets, tunnels, B-factors, conservation, and back-to-consensus mutations. The structure, if available, is shown using the Mol* viewer [33] (Fig. 7). Mutated positions can be highlighted in the structure by clicking on the eye icons in the data table.

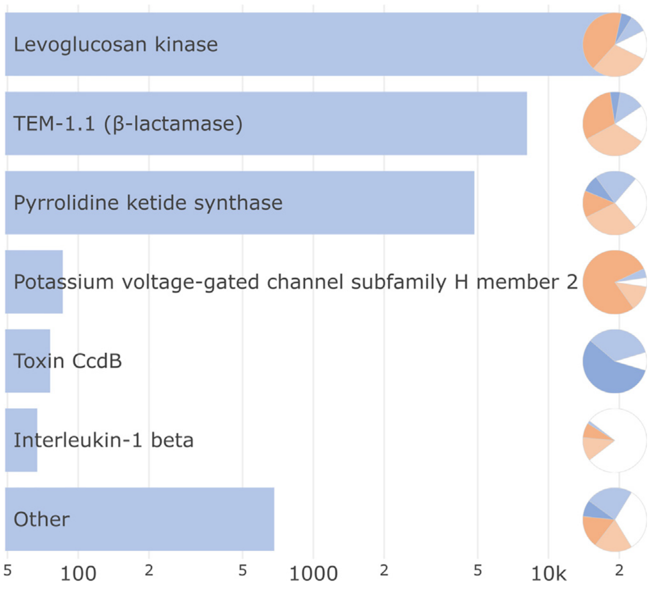
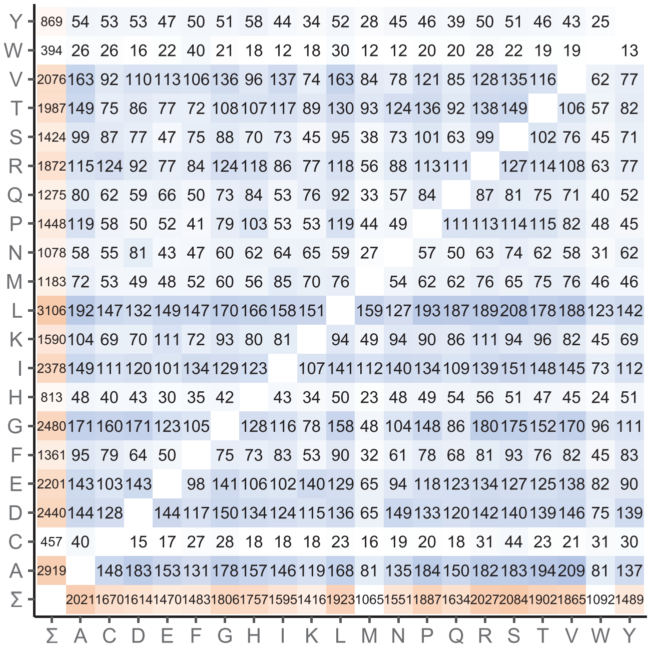


Fig. 3. The six most represented proteins in the database by their entry counts. The data for the first three proteins come from deep mutational scanning experiments. The 'Other' category contains the remaining 97 proteins. The horizontal axis has a logarithmic scale, and the pie charts on the right display mutational effect distribution per category with the same color coding as in Fig. 2: neutral, desolubilizing and solubilizing mutations.



	Σ	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Y	869	54	53	53	47	50	51	58	44	34	52	28	45	46	39	50	51	46	43	25	
W	394	26	26	16	22	40	21	18	12	18	30	12	12	20	20	28	22	19	19		13
V	2076	163	92	110	113	106	136	96	137	74	163	84	78	121	85	128	135	116		62	77
T	1987	149	75	86	77	72	108	107	117	89	130	93	124	136	92	138	149		106	57	82
S	1424	99	87	77	47	75	88	70	73	45	95	38	73	101	63	99		102	76	45	71
R	1872	115	124	92	77	84	124	118	86	77	118	56	88	113	111		127	114	108	63	77
Q	1275	80	62	59	66	50	73	84	53	76	92	33	57	84		87	81	75	71	40	52
P	1448	119	58	50	52	41	79	103	53	53	119	44	49		111	113	114	115	82	48	45
N	1078	58	55	81	43	47	60	62	64	65	59	27		57	50	63	74	62	58	31	62
M	1183	72	53	49	48	52	60	56	85	70	76		54	62	62	76	65	75	76	46	46
L	3106	192	147	132	149	147	170	166	158	151		159	127	193	187	189	208	178	188	123	142
K	1590	104	69	70	111	72	93	80	81		94	49	94	90	86	111	94	96	82	45	69
I	2378	149	111	120	101	134	129	123		107	141	112	140	134	109	139	151	148	145	73	112
H	813	48	40	43	30	35	42		43	34	50	23	48	49	54	56	51	47	45	24	51
G	2480	171	160	171	123	105		128	116	78	158	48	104	148	86	180	175	152	170	96	111
F	1361	95	79	64	50		75	73	83	53	90	32	61	78	68	81	93	76	82	45	83
E	2201	143	103	143		98	141	106	102	140	129	65	94	118	123	134	127	125	138	82	90
D	2440	144	128		144	117	150	134	124	115	136	65	149	133	120	142	140	139	146	75	139
C	457	40		15	17	27	28	18	18	18	23	16	19	20	18	31	44	23	21	31	30
A	2919		148	183	153	131	178	157	146	119	168	81	135	184	150	182	183	194	209	81	137
Σ		2021	1670	1614	1470	1483	1806	1757	1595	1416	1923	1065	1551	1887	1634	2027	2084	1902	1865	1092	1489

Fig. 4. A matrix showing the numbers of mutation occurrences in the database 'from' (rows) and 'to' (columns) specific amino acids. The Σ column and row represent sums of mutations 'from' and 'to' given amino acids, respectively. A cell color saturation shows the abundance of the corresponding combination.

The Datasets page lists the known mutational solubility datasets. Further details, including the authors and the links to the publication and the raw dataset, can be obtained by clicking on a dataset name. Furthermore, the dataset page contains statistics on the overall distribution of solubility effects in each dataset and the similarity to the other datasets.

3.2. Data export

The complete database can be downloaded as a MariaDB database server dump in the SQL format. In addition to this option, we developed Export Wizard for user-friendly exporting a currently browsed subset of the database, e.g., defined by the active search filter, as a tabular dataset in the CSV format. This functionality is specifically aimed at data scientists and machine learning developers to allow them to analyze or use the data with minimum processing effort. Optionally, additional filtering/labeling and data augmentation may be applied before data export.

The filtering also allows selecting only the results measured in continuous values, suitable for a regression analysis and modeling. The alternative is the labeling that adapts the data to a specific model according to the number of bins distinguished by effects on solubility: after selecting a model from Table 1, each exported datapoint is assigned a label from that system. If a reported effect is not present in the selected system, it is either converted to a partially compatible label or dropped. The process may be adjusted by selecting one of the abundance, reliability, or compromise modes. The first option converts as many values as possible; the second option leaves out all incompatible values; and the third option compromises on the significance, i.e., all converted labels are marked defensively as a slight change. Users can display the active conversion table by clicking 'See details'. The user interface for this step is shown in Fig. A.6.

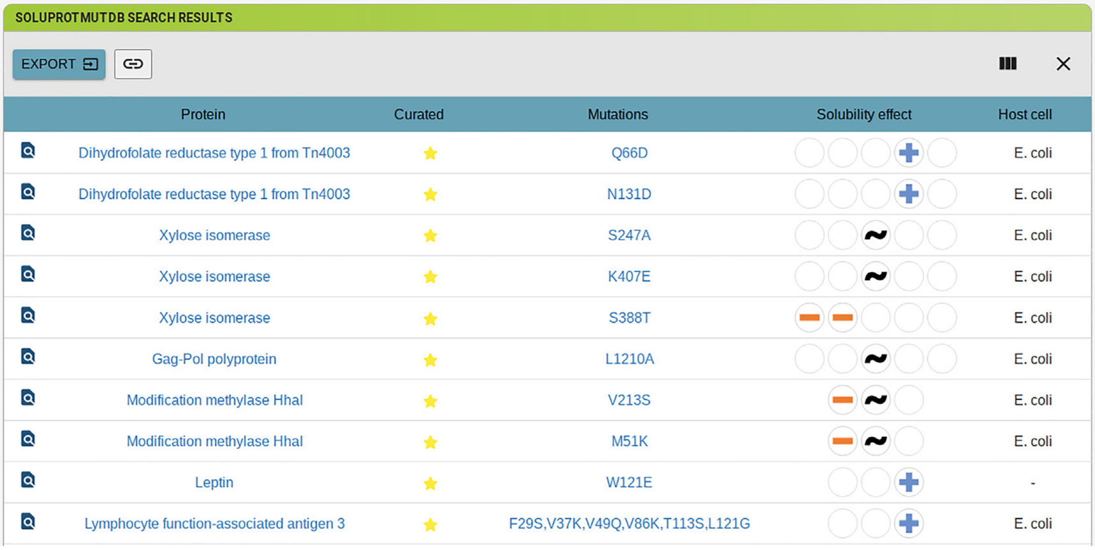
Finally, in the case of ML-dataset creation, users may want to use the data-augmentation (data-symmetrization) function, which adds the reverse mutations to the dataset, i.e., datapoints with substituted and target residues swapped and inverse solubility effects. This will resolve the likely problem of the imbalance between the counts of deteriorating and enhancing mutations (Fig. 2), which has often been reported to decrease the performance of predictors for other mutational data types [34–36].

4. Discussion

SoluProtMut^{DB} is the first mutational database of solubility data and is ready to serve as a central depository for data from mutagenesis experiments targeting protein solubility. To date, our database contains almost 33 000 experimental results of solubility effects upon mutations, thereby representing an essential digital resource for this type of data. The database comprises the previously published datasets and new data from the more recent literature. We have improved the reliability of these datasets by manual curation and overlap checks. We examined over a hundred original publications from which the data were gathered, including a few studies that produced hundreds to thousands of datapoints each, thanks to the use of such high-throughput experimental techniques as FACS. Lastly, we deposited the solubility data measured in our group. We will maintain the database, add new data, and continue with the curation process.

We believe the database is of great value for data scientists and will help to understand the mechanisms controlling solubility. With this in mind, we also focused on the ML potential of the database by making our database friendly for the ML community: (i) we ensured the data are reliable; (ii) we systematized the solubility effects reported in the literature to be easily understood by the experts outside biology; and (iii) we created Export Wizard to facilitate adaptation of the data for ready-made ML models. As a result, we expect that the user-friendly web interface and the other steps taken will broaden the audience and user community. The data can now be analyzed or modeled, even without a deep understanding of the underlying technical or biological details.

Thanks to the new data published in recent years, the database is an order of magnitude larger than an average solubility dataset. This abundance comes from recent high-throughput experiments, generating a more realistic distribution of target amino acids and observed effects compared to the previous datasets owing to the possibility of covering all possible single-point mutants.

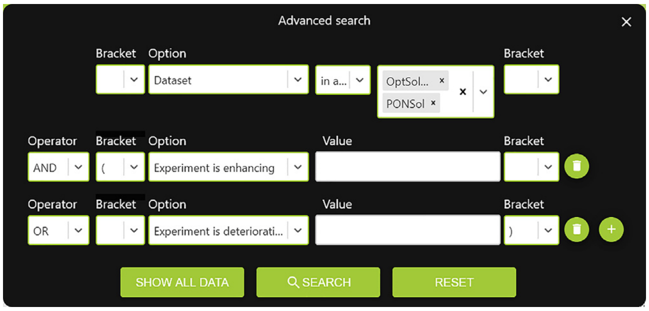


SOLUPROTMUTDB SEARCH RESULTS

EXPORT

Protein	Curated	Mutations	Solubility effect	Host cell
Dihydrofolate reductase type 1 from Tn4003	★	Q66D	○○○+○	E. coli
Dihydrofolate reductase type 1 from Tn4003	★	N131D	○○○+○	E. coli
Xylose isomerase	★	S247A	○○~○○	E. coli
Xylose isomerase	★	K407E	○○~○○	E. coli
Xylose isomerase	★	S388T	−−○○○	E. coli
Gag-Pol polyprotein	★	L1210A	○○~○○	E. coli
Modification methylase HhaI	★	V213S	−~○	E. coli
Modification methylase HhaI	★	M51K	−~○	E. coli
Leptin	★	W121E	○○+	-
Lymphocyte function-associated antigen 3	★	F29S,V37K,V49Q,V86K,T113S,L121G	○○+	E. coli

Fig. 5. An example of a result table. For clarity, only the most important columns are displayed by default: protein names, curation flags, mutations, solubility effects, and host cells. The table is paginated to avoid performance issues. A solubility effect graphic depicts both an effect and a value system provided in Table 1. The binning system is given by the number of circles, whereas the effect is given by one of the signs: **orange minus** (−) for deteriorating, black tilde (∼) for neutral and ***blue plus (+)*** for enha.ncing mutations.



Advanced search

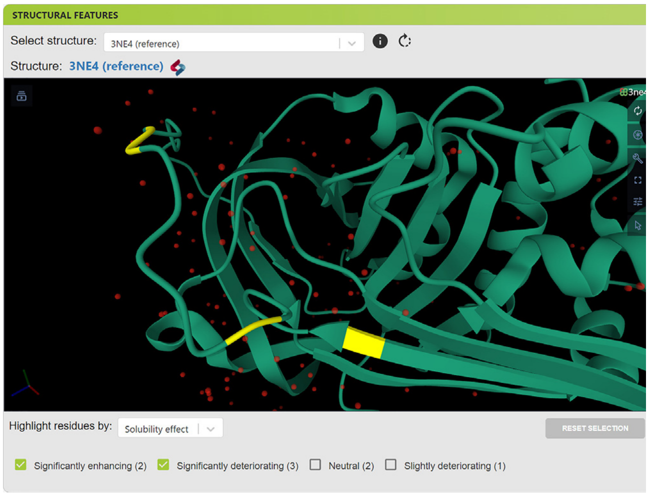
Bracket: — Option: Dataset — in a... — OptSol... ×, PONSol × — Bracket: —

Operator: AND — Bracket: (— Option: Experiment is enhancing — Value: — Bracket: —

Operator: OR — Bracket: — Option: Experiment is deteriorati... — Value: — Bracket:)

SHOW ALL DATA　SEARCH　RESET

Fig. 6. The advanced search with an example of a filtering protocol. In this example, the database will find measurements from OptSolMut and PON-Sol datasets with enhancing or deteriorating solubility effect.



STRUCTURAL FEATURES

Select structure: 3NE4 (reference)

Structure: 3NE4 (reference)

Highlight residues by: Solubility effect

RESET SELECTION

☑ Significantly enhancing (2)　☑ Significantly deteriorating (3)　☐ Neutral (2)　☐ Slightly deteriorating (1)

Fig. 7. The visualization of mutations in a protein with a known 3D structure. User-selected mutations can be highlighted in the structure. In this example, the mutated positions resulting in a significant change in solubility are highlighted in yellow.

Specifically, the DMS experiments manifest their strength as they show no extreme per-protein deviation of the effect distribution (Fig. 3) from the overall distribution (Fig. 2), which is of particular importance for ML applications. The DMS data are highly representative as they lack a selection bias in introduced mutations (Fig. A.3). Moreover, the substituted amino acids in the database follow the distribution of amino acids seen in nature (Fig. A.2). In contrast, the selection bias is apparent in the small-scale experiments, even when all their data are merged (Fig. A.4). In terms of effect distribution, the DMS data display more desolubilizing mutations (Fig. A.5). And since the DMS data are measured indirectly and a systematic error of a measurement may be present, we suggest using non-DMS data for ML model evaluation.

In order not to miss any important factor possibly affecting solubility, we track many conditions of experiments. Yet, several factors known or suspected to influence protein expression or solubility are not stored explicitly in the current version of the database. Some of these factors are silent mutations, i.e., mutations on the nucleotide-sequence level that do not propagate into the amino-acid sequence but may strongly influence soluble expression, especially heterologous [37]. Another factor is the time of expression, often not reported clearly, e.g., due to a possible complexity of the assay. Timings of different steps of an experiment may influence soluble expression, for example, through expression rate or by providing a different time for molecular interactions (precipitation, aggregation) [38].

Finally, the database promotes the FAIR principles not only by making the solubility data more accessible but also by allowing negative reporting. Currently, many negative findings in solubility experiments remain unreported as they do not bring the desired outcome to the scientists. We encourage the deposition of negative solubility data in SoluProtMut^{DB} to meet the obligations to publish results and reach FAIRness, often imposed by grant agencies. At the same time, these data are of considerable value for the field of ML, even to the extent comparable to that of positive results. Last but not least, non-reporting of negative findings may lead to repeating the same experiments and result in wasting human and material resources. Results of mutational solubility experiments can be sent to soluprot@sci.muni.cz to be deposited in the database.

CRediT authorship contribution statement

Jan Velecký: Software, Visualization, Data curation, Writing – original draft, Writing – review & editing. **Marie Hamsikova:** Software, Visualization, Writing – original draft, Writing – review & editing. **Jan Stourac:** Writing – review & editing, Supervision. **Milos Musil:** Software, Writing – original draft. **Jiri Damborsky:** Writing – review & editing, Supervision. **David Bednar:** Writing – review & editing, Supervision. **Stanislav Mazurenko:** Data curation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

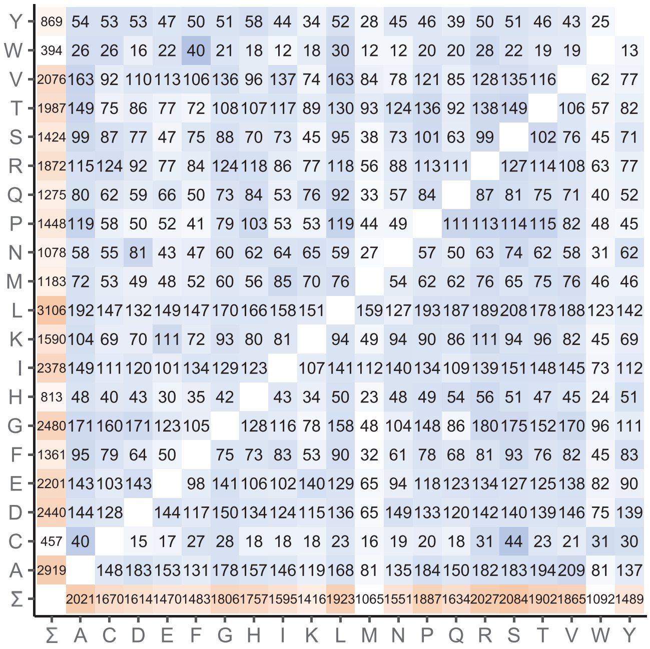
This work was supported by Czech Ministry of Education, Youth and Sports [INBIO CZ.02.1.01/0.0/0.0/16_026/0008451, CETOCOEN Excellence CZ.02.1.01/0.0/0.0/17_043/0009632, CETOCOEN Plus CZ.02.1.01/0.0/0.0/15_003/0000469, ESFRI RECETOX RI LM2018121, ESFRI ELIXIR LM2018131]; Technology Agency of the Czech Republic [FW03010208]; Brno University of Technology [FIT-S-20–6293 to M. M.]; CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 of the European Union [857560]; National Institute for Cancer Research [LX22NPO5102] funded by EU - Next Generation EU; and Czech Science Foundation [20-15915Y to D.B.].

This publication reflects only the authors' view, and the European Commission is not responsible for any use that may be made of the information it contains.

The authors thank the anonymous reviewers for their valuable suggestions.

Appendix A. Supportive information

Figs. A.1, A.2, A.3, A.4, A.5, A.6



	Σ	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Y	869	54	53	53	47	50	51	58	44	34	52	28	45	46	39	50	51	46	43	25	
W	394	26	26	16	22	40	21	18	12	18	30	12	12	20	20	28	22	19	19		13
V	2076	163	92	110	113	106	136	96	137	74	163	84	78	121	85	128	135	116		62	77
T	1987	149	75	86	77	72	108	107	117	89	130	93	124	136	92	138	149		106	57	82
S	1424	99	87	77	47	75	88	70	73	45	95	38	73	101	63	99		102	76	45	71
R	1872	115	124	92	77	84	124	118	86	77	118	56	88	113	111		127	114	108	63	77
Q	1275	80	62	59	66	50	73	84	53	76	92	33	57	84		87	81	75	71	40	52
P	1448	119	58	50	52	41	79	103	53	53	119	44	49		111	113	114	115	82	48	45
N	1078	58	55	81	43	47	60	62	64	65	59	27		57	50	63	74	62	58	31	62
M	1183	72	53	49	48	52	60	56	85	70	76		54	62	62	76	65	75	76	46	46
L	3106	192	147	132	149	147	170	166	158	151		159	127	193	187	189	208	178	188	123	142
K	1590	104	69	70	111	72	93	80	81		94	49	94	90	86	111	94	96	82	45	69
I	2378	149	111	120	101	134	129	123		107	141	112	140	134	109	139	151	148	145	73	112
H	813	48	40	43	30	35	42		43	34	50	23	48	49	54	56	51	47	45	24	51
G	2480	171	160	171	123	105		128	116	78	158	48	104	148	86	180	175	152	170	96	111
F	1361	95	79	64	50		75	73	83	53	90	32	61	78	68	81	93	76	82	45	83
E	2201	143	103	143		98	141	106	102	140	129	65	94	118	123	134	127	125	138	82	90
D	2440	144	128		144	117	150	134	124	115	136	65	149	133	120	142	140	139	146	75	139
C	457	40		15	17	27	28	18	18	18	23	16	19	20	18	31	44	23	21	31	30
A	2919		148	183	153	131	178	157	146	119	168	81	135	184	150	182	183	194	209	81	137
Σ		2021	1670	1614	1470	1483	1806	1757	1595	1416	1923	1065	1551	1887	1634	2027	2084	1902	1865	1092	1489

Fig. A.1. A matrix showing the numbers of mutation occurrences in the database 'from' (rows) and 'to' (columns) specific amino acids. The Σ column and row represent the total numbers of mutations 'from' and 'to' given amino acids, respectively. The matrix is *row-weighted* – blue saturation corresponds to the relative abundance of the given 'to' amino acid in the corresponding row. This is to avoid accentuation of differences naturally caused by the uneven distribution of amino acids in natural sequences.

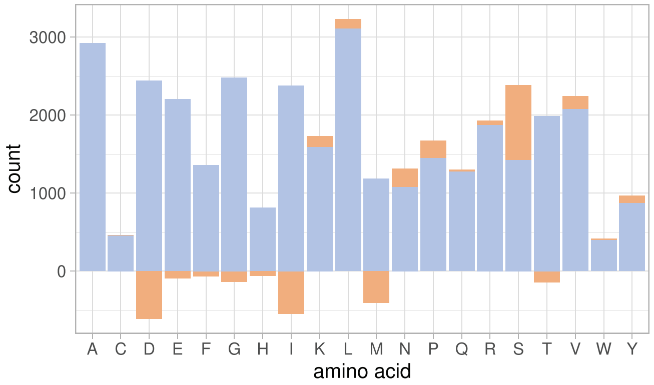
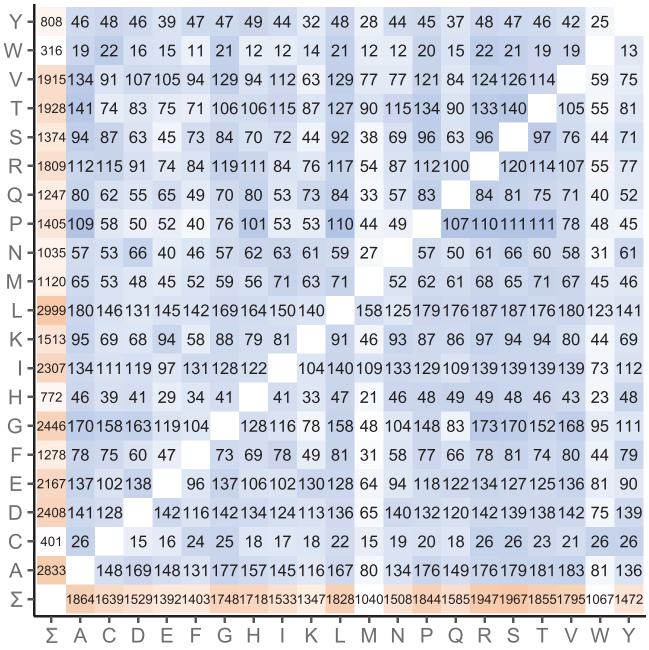
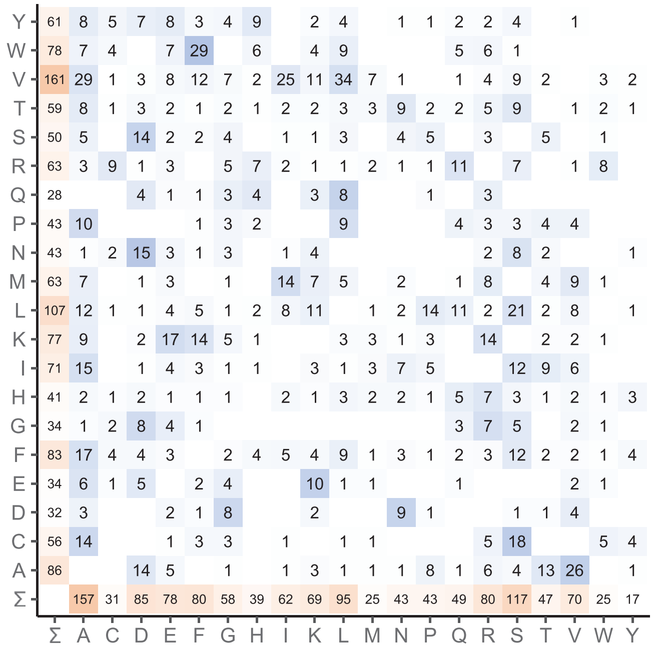


Fig. A.2. The histogram of the substituted amino acids in the database. The red bar shows the deviation from the natural distribution of amino acids, as in sequences of all kingdoms of life [39].



	Σ	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Y	808	46	48	46	39	47	47	49	44	32	48	28	44	45	37	48	47	46	42	25	
W	316	19	22	16	15	11	21	12	12	14	21	12	12	20	15	22	21	19	19		13
V	1915	134	91	107	105	94	129	94	112	63	129	77	77	121	84	124	126	114		59	75
T	1928	141	74	83	75	71	106	106	115	87	127	90	115	134	90	133	140		105	55	81
S	1374	94	87	63	45	73	84	70	72	44	92	38	69	96	63	96		97	76	44	71
R	1809	112	115	91	74	84	119	111	84	76	117	54	87	112	100		120	114	107	55	77
Q	1247	80	62	55	65	49	70	80	53	73	84	33	57	83		84	81	75	71	40	52
P	1405	109	58	50	52	40	76	101	53	53	110	44	49		107	110	111	111	78	48	45
N	1035	57	53	66	40	46	57	62	63	61	59	27		57	50	61	66	60	58	31	61
M	1120	65	53	48	45	52	59	56	71	63	71		52	62	61	68	65	71	67	45	46
L	2999	180	146	131	145	142	169	164	150	140		158	125	179	176	187	187	176	180	123	141
K	1513	95	69	68	94	58	88	79	81		91	46	93	87	86	97	94	94	80	44	69
I	2307	134	111	119	97	131	128	122		104	140	109	133	129	109	139	139	139	139	73	112
H	772	46	39	41	29	34	41		41	33	47	21	46	48	49	49	48	46	43	23	48
G	2446	170	158	163	119	104		128	116	78	158	48	104	148	83	173	170	152	168	95	111
F	1278	78	75	60	47		73	69	78	49	81	31	58	77	66	78	81	74	80	44	79
E	2167	137	102	138		96	137	106	102	130	128	64	94	118	122	134	127	125	136	81	90
D	2408	141	128		142	116	142	134	124	113	136	65	140	132	120	142	139	138	142	75	139
C	401	26		15	16	24	25	18	17	18	22	15	19	20	18	26	26	23	21	26	26
A	2833		148	169	148	131	177	157	145	116	167	80	134	176	149	176	179	181	183	81	136
Σ		1864	1639	1529	1392	1403	1748	1718	1533	1347	1828	1040	1508	1844	1585	1947	1967	1855	1795	1067	1472

Fig. A.3. A row-weighted substitution matrix for Whitehead's data. It shows some anomalies, such as visible under-representation of substitutions to methionine (M) or tryptophan (W).



	Σ	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Y	61	8	5	7	8	3	4	9		2	4		1	1	2	2	4		1		
W	78	7	4		7	29		6		4	9				5	6	1				
V	161	29	1	3	8	12	7	2	25	11	34	7	1		1	4	9	2		3	2
T	59	8	1	3	2	1	2	1	2	2	3	3	9	2	2	5	9		1	2	1
S	50	5		14	2	2	4		1	1	3		4	5		3		5		1	
R	63	3	9	1	3		5	7	2	1	1	2	1	1	11		7		1	8	
Q	28			4	1	1	3	4		3	8			1		3					
P	43	10				1	3	2			9				4	3	3	4	4		
N	43	1	2	15	3	1	3		1	4						2	8	2			1
M	63	7		1	3		1		14	7	5		2		1	8		4	9	1	
L	107	12	1	1	4	5	1	2	8	11		1	2	14	11	2	21	2	8		1
K	77	9		2	17	14	5	1			3	3	1	3		14		2	2	1	
I	71	15		1	4	3	1	1		3	1	3	7	5			12	9	6		
H	41	2	1	2	1	1	1		2	1	3	2	2	1	5	7	3	1	2	1	3
G	34	1	2	8	4	1									3	7	5		2	1	
F	83	17	4	4	3		2	4	5	4	9	1	3	1	2	3	12	2	2	1	4
E	34	6	1	5		2	4			10	1	1			1				2	1	
D	32	3			2	1	8			2			9	1			1	1	4		
C	56	14			1	3	3		1		1	1				5	18			5	4
A	86			14	5		1		1	3	1	1	1	8	1	6	4	13	26		1
Σ		157	31	85	78	80	58	39	62	69	95	25	43	43	49	80	117	47	70	25	17

Fig. A.4. A row-weighted substitution matrix for all but Whitehead's data. It shows the selection bias in the small-scale experiments. For example, alanine (A) or serine (S) is chosen as a substituent more frequently than other amino acids. Some of the biases are apparently due to avoidance of introducing a different functional group by a mutation, e.g., tryptophan (W) is mostly replaced with phenylalanine (F).

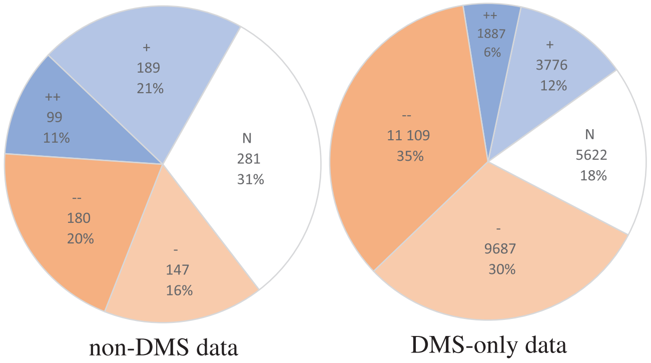
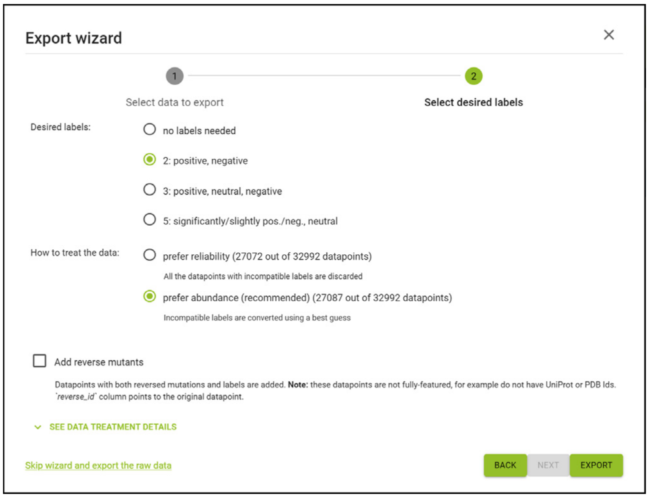


Fig. A.5. A comparison between the distributions of effects in the non-DMS and DMS-only datasets. The latter is skewed towards mutations having desolubilizing effect.



Export wizard

Desired labels:

○ no labels needed

● 2: positive, negative

○ 3: positive, neutral, negative

○ 5: significantly/slightly pos./neg., neutral

How to treat the data:

○ prefer reliability (27072 out of 32992 datapoints)

All the datapoints with incompatible labels are discarded

● prefer abundance (recommended) (27087 out of 32992 datapoints)

Incompatible labels are converted using a best guess

□ Add reverse mutants

Datapoints with both reversed mutations and labels are added. **Note:** these datapoints are not fully-featured, for example do not have UniProt or PDB Ids. 'reverse_id' column points to the original datapoint.

SEE DATA TREATMENT DETAILS
Skip wizard and export the raw data
BACK
NEXT
EXPORT

Fig. A.6. An example of the 2nd step of Export Wizard. Here, the solubility effect of all selected entries will be converted into the 2-value system using a best guess, and datapoints will be exported into a CSV file upon clicking on 'Export'. There is also an option to skip the wizard and export the raw data.

References

[1] Stourac J, Dubrava J, Musil M, Horackova J, Damborsky J, Mazurenko S, Bednar D. FireProtDB: database of manually curated protein stability data. Nucleic Acids Res 2020;49(D1):D319–24. https://doi.org/10.1093/nar/gkaa981.

[2] Kulandaisamy A, Sakthivel R, Gromiha MM. MPTherm: database for membrane protein thermodynamics for understanding folding and stability. Briefings Bioinform 2020;22(2):2119–25. https://doi.org/10.1093/bib/bbaa064.

[3] Wang X, Zhang X, Peng C, Shi Y, Li H, Xu Z, Zhu W. D3distalmutation: a database to explore the effect of distal mutations on enzyme activity. J Chem Inf Model 2021;61(5):2499–508. https://doi.org/10.1021/acs.jcim.1c00318.

[4] Shire SJ, Shahrokh Z, Liu J. Challenges in the development of high protein concentration formulations. J Pharm Sci 2004;93(6):1390–402. https://doi.org/10.1002/jps.20079. URL https://www.sciencedirect.com/science/article/pii/S0022354916315234.

[5] Vázquez-Rey M., Lang D.A. Aggregates in monoclonal antibody manufacturing processes, Biotechnol Bioeng 108 (7) (2011) 1494–1508, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.23155. doi:10.1002/bit.23155. https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.23155.

[6] W. Chen, X. Chen, Z. Hu, H. Lin, F. Zhou, L. Luo, X. Zhang, X. Zhong, Y. Yang, C. Wu, Z. Lin, S. Ye, Y. Liu, F. t. S.G.O. Ccpmoh, A Missense Mutation in CRYBB2 Leads to Progressive Congenital Membranous Cataract by Impacting the Solubility and Function of βB2-Crystallin, PLOS ONE 8 (11) (2013) e81290, publisher: Public Library of Science. doi:10.1371/journal.pone.0081290. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081290.

[7] Tian Y, Deutsch C, Krishnamoorthy B. Scoring function to predict solubility mutagenesis. Algorith Mol Biol 2010;5(1):33. https://doi.org/10.1186/1748-7188-5-33.

[8] Sormanni P, Aprile FA, Vendruscolo M. The camsol method of rational design of protein mutants with enhanced solubility. J Mol Biol 2015;427(2):478–90. https://doi.org/10.1016/j.jmb.2014.09.026.

[9] 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(W1):W306–13. https://doi.org/10.1093/nar/gkv359.

[10] Yang Y, Niroula A, Shen B, Vihinen M. PON-sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics 2016;32(13):2032–4. https://doi.org/10.1093/bioinformatics/btw066.

[11] Yang Y, Zeng L, Vihinen M. Pon-sol2: Prediction of effects of variants on protein solubility. Int J Mol Sci 2021;22(15). https://doi.org/10.3390/ijms22158027. URL https://www.mdpi.com/1422-0067/22/15/8027.

[12] Klesmith J.R., Bacik J.-P., Wrenbeck E.E., Michalczyk R., Whitehead T.A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning, Proc of the Natl Acad of Sci USA 114 (9) (2017) 2265–2270. arXiv: https://www.pnas.org/content/114/9/2265.full.pdf, doi:10.1073/pnas.1614437114. https://www.pnas.org/content/114/9/2265.

[13] Wrenbeck E, Bedewitz M, Klesmith J, Noshin S, Barry C, Whitehead T. An automated data-driven pipeline for improving heterologous enzyme expression. ACS Synthet Biol 2019;8(02). https://doi.org/10.1021/acssynbio.8b00486.

[14] Mazurenko S, Prokop Z, Damborsky J. Machine Learning in Enzyme Engineering. In: ACS Catal, 10. publisher: American Chemical Society; 2020. p. 1210–23. https://doi.org/10.1021/acscatal.9b04321.

[15] T.U. Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res 49 (D1) (2020) D480–D489. doi:10.1093/nar/gkaa1100. URL https://doi.org/10.1093/nar/gkaa1100.

[16] Sumbalova L., Stourac J., Martinek T., Bednar D., Damborsky J. HotSpot wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information, Nucleic Acids Res 46 (W1) (2018) W356–W362. https://doi.org/10.1093/nar/gky417.

[17] Kaur J, Kumar A, Kaur J. Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. Int J Biol Macromol 2018;106:803–22. https://doi.org/10.1016/j.ijbiomac.2017.08.080.

[18] Slanská K. Study of protein solubility [online] Master's thesis, Faculty of Science, Masaryk University, Brno (2021). URL Availableat<https://is.muni.cz/th/e3jlf/>

[19] Bendl J., Stourac J., Sebestova E., Vavra O., Musil M., Brezovsky J., Damborsky J. HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering, Nucleic Acids Res 44 (Web Server issue) (2016) W479–W487. doi:10.1093/nar/gkw416. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987947/.

[20] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.

[21] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics (Oxford, England) 2015;31(6):926–32. https://doi.org/10.1093/bioinformatics/btu739.

[22] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 2010;26(19):2460–1. https://doi.org/10.1093/bioinformatics/btq461.

[23] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539. https://doi.org/10.1038/msb.2011.75.

[24] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics (Oxford, England) 2007;23(15):1875–82. https://doi.org/10.1093/bioinformatics/btm270.

[25] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.

[26] Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79(2):351–71. https://doi.org/10.1016/0022-2836(73)90011-9.

[27] Reetz M.T., Carballeira J.D., Vogel A. Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability, Angewandte Chem Int Ed 45(46) (2006) 7745–7751, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200602795. doi:10.1002/anie.200602795. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200602795.

[28] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinform 2009;10:168. https://doi.org/10.1186/1471-2105-10-168.

[29] Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, Gora A, Sustr V, Klvana M, Medek P, Biedermannova L, Sochor J, Damborsky J. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 2012;8(10):. https://doi.org/10.1371/journal.pcbi.1002708e1002708.

[30] Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin M-J, Kleywegt GJ. SIFTS: Structure integration with function, taxonomy and sequences resource. Nucleic Acids Res 2012;41(D1): D483–9. https://doi.org/10.1093/nar/gks1258.

[31] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Sci Data 3(1) (Mar. 2016). doi:10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.

[32] Watkins X, Garcia LJ, Pundir S, Martin MJ. the UniProt Consortium, Protvista: visualization of protein sequence annotations. Bioinformatics 2017;33 (13):2040–1. https://doi.org/10.1093/bioinformatics/btx120.

[33] Sehnal D., Bittrich S., Deshpande M., Svobodova R., Berka K., Bazgier V., Velankar S., Burley S.K., Koca J., Rose A.S. Mol* viewer: modern web app for 3d visualization and analysis of large biomolecular structures, Nucleic Acids Res 49(W1) (2021) W431–W437. https://doi.org/10.1093/nar/gkab314.

[34] Pucci F, Schwersensky M, Rooman M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr Opin Struct Biol 2022;72:161–8. https://doi.org/10.1016/j.sbi.2021.11.001. URL https://www.sciencedirect.com/science/article/pii/S0959440X21001445.

[35] Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Briefings Bioinform 2020;21(4):1285–92. https://doi.org/10.1093/bib/bbz071.

[36] Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotechnol J 2020;18:1968–79. https://doi.org/10.1016/j.csbj.2020.07.011.

[37] Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. Trends Biotechnol 2004;22(7):346–53. https://doi.org/10.1016/j.tibtech.2004.04.006. URL https://www.sciencedirect.com/science/article/pii/S0167779904001118.

[38] Kuroda Y. Biophysical studies of protein solubility and amorphous aggregation by systematic mutational analysis and a helical polymerization model. Biophys Rev 2018;10(2):473–80. https://doi.org/10.1007/s12551-017-0342-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5899702/.

[39] Kozlowski LP. Proteome-pI: proteome isoelectric point database. Nucleic Acids Res 2017;45(D1):D1112–6. https://doi.org/10.1093/nar/gkw978.