BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INFORMATION SYSTEMS
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

MINING OF SOLUBLE ENZYMES FROM GENOMIC DATABASES

DOLOVÁNÍ ROZPUSTNÝCH ENZYMŮ Z GENOMICKÝCH DATABÁZÍ

PHD THESIS
DISERTAČNÍ PRÁCE

AUTHOR Ing. JIŘÍ HON
AUTOR PRÁCE

SUPERVISOR doc. Ing. JAROSLAV ZENDULKA, CSc.
ŠKOLITEL

BRNO 2021

Abstract

Enzymes are proteins accelerating chemical reactions, which makes them attractive targets for both pharmaceutical and industrial applications. The enzyme function is mediated by several essential amino acids which form the optimal chemical environment to catalyse the reaction. In this work, two integrated bioinformatics tools for mining and rational selection of novel soluble enzymes, EnzymeMiner and SoluProt, are presented.

EnzymeMiner uses one or more enzyme sequences as input along with a description of essential residues to search the protein database. The description of essential amino acids is used to increase the probability of similar enzymatic function. EnzymeMiner output is a set of annotated database hits. EnzymeMiner integrates taxonomic, environmental, and protein domain annotations to facilitate selection of promising targets for experiments. The main prioritization criterion is solubility predicted by the second tool being presented, SoluProt.

SoluProt is a machine-learning method for the prediction of soluble protein expression in *Escherichia coli*. The input is a protein sequence and the output is the probability of such protein to be soluble. SoluProt exploits a gradient boosting machine to decide on the output prediction class. The tool was trained on TargetTrack database. When evaluated against a balanced independent test set derived from the NESG database, SoluProt accuracy was 58.5% and its AUC 0.62, slightly exceeding those of a suite of alternative solubility prediction tools. Both EnzymeMiner and SoluProt are frequently used by the protein engineering community to find novel soluble biocatalysts for chemical reactions. These have a great potential to decrease energetic consumption and environmental burden of many industrial chemical processes.

Abstrakt

Enzymy jsou proteiny urychlující chemické reakce s velkým potenciálem pro farmaceutický a obecně chemický průmysl. Enzymatická funkce je obvykle zajištěna několika nepostradatelnými aminokyselinami, které tvoří tzv. aktivní místo, kde se odehrává chemická reakce. V této práci jsou prezentovány dva integrované softwarové nástroje pro dolování a racionální výběr nových rozpustných enzymů – EnzymeMiner a SoluProt.

EnzymeMiner slouží k hledání nových enzymů. Na vstupu vyžaduje jednu nebo více sekvencí zvoleného enzymu spolu se seznamem klíčových aminokyselin. Tento seznam slouží k zvýšení pravděpodobnosti, že nalezený enzym bude mít podobnou funkci jako vstupní enzym. Výstupem EnzymeMineru je množina anotovaných sekvencí nalezených v databázi. Za účelem ulehčení výběru několika málo kandidátů pro experimentální ověření v laboratoři integruje EnzymeMiner anotace z dostupných databází – informaci o zdrojovém organismu a prostředí, ve kterém se vyskytuje, a informaci o proteinových doménách, ze kterých se enzym skládá. Hlavním kritériem pro výběr kandidátů je rozpustnost predikovaná druhým prezentovaným nástrojem, SoluProtem.

SoluProt je metoda založená na strojovém učení, která predikuje heterologní rozpustnou expresi proteinu v organismu *Escherichia coli*. Vstupem je sekvence a výstupem je pravděpodobnost, že protein bude exprimován v rozpustné formě. SoluProt využívá model gradient boosting machine a byl trénován na datové sadě odvozené od databáze TargetTrack. Při srovnání na vyvážené nezávislé datové sadě odvozené z databáze NESG dosáhl SoluProt přesnosti 58,5% a hodnoty AUC 0,62, čímž lehce převyšuje ostatní existující nástroje. Nástroje EnzymeMiner i SoluProt jsou často využívány řadou uživatelů z oblasti proteinového inženýrství za účelem hledání nových rozpustných biokatalyzátorů chemických reakcí. Ty mají velký potenciál snížit energetickou náročnost a ekologickou zátěž mnoha průmyslových procesů.

Keywords

enzyme mining, protein solubility, protein engineering, machine-learning

Klíčová slova

dolování enzymů, rozpustnost proteinů, proteinové inženýrství, strojové učení

Reference

HON, Jiří. *Mining of soluble enzymes from genomic databases.* Brno, 2021. PhD thesis. Brno University of Technology, Faculty of Information Technology. Supervisor doc. Ing. Jaroslav Zendulka, CSc.

Mining of soluble enzymes from genomic databases

Declaration

I hereby declare that this Thesis was prepared as an original work by the author under the supervision of doc. Ing. Jaroslav Zendulka, CSc. The supplementary information was provided by Ing. Tomáš Martínek, Ph.D. and prof. Mgr. Jiří Damborský, Dr. I have listed all the sources, which were used during the preparation of this Thesis.

....
Jiří Hon
August 31, 2021

Acknowledgements

I would like to thank everyone who helped me to solve all the challenging problems related to this Thesis. I especially thank Martin Marušiak and Simeon Borko for being excellent partners in the development process of the presented software and for being inspiring debaters raising fundamental questions, not only scientific ones. I thank Jaroslav Zendulka, Tomáš Martínek, David Bednář and Jiří Damborský for their supporting guidance, sincere comments and knowledgeable suggestions improving the developed software and the text of the Thesis. I thank Joan Planas-Iglesias for his insightful proofreading. I thank Tomáš Martínek for showing me the beauty of the bioinformatics field and for patient guidance through the pitfalls of doctoral studies. I thank my wife for her unlimited support. I thank God for his bountiful gifts.

Contents

Chapter 1

Introduction

Proteins are molecules that play essential roles in all living organisms. They provide structure to cells and perform key functions, such as DNA replication, molecule transportation, regulation, cell signaling, or catalysis of biochemical reactions. Proteins catalysing chemical reactions are called enzymes or biocatalysts. The catalytic function of enzymes is mediated by several essential amino acids which form the optimal chemical environment to accelerate the reaction. Enzymes are attractive targets for pharmaceutical and industrial applications because of reduced process time, intake of low energy input, cost effective, nontoxic and eco-friendly characteristics [18]. They are successfully used in drug design, biofuel production, detergents, waste treatment, food processing, paper industry, and many others [16].

There are currently more than 395 million non-redundant proteins[1] in protein sequence databases [73], approximately 16% of which are assumed to be enzymes[2]. Despite their enormous promise for biological and biotechnological discovery, a thorough experimental characterization has been only performed for 0.3% of the proteins available[3] because current biochemical characterization techniques are time- and resource-demanding. Therefore, computational methods are currently more convenient to explore the immense protein diversity available among the millions of uncharacterised protein entries. However, existing computational mining approaches for large databases usually yield hundreds or thousands of hits. Production and experimental testing of all of the hits would be extremely resource demanding and cost-ineffective. Therefore, prioritization methods narrowing down the selection to just the most promising hits based on computational analysis are needed.

An important prioritization criterion is protein solubility which is one of the most critical factors limiting the success of protein production. Insufficient protein solubility is probably the most common cause of failure of protein production and experimental characterization pipelines as evidenced by the large-scale Protein Structure Initiative (PSI) project [8]. PSI sought to produce thousands of different protein sequences. In 81% of cases[4], it was probably impossible to produce the target proteins in soluble form. Although many computational tools were developed to predict solubility, obtaining an accurate estimation of protein solubility is still an open problem.

[1] NCBI NR database release 2021-05-14.

[2] Number of records with an assigned enzyme class (https://www.uniprot.org/uniprot/#enzymesViewBy) to the total number of records in UniProtKB.

[3] Number of reviewed records in Swiss-Prot database to the total number of records in UniProtKB.

[4] Based on the total number of purified targets to the total number of targets in the TargetTrack database (http://sbkb.org/metrics/)

The goal of the Thesis is to develop integrated tools for mining soluble enzymes from protein databases. The only required input for the analysis should be one or more enzyme sequences and a description of enzyme's essential residues to keep the analysis easy to set up, thus enabling broad applicability of the tools. The essential residues should be used in additional filtering steps in order to increase the specificity of the results. The output of the analysis should be a list of protein sequences annotated with the predicted solubility, sequence similarity to the input sequences and living environment of the source organism. The source environment is of particular interest as enzymes from organisms adapted to harsh conditions—e.g. extreme temperatures, broad range of pH, and extreme salinity, could also show higher adaptation to such extreme conditions. This would be beneficial for practical applications in industry where enzymes are required to sustain more demanding operating conditions. The predicted solubility and sequence similarity will be the primary prioritization criteria to select hits for further experimental characterization. A new method should be developed for more accurate sequence-based solubility prediction.

1.1 Specific objectives of the Thesis

1. To develop a method for mining enzymes from NCBI Protein database [73] based on the input protein sequence and the description of the essential amino acids.

2. To integrate protein domain annotations [33] and environment annotations [7] into the mining method to facilitate a rational selection of promising hits for experimental characterization.

3. To compile training set and independent test set for sequence-based protein solubility prediction.

4. To develop state-of-the-art sequence-based solubility prediction method.

5. To integrate the solubility prediction method into the protein mining method for prioritization of hits by predicted solubility.

1.2 Organization of the Thesis

The Thesis is organised into six chapters. In the Chapter 2, the field of protein engineering is introduced, its methods are described, and a newly emerging approach—rational selection of enzymes, is presented. In the Chapter 3, existing methods for identification of proteins in databases and visualization of results are discussed along with three main databases used for mining of enzymes. In the Chapter 4, the current state of the art in the field of protein solubility prediction is presented. The Chapter 5 presents published works and describe author's participation on each of the result. The last Chapter 6 summarises important points about the presented research. Three articles published in peer-reviewed journals as a result of the Thesis are attached in the Appendices. The first article (Appendix A) describes EnzymeMiner—a method for mining enzymes. The second article (Appendix B) presents SoluProt—a state-of-the-art sequence-based solubility prediction method. The third article (Appendix C) provides a summary and critical assessment of existing protein solubility prediction methods.

Chapter 2

Protein engineering

Protein engineering [11] is the process of developing proteins with novel properties by modifying the sequences of naturally existing proteins, so-called wild-type proteins. Protein sequences are chains of amino acids, which can be adjusted by substituting, adding, or removing specific amino acids. To design these modifications, protein engineering requires understanding of molecular mechanisms and biological processes governing the genesis and evolution of proteins and their interactions with other molecules. Such understanding is then used to intentionally design novel proteins applicable for therapeutic or industrial purposes.

There exists two well established protein engineering methods: directed evolution [63] and rational design [39]. These are not mutually exclusive but instead might be combined to obtain even better design products. Rational selection [85] is a third emerging method complementing the previous two by suggesting alternative wild-type proteins. All three protein engineering methods can be described in terms of seven engineering steps: (i) computer aided analysis and application of expert knowledge, (ii) generation of gene library, (iii) transformation of the genes into target expression system, (iv) protein expression, (v) protein purification, (vi) broad screening and selection for target biochemical property and (vii) in-depth biochemical characterization (Figure 2.1).

2.1 Directed evolution

Directed evolution is a method that mimics natural evolutionary processes that generated the current set of proteins present in nature. The central idea of directed evolution is to rapidly mutate genes encoding the target protein at random positions. The best gene variants are then selected by fitness value screening, which is derived from biochemical testing of the resulting proteins.

The key biochemical process used for directed evolution is a random mutagenesis based on an error-prone polymerase chain reaction (PCR) [57] that allows introduction of random nucleotide substitutions into the target gene. Such mutagenesis results in a large library of mutated genes. The gene-encoding DNA is then transferred into either an *in vivo* or *in vitro* expression system to produce the corresponding proteins. *In vivo* systems use the natural protein production machinery of living organisms which are coerced to accept and translate the modified DNA. *In vitro* expression systems do not require a living organism but consists of all chemicals necessary for initiation of gene transcription and RNA translation [75].

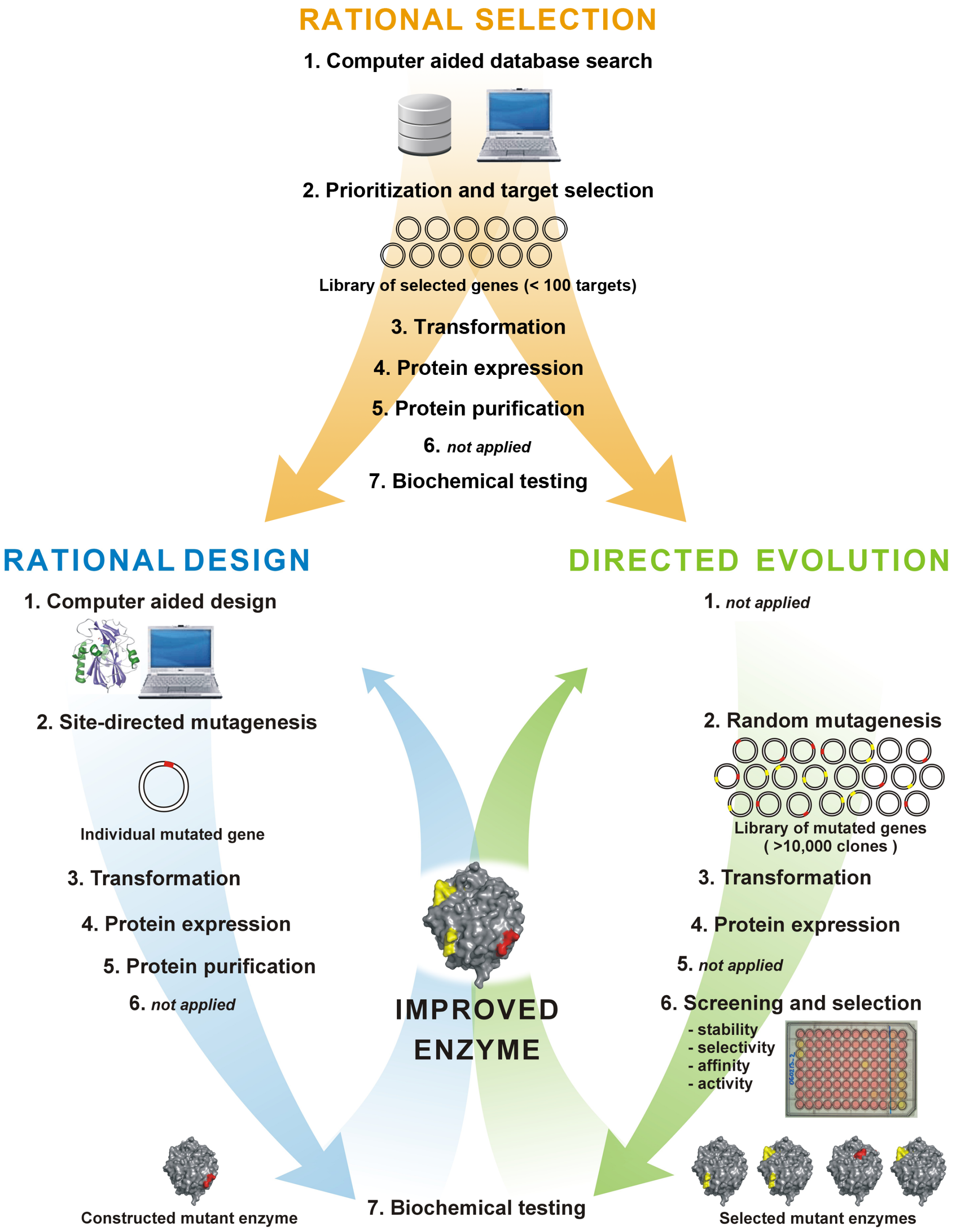


Figure 2.1: Protein engineering methods. The goal of protein engineering is to design a protein, usually an enzyme for catalysis of biochemical reactions, with improved properties. Rational design uses previous expert knowledge and computational simulations to design individual improved protein variants. Directed evolution relies on random mutagenesis and high-throughput screening of generated gene libraries. Rational selection provides alternative starting proteins based on computer aided database mining for both rational design and directed evolution. The figure was adapted from the previous work by Damborský [24].

2.2 Rational design

Rational design uses knowledge of the target protein and extensive computational analysis to design specific variants of the initial protein. A small set of designed genes is then synthesised biochemically and produced using preferred expression system. However, the three-dimensional structure of the protein is required to reliably design specific variants. Conveniently, such information is available for many proteins thanks to the advent of X-ray crystallography [79], which helped to solve the 3D structure of proteins with a resolution even below one angstrom (0.1 nanometers). The set of reliable protein structures can be expanded by currently available protein structure predictors that can reach accuracy of values close to those of X-ray crystallography [53]. In this context, the protein structure is used to infer important sites, responsible for binding other molecules or facilitating biochemical reactions (catalytic sites). To understand and predict the effect of an amino acid substitution, extensive computational analysis and simulations are essential [45, 3].

The combination of rational design and directed evolution is often favourable, leading to a semi-rational design. The sequence space of a designed protein is reduced by computational approaches to several amino acids and these positions are subjected to site-directed or saturation mutagenesis to generate small-sized mutant libraries. These smart libraries are then screened for desired function similarly as in the process of directed evolution. A distinctive discipline of rational design is *de novo* design [71]. The target protein is constructed from scratch using small building blocks. On the one hand, this is the most demanding approach and requires large amount of computational power and extensive laboratory work in multi-step iterative process. On the other hand, it might produce proteins displaying properties never observed before in nature.

2.3 Rational selection

Both directed evolution and rational design require the knowledge of the input protein sequence. However, the question about which protein is the best target for protein engineering needs to be answered by an expert on the field. Usually, protein engineers select from a set of proteins for which exists a previous knowledge of their function and biological context based on available experimental characterization. However, there are cases, where such characterization is not available or where the characterised proteins display poor biochemical properties, such as solubility, thermodynamic stability, activity, or substrate specificity. That kind of proteins make for a poor starting point for engineering. As protein engineering methods are most effective in making gradual improvements rather than great leaps, it is advisable to identify the protein with better starting properties and, if necessary, optimise it by protein engineering.

The current knowledge about existing proteins is accumulated in large protein databases (Protein Data Bank—PDB [9], UniProtKB [83], NCBI Protein [73]). UniProtKB and NCBI Protein databases contain both experimentally confirmed proteins and hypothetical proteins computationally extracted and annotated from known genomes. According to the UniProtKB database—a comprehensive, high-quality, and freely accessible resource of protein sequence and functional information, approximately 500,000 confirmed proteins represent less than 0.3% of all deposited proteins. The remaining 209 million uncharacterised proteins represent a great wealth of potentially interesting and diverse proteins for both basic science and industrial applications.

Rational selection methods help to search in the millions of uncharacterised protein sequences and to suggest the most promising targets. Here, the key hypothesis is that similar proteins from different organisms perform similar function, but encompass different interesting properties because they adapted to different environmental conditions. Some specially adapted organisms—so called extremophiles (Table 2.1), live in otherwise unbearable conditions such as extreme temperature, wide pH range, and extreme salinity [28]. Proteins from such organisms had to evolve to withstand these harsh conditions and preserve their function. The rational selection methods combine expert knowledge with database searches based on sequence similarity, identification of protein domains, identification of essential residues, computational annotations, prediction of 3D structure and analysis of structural properties [85, 55].

Table 2.1: Classification of extremophilic organisms. Extremophiles might be a good source of proteins suitable for industrial applications which often requires resilience to non-standard conditions. The table was adapted from the previous work by Vaňáček [84].

Name	**Growth characteristics**
Psychrophile	Growing at low-temperature optimum (∼0–20°C)
Thermophile	Growing at high-temperature optimum (∼60–80°C)
Hyperthermophile	Growing at extremely temperature optimum (∼80–120°C)
Acidophile	Growing at acidic pH optimum (pH <4)
Alkalophile	Growing at alcalic pH optimum (pH >8)
Halophile	Growing at high salt concentration (>1 M NaCl)
Piezophile	Tolerating high hydrostatic pressures (p >40 MPa)
Metalophile	Tolerating the presence of high level of heavy metals
Radiophile	Resisting to high ionizing and ultraviolet levels
Oligotroph	Growing in nutritionally deplete habitats
Toxitolerant	Resisting to high levels of damaging agents
Xerophile	Tolerating the low water level and resisting to high desiccation

Chapter 3

Searching for proteins in databases

The basic computer representation of a protein is a protein sequence—a string of letters from the standardised amino acid alphabet [44]. Other metadata describing the current knowledge about the protein are usually available. The protein sequence and associated metadata are deposited in protein sequence databases.

Both protein sequence and metadata may be trusted to a different extent based on the source of each data. The annotation can be either experimentally determined, manually added by a domain expert, predicted computationally, or obtained by combination of both experiments and predictions. A typical path through which a new protein sequence gets to a protein database is as follows: (i) DNA is read using DNA sequencing method, (ii) protein coding genes are identified in the DNA sequence, (iii) the genes are translated into their amino acid sequence, and (iv) the protein sequence is deposited into the database.

Millions of novel protein sequences are being added to databases every year thanks to efficiency of sequencing technologies [13]. However, most of the novel gene products are only computationally predicted and automatically translated to proteins using standard codon tables. Few of such novel proteins are manually reviewed or experimentally characterised because of the vast human and laboratory resources such tasks demand. The fact that most proteins in databases are uncharacterised makes searching for proteins with desired biochemical properties a difficult task.

3.1 Databases

In this section, three main protein sequence databases are introduced which can be used as a data source for identification of novel enzymes.

3.1.1 UniProtKB

UniProtKB [83] is a joint protein database of three collaborating institutions associated in the UniProt consortium—the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). It is the most comprehensive protein database available. UniProtKB provides information on protein function, involvement in regulatory paths, cell location, and pathological associations connected to protein variants. It also provides annotation of post-translational protein modifications and interactions of the protein. UniProtKB is linked to the Protein Data Bank (PDB) of 3D structures of available macromolecules.

UniProtKB has two parts: (i) TrEMBL (>209,000,000 records) which contains automatically annotated and not reviewed records (ii) Swiss-Prot (500,000 records) which provides curated annotations. The enormous size-difference of these two UniProtKB parts highlights the extreme contrast between the fast and easy prediction of novel protein sequences and their slow and tedious experimental characterization.

The UniProt consortium maintains three other large databases: UniParc—an archive of all available protein sequences containing one record for each unique protein sequence; UniRef—sets of protein sequences clustered at various sequence identity thresholds; and MGnify—a repository of metagenomic and environmental data.

3.1.2 NCBI Protein

The NCBI Protein database [73] is the protein sequence database of National Center for Biotechnology Information (NCBI) of the National Library of Medicine. The database is mainly based on coding sequence translations from two other NCBI databases of genomic assemblies—GenBank and RefSeq. It also cross-references other protein databases like Swiss-Prot or PDB. Beside the Protein database, NCBI maintains a compilation of non-identical protein sequences—the non-redundant (NR) database, being similar to the UniParc database. Currently, the NR database is the largest database of non-redundant genomic protein sequences.

3.1.3 BRENDA

BRENDA [14] is a database from the Technische Universität Braunschweig specialised on enzymes—the proteins involved in catalysis of biochemical reactions. BRENDA provides information on enzyme and ligand nomenclature, source organism, reaction and specificity, kinetic properties, structure and role of the ligands, stability information, ligand-enzyme information, enzyme sequence and structure, mutants, connection to diseases, and biological pathways. BRENDA is the most comprehensive source of enzyme-related information attempting to map entire enzyme classes and families.

3.2 Existing approaches

There are two basic approaches to search in protein sequence databases: (i) metadata-based search and (ii) sequence-based search. Both methods can be combined to achieve more refined results.

Metadata-based search uses protein metadata like protein name, genomic location, source organism, domain and functional annotations, and references to other databases to find relevant hits. This type of search is supported by UniProtKB, NCBI Protein database, and BRENDA. The user can construct arbitrary complex queries by combining terms using the logical operators AND, OR, NOT. The metadata-based search has two major advantages. First, it is fast as search indexes can be precomputed for the terms to speed up the query evaluation. Second, the query can target specific protein properties such as genomic context, ligand-enzyme relation, or protein function. The main disadvantage of this approach is the low confidence of most annotations. As mentioned previously, most of the protein sequences are automatically annotated and the accuracy of such annotations is still modest [70]. Therefore, the results need additional evaluation. Metadata-based methods are herein not further discussed as the Thesis is focused on sequence-based search methods.

In contrast, sequence-based search uses the protein sequence as a query to perform the search in the database. Due to a natural evolutionary variation in proteins, algorithms based on sequence similarity are required. These algorithms reflect biological homology of the protein sequences, defined in terms of shared ancestry in the evolutionary history. Notably, the substitution of biologically similar amino acids is less penalised by this strategy than the substitution of biologically distant ones. The key advantage of sequence-based search methods over metadata-based ones is the specificity of the search. Proteins showing high sequence similarity tend to have similar biochemical properties. The output of the sequence-based search is the alignment of the hit proteins with the query itself. Thus, the user can further check if important functional parts of the protein are correctly aligned. The disadvantage of the approach is its high computational complexity in comparison to metadata search. An alignment needs to be computed for many protein sequence pairs which is time-demanding. The alignment could be computed using optimal deterministic algorithms, such as those designed by Needleman and Wunsch [60] or by Smith and Waterman [78], but in practice, faster heuristic approaches are used. There are many existing algorithms differing in speed, memory usage, and sensitivity of the search, including BLAST [4], DIAMOND [12], MMseqs2 [82], UBLAST [32], RAPSearch2 [90], FASTA [64], and HMMER [31]. Here, two most-frequently used ones—BLAST and HMMER, are discussed as well as other enzyme-specific methods that consider the catalytic function of the enzyme.

3.2.1 BLAST

Basic Local Alignment Search Tool (BLAST) [4] is one of the oldest and most widely used heuristic algorithms for sequence-based search in protein databases. BLAST nucleates regions of similarity from minimal alignments, whose length is determined by the word-size parameter and extends them to produce local alignments. Due to the modular nature of proteins, this local alignment approach outperforms optimal methods in finding shorter stretches of sequence similarity while producing results in much shorter time. BLAST outputs an alignment bit score for each hit which is based on selected scoring matrix, typically some matrix from the BLOSUM family [40], although PAM matrices [26] can be used as well. These matrices are position-independent, giving always the same score for a particular amino acid substitution regardless of the position in the query sequence.

To assess statistical relevance of the hits, BLAST provides a measure called expectation value (E-value). It gives the expected number of how many times an alignment with a better or equal score could be found by chance in a database of a particular size. Hence, the lower the E-value the more statistically significant the found solution is. The E-value is given by Equation 3.1, where m is the length of the query sequence, N is the length of all sequences in the database (total number of residues) and S' is a bit score.

$$E = mN2^{-S'} \tag{3.1}$$

The E-value can be transformed to a P-value (Equation 3.2) which gives probability that a hit with a greater or equal score is found in a database of particular size.

$$P = 1 - e^{-E} \tag{3.2}$$

To improve the results of the search, BLAST can use a position specific scoring matrix (PSSM) as an alternative to a single query sequence and the position independent scoring matrix. The PSSM is based on alignment of multiple sequences similar to the query. The

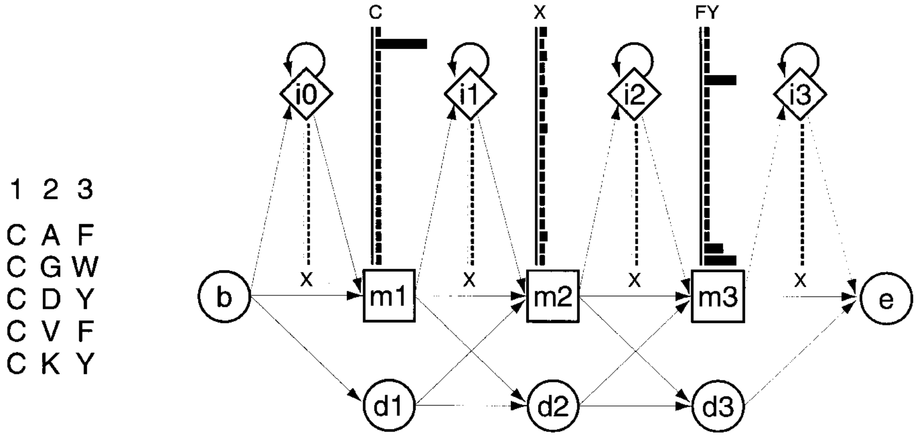


Figure 3.1: A profile HMM (right) representing a short multiple alignment of five sequences (left) with three consensus columns. The three columns are modelled by three match states M, each of which has 20 residue emission probabilities, shown with black bars. Insert states I also have 20 emission probabilities each. Delete states D have no emission probabilities. A begin and end state are included (b and e, respectively). State transition probabilities are shown as arrows. The figure was adapted from the previous work by Eddy [30].

matrix represents probability of the 20 different amino acids at each position in the query protein and it is of size $L \times 20$, where L is the length of the query sequence. The score for aligning an amino acid with a PSSM position is given by the matrix itself, rather than by reference to a fixed scoring matrix. The PSSM accounts for the allowed variability in the protein and can express that certain parts of the protein are evolutionary more conserved and thus do not allow substitutions.

The PSSM approach is used for iterative search by the PSI-BLAST tool which enhances the sensitivity of the search. In the first iteration, a classic BLAST search is performed using a single query sequence. Then a PSSM is constructed based on the search results. In the next iterations, the search is performed using the PSSM from the previous steps. This approach allows detection of more distant homologous sequences (higher sensitivity) while not introducing excessive amount of false positive hits.

3.2.2 HMMER

HMMER [31] is a suite of tools for searching similar biological sequences based on profile Hidden Markov Models (profile HMM) [30]. A profile HMM is a variant of an HMM designed specifically to biological sequences to model evolutionary variation in the sequences. Similarly to PSSM, profile HMM is calculated using a multiple sequence alignment. A profile HMM defines three types of states: match (M), insertion (I), and deletion (D) (Figure 3.1). M state emits a single amino acid and the probability of emitting is determined by the frequency at which that residue has been observed in the corresponding column of the alignment. The sequence of match states is analogous to the PSSM. A profile HMM captures multiple sequence alignments better than PSSM by modelling insertions and deletions using D and I states. M, I, and D states are connected by state transition probabilities, which reflect different rate of insertions and deletions along the sequence alignment.

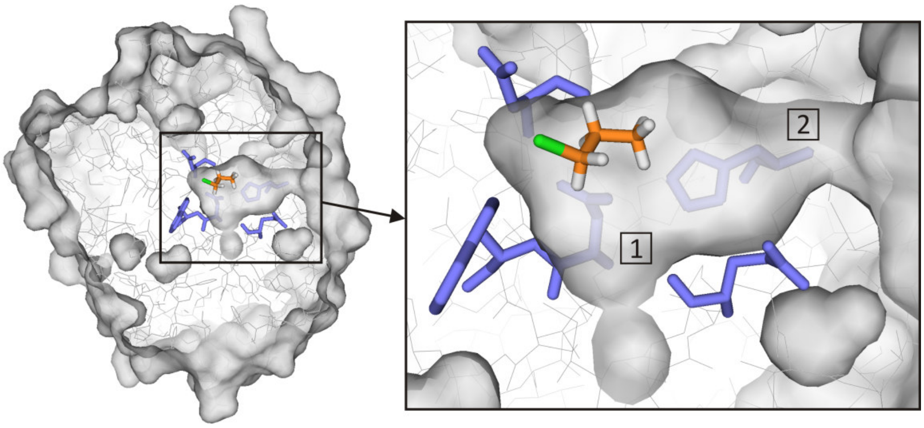


Figure 3.2: A cross-section of a 3D structure of haloalkane dehalogenase with a detail of the active site (1) and main access tunnel (2). The catalytic pentad residues and the substrate are highlighted by blue and orange sticks, respectively. The figure was adapted from the previous work by Chovancová et al. [19].

3.2.3 Enzyme-specific methods

Distinctive feature of enzymes is the presence of essential residues, which are the key amino acids involved in the catalysis. For example, enzymes from the family of haloalkane dehalogenases encompass so-called catalytic pentad—five essential residues accelerating hydrolytic conversion of halogenated alkanes into alcohols (Figure 3.2). Removing a single essential residue from the protein usually results in a detrimental impact on the enzymatic function.

Therefore, enzyme mining methods must consider essential residues. A common approach to validate essential residues calculates the optimal pair-wise sequence alignment of the query and the target sequence using one of Smith-Waterman or Needleman-Wunsch algorithms. The validation succeeds when the essential residues on the query are aligned to identical residues on the target. An alternative approach uses active site profiles of the query taking into account not only the essential residues themselves but also flanking amino acids upstream and downstream, respectively [55].

Another step to validate enzyme hits is the constitution of protein domains. Protein domains are structurally distinct parts of proteins having a specific role in the protein's architecture. Enzymes usually require a specific arrangement of protein domains to perform their function. The largest database of protein domains is Pfam [33], currently containing 18,259 unique domains[1]. The domains are represented by multiple sequence alignments and profile HMMs, and their detection can be performed using InterProScan software [67].

3.3 Sequence-space visualization

As protein search methods yield thousands of hits, an easy-to-interpret visualization showing the most significant relationships between the sequences could be of great help to select targets. A sequence-similarity network (SSN, Figure 3.3) is suitable for such task [6]. In SSN, sequences are represented by nodes. Edges between nodes indicate substantial sequence similarity between the sequences. The network is arranged in a two-dimensional

[1] Pfam 33.1 (05/2020) http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/relnotes.txt

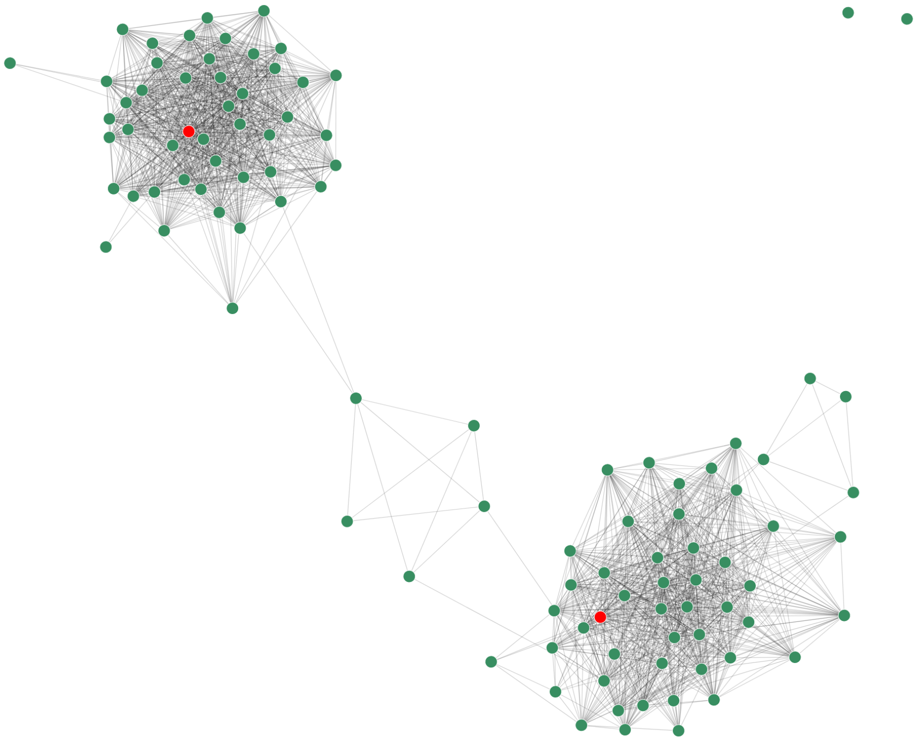


Figure 3.3: Representative sequence similarity network (SSN) of 3675 haloalkane dehalogenases. The SSN was generated by the EnzymeMiner web server [42] which uses Cytoscape [74] to lay out the network. Sequences sharing greater than 50% identity were consolidated in a single node. Edges indicate sequence identity between representative sequences of the connected nodes. Red nodes represent clusters that contain the query sequences used for the EnzymeMiner analysis.

space such that the edge length is proportional to the sequence similarity. The SSN construction is done in several steps:

1. All-to-all BLAST similarity search is performed to calculate similarity between sequences which will be used as edge weights.
2. A minimum similarity threshold is applied to remove irrelevant edges.
3. A layout algorithm is applied to spread the nodes in a 2D space.

If the SSN is colour-coded, it helps to analyse sequence relationships between the input and the identified sequences. Sequence groups and outliers can be easily spotted. The SSN is especially useful for selecting promising targets across to whole sequence space to increase the sequence variability of the selection and increase the chance of finding enzymes with novel biochemical properties. SSNs can be visualised and interactively analysed, for example using Cytoscape software [74].

3.4 Summary

Existing metadata- and sequence-based approaches are well developed and widely used for mining enzyme sequences in protein databases. The fundamental unsolved problem is how to deal with the overwhelming number of sequence entries identified by these methods and select a small number of relevant hits for in-depth experimental characterization. For example, a metadata-based search for members of the haloalkane dehalogenase model family using the UniProt web interface yields 3598 sequences[2]. It is impossible to rationally select several tens of targets for experimental testing without additional bioinformatics analyses that would help to prioritize such a large pool of sequences. To address the challenge of exploring the unmapped enzyme sequence space and rational selection of attractive targets, a novel tool is needed that would identify enzyme family members, comprehensively annotate the targets and visualise them to facilitate efficient prioritization and selection of representative hits for experimental characterization. Since solubility is a critical parameter for successful production of proteins in laboratory conditions, it would be a convenient criterion for such prioritization.

[2] UniProtKB release 2020_01

Chapter 4

Protein solubility prediction

Protein solubility is a key attribute when choosing protein targets for experimental characterization [85]. Its accurate computational prediction based on protein sequence would save high amount of resources wasted on difficult-to-produce proteins. The sequence-based protein solubility prediction task is a problem of finding a mapping between protein sequences Σ^+ and solubility values S (Equation 4.1).

$$\Sigma^+ \longrightarrow S \tag{4.1}$$

$$\Sigma = \{\mathrm{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}\} \tag{4.2}$$

Σ^+ is a set of all possible non-empty protein sequences composed from letters of the amino acid alphabet Σ (Equation 4.2). The solubility values S might be defined as any of the following sets depending on the data and model used for the prediction: (i) discrete classes (soluble/insoluble), (ii) numeric values from 0 to 1 representing probability of the protein being soluble, or (iii) real-scale numeric values representing experimental quantitative solubility measurement. In the context of sequence-based solubility prediction, the second definition is the most frequent by existing sequence-based methods [43].

This chapter is based on one of the three main results of the Thesis—a summary and critical assessment of computational tools and databases for predicting protein stability and solubility [59] (Appendix C). While this chapter focuses on sequence-based solubility prediction only, the summary also discusses other types of methods and databases.

4.1 Biochemical background

Protein solubility is a thermodynamic parameter defined as the concentration of protein in a saturated solution that is in equilibrium with a solid phase, either crystalline or amorphous, under a given set of conditions [5]. However, it is challenging to quantitatively measure the solubility of large sets of proteins [51], so there is little quantitative experimental data on protein solubility. More often, protein solubility is recorded as binary value in existing datasets—1 for soluble proteins and 0 for insoluble proteins. The exact understanding of the two solubility classes and their relation to the formal definition of protein solubility can differ greatly between datasets.

The key biochemical process limited by protein solubility is a *recombinant protein expression* (RPE) [22]. The main goal of this process is to manipulate living organisms, usually bacteria, to produce the desired recombinant (artificial) protein. RPE is widely

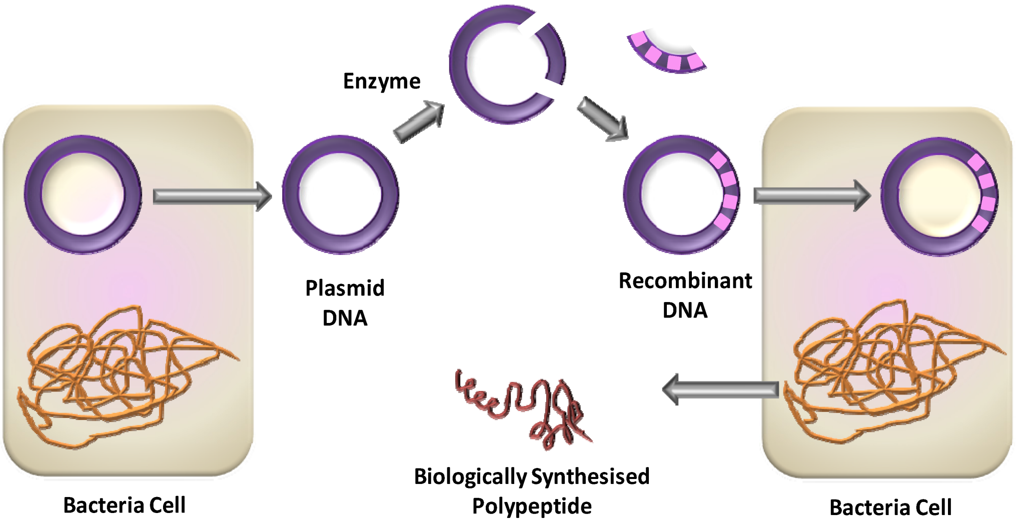


Figure 4.1: Recombinant protein expression (RPE). The gene of interest coding a required protein (enzyme) is synthesised and cloned into a plasmid (expression vector). Recombinant plasmid DNA is then transformed into a host bacteria cell, cells are cultured, the expression is induced, and the protein produced. As the protein is produced on the basis of a recombinant DNA, it is referred to as a *recombinant* protein. The figure was adapted from the previous work by Duro-Castano et al. [29].

used in both research and industry. At the theoretical level, the steps needed for obtaining a recombinant protein are straightforward. First of all, a gene of interest is selected, cloned into an expression vector, transformed into the host cell of choice, induced and then, the protein is ready for purification and characterization (Figure 4.1). In practice, however, several problems often arise. Apart from potential difficulties with the host organism cultivation, the recombinant protein might result to be insoluble. Protein insolubility can arise from various reasons at multiple stages of RPE process:

- Protein toxicity: the host cell does not tolerate the recombinant protein and eliminate it using the self-protective cell mechanisms.
- Inability to fold: the protein in its primary form can not be properly folded and requires additional post-translational modifications of the protein chain.
- Inability to self-fold: the protein needs the assistance of auxiliary molecules (chaperones) to properly fold into an active conformation.
- Inappropriate cellular environment: the host cell is not able to provide a suitable cellular environment that is essential for the given protein to fold.
- Aggregation: the protein aggregates and forms inclusion bodies in where it is deposited in alternate non-native conformations rather than folding into its natural conformation.

To conclude, the reasons for the protein insolubility are usually tightly related to the specific conditions in the host cell. This has one important implication. A protein that is soluble in one expression system is not necessarily soluble in the other systems and vice

Table 4.1: Summary of protein solubility databases. The proteins in PDB were counted on April 22, 2021. TT—TargetTrack. *Refined numbers after the chaperones were added to the PURE system.

DB	Format	Proteins	Soluble	Insoluble	Comments
eSol	CSV	3,173	2,385 *2,911	788 *262	• solubility from 0% to ~100% • *in vitro* system PURE • *E. coli* proteins only • might be over-estimated
TT	XML	339,354	87,854	251,500	• binary solubility information • heterogenous protocols • insolubility derived indirectly
NESG	CSV	9,478	5,773	3,705	• discrete solubility from 0 to 5 • repeated experiments • *E. coli* expression system
PDB	PDB	155,045	155,045	0	• only soluble proteins • expression system annotations

versa. Therefore, solubility is not exclusively an intrinsic property of the protein sequence. Each prediction tool, especially those based on statistical or machine-learning methods, are reliable only when applied on proteins expressed in similar conditions as those from the training dataset.

4.2 Databases

In this section, four major databases of protein solubility—eSol [62], TargetTrack [8], NESG [66], and PDB [9], are discussed. All these databases are valuable resources for designing protein solubility prediction tools, but each have its own specific properties that have to be taken into consideration before constructing a derived dataset and applying statistical models or machine learning algorithms. In the Table 4.1, important facts about each database are summarised.

4.2.1 eSol

Solubility database eSol [62] is very specific. The solubility data was obtained using an *in vitro* cell-free translation system–PURE [75], that is notably different from the conventional *in vivo* methods (Figure 4.2). In the first study from 2009 [62], eSol authors successfully quantified solubility for 70% of the *E. coli* ASKA library [48] of putative protein coding sequences (3173 from 4132) without the presence of the chaperones. 788 proteins showed to be insoluble. Later, in 2012, the same authors published next version of eSol [61]. They newly evaluated the effects of the major *E. coli* chaperones—trigger factor, DnaK/DnaJ/-GrpE, and GroEL/GroES, on the 788 insoluble proteins using the same PURE system. As a result, approximately 600 of the previously insoluble proteins turned up to be soluble with

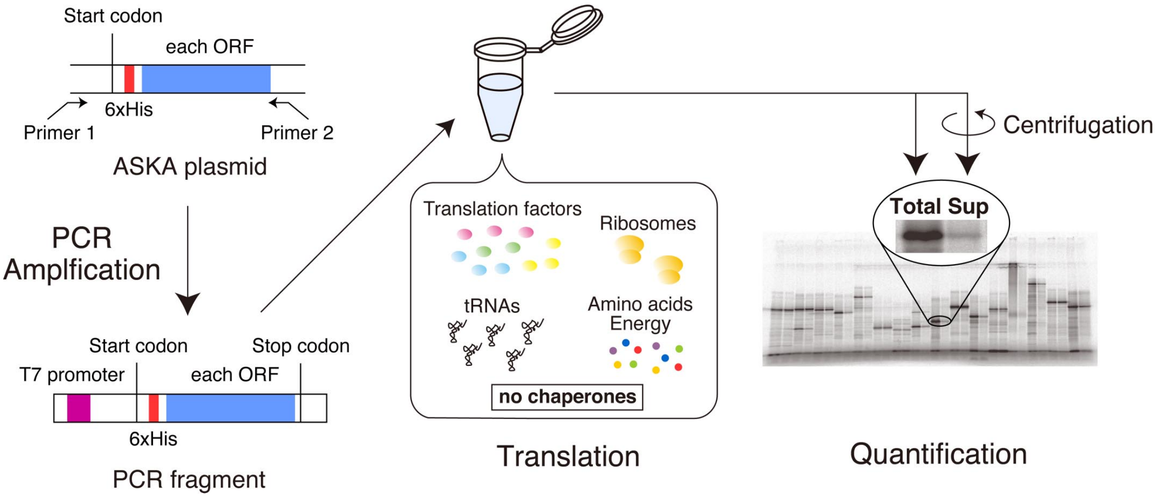


Figure 4.2: Cell-free translation system PURE [62, 75]. Each open reading frame (ORF—DNA sequence coding a protein) is amplified and translated in the presence of translation factors, ribosomes, tRNAs, amino acids and energetic molecules. Solubility was defined as the proportion of the supernatant fraction (which was obtained after the centrifugation of the translation mixture) over the uncentrifuged total protein. The figure was adapted from the previous work by Niwa et al. [62].

the help of at least one of the *E. coli* chaperones. The recent version of the eSol database is available for a download in the form of an annotation table (CSV file).

Several caveats should be stated regarding the interpretation of the eSol data. First, because the solubility analysis completely depends on a centrifugation process (Figure 4.2), it is possible that soluble fractions include oligomeric assemblies that act as aggregation precursors and, thus, the resulting solubility values could be overestimated. Second, the set of quantified proteins is limited to the *E. coli* proteins only. Machine-learning methods that would be trained solely on the eSol database, would certainly introduce a significant bias towards typical *E. coli* proteins and would not account for protein toxicity since the eSol dataset contains only proteins that are naturally occurring in *E. coli*. Third, after the evaluation of insoluble proteins in the presence of chaperones, only about 200 insoluble proteins are left. This makes eSol a highly-imbalanced database. On the other hand, the eSol database offers unique information about the effects of typical *E. coli* chaperones. This may be used to find a novel protein feature that would discriminate the proteins that need chaperones from those that do not. Moreover, the eSol experimental data are homogenous—measured under standard reproducible conditions.

4.2.2 TargetTrack

Despite the fact that TargetTrack [8] was not originally designed as a solubility database, it is now probably the most valuable resource regarding this property. Primarily, TargetTrack provides information on the experimental progress and status of protein targets selected for structural determination by the Protein Structure Initiative and other worldwide high-throughput structural biology projects. However, before the structural determination is performed, sufficient amount of a pure protein has to be first produced and this is usually achieved by the recombinant protein expression (Section 4.1). As a secondary outcome, pro-

Table 4.2: TargetTrack protocol statistics across research centers. JCSG—Joint Center for Structural Genomics, MCSG—Midwest Center for Structural Genomics, NYSGRC—New York Structural Genomics Research Consortium, SECSG—Southeast Collaboratory for Structural Genomics, SSGCID—Seattle Structural Genomics Center for Infectious Disease.

Center	Selection	Cloning	Expression	Purification	Crystallization
JCSG	127	1	1	1	2
MCSG	3	8	6	11	9
NYSGRC	1	17	36	28	5
SECSG	1	9	85	47	2
SSGCID	26	17	17	21	3
...	...	...	...	...	...
Total	**251**	**232**	**291**	**235**	**84**

tein solubility can be either directly or indirectly derived from the status of the structural determination process. Each status is defined by a keyword from a controlled dictionary and follows a particular experimental protocol used by a certain research centre which conducted the experiment. A typical protocol involves steps like selection, cloning, expression, purification and crystallization.

Extraction of soluble proteins from TargetTrack is then straightforward. Those proteins that reached the experimental status *soluble*, or any subsequent status (*diffraction*, *crystal structure*, *in PDB*, ...) are considered to be soluble. On the other hand, extraction of insoluble proteins is rather indirect. All proteins that did not reach *soluble* status or any subsequent one and at the same time reached a *work stopped* status, are most likely to be insoluble. The explicit *work stopped* status is required to decrease the chance of false negatives.

Unlike eSol, TargetTrack is hierarchically organised and compiled as an XML file with a controlled schema. Nowadays, it contains 339,354 target protein records. If the procedure described in the previous paragraph is applied, nearly 75% of all targets turn out to be insoluble. Most of the target proteins come from Bacteria (66%), then from Eukaryota (27%) and Archea (7%).

Several important facts should be considered before TargetTrack is used in practise for solubility prediction. First, TargetTrack records were created by a varied group of research centres and thus also the set of used protocols is very diverse (Table 4.2). This can be taken as both an advantage and a disadvantage. It is a significant complication for all prediction tools that would try to derive their models from the entire TargetTrack. The expression conditions of the individual targets are too different and it is nearly impossible to reasonably include them directly in the model. However, TargetTrack could be possibly divided into disjoint subsets, one for each specific protocol, and a model could be built separately on each subset. Second, as described above, the insolubility of the target protein is derived from the indirect signal that the work on the particular protein was stopped and no soluble status was reached. Nevertheless, the exact reason for the interruption of the process does not have to be the protein insolubility. Some of the protocols might be inherently error-prone and the human factor can also play an important role. This might lead to solubility underestimation.

4.2.3 NESG

NESG dataset was generated by the North East Structural Consortium (NESG), which represents 9644 proteins expressed in *E. coli* using a unified production pipeline [66]. The dataset contains two integer scores ranging from 0 to 5 for each target, indicating the protein's level of expression and the soluble fraction recovery. The reproducibility of the experimental results in the dataset was validated by performing repeated measurements for selected targets and achieving very similar results. The NESG dataset targets are included in the TargetTrack database because the NESG participated in the Protein Structure Initiative project. However, the expression and solubility levels from the NESG dataset were not included in the TargetTrack database. The high consistency and quality of the NESG dataset make it suitable for benchmarking purposes.

4.2.4 PDB

The Protein Data Bank (PDB) [9] archives primarily information about experimentally-determined 3D structures of proteins. However, from the perspective of protein solubility, all proteins that have a determined 3D structure are assumed to be soluble because sufficient solubility is a necessary condition for the structure determination. Thus, PDB can be used as a reliable source of soluble proteins. Moreover, PDB records often contain additional annotations, e.g., information about the expression system or expression vector type, that might be exploited in the construction of solubility datasets for a specific expression system.

4.3 Sequence features

Computational prediction models typically expect a fixed number of inputs. However, a protein might be encoded by an arbitrary-long sequence. Therefore, the protein sequences have to be transformed first into a fixed-size vector of numerical (Equation 4.3) or categorical values. This procedure is referred to as sequence encoding or feature extraction.

$$\Sigma^{+} \longrightarrow \mathbb{R}^{N} \tag{4.3}$$

The term encoding is used for reversible transformation of the sequence. One-hot protein encoding is a typical example occurring mainly in deep-learning methods. The sequence is represented by a matrix of size $N \times 20$, where N is the length of the longest expected sequence. Each row represents one position in the sequence. Columns represent the 20 amino acids. There is 1 in the matrix at position (i, k) if there is amino acid k on position i in the sequence. Otherwise there is 0. The feature extraction, on the other hand, is irreversible operation. The sequence can not be inferred just from the value of the feature. In this section, several common extracted features used in existing models are discussed.

4.3.1 Amino acid content

The amino acid content is a relative frequency of the amino acid in the protein sequence. It is the most common feature used for sequence-based solubility prediction. There is a major concern about this feature—random shuffling of the amino acids in the sequence do not change the amino acid content. Biologically, random shuffling of amino acids leads to low solubility and the loss of protein function as important residue interactions forming secondary and tertiary structures are disrupted. The invariability to random shuffling can be avoided by using higher orders of amino acid content, for example the content of amino acid dimers or trimers.

4.3.2 Physico-chemical properties

Physico-chemical properties of amino acids are frequently used for solubility prediction. The typical approach is to use the average of a specific physico-chemical property for all amino acids in the sequence in the prediction model, for example average hydrophobicity, hydropathy (GRAVY) or charge. The largest source of possible physico-chemical properties is AAindex database [46]. There are also physico-chemical properties requiring additional calculation—isoelectric point, flexibility, instability index, or molar extinction coefficient. A similar concern that is mentioned for the amino acid content—invariability to random shuffling, applies for most of the physico-chemical properties.

4.3.3 Sequence similarity

Sequence similarity introduces basic evolutionary information in the prediction model. The feature is usually defined as similarity to a fixed set of sequences, often to a set of insoluble or soluble sequences. The similarity value might be a BLAST score or proportion of identical amino acids in the pair-wise sequence alignment, which is then referred to as sequence identity.

4.3.4 Predicted features

Many existing models include features predicted by other predictors. The most frequent example is the predicted secondary structure of a protein which gives estimation of helix, sheet and coil structure in the folded protein. Other predicted features are protein disorder or content of transmembrane helices.

4.4 Performance evaluation

Sequence-based solubility predictors are usually binary classifiers predicting soluble or insoluble class based on a numeric decision threshold which can be tuned. At a specific threshold, the classifier's performance is fully described by confusion matrix (Table 4.3) which is 2×2 contingency table of positive and negative predictions.

Table 4.3: Confusion matrix for solubility classification.

		True class	
		Soluble	**Insoluble**
Prediction	**Soluble**	true positives (TP)	false positives (FP)
	Insoluble	false negatives (FN)	true negatives (TN)

Many derived metrics are based on the confusion matrix. There are metrics independent on the prevalence (which is how often each category occurs in the population), and metrics

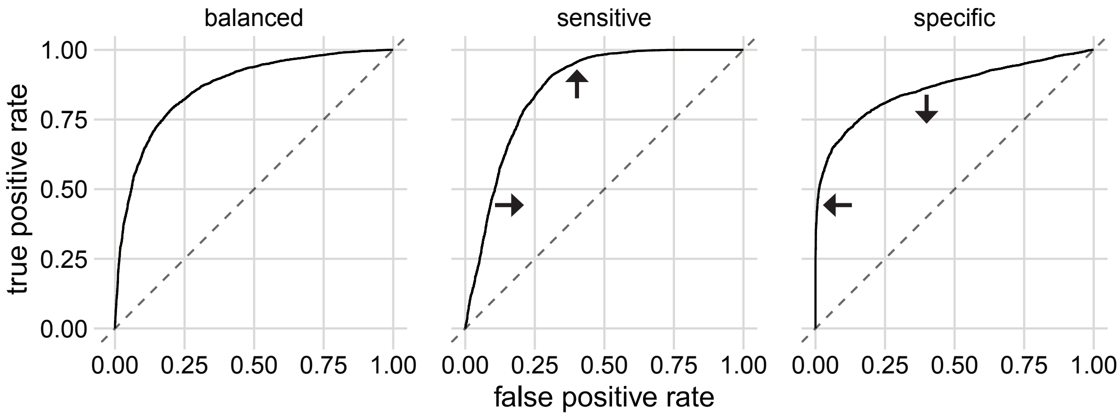


Figure 4.3: Receiver operating characteristic curve (ROC). The ROC reveals to which type of error is the classifier more predisposed. The figure shows three ROC curves of three different hypothetical classifiers with equal area under the curve (AUC) of 0.86. The balanced classifier makes similar error in predicting soluble and insoluble class. The sensitive classifier makes fewer errors in prediction of the soluble class. The specific classifier makes fewer errors in prediction of the insoluble class. The dashed line shows a ROC of a baseline random classifier with AUC of 0.5. The specific classifier is preferred for tasks where low false positive rate is required.

that depend on the prevalence. Accuracy (ACC, Equation 4.4), sensitivity (TPR—true positive rate, Equation 4.5) and specificity (TNR—true negative rate, Equation 4.6) are dependent on the prevalence, whereas Matthew's correlation coefficient (MCC, Equation 4.7) is not. The latter is therefore preferred when comparing performance using imbalanced datasets.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{4.4}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.5}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4.6}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{4.7}$$

The performance of the classifier over different prediction thresholds is typically evaluated using Receiver operating characteristic curve (ROC) and the area under the ROC (AUC). The ROC curve is created by plotting the true positive rate (TPR, Equation 4.5) as a function of the false positive rate (FPR, Equation 4.8) at all possible threshold settings (Figure 4.3). Models predicting quantitative solubility measurements are usually compared using Pearson correlation coefficient (PCC, Equation 4.9, where n is test set size, x_i, y_i are the individual predicted and actual values of the i-th element of the series, and $\bar{x}$, $\bar{y}$ are means of predicted and actual values).

$$\mathrm{FPR} = 1 - \mathrm{TNR} \tag{4.8}$$

$$\text{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.9}$$

4.5 Existing methods

Up to the present, many protein solubility prediction methods have been created. In this section, methods based on protein sequence features providing single solubility score are discussed (Table 4.4). Each method is described at different level of detail depending on the complexity of the method and available information. Simple methods are described fully.

4.5.1 Wilkinson-Harrison and RPSP

One of the first methods described is a six-parameter model by Wilkinson and Harrison [89] published in 1991 combining residue charge average, turn-forming residue fraction, cysteine and proline content, residue hydrophilicity, and total sequence length. These parameters emerged from a correlation analysis on a small data set of 81 proteins. The coefficients for the model equation were obtained by Fisher's linear discriminant analysis [35].

In 1999, Davis et al. [25] published a revised model of Wilkinson and Harrison (rWH). On a larger set of *E. coli* proteins (4,000 sequences), they discovered that only two of the original parameters were critical for distinguishing between soluble and insoluble proteins. The critical parameters are (i) the residue charge average, which accounts for differences in the numbers of aspartic acid plus glutamic acid vs. lysine plus arginine residues, and (ii) turn-forming residue content, which accounts for the total number of asparagine, glycine, proline, and serine residues. The core of the solubility prediction is the canonical variable expressed by the Equation 4.10.

$$CV = 15.43\left(\frac{N+G+P+S}{n}\right) - 29.56\left(\frac{(R+K)-(D+E)}{n} - 0.03\right) \tag{4.10}$$

Here, n is the protein sequence length and N, G, P, S, R, K, D, E are the counts of the specific amino acids denoted by the IUPAC codes. The final prediction is computed as $CV - CV'$, where CV' is 1.71. In case the difference is positive, protein is predicted as insoluble. Otherwise, it is considered to be soluble. The probability of solubility is expressed by the Equation 4.11.

$$P(S) = 0.4934 + 0.276|CV - CV'| - 0.0392(CV - CV')^2 \tag{4.11}$$

In 2009, Harrisson's research group published a different model which is commonly referred to as RPSP (Recombinant Protein Solubility Prediction) [27]. The model uses logistic regression of 32 possible parameters. The protein database used to create this model consisted of 212 proteins. The parameters used for the model include molecular weight, amino acid fractions, aliphatic index, alpha-helix propensity, beta-sheet propensity, average pI, approximate charge average, and hydrophilicity index. The authors reported an accuracy of 87% on cross-validation.

4.5.2 SOLpro

One of the first well-established machine learning prediction tools based on the global protein features is SOLpro [56]. It is designed as a two-stage support vector machine classifier. In the first stage, 20 SVMs are trained on different sets of features. 18 from the 20 feature sets represent frequencies of mono-, di- and trimers derived from 7 distinct amino acid alphabets. One feature set contains features directly computed from protein sequence, similar to those used by Wilkinson and Harrison—sequence length, turn-forming residues fraction, absolute charge per residue, molecular weight, GRAVY index (averaged hydropathy value) [54], and aliphatic index. The last feature set is composed of the features predicted by other application-specific prediction tools—alpha and beta secondary structure forming residues fraction, the number of domains, and exposed residues fraction.

In the second stage, the outputs of the 20 SVMs are used as the inputs for a final SVM. To conclude, SOLpro introduced some novel sequence features, while others appeared to be in good agreement with the previous research. SOLpro performance was evaluated by a ten-fold cross-validation for which the authors stated an average accuracy of 74% and a Matthews correlation coefficient (MCC) of 0.487.

4.5.3 PROSO II

PROSO II [77] has a two-layered structure where the output of a primary Parzen window model and a logistic regression classifier serve as the input of a second-level logistic regression classifier. PROSO II uses the best performing mono- and dimer frequencies that were selected using a wrapper feature selection method [50]. Eighteen out of 20 monopeptide frequencies and thirteen out of 400 dipeptide frequencies were selected as the most important for the model performance. Eight selected dimers contain electrically charged side chains, which is in good agreement with Wilkinson–Harrison model [89]. Other frequently occurring amino acid groups include hydrophobic aromatic and hydrophobic aliphatic residues. Five selected dimers contain aromatic amino acids. As demonstrated before in the study of Christendat et al. [20], a high percentage of aromatic residues is a good indicator of insolubility. Also, a high content of hydrophobic dimers seems to be an important factor for protein solubility.

As an additional feature in the first layer of the method, a sequence-similarity-based model using an adapted Parzen window approach was used to capture the differences between sequence patterns of soluble and insoluble proteins. It relies on a BLAST [4] score to calculate similarity values to two sets of proteins (soluble vs. insoluble) using a modified Cauchy kernel. Additionally, the aliphatic index, fold index, GRAVY index, and isoelectric point were also used as features. PROSO II was tested on a separate holdout set not used at any point of the method development and the authors declare an accuracy of 75.4% and MCC of 0.39.

4.5.4 ccSOL

Beyond other features that, to some extent, are employed also in the other tools, ccSOL [2] introduced a coil and disorder proneness as a novel type of feature. It was shown that disorder prediction correlates ($\rho = 0.45$) with the experimental solubility. The 6 final features were selected by constructing a SVM for each subset of the initial 11 promising features and evaluating each of them by a ten-fold cross-validation. The 6 final features (coil and disorder propensities, hydrophobicity, hydrophilicity, alpha and beta secondary structure forming residues fractions) were then associated with the best performing SVM. The overall published ccSOL accuracy is around 76%.

ccSOL omics [1] is a variant of ccSOL method where a solubility score is predicted for each amino acid in the sequence. The method uses a sliding window of 21 amino acids that is moved one residue at a time until the C-terminus is reached. The solubility propensity of each fragment is calculated by the previously published ccSOL method. The final overall solubility prediction is expressed as a solubility score that is computed using Fourier transform of the solubility profile and a neural network. Unfortunately, the authors give very little information about the design of the neural network. The published accuracy of this method on an independent test set is 74%.

4.5.5 ESPRESSO

ESPRESSO [41] implements two methods for protein solubility prediction—property-based and pattern-based. The property-based approach relies on SVM and unlike the previous tools, ESPRESSO uses information also from the protein coding DNA sequence. The set of features contains protein length, single nucleotide frequencies, GC content, codon frequencies, amino acid frequencies and amino acid group frequencies. The sequence information, except for the protein length, was computed for the entire chain and both terminal regions, which are defined as 60 bases (meaning 20 amino acid residues). Additionally, several pieces of predicted structure information like secondary structure ratios, trans-membrane elements, disordered regions, and accessible surface area are also used. For each of the features, the statistically significant difference between the positive and negative datasets was computed by the Student's t-test. The features with $p < 0.05$ were considered to be associated with protein solubility.

The pattern-based method uses the occurrence frequencies of highly frequent sequence patterns. In the first step, the authors defined a set of sequence patterns as all combinations of ten amino acid groups (based on the physicochemical properties) with the length that exhibited the highest prediction performance (six or seven amino acids). In the second step, they searched for the sequence patterns that appeared exclusively in either the positive or negative data. The counts of the most significant patterns were then used in a simple linear discrimination function to get the final prediction. Moreover, the locations of the sequence patterns can be easily mapped to a query sequence, therefore an additional benefit of the pattern-based method is to provide candidate regions, matching either positive or negative sequence patterns, that the researcher can modify, to change protein solubility. Property- and pattern-based methods for solubility prediction reached an accuracy of 68% and 63%, respectively.

4.5.6 CamSol

CamSol [81] was originally designed for a slightly different purpose than sequence-based solubility prediction. It identifies protein hot spots that could be mutated to improve protein solubility. It also employs a 3D protein structure as an additional input besides protein sequence.

CamSol, in order to obtain a solubility profile, employs a linear combination of four physicochemical properties of amino acids—hydrophobicity, charge (at neutral pH), alpha-helix propensity, and beta-strand propensity. The linear combination is then averaged over a window of seven residues to account for the effect of the neighbouring residues. A correction is added to consider the possible presence of hydrophobic–hydrophilic patterns and the influence of charges of the same sign.

In the next step, the intrinsic solubility profile is modified to account for the proximity of the amino acids in the three-dimensional structure of the input protein and for their solvent exposure. These modified profiles are used to identify residues unlikely to be soluble. Such residues are usually required to prompt a fast and correct folding of a protein and are typical constituents of the hydrophobic core of the protein native state. In contrast, more soluble residues are normally exposed to the solvent and thus, are more likely to elicit the aggregation process.

In 2017, CamSol authors published a second version of their method [80] that allows for calculating an overall solubility score based on the previously described solubility profile using the Equation 4.12,

$$S_P = \frac{\sum_{i=1}^{N} \begin{cases} \omega_{up}(S_i - th_{up}) & \text{if } S_i > th_{up} \\ \omega_{low}(S_i - th_{low}) & \text{if } S_i < th_{low} \\ 0 & \text{otherwise} \end{cases}}{\gamma N^{\delta}} \tag{4.12}$$

where S_i is the value of the intrinsic solubility profile for the amino acid i and N the length of the input sequence. The upper and lower thresholds th_{up} and th_{low}, as well as the coefficients ω_{up}, ω_{low}, γ, and δ were fitted with a Monte Carlo procedure maximising both the S_P correlation with measurements of aggregation rates from the literature and the ability of S_P to discriminate between non-aggregating and aggregating peptides and proteins [80]. The CamSol score was validated experimentally using 9 monoclonal antibodies achieving a Pearson correlation of 0.79 ($p < 0.05$) with experimental results. However, the performance on larger datasets was not commented by the authors.

4.5.7 Protein-Sol

Protein-Sol [38] is a linear model based on 10 features (six amino acid propensities, sequence length, absolute charge, fold propensity, sequence entropy) which was trained using eSOL dataset. Feature weights were determined from separation of low and high solubility subsets of eSOL database. Protein-Sol achieved Pearson correlation of 0.62 with the eSOL dataset.

4.5.8 DeepSol

DeepSol [47] is one of the first solubility predictors using deep-learning methods. DeepSol consists of a convolutional neural network (CNN) with multiple convolution blocks using a one-hot-encoded raw sequence as input. The usage of a raw sequence allows to learn feature representations that best encode the information essential for solubility prediction. Fifty-seven additional sequence- and structure-related features are used to complement the raw sequence input.

DeepSol is trained using TargetTrack data compiled by the authors of PROSO II. DeepSol authors then performed two major pre-processing steps (as in their previous tool PaRSnIP [69]) to avoid any unwanted bias and to ensure heterogeneity of sequences within the training set. As an independent test set, they used a balanced dataset compiled by Chang et al. [15] consisting of 2001 sequences. DeepSol attained an accuracy of 77% and a Matthew's correlation coefficient of 0.55 using the independent test set.

4.5.9 SKADE

SKADE [68] is the latest addition to deep learning based solubility predictors. It uses a neural network model of two sub-networks: the predictor network and the attention network. The final prediction is a scalar product of outputs of both networks. The final model has 25462 trainable parameters. SKADE used the same datasets for both training and testing as DeepSol and used one-hot encoding for sequence input.

The attention network enabled SKADE authors to provide some insight into the prediction. The attention profiles suggest that N- and C-termini are the most relevant regions for solubility prediction and are predictive for complex emergent properties such as aggregation-prone regions involved in beta-amyloidosis and contact density. DeepSol achieved an accuracy of 73% and a Matthew's correlation coefficient of 0.47 using the independent test set.

4.5.10 Solubility-weighted index

The Solubility-weighted index (SWI) method [10] is surprisingly simple. It uses optimized normalized B-factors [78] to calculate the solubility-weighted index (Equation 4.13 and a logistic regression formula (Equation 4.14) to calculate the probability of solubility. W_i is the optimized B-factor for i-th residue and N is the sequence length.

$$\mathrm{SWI} = \frac{\sum W_i}{N} \tag{4.13}$$

$$P(S) = \frac{1}{1 - \exp(-(81.05812 \cdot \mathrm{SWI} - 62.7775))} \tag{4.14}$$

SWI used binary solubility data from the DNASU database [23] to optimize the weights and find the coefficients of the logistic function. SWI showed a Pearson correlation of 0.50 with the eSOL dataset which was left as an independent test set.

4.6 Summary

There are many existing tools that address the problem of sequence-based protein solubility prediction [59]. However, Chang et al. reported large drop of 10–20% in accuracy of existing tools when evaluated using a larger test set [15]. This suggests that the authors of the solubility predictors overestimate the performance of their methods. The main reason for the overestimation might be that the training and testing data are not independent—they might be similar in terms of sequence similarity or they might share similar bias. The existing solubility datasets are very small in comparison to the number of all known proteins. If there is no strong effort to decrease the similarity between the training and test set and to keep the number of model parameters at a reasonable level in comparison to the size of the dataset, the overestimation of the performance is inevitable [88]. Although the field of sequence-based protein solubility prediction has been already thoroughly explored, there are several aspects that could be still improved.

First, surprisingly, there has been no attempt to partition TargetTrack database more carefully, for instance from the perspective of experimental protocols. This suggests that recently published tools might be trained on too heterogeneous datasets that mix solubility information for very different expression systems and even different host organisms. The reason why TargetTrack was not carefully partitioned yet might be due to its organization. Such an effort would require manual analysis of all protocols in TargetTrack, a task that might be time-consuming. Moreover, there has been a little effort on correcting the data using the available knowledge and technology. Some of the unexpressed proteins might be produced now thanks to the advance in the technology.

Table 4.4: Sequence-based protein solubility prediction methods. SVM—support vector machine, GRAVY—grand averaged hydropathy value, pI—isoelectric point

Method	Model	Features
rWH	discriminant analysis	residue charge average, turn-forming residue content
RPSP	logistic regression	molecular weight, frequencies of monomers, aliphatic index, alpha-helix propensity, beta-sheet propensity, average pI, approximate charge average, hydrophilicity index
SOLpro	two-stage SVM	frequencies of mono-, di- and trimers derived from seven amino acid alphabets, sequence length, turn-forming residues fraction, absolute charge per residue, molecular weight, GRAVY index, aliphatic index, alpha and beta residues fractions, the number of domains and exposed residues fraction
PROSSO II	logistic regression, Parzenov window	frequencies of mono- and dimers, GRAVY index, aliphatic index, fold index, isoelectric point and sequence similarity to both insoluble and soluble protein sets
ccSOL	SVM	coil and disorder propensities, hydrophilicity, hydrophobicity, alpha and beta residues fractions
ESPRESSO	SVM, pattern discrimination function	protein length, single nucleotide frequencies, GC content, codon frequencies, amino acid frequencies, amino acid group frequencies, secondary structure ratios, trans-membrane elements, disordered regions, accessible surface area, sequence patterns
CamSol	linear regression	hydrophobicity, charge at neutral pH, alpha-helix propensity, and beta-strand propensity
Protein-Sol	linear regression	frequencies of monomers, sequence length, absolute charge, fold propensity, sequence entropy
DeepSol	convolutional neural network	one-hot encoding
SKADE	predictor and attention deep neural network	one-hot encoding
SWI	logistic regression	normalized B-factors

Second, the balancing of the training and test sets could be improved. A class-based balancing is usually used to avoid overtraining on the majority class. However, length-based balancing tends to be overlooked. It is important to balance the sequence length distribution in the datasets so that length alone would not play a dominant role in the predictions [77]. Although the sequence length was shown to correlate with protein solubility—larger proteins are usually less soluble, the expected major use case for sequence-based solubility predictors is the prioritization of proteins of similar lengths, usually from a single protein family. A prediction model relying heavily on sequence length would not perform well in this use case.

Third, performing extensive feature selection among all of the known features could improve the performance significantly [72]. The set of features used among the prediction tools is quite rich and it mainly relies on three types of features—physicochemical properties derived directly from the protein sequence, predicted or direct structural features, and sequence patterns. However, the pool of the relevant features could still be extended. As the performance of computers increases, more computationally demanding feature extractions are feasible.

In addition to proposing a better solubility predictor, the contribution of a novel method could be a comprehensive comparison of existing methods. The comparison itself is a challenging task as there is little data that is not used by the existing methods for training and, thus, could be potentially used to construct a fully-independent test set. However, at least the overlap of the training set with the test data can be quantified to indicate the level of overestimation. An additional challenge when comparing different methods arises from the variety of working definitions used for „solubility". The existing variety of such working definitions and their relation to the formal definition of solubility (Section 4.1) has not been discussed yet.

Chapter 5

Results

This Thesis addresses the challenge of mining and selecting soluble enzymes from protein databases by developing and publishing two novel tools: (i) EnzymeMiner [42] for mining of enzymes in protein databases and (ii) SoluProt [43] for sequence-based protein solubility prediction. Additionally, a summary and critical assessment of existing computational methods and databases for protein stability and solubility prediction is presented [59]. Other results that were under consideration for acceptance at the time of writing or that are not directly related to the topic of the Thesis are mentioned only briefly.

5.1 EnzymeMiner

EnzymeMiner is an enzyme sequence search tool addressing the challenge of selecting a small number of relevant proteins from a large pool of database hits. It has several distinctive features which are not available in existing tools. First, it checks the presence of user-specified essential amino acids in the protein sequence which allows to target the search to a very specific set of enzymes performing the required function. Second, it integrates available environmental information which enables selection of hits from extremophilic organisms that might be resilient to harsh conditions. Third, it generates sequence similarity network which can be used to select hits with higher sequence diversity. Fourth, it provides solubility prediction which can be used for prioritization and for increasing the success rate of protein production.

EnzymeMiner requires two inputs: (i) query sequences and (ii) essential residue templates. The essential residue template is defined as a pair of a protein sequence and a set of essential residues in that sequence. The output is an interactive selection table containing the annotated identified sequences that can be prioritized based on various criteria. The table helps to select a small diverse set of enzyme sequences with a putative function for experimental characterization.

EnzymeMiner implements a three-step bioinformatics workflow: (i) homology search, (ii) essential residues based filtering and (iii) annotation of hits. In the first step, the input sequence is used as a query for a PSI-BLAST [4] two-iteration search in the NCBI nr database [73]. In the second step, the obtained hits are filtered using the input essential residue templates. Essential residues are checked using a global pairwise alignment with the template calculated by USEARCH [32] and a multiple sequence alignment calculated by Clustal Omega [76]. In the third step, the identified sequences are annotated using several databases and predictors: (i) transmembrane regions are predicted by TMHMM

[52], (ii) Pfam domains are predicted by InterProScan [67], (iii) source organism annotation is extracted from the NCBI Taxonomy [34] and the NCBI BioProject database [7], (iv) sequence identities to queries, hits or other optional sequences are calculated by USEARCH [32] and (v) solubility is predicted by the solubility predictor SoluProt (Section 5.2). More details on EnzymeMiner method are described in the corresponding publication.

EnzymeMiner workflow is to some extent based on two previous studies [85, 86]. The main differences between the previous computational pipelines and the EnzymeMiner workflow are described herein. First, the pipeline was simplified and generalised to work with any enzyme families and not just haloalkane dehalogenase family. The essential residue based filtering was improved to effectively replace the original hierarchical clustering by global pairwise alignment with the template. This step reduced the calculation time and eliminated parameters needed for the hierarchical clustering. Second, SoluProt (Section 5.2 was used instead of the revised Wilkinson-Harrison model to predict solubility. Third, the calculation and visualization of the sequence similarity network was integrated in the tool. Fourth, the pipeline was automatised and made accessible as a publicly available web server.

Publication: Hon J*, Borko S*, Štourač J, Prokop Z, Zendulka J, Bednář D, Martínek T, Damborský J. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Research.* 2020, 48(W1), W104–W109. *These authors contributed equally. The article is attached in Appendix A.

Journal Impact Factor 2019: 11.5 (Q1), citations by Scopus: 7 (self citations excluded)

Author contribution: design of the updated computational workflow and its initial implementation, leading the development of the web server, design of the user interface, contributing to the implementation of the user interface, writing of the manuscript and documentation (45%)

Abstract: Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Despite genomic databases growing exponentially, classical biochemical characterization techniques are time-demanding, cost-ineffective and low-throughput. Therefore, computational methods are being developed to explore the unmapped sequence space efficiently. Selection of putative enzymes for biochemical characterization based on rational and robust analysis of all available sequences remains an unsolved problem. To address this challenge, we have developed EnzymeMiner—a web server for automated screening and annotation of diverse family members that enables selection of hits for wet-lab experiments. EnzymeMiner prioritizes sequences that are more likely to preserve the catalytic activity and are heterologously expressible in a soluble form in *Escherichia coli*. The solubility prediction employs the in-house SoluProt predictor developed using machine learning. EnzymeMiner reduces the time devoted to data gathering, multi-step analysis, sequence prioritization and selection from days to hours. The successful use case for the haloalkane dehalogenase family is described in a comprehensive tutorial available on the EnzymeMiner web page. EnzymeMiner is a universal tool applicable to any enzyme family that provides an interactive and easy-to-use web interface freely available at https://loschmidt.chemi.muni.cz/enzymeminer/.

5.2 SoluProt

SoluProt is a predictor of soluble protein expression in *Escherichia coli* used in EnzymeMiner to prioritize hits. SoluProt uses a gradient boosting machine [36] and its only input is a protein sequence. The output is the predicted probability of soluble class.

SoluProt addressed several areas which the current predictors did not solve properly. First, the TargetTrack database used for training was partitioned more carefully than in other methods. Most importantly, keyword matching combined with manual checking of TargetTrack annotations was performed to extract only proteins expressed in the most common host organism, *E. coli*.

Second, the sequence redundancy in the training set was reduced by clustering to 25% identity using MMseqs2 [82] and retaining only representative sequences from each cluster. This was done separately for positive and negative samples to avoid simplifying the prediction problem. The number of soluble and insoluble samples was balanced such that both classes were equally represented. Additionally, the sequence length distribution was balanced so that length alone would not play a dominant role in the predictions.

Third, the SoluProt test set was built from a consistent dataset generated by the North East Structural Consortium (NESG) [66] whereas existing tools usually uses part of the TargetTrack database as a test set. The advantage of using NESG over TargetTrack for testing is a higher quality of solubility data which were measured using a unified pipeline and the measurements were shown to be reproducible.

Fourth, extensive feature selection was performed using a set of 251 sequence characteristics that were divided into eight groups: (i) single amino acid content (20 features), (ii) amino acid dimer content (210 features), (iii), sequence physicochemical features (12 features), (iv) average flexibility as computed by DynaMine [21] (1 feature), (v) secondary structure content as predicted by FESS [65] (3 features), (vi) average disorder as predicted by ESPRITZ [87] (1 feature), (vii) content of amino acids in transmembrane helices as predicted by TMHMM [52] (3 features) and (viii) maximum identity to a specific *E. coli* subset of Protein Data Bank [9] as calculated using USEARCH [32] (1 feature). In the end, 96 features were selected for inclusion in the predictive model.

Fifth, SoluProt model and existing tools were evaluated using the SoluProt test set to an extent which was not done before. The overlap of the training sets with the test set was considered and also different understandings of solubility classes among existing tools were pointed out. SoluProt achieved a slightly higher accuracy (58.5%) and AUC (0.62) than other available tools. Surprisingly, some recently reported tools, which are based on deep learning methods, performed worse than simpler methods in the comparison. More details on the comparison are provided in the corresponding publication.

Publication: Hon J, Marušiak M, Martínek T, Kunka A, Zendulka J, Bednář D, Damborský J. SoluProt: prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics*. 2021, 31(1), 23–28. The article is attached in Appendix B.

Journal Impact Factor 2019: 5.6 (Q1), citations by Scopus: not available

Author contribution: design and analysis of most of the experiments, feature calculations, design of dataset construction, leading of predictor's implementation, implementation of web interface, performance comparison of existing tools, writing of the manuscript (60%)

Abstract: Poor protein solubility hinders the production of many therapeutic and industrially useful proteins. Experimental efforts to increase solubility are plagued by low success rates and often reduce biological activity. Computational prediction of protein expressibility and solubility in *Escherichia coli* using only sequence information could reduce the cost of experimental studies by enabling prioritization of highly soluble proteins. A new tool for sequence-based prediction of soluble protein expression in *Escherichia coli*, SoluProt, was created using the gradient boosting machine technique with the TargetTrack database as a training set. When evaluated against a balanced independent test set derived from the NESG database, SoluProt's accuracy of 58.5% and AUC of 0.62 exceeded those of a suite of alternative solubility prediction tools. There is also evidence that it could significantly increase the success rate of experimental protein studies. SoluProt is freely available as a standalone program and a user-friendly webserver at https://loschmidt.chemi.muni.cz/soluprot/.

5.3 Computational design of stable and soluble biocatalysts

The progress in the development of computational tools and databases for predicting protein stability and solubility is summarised in this publication. Strengths and weaknesses of the methods were critically assessed. The solubility prediction methods and databases are presented in the second part of the paper. Section 5.2.1 of the paper is dedicated to the sequence-based solubility prediction methods which are also discussed here in the Thesis (Section 4.5). Additionally, two other groups of solubility prediction methods are included: (i) tools predicting solubility or aggregation profile, and (ii) tools predicting the effect of an amino acid mutation on the solubility. In the last section of the paper, perspectives on the computational design of stable and soluble biocatalysts are presented.

Publication: Musil M*, Konegger H*, **Hon J***, Bednář D, Damborský J. Computational Design of Stable and Soluble Biocatalysts. *ACS Catalysis*, 2018, 9(2), 1033–1054. *These authors contributed equally. The article is attached in Appendix C.

Journal Impact Factor 2019: 12.4 (Q1), citations by Scopus: 21 (self citations excluded)

Author contribution: critical assessment of protein solubility prediction tools and databases, writing of solubility-related parts of the manuscript (30%)

Abstract: Natural enzymes are delicate biomolecules possessing only marginal thermodynamic stability. Poorly stable, misfolded, and aggregated proteins lead to huge economic losses in the biotechnology and biopharmaceutical industries. Consequently, there is a need to design optimized protein sequences that maximize stability, solubility, and activity over a wide range of temperatures and pH values in buffers of different composition and in the presence of organic cosolvents. This has created great interest in using computational methods to enhance biocatalysts' robustness and solubility. Suitable methods include (i) energy calculations, (ii) machine learning, (iii) phylogenetic analyses, and (iv) combinations of these approaches. We have witnessed impressive progress in the design of stable enzymes over the last two decades, but predictions of protein solubility and expressibility are scarce. Stabilizing mutations can be predicted accurately using available force fields, and the number of sequences available for phylogenetic analyses is growing. In addition, complex computational workflows are being implemented in intuitive web tools, enhancing the quality of protein stability predictions. Conversely, solubility predictors are limited by the lack of robust and balanced experimental data, an inadequate understanding of fundamental principles of protein aggregation, and a dearth of structural information on folding intermediates. Here we summarize recent progress in the development of compu-

tational tools for predicting protein stability and solubility, critically assess their strengths and weaknesses, and identify apparent gaps in data and knowledge. We also present perspectives on the computational design of stable and soluble biocatalysts.

5.4 Other results

Functional annotation of enzyme family

In this work, uncharacterised members of haloalkane dehalogenase enzyme family were functionally annotated using a combined computational and experimental approach, and novel enzymes with activities that exceed activities of most published haloalkane dehalogenases were identified. This work is closely related to the topic of the Thesis and especially to the EnzymeMiner software. This study is the second of the two works on which the EnzymeMiner is based and it was performed before the EnzymeMiner and SoluProt tools were developed. The computational workflow herein presented is based on the first study [85], expanding it by: (i) applying EFI-EST [37] and Cytoscape [74] for calculating and visualizing the sequence similarity network, (ii) extracting biotic relationships and disease annotations of the source organisms from the BioProject [7] database, and (iii) quantitatively assessing the quality of all homology models by MolProbity [17]. Because of the difficulty and time demands of the experimental work, only a preprint has been published and it still awaits the full peer-review process. Therefore, it is mentioned here and not included in the main results of the Thesis.

Preprint: Vaňáček P, Vašina M, **Hon J**, Kovář D, Faldýnová H, Kunka A, Buryška T, Badenhorst C, Mazurenko S, Bednář D, Bornscheuer U, Damborský J, Prokop Z. Functional annotation of an enzyme family by integrated strategy combining bioinformatics with microanalytical and microfluidic technologies. *ChemRxiv.* DOI: 10.26434/chemrxiv.13621517.v1

Author contribution: reimplementation and extension of the previous computational pipeline [85], performing all calculations, analysis of results, sequence-space visualization, contribution to writing of the manuscript (20%)

Pqsfinder

Pqsfinder is an algorithm for efficient detection of imperfect potential quadruplex-forming sequences in DNA. It is not related to the topic of the Thesis.

Publication: **Hon J**, Martínek T, Zendulka J, Lexa M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics.* 2017, 33(21), 3373–3379.

Journal Impact Factor 2019: 5.6 (Q1), citations by Scopus: 35 (self citations excluded)

Author contribution: design and implementation of the method, writing majority of the manuscript (70%)

Pqsfinder web

Pqsfinder web is an easy-to-use web interface for the pqsfinder algorithm. It is not related to the topic of the Thesis.

Publication: Labudová D*, **Hon J***, Lexa M. pqsfinder web: G-quadruplex prediction using optimized pqsfinder algorithm. *Bioinformatics.* 2020, 36(8), 2584–2586. *These authors contributed equally.

Journal Impact Factor 2019: 5.6 (Q1), citations by Scopus: 5 (self citations excluded)

Author contribution: design and implementation of speed optimizations, writing of the manuscript (45%)

Chapter 6

Concluding remarks

Two tightly integrated computational methods—EnzymeMiner [42] and SoluProt [43], for mining of soluble enzymes from protein sequence databases are introduced in the Thesis. EnzymeMiner identifies putative members of enzyme families and facilitate their prioritization and well-informed manual selection for experimental characterization to reveal novel biocatalysts. Such a task is difficult to address using the web interfaces of the available protein databases, e.g. UniProtKB/TrEMBL [83] and NCBI Protein [73], since additional analyses are often required. The major advantage of EnzymeMiner over existing protein sources is the flexibility of input and concise annotation-rich interactive presentation of results. The user can input custom queries and a custom description of essential residues to focus the search on specific protein families or subfamilies. The output of EnzymeMiner is an interactive selection table containing the annotated sequences that can be prioritized based on various selection criteria. The table helps to select a diverse set of sequences for experimental characterization. Two key prioritization criteria are (i) the solubility score predicted by SoluProt, which can be used to prioritize the identified sequences and increase the chance of finding enzymes with soluble protein expression in *E. coli*, and (ii) the sequence identity to query sequences complemented with an interactive sequence similarity network visualization, which can be used to explore diverse sequences. Additionally, source organism and domain annotations help to select the sequences with diverse properties. EnzymeMiner is a universal tool applicable to any enzyme family. It reduces the time needed for data gathering, multi-step analysis and sequence prioritization from days to hours and provides this analysis to non-experienced users. EnzymeMiner web server is available at `https://loschmidt.chemi.muni.cz/enzymeminer/`.

SoluProt is a sequence-based predictor of soluble protein expression in *E. coli*, which was created using the gradient boosting machine technique with manually curated TargetTrack database as a training set. SoluProt achieved a slightly higher accuracy (58.5%) and AUC (0.62) than a suite of alternative solubility prediction tools when evaluated using balanced independent test set of 3100 sequences. PROSO II [77], SWI [10] and CamSol [81] were the next best tools, achieving accuracies of 58.0%, 55.9% and 54.1%, respectively. Surprisingly, the recently reported DeepSol [47] and SKADE [68] tools, which are based on deep learning methods, performed worse than simpler methods PROSO II, SWI and CamSol in this comparison. SoluProt also performed well in protein prioritization. The main strengths of SoluProt are that it was trained using a dataset generated by thorough pre-processing of the noisy TargetTrack data, and was validated using a high-quality independent test set. The SoluProt predictor is available via a user-friendly web server or as a standalone software package at `https://loschmidt.chemi.muni.cz/soluprot/`.

In the future, improvements for both EnzymeMiner and SoluProt methods could be implemented. EnzymeMiner can be improved in three aspects. First, metagenomic database MGnify [58] of more than 267 million protein sequences could be included as an additional database for sequence search. The MGnify database contains proteins from organisms which were not yet identified or can not be cultivated under laboratory conditions, such as organisms living deep in ocean in hot springs or in digestion systems. However, greater attention and expert validation is needed for the proteins found in the metagenomic databases as they may include erroneously assembled chimeric sequences which are just an artefact of the whole-genome shot-gun sequencing and subsequent data processing. Second, automated tertiary structure prediction based on homology modelling and threading could be implemented for all identified sequences. The structural predictions will allow a subsequent analysis of active site pockets/cavities and access tunnels. Structural features will significantly enrich the set of annotations and help to identify additional attractive targets for experimental characterization. Third, automated periodical mining could be implemented. When enabled, EnzymeMiner will rerun the analysis periodically and inform the user about novel sequences found since the last search.

SoluProt most probably reached the prediction limit due to both the quality and the quantity of the available protein sequence solubility data. Thus, to improve its applicability, a new prediction task could be considered—the prediction of the effect of an amino acid mutation on protein solubility, specifically a prediction of the difference in solubility between wild-type protein and a variant of the same protein with a change in a single amino acid (single-point mutant). This effort has two motivations. First, the prediction of the effect of a mutation would be extremely useful for engineering solubility of proteins. It would allow to design novel protein variants with improved solubility or just to avoid mutations decreasing solubility. Second, novel experimental data for protein solubility change upon a single-point mutation are emerging rapidly thanks to the advent of deep mutational scanning technology [49]. The data usually contains thousands of samples covering nearly all possible point mutations in a selected protein which makes them well suited for understanding the key mechanisms influencing protein solubility.

EnzymeMiner and SoluProt have promising applicability prospects. EnzymeMiner identifies novel putative enzymes and facilitates selection of several targets for experimental characterization. In industry, these enzymes have a great potential to decrease energetic consumption and environmental burden of many chemical processes. SoluProt indicates the probability of soluble expression for a given protein sequence which helps to prioritize proteins that are easier to produce. This will accelerate the discovery of novel proteins or enzymes which can be produced with high yields. The tight integration of EnzymeMiner and SoluProt enables easy-to-use mining of soluble enzymes, which makes them unique and powerful tools for the protein engineering community. Notably, both tools have already caught the attention of the community, as shown by the number of requests satisfied by both services at this time—more than 1400 jobs calculated by EnzymeMiner and more than 8700 jobs calculated by SoluProt.

Bibliography

[1] Agostini, F., Cirillo, D., Livi, C. M., Delli Ponti, R. and Tartaglia, G. G. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics.* 2014, vol. 30, no. 20, p. 2975–2977.

[2] Agostini, F., Vendruscolo, M. and Tartaglia, G. G. Sequence-Based Prediction of Protein Solubility. *Journal of Molecular Biology.* 2012, vol. 421, 2-3, p. 237–241.

[3] Alford, R. F., Leaver Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. and Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation.* 2017, vol. 13, no. 6, p. 3031–3048.

[4] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 1997, vol. 25, no. 17, p. 3389–3402.

[5] Arakawa, T. and Timasheff, S. N. Theory of protein solubility. *Methods in Enzymology.* 1985, vol. 114, p. 49–77.

[6] Atkinson, H. J., Morris, J. H., Ferrin, T. E. and Babbitt, P. C. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS ONE.* 2009, vol. 4, no. 2.

[7] Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch Mizrachi, I., Kimelman, M., Pruitt, K. D., Resenchuk, S., Tatusova, T., Yaschenko, E. and Ostell, J. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research.* 2012, vol. 40, D1, p. D57–D63.

[8] Berman, H. M., Ganayi, M. J., Kouranov, A., Micallef, D. I., Westbrook, J. and investigators, P. S. I. n. o. *Protein Structure Initiative - TargetTrack 2000-2017 - all data files.* 2017. Type: dataset.

[9] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research.* 2000, vol. 28, no. 1, p. 235–242.

[10] Bhandari, B. K., Gardner, P. P. and Lim, C. S. Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics.* 2020.

[11] Brannigan, J. A. and Wilkinson, A. J. Protein engineering 20 years on. *Nature Reviews Molecular Cell Biology.* 2002, vol. 3, no. 12, p. 964–970.

[12] Buchfink, B., Reuter, K. and Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods.* 2021, vol. 18, no. 4, p. 366–368.

[13] Buermans, H. P. J. and Dunnen, J. T. den. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease.* 2014, vol. 1842, no. 10, p. 1932–1941. From genome to function.

[14] Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann Schaal, M., Jahn, D. and Schomburg, D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research.* 2021, vol. 49, D1, p. D498–D508.

[15] Chang, C. C. H., Song, J., Tey, B. T. and Ramanan, R. N. Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. *Briefings in Bioinformatics.* 2014, vol. 15, no. 6, p. 953–962.

[16] Chapman, J., Ismail, A. E. and Dinu, C. Z. Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks. *Catalysts.* 2018, vol. 8, no. 6, p. 238.

[17] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. and Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica. Section D, Biological Crystallography.* 2010, vol. 66, Pt 1, p. 12–21.

[18] Choi, J.-M., Han, S.-S. and Kim, H.-S. Industrial applications of enzyme biocatalysis: Current status and future aspects. *Biotechnology Advances.* 2015, vol. 33, no. 7, p. 1443–1454.

[19] Chovancová, E. *Bioinformatic analysis and design of haloalkane dehalogenases.* Brno, 2011. Dissertation. Masaryk University.

[20] Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M. and Arrowsmith, C. H. Structural proteomics of an archaeon. *Nature Structural Biology.* 2000, vol. 7, no. 10, p. 903–909.

[21] Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T. and Vranken, W. F. The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Research.* 2014, vol. 42, Web Server issue, p. W264–270.

[22] Cohen, S. N., Chang, A. C. Y., Boyer, H. W. and Helling, R. B. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences.* 1973, vol. 70, no. 11, p. 3240–3244.

[23] Cormier, C. Y., Park, J. G., Fiacco, M., Steel, J., Hunter, P., Kramer, J., Singla, R. and LaBaer, J. PSI:Biology-materials repository: a biologist's resource for protein expression plasmids. *Journal of Structural and Functional Genomics.* 2011, vol. 12, no. 2, p. 55–62.

[24] Damborsky, J. Meeting Report: Protein design and evolution for biocatalysis August 30 – September 1, 2006, Greifswald, Germany. *Biotechnology Journal.* 2007, vol. 2, no. 2, p. 176–179.

[25] Davis, G. D., Elisee, C., Newham, D. M. and Harrison, R. G. New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnology and Bioengineering.* 1999, vol. 65, no. 4, p. 382–388.

[26] Dayhoff, M. O. and Schwartz, R. M. Chapter 22: A model of evolutionary change in proteins. In: *In Atlas of Protein Sequence and Structure.* 1978.

[27] Diaz, A. A., Tomba, E., Lennarson, R., Richard, R., Bagajewicz, M. J. and Harrison, R. G. Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnology and Bioengineering.* 2010, vol. 105, no. 2, p. 374–383.

[28] Dumorné, K., Córdova, D. C., Astorga Eló, M. and Renganathan, P. Extremozymes: A Potential Source for Industrial Applications. *Journal of Microbiology and Biotechnology.* 2017, vol. 27, no. 4, p. 649–659.

[29] Duro Castano, A., Conejos Sánchez, I. and Vicent, M. J. Peptide-Based Polymer Therapeutics. *Polymers.* 2014, vol. 6, no. 2, p. 515–551.

[30] Eddy, S. R. Profile hidden Markov models. *Bioinformatics.* 1998, vol. 14, no. 9, p. 755–763.

[31] Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics.* 2009, vol. 23, no. 1, p. 205–211.

[32] Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010, vol. 26, no. 19, p. 2460–2461.

[33] El Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. and Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Research.* 2019, vol. 47, D1, p. D427–D432.

[34] Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Research.* 2012, vol. 40, D1, p. D136–D143.

[35] Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics.* 1936, vol. 7, no. 2, p. 179–188.

[36] Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics.* 2001, vol. 29, no. 5, p. 1189–1232.

[37] Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R. and Whalen, K. L. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.* 2015, vol. 1854, no. 8, p. 1019–1037.

[38] Hebditch, M., Carballo Amador, M. A., Charonis, S., Curtis, R. and Warwicker, J. Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics.* 2017, vol. 33, no. 19, p. 3098–3100.

[39] Hellinga, H. W. Rational protein design: Combining theory and experiment. *Proceedings of the National Academy of Sciences.* 1997, vol. 94, no. 19, p. 10015–10017.

[40] Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences.* 1992, vol. 89, no. 22, p. 10915–10919.

[41] Hirose, S. and Noguchi, T. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics.* 2013, vol. 13, no. 9, p. 1444–1456.

[42] Hon, J., Borko, S., Stourac, J., Prokop, Z., Zendulka, J., Bednar, D., Martinek, T. and Damborsky, J. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Research.* 2020, vol. 48, W1, p. W104–W109.

[43] Hon, J., Marusiak, M., Martinek, T., Kunka, A., Zendulka, J., Bednar, D. and Damborsky, J. SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics.* 2021, vol. 37, no. 1, p. 23–28.

[44] JCBN. Nomenclature and Symbolism for Amino Acids and Peptides. *European Journal of Biochemistry.* 1984, vol. 138, no. 1, p. 9–37.

[45] Karplus, M. and McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology.* 2002, vol. 9, no. 9, p. 646–652.

[46] Kawashima, S. and Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Research.* 2000, vol. 28, no. 1, p. 374.

[47] Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., Bensmail, H. and Mall, R. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics.* 2018, vol. 34, no. 15, p. 2605–2613.

[48] Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka Nakamichi, T., Inamoto, E., Toyonaga, H. and Mori, H. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA research: an international journal for rapid publication of reports on genes and genomes.* 2005, vol. 12, no. 5, p. 291–299.

[49] Klesmith, J. R., Bacik, J.-P., Wrenbeck, E. E., Michalczyk, R. and Whitehead, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences.* 2017, vol. 114, no. 9, p. 2265–2270.

[50] Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence.* 1997, vol. 97, 1-2, p. 273–324.

[51] Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N. and Scholtz, J. M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophysical Journal*. 2012, vol. 102, no. 8, p. 1907–1915.

[52] Krogh, A., Larsson, B., Heijne, G. von and Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*. 2001, vol. 305, no. 3, p. 567–580.

[53] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. and Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*. 2019, vol. 87, no. 12, p. 1011–1020.

[54] Kyte, J. and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*. 1982, vol. 157, no. 1, p. 105–132.

[55] Leuthaeuser, J. B., Morris, J. H., Harper, A. F., Ferrin, T. E., Babbitt, P. C. and Fetrow, J. S. DASP3: identification of protein sequences belonging to functionally relevant groups. *BMC Bioinformatics*. 2016, vol. 17, no. 1, p. 458.

[56] Magnan, C. N., Randall, A. and Baldi, P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*. 2009, vol. 25, no. 17, p. 2200–2207.

[57] McCullum, E. O., Williams, B. A. R., Zhang, J. and Chaput, J. C. Random mutagenesis by error-prone PCR. *Methods in Molecular Biology (Clifton, N.J.)*. 2010, vol. 634, p. 103–109.

[58] Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A. and Finn, R. D. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*. 2020, vol. 48, D1, p. D570–D578.

[59] Musil, M., Konegger, H., Hon, J., Bednar, D. and Damborsky, J. Computational Design of Stable and Soluble Biocatalysts. *ACS Catalysis*. 2019, vol. 9, no. 2, p. 1033–1054.

[60] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970, vol. 48, no. 3, p. 443–453.

[61] Niwa, T., Kanamori, T., Ueda, T. and Taguchi, H. Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proceedings of the National Academy of Sciences*. 2012, vol. 109, no. 23, p. 8937–8942.

[62] Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T. and Taguchi, H. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences*. 2009, vol. 106, no. 11, p. 4201–4206.

[63] Packer, M. S. and Liu, D. R. Methods for the directed evolution of proteins. *Nature Reviews Genetics*. 2015, vol. 16, no. 7, p. 379–394.

[64] Pearson, W. R. and Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences.* 1988, vol. 85, no. 8, p. 2444–2448.

[65] Piovesan, D., Walsh, I., Minervini, G. and Tosatto, S. C. E. FELLS: fast estimator of latent local structure. *Bioinformatics.* 2017, vol. 33, no. 12, p. 1889–1891.

[66] 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. and Hunt, J. F. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microbial Informatics and Experimentation.* 2011, vol. 1, no. 1, p. 6.

[67] Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. InterProScan: protein domains identifier. *Nucleic Acids Research.* 2005, vol. 33, Web Server, p. W116–W120.

[68] Raimondi, D., Orlando, G., Fariselli, P. and Moreau, Y. Insight into the protein solubility driving forces with neural attention. *PLoS Computational Biology.* 2020, vol. 16, no. 4, p. e1007722.

[69] Rawi, R., Mall, R., Kunji, K., Shen, C.-H., Kwong, P. D. and Chuang, G.-Y. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics.* 2018, vol. 34, no. 7, p. 1092–1098.

[70] Rembeza, E. and Engqvist, M. K. Experimental investigation of enzyme functional annotations reveals extensive annotation error. *BioRxiv.* 2020, p. 2020.12.18.423474.

[71] Richardson, J. S. and Richardson, D. C. The de novo design of protein structures. *Trends in Biochemical Sciences.* 1989, vol. 14, no. 7, p. 304–309.

[72] Saeys, Y., Inza, I. and Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007, vol. 23, no. 19, p. 2507–2517.

[73] Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., Holmes, J. B., Kim, S., Kimchi, A., Kitts, P. A., Lathrop, S., Lu, Z., Madden, T. L., Marchler Bauer, A., Phan, L., Schneider, V. A., Schoch, C. L., Pruitt, K. D. and Ostell, J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research.* 2019, vol. 47, D1, p. D23–D28.

[74] Shannon, P. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research.* 2003, vol. 13, no. 11, p. 2498–2504.

[75] Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K. and Ueda, T. Cell-free translation reconstituted with purified components. *Nature Biotechnology.* 2001, vol. 19, no. 8, p. 751–755.

[76] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. and Higgins,

D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology.* 2011, vol. 7, no. 1, p. 539.

[77] Smialowski, P., Doose, G., Torkler, P., Kaufmann, S. and Frishman, D. PROSO II - a new method for protein solubility prediction. *FEBS journal.* 2012, vol. 279, no. 12, p. 2192–2200.

[78] Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology.* 1981, vol. 147, no. 1, p. 195–197.

[79] Smyth, M. S. and Martin, J. H. J. X Ray crystallography. *Molecular Pathology.* 2000, vol. 53, no. 1, p. 8–14.

[80] Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M. and Popovic, B. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Scientific Reports.* 2017, vol. 7, no. 1, p. 8200.

[81] Sormanni, P., Aprile, F. A. and Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology.* 2015, vol. 427, no. 2, p. 478–490.

[82] Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology.* 2017, vol. 35, no. 11, p. 1026–1028.

[83] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research.* 2019, vol. 47, D1, p. D506–D515.

[84] Vanacek, P. *Automated exploration and characterization of novel haloalkane dehalogenases.* 2020. Dissertation. Masaryk University.

[85] Vanacek, P., Sebestova, E., Babkova, P., Bidmanova, S., Daniel, L., Dvorak, P., Stepankova, V., Chaloupkova, R., Brezovsky, J., Prokop, Z. and Damborsky, J. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catalysis.* 2018, vol. 8, no. 3, p. 2402–2412.

[86] Vanacek, P., Vasina, M., Hon, J., Kovar, D., Faldynova, H., Kunka, A., Buryska, T., Badenhorst, C. P. S., Mazurenko, S., Bednar, D., Bornscheuer, U. T., Damborsky, J. and Prokop, Z. Functional Annotation of an Enzyme Family by Integrated Strategy Combining Bioinformatics with Microanalytical and Microfluidic Technologies. *ChemRxiv.* 2021.

[87] Walsh, I., Martin, A. J. M., Di Domenico, T. and Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics.* 2012, vol. 28, no. 4, p. 503–509.

[88] Walsh, I., Pollastri, G. and Tosatto, S. C. E. Correct machine learning on protein sequences: a peer-reviewing perspective. *Briefings in Bioinformatics.* 2016, vol. 17, no. 5, p. 831–840.

[89] Wilkinson, D. L. and Harrison, R. G. Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/Technology.* 1991, vol. 9, no. 5, p. 443–448.

[90] Zhao, Y., Tang, H. and Ye, Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics.* 2012, vol. 28, no. 1, p. 125–126.

Part I

Appendices

Appendix A

EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities

W104–W109 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 11 May 2020*
doi: 10.1093/nar/gkaa372

EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities

Jiri Hon[1,2,3,†], Simeon Borko[1,2,†], Jan Stourac[1,3], Zbynek Prokop[1,3], Jaroslav Zendulka[2], David Bednar[1,3], Tomas Martinek[2] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and Research Center for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic, [2]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Bozetechova 2, Brno, Czech Republic and [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

Received March 08, 2020; Revised April 13, 2020; Editorial Decision April 27, 2020; Accepted April 29, 2020

ABSTRACT

Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Despite genomic databases growing exponentially, classical biochemical characterization techniques are time-demanding, cost-ineffective and low-throughput. Therefore, computational methods are being developed to explore the unmapped sequence space efficiently. Selection of putative enzymes for biochemical characterization based on rational and robust analysis of all available sequences remains an unsolved problem. To address this challenge, we have developed EnzymeMiner—a web server for automated screening and annotation of diverse family members that enables selection of hits for wet-lab experiments. EnzymeMiner prioritizes sequences that are more likely to preserve the catalytic activity and are heterologously expressible in a soluble form in *Escherichia coli*. The solubility prediction employs the in-house SoluProt predictor developed using machine learning. EnzymeMiner reduces the time devoted to data gathering, multi-step analysis, sequence prioritization and selection from days to hours. The successful use case for the haloalkane dehalogenase family is described in a comprehensive tutorial available on the EnzymeMiner web page. EnzymeMiner is a universal tool applicable to any enzyme family that provides an interactive and easy-to-use web interface freely available at https://loschmidt.chemi.muni.cz/enzymeminer/.

INTRODUCTION

There are currently >259 million non-redundant protein sequences in the NCBI nr database (release 2020-02-10) (1). Despite their enormous promise for biological and biotechnological discovery, experimental characterization has been performed on only a small fraction of the available sequences. Currently, there are about 560 000 protein sequences reliably curated in the UniProtKB/Swiss-Prot database (release 2020_01) (2).

The low ratio of characterized to uncharacterized sequences reflects the sharp contrast in time-demanding/low-throughput biochemical techniques versus fast/high-throughput next-generation sequencing technology. Although more efficient biochemical techniques employing miniaturization and automation have been developed (3–5), the most widely used experimental methods do not provide sufficient capacity for biochemical characterization of proteins spanning the ever-increasing sequence space. Therefore, computational methods are currently the only way to explore the immense protein diversity available among the millions of uncharacterized sequence entries.

Two different computational strategies are generally used for exploration of the unknown sequence space. The first strategy takes a novel uncharacterized sequence as input and predicts functional annotations. The method involves annotating the unknown input sequences by predicting protein domains (6), Enzyme Commission (EC) number (7) or Gene Ontology terms that are a subject of the initiative named the Critical Assessment of Functional Annotation (8). These methods are often universal and applicable to any protein sequence. However, they often lack specificity as the automatic annotation rules or statistical models need to be substantially general. A significant advantage of these methods is their seamless integration into available

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Email: jiri@chemi.muni.cz
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

databases. Submission of a query sequence to a database is sufficient, with no need for running computation- and memory-intensive bioinformatics pipelines locally. A model example of this approach is the automatic annotation workflow of the UniProtKB/TrEMBL database (2).

The second strategy takes a well-known characterized sequence as an input and applies a computational workflow, typically based on a homology search, to identify novel uncharacterized entries in genomic databases that are related to the input query sequence (5,9). The homology search is often followed by a filtration step, which checks the essential sequence properties, e.g. domain structure or presence of catalytic residues. The main advantage of these methods is the higher specificity of the analysis. A disadvantage is that it may be complicated to apply the developed workflow to protein families other than those for which it was designed. Moreover, these workflows typically require running complex bioinformatics pipelines and are usually not available through a web interface.

The fundamental unsolved problem is how to deal with the overwhelming number of sequence entries identified by these methods and select a small number of relevant hits for in-depth experimental characterization. For example, a database search for members of the haloalkane dehalogenase model family using the UniProt web interface yields 3598 sequences (UniProtKB release 2020_01). It is impossible to rationally select several tens of targets for experimental testing without additional bioinformatics analyses to help prioritize such a large pool of sequences.

To address the challenge of exploring the unmapped enzyme sequence space and rational selection of attractive targets, we have developed the EnzymeMiner web server. EnzymeMiner identifies novel enzyme family members, comprehensively annotates the targets and facilitates efficient prioritization and selection of representative hits for experimental characterization. To the best of our knowledge, there is currently no other tool available that allows such a comprehensive analysis in a single easy-to-run integrated workflow on the web.

MATERIALS AND METHODS

EnzymeMiner implements a three-step workflow: (i) homology search, (ii) essential residue based filtering and (iii) hits annotation (Figure 1). To execute these tasks, the server requires two different types of input information: (i) query sequences and (ii) essential residue templates. The query sequences serve as seeds for the initial homology search. The essential residue templates, defined as pairs of a protein sequence and a set of essential residues in that sequence, allow the server to prioritize hits that are more likely to display the enzyme function. Therefore, the essential residues may be the catalytic and ligand- or cofactor-binding residues that are indispensable for proper catalytic function. Each essential residue is defined by its name, position and a set of allowed amino acids for that position.

In the first *homology search step*, a query sequence is used as a query for a PSI-BLAST (10) two-iteration search in the NCBI nr database (1). If more than one query sequence is provided, a search is conducted for each sequence separately. Besides a minimum *E*-value threshold 10^{-20}, the PSI-BLAST hits must share a minimum of 25% global sequence identity with at least one of the query sequences. Artificial protein sequences, i.e. sequences described by the term artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag or replicon, are removed. EnzymeMiner sorts the PSI-BLAST hits by *E*-value and passes a maximum of 10,000 best hits to the next steps in the workflow. The default parameters for the homology search step, as well as the other steps, can be modified using advanced options in the web server.

In the second *essential residue based filtering step*, the homology search hits are filtered using the essential residue templates. First, the hits are divided into template clusters. Each cluster contains all hits matching essential residues of a particular template. Essential residues are checked using global pairwise alignment with the template calculated by USEARCH (11). When multiple essential residue templates match, the hit is assigned to the template with the highest global sequence identity. Second, for each cluster, an initial multiple sequence alignment (MSA) is constructed using Clustal Omega (12). The MSA is used to revalidate the essential residues of identified hits by checking the corresponding column in the MSA. Sequences not matching essential residues of the template are removed from the cluster. Third, the MSA is constructed again for each template cluster and the essential residues are checked for the last time. The final set of identified sequences reported by EnzymeMiner contains all sequences left in the template clusters.

In the third *annotation step*, the identified sequences are annotated using several databases and predictors: (i) transmembrane regions are predicted by TMHMM (13), (ii) Pfam domains are predicted by InterProScan (14), (iii) source organism annotation is extracted from the NCBI Taxonomy (15) and the NCBI BioProject database (16), (iv) protein solubility is predicted by the in-house tool SoluProt for prediction of soluble protein expression in *Escherichia coli* and (v) sequence identities to queries, hits or other optional sequences are calculated by USEARCH (11). SoluProt is based on a random forest regression model that employs 36 sequence-based features (https://loschmidt.chemi.muni.cz/soluprot/). It has been shown to achieve an accuracy of 58%, specificity of 73% and sensitivity of 44% on a balanced independent test set of 3788 sequences (Hon et al., manuscript in preparation). Alternative solubility prediction tools are summarised in a recently published review (17). It is not advised to use the solubility score for other expression systems because it was trained solely on *E. coli* data. We expect further intensive development of protein solubility predictors in coming years and will ensure that the solubility score in the EnzymeMiner stays at the cutting-edge in terms of its accuracy and reproducibility.

The sequence space of the identified hits is visualized using representative sequence similarity networks (SSNs) generated at various clustering thresholds using MMseqs2 (18) and Cytoscape (19). SSNs provide a clean visual approach to identify clusters of highly similar sequences and rapidly spot sequence outliers. SSNs proved to facilitate identification of previously unexplored sequence and function space (20). The SSN generation method used in EnzymeMiner is inspired by the EFI-EST tool (21). The minimum align-

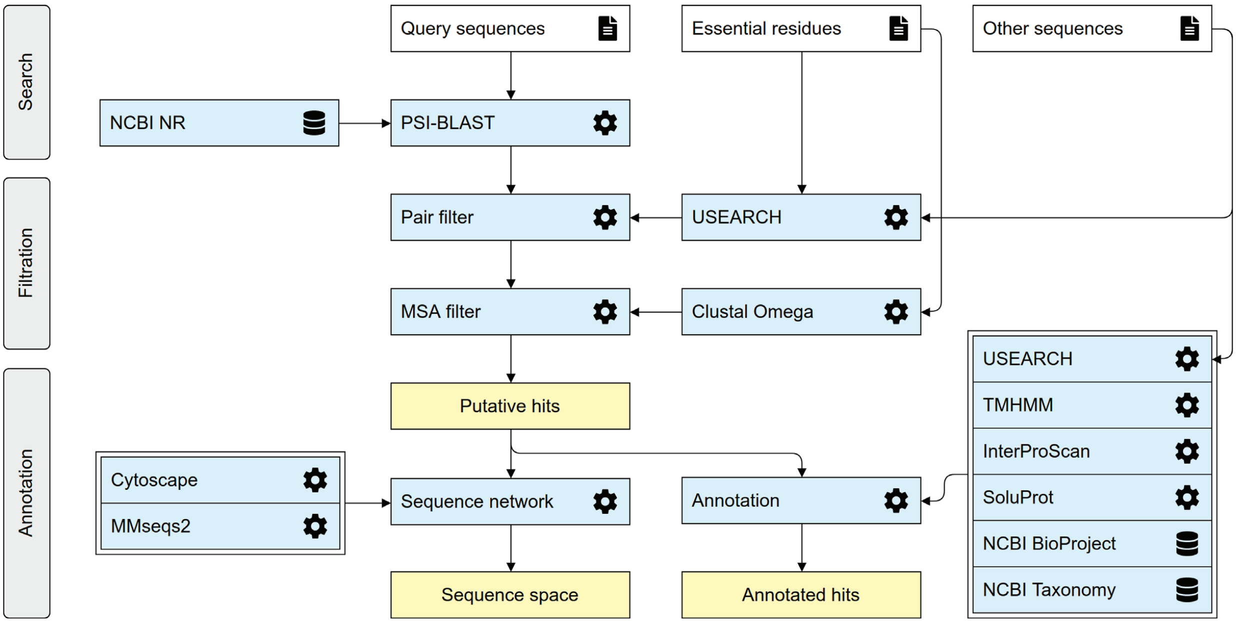


Figure 1. The EnzymeMiner workflow. The workflow consists of three distinct steps: (i) sequence homology search, (ii) filtration of functional sequences, and (iii) annotation of hits. These steps are executed consecutively and automatically. EnzymeMiner has only two required inputs: (i) query sequences, and (ii) essential residue templates. The *Other sequences* are optional inputs that allow EnzymeMiner to calculate the sequence identity between these sequences and all the hits. Input files are highlighted by a white background, tools and databases have a light blue background, outputs are highlighted by a yellow background.

ment score to include an edge between two representative sequences in an SSN is 40.

DESCRIPTION OF THE WEB SERVER

Job submission

New jobs can be submitted from the EnzymeMiner homepage. EnzymeMiner provides two conceptually different ways to define the input of the workflow: (i) using curated sequences from the UniProtKB/Swiss-Prot database and (ii) using custom sequences. We recommend the UniProtKB/Swiss-Prot option for users who do not have in-depth knowledge of the enzyme family. In contrast, the *Custom sequences* tab gives full control over the EnzymeMiner input—query sequences and essential residue templates are specified manually by the user. This is recommended for users who have good knowledge about the enzyme family and want to provide additional starting information to obtain refined results. The last option is a combination of both approaches, where Swiss-Prot sequences can be pre-selected first and then the input can be modified in the *Custom sequences* tab.

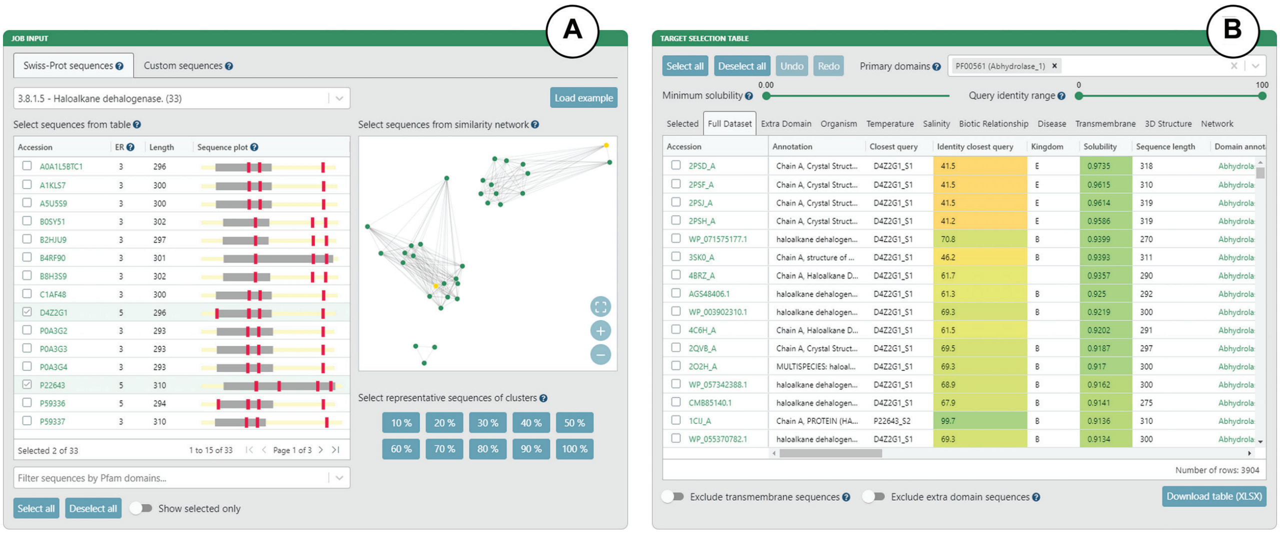
In the *Swiss-Prot sequences* tab (Figure 2A), sequences from the Swiss-Prot database can be queried by Enzyme Commission (EC) number. As a result, a table of all sequences annotated by the EC number and corresponding SSN is generated. The table has four columns: (i) sequence accessions hyperlinked to the UniProt database, (ii) number of essential residues, (iii) sequence length and (iv) sequence plot. The sequence plot summarizes two important features of the sequence – positions of essential residues and identified Pfam domains. The positions of essential residues are obtained from the Swiss-Prot database. The SSN visualizes the sequence space of all the sequences in the current EC group. Nodes represent Swiss-Prot sequences, whereas edge lengths are proportional to the pairwise sequence identities. Similar sequences are close to each other, whereas more distant sequences are not connected at all.

There are three strategies possible for selecting Swiss-Prot sequences as the EnzymeMiner query: (i) select a row from the sequence table, (ii) select a node in the SSN and (iii) select cluster representatives by defining a sequence identity threshold. The sequence identity threshold buttons select cluster representatives at the given percentage threshold. Using this feature, the user can automatically select a small set of sequences that cover the whole known sequence space of the current EC group. All selected Swiss-Prot sequences are used as a query in the homology search step and also as essential residue templates for the filtration step. To modify the selected sets of queries and essential residue templates, the user can switch to the *Custom sequences* tab and refine the selection manually.

EnzymeMiner results

The results page is organized into four sections: (i) *job information* box, (ii) *download results* box, (iii) *target selection table* and (iv) *sequence similarity network*.

In the *job information* box, the user can find the job ID, title, start time and status of the job. There is also a rerun button for rerunning the same analysis without the need for re-entering the same input. This feature is handy for periodically mining new sequences as the sequence databases



3.8.1.5 - Haloalkane dehalogenase. (33)

Select sequences from table

	Accession	ER	Length	Sequence plot
□	A0A1L5BTC1	3	296	
□	A1KLS7	3	300	
□	A5U5S9	3	300	
□	B0SY51	3	302	
□	B2HJU9	3	297	
□	B4RF90	3	301	
□	B8H3S9	3	302	
□	C1AF48	3	300	
☑	D4Z2G1	5	296	
□	P0A3G2	3	293	
□	P0A3G3	3	293	
□	P0A3G4	3	293	
☑	P22643	5	310	
□	P59336	5	294	
□	P59337	3	310	

Selected 2 of 33   1 to 15 of 33   Page 1 of 3

Filter sequences by Pfam domains...

Select all   Deselect all   Show selected only

Select representative sequences of clusters

10 %   20 %   30 %   40 %   50 %
60 %   70 %   80 %   90 %   100 %

Primary domains: PF00561 (Abhydrolase_1)

Minimum solubility: 0.00

Query identity range: 0 – 100

Selected | Full Dataset | Extra Domain | Organism | Temperature | Salinity | Biotic Relationship | Disease | Transmembrane | 3D Structure | Network

	Accession	Annotation	Closest query	Identity closest query	Kingdom	Solubility	Sequence length	Domain annot
□	2PSD_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.5	E	0.9735	318	Abhydrola
□	2PSF_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.5	E	0.9615	310	Abhydrola
□	2PSJ_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.5	E	0.9614	319	Abhydrola
□	2PSH_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.2	E	0.9586	319	Abhydrola
□	WP_071575177.1	haloalkane dehalogen...	D4Z2G1_S1	70.8	B	0.9399	270	Abhydrola
□	3SK0_A	Chain A, structure of ...	D4Z2G1_S1	46.2	B	0.9393	311	Abhydrola
□	4BRZ_A	Chain A, Haloalkane D...	D4Z2G1_S1	61.7		0.9357	290	Abhydrola
□	AGS48406.1	haloalkane dehalogen...	D4Z2G1_S1	61.3	B	0.925	292	Abhydrola
□	WP_003902310.1	haloalkane dehalogen...	D4Z2G1_S1	69.3	B	0.9219	300	Abhydrola
□	4C6H_A	Chain A, Haloalkane D...	D4Z2G1_S1	61.5		0.9202	291	Abhydrola
□	2QVB_A	Chain A, Crystal Struct...	D4Z2G1_S1	69.5	B	0.9187	297	Abhydrola
□	2O2H_A	MULTISPECIES: haloal...	D4Z2G1_S1	69.3	B	0.917	300	Abhydrola
□	WP_057342388.1	haloalkane dehalogen...	D4Z2G1_S1	68.9	B	0.9162	300	Abhydrola
□	CMB85140.1	haloalkane dehalogen...	D4Z2G1_S1	67.9	B	0.9141	275	Abhydrola
□	1CIJ_A	Chain A, PROTEIN (HA...	P22643_S2	99.7	B	0.9136	310	Abhydrola
□	WP_055370782.1	haloalkane dehalogen...	D4Z2G1_S1	69.3	B	0.9134	300	Abhydrola

Number of rows: 3904

Exclude transmembrane sequences   Exclude extra domain sequences   Download table (XLSX)

Figure 2. The EnzymeMiner graphical user interface showing example inputs and results for the haloalkane dehalogenase family (EC 3.8.1.5). (**A**) Inputs based on curated sequences from the UniProtKB/Swiss-Prot database. The input sequences can be selected using: (i) the sequence table, (ii) the SSN or (iii) the sequence identity threshold. (**B**) Target selection table. The table is organized into eleven sheets that summarize the results from different perspectives. The table can be filtered using solubility and identity sliders, and transmembrane and extra domain exclusion switches.

grow. For example, there are hundreds of new hits for the haloalkane dehalogenase family every year. In the *download results* box, the user can download the results table in XLSX format or tab-separated text format. A ZIP archive containing all output files from the EnzymeMiner workflow can also be downloaded.

The *target selection table* is the most important component of the EnzymeMiner results (Figure 2B). It presents all the putative enzyme sequences identified by EnzymeMiner and helps to select targets for experimental characterization. The table is organized into eleven sheets summarizing the results from different perspectives. (i) The *Selected* sheet shows all the sequences selected from individual sheets. It contains an extra column to track the argument used for the selection. By default, it is prefilled by the name of the sheet from which the sequence was selected, but it can be freely changed. (ii) The *Full Dataset* sheet shows all identified sequences. (iii) The *Extra domain* sheet shows sequences with extra Pfam domains found in the sequence but not listed in the *Primary domains* selection box. (iv) The *Organism* sheet shows sequences with known source organisms. (v) The *Temperature* sheet shows sequences from organisms having extreme optimum temperature annotation in the NCBI BioProject database, including sequences from thermophilic or cryophilic organisms. (vi) The *Salinity* sheet shows sequences from organisms having extreme salinity annotation in the NCBI BioProject database. (vii) The *Biotic Relationship* sheet shows sequences from organisms having biotic relationship annotation in the NCBI BioProject database. (viii) The *Disease* sheet shows sequences from organisms having disease annotation in the NCBI BioProject database. (ix) The *Transmembrane* sheet shows sequences with transmembrane regions predicted by the TMHMM tool. (x) The *3D Structure* sheet shows sequences with an available 3D structure in the Protein Data Bank (22). (xi) The *Network* sheet shows sequences clustered into a selected sequence similarity network node.

There are four options for filtering the identified sequences displayed in the target selection table. The first option is the minimum solubility slider. Sequences with lower predicted solubility will be hidden. We recommend setting the solubility threshold to >0.5 to increase the probability of finding soluble protein expression in *E. coli*. We do not recommend to set the solubility threshold too high because of possible trade-off between enzyme solubility and activity (23). The second option is the identity range bar. Only sequences with identity to query sequences in the specified range will be visible. The third option is to exclude transmembrane proteins. We recommend removing these sequences as they are usually difficult to produce and tend to have lower predicted solubility. The fourth option is to exclude proteins with an extra domain. Extra domains are defined as domains found in the sequence but not listed in the *Primary domains* selection box. We recommend avoiding sequences with extra domains, but these sequences may also show interesting and unusual biological properties. The selection table can be sorted by clicking on a column header. Holding 'Shift' while clicking on the column headers allows sorting by multiple columns.

The SSN visualizes the sequence space of all identified sequences. Both clusters of similar sequences and sequence outliers can be easily identified. As there might be thousands of sequences, the sequences are clustered at the identity threshold and only an SSN of the representative sequences is shown for performance reasons. Sequences having greater sequence identity are consolidated into a single metanode. Edges indicate high sequence identity between representative sequences of the connected metanodes. Clicking on a metanode displays the *Network* sheet showing which sequences are represented by a particular metanode. The SSN can be downloaded as a Cytoscape session file for further analysis and custom visualization. Networks clustered at different identities are available. The numbers of nodes and edges are indicated for each identity threshold. The SSN is interactively linked to the target selection table. All nodes representing selected sequences are automatically highlighted in the SSN.

Target selection

The target selection table and SSN facilitate the selection of a diverse set of soluble putative enzyme sequences for experimental validation. First, we recommend setting the maximum sequence identity to queries to 90%. This will remove all hits that are very similar to already known proteins. Second, we recommend selecting a few sequences from individual sheets to cover different phyla from the domains Archea, Bacteria and Eukarya. The most exciting enzymes might be from extremophilic organisms. Third, the SSN can be used to check that the selection covers all sequence clusters. Fourth, users can select sequences from all subfamilies of the enzyme family of interest. The members of different subfamilies can be easily recognized by the *Closest query* or *Closest known* column in the selection table (note: requires representative sequences of subfamilies as job input). Fifth, the available filtering options can be used to (i) prioritize sequences with the highest predicted solubility, (ii) prioritize sequences with known tertiary structures, (iii) eliminate proteins with predicted transmembrane regions and (iv) eliminate sequences with extra domains.

EXPERIMENTAL VALIDATION OF THE EnzymeMiner WORKFLOW

The EnzymeMiner workflow has been thoroughly experimentally validated using the model enzymes haloalkane dehalogenases (5). The sequence-based search identified 658 putative dehalogenases. The subsequent analysis prioritized and selected 20 candidate genes for exploration of their protein structural and functional diversity. The selected enzymes originated from genetically unrelated Bacteria, Eukarya and, for the first time, also Archaea and showed novel catalytic properties and stabilities. The workflow helped to identify novel haloalkane dehalogenases, including (i) the most catalytically efficient enzyme ($k_{\mathrm{cat}}/\mathrm{K}_{0.5} = 96.8\ \mathrm{mM}^{-1}\ \mathrm{s}^{-1}$), (ii) the most thermostable enzyme showing a melting temperature of 71°C, (iii) three different cold-adapted enzymes active at near to 0°C, (iv) highly enantioselective enzymes, (v) enzymes with a wide range of optimal operational temperature from 20 to 70°C and an unusually broad pH range from 5.7–10 and (vi) biocatalysts degrading the warfare chemical yperite and various environmental pollutants. The sequence mining, annotation, and visualization steps from the workflow published by Vanacek and coworkers (5) were fully automated in the EnzymeMiner web server. The successful use case for the haloalkane dehalogenase family is described in an easy-to-follow tutorial available on the EnzymeMiner web page. Additional extensive validation of the fully automated version of EnzymeMiner, experimentally testing the properties of another 45 genes of the haloalkane dehalogenases, is currently ongoing in our laboratory.

CONCLUSIONS AND OUTLOOK

The EnzymeMiner web server identifies putative members of enzyme families and facilitates their prioritization and well-informed manual selection for experimental characterization to reveal novel biocatalysts. Such a task is difficult using the web interfaces of the available protein databases, e.g. UniProtKB/TrEMBL and NCBI Protein, since additional analyses are often required. The major advantage of EnzymeMiner over existing protein sources is the flexibility of input and concise annotation-rich interactive presentation of results. The user can input custom queries and a custom description of essential residues to focus the search on specific protein families or subfamilies. The output of EnzymeMiner is an interactive selection table containing the annotated sequences divided into sheets based on various criteria. The table helps to select a diverse set of sequences for experimental characterization. Two key prioritization criteria are (i) the predicted solubility score, which can be used to prioritize the identified sequences and increase the chance of finding enzymes with soluble protein expression, and (ii) the sequence identity to query sequences complemented with an interactive SSN displayed directly on the web, which can be used to find diverse sequences. Additionally, source organism and domain annotations help to select sequences with diverse properties. EnzymeMiner is a universal tool applicable to any enzyme family. It reduces the time needed for data gathering, multi-step analysis and sequence prioritization from days to hours. All the EnzymeMiner features are implemented directly on the web server and no external tools are required. The web server was optimized for modern browsers including Chrome, Firefox and Safari. An EnzymeMiner job can take a few hours or days to compute, depending on the current load of the server. In the next EnzymeMiner version, we plan three major improvements. First, we will implement automated tertiary structure prediction based on homology modeling and threading for all identified sequences. The structural predictions will allow subsequent analysis of active site pockets/cavities and access tunnels. Structural features will significantly enrich the set of annotations and help to identify additional attractive targets for experimental characterization. Second, we will implement automated periodical mining. When enabled, EnzymeMiner will rerun the analysis periodically and inform the user about novel sequences found since the last search. Finally, we will implement a wizard for automated selection of hits based on input criteria provided by a user.

ACKNOWLEDGEMENTS

We thank the participants of the 1st Hands-on Computational Enzyme Design Course (Brno, Czech Republic) for giving valuable feedback on the EnzymeMiner user interface. Their comments inspired us to make the sequence similarity network visualization more interactive.

FUNDING

Czech Ministry of Education [857560, 02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16_026/0008451, CZ.02.1.01/0.0/0.0/16_019/0000868, LQ1602]; European Commission [720776, 814418]; AI Methods for Cybersecurity and Control Systems project of the Brno University of Technology [FIT-S-20-6293]; Computational resources were supplied by the project 'e-Infrastruktura CZ' [e-INFRA LM2018140] provided within the program Projects of Large Research, Development and Innovations Infrastructures, and by the ELIXIR-CZ project [LM2015047], part of the international ELIXIR infrastructure. Funding for open access charge: Czech Ministry of Education.
Conflict of interest statement. None declared.

REFERENCES

1. Sayers,E.W., Agarwala,R., Bolton,E.E., Brister,J.R., Canese,K., Clark,K., Connor,R., Fiorini,N., Funk,K., Hefferon,T. *et al.* (2019) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **47**, D23–D28.
2. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
3. Colin,P.-Y., Kintses,B., Gielen,F., Miton,C.M., Fischer,G., Mohamed,M.F., Hyvönen,M., Morgavi,D.P., Janssen,D.B. and Hollfelder,F. (2015) Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.*, **6**, 1–12.
4. Beneyton,T., Thomas,S., Griffiths,A.D., Nicaud,J.-M., Drevelle,A. and Rossignol,T. (2017) Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica. *Microb. Cell Fact.*, **16**, 18.
5. Vanacek,P., Sebestova,E., Babkova,P., Bidmanova,S., Daniel,L., Dvorak,P., Stepankova,V., Chaloupkova,R., Brezovsky,J., Prokop,Z. *et al.* (2018) Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization. *ACS Catal.*, **8**, 2402–2412.
6. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
7. Li,Y., Wang,S., Umarov,R., Xie,B., Fan,M., Li,L. and Gao,X. (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
8. Zhou,N., Jiang,Y., Bergquist,T.R., Lee,A.J., Kacsoh,B.Z., Crocker,A.W., Lewis,K.A., Georghiou,G., Nguyen,H.N., Hamid,M.N. *et al.* (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, **20**, 244.
9. Mak,W.S., Tran,S., Marcheschi,R., Bertolani,S., Thompson,J., Baker,D., Liao,J.C. and Siegel,J.B. (2015) Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nat. Commun.*, **6**, 1–10.
10. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
12. 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.
13. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
14. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
15. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
16. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
17. Musil,M., Konegger,H., Hon,J., Bednar,D. and Damborsky,J. (2019) Computational design of Stable and Soluble Biocatalysts. *ACS Catal.*, **9**, 1033–1054.
18. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
19. Shannon,P. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
20. Copp,J.N., Akiva,E., Babbitt,P.C. and Tokuriki,N. (2018) Revealing unexplored sequence-function space using sequence similarity networks. *Biochemistry*, **57**, 4651–4662.
21. Gerlt,J.A., Bouvier,J.T., Davidson,D.B., Imker,H.J., Sadkhin,B., Slater,D.R. and Whalen,K.L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta (BBA) - Proteins Proteomics*, **1854**, 1019–1037.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. Klesmith,J.R., Bacik,J.-P., Wrenbeck,E.E., Michalczyk,R. and Whitehead,T.A. (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl Acad. Sci. U.S.A.*, **114**, 2265–2270.

Appendix B

SoluProt: prediction of soluble protein expression in *Escherichia coli.*

Bioinformatics, 37(1), 2021, 23–28
doi: 10.1093/bioinformatics/btaa1102
Advance Access Publication Date: 8 January 2021
Original Paper

Sequence analysis

SoluProt: prediction of soluble protein expression in *Escherichia coli*

Jiri Hon[1,2,3], Martin Marusiak[3], Tomas Martinek[3], Antonin Kunka[1,2], Jaroslav Zendulka[3], David Bednar[1,2,*] and Jiri Damborsky[1,2,*]

[1]Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic, [2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic and [3]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Brno 612 66, Czech Republic

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on October 3, 2020; revised on December 5, 2020; editorial decision on December 24, 2020; accepted on December 28, 2020

Abstract

Motivation: Poor protein solubility hinders the production of many therapeutic and industrially useful proteins. Experimental efforts to increase solubility are plagued by low success rates and often reduce biological activity. Computational prediction of protein expressibility and solubility in *Escherichia coli* using only sequence information could reduce the cost of experimental studies by enabling prioritization of highly soluble proteins.

Results: A new tool for sequence-based prediction of soluble protein expression in *E.coli*, SoluProt, was created using the gradient boosting machine technique with the TargetTrack database as a training set. When evaluated against a balanced independent test set derived from the NESG database, SoluProt's accuracy of 58.5% and AUC of 0.62 exceeded those of a suite of alternative solubility prediction tools. There is also evidence that it could significantly increase the success rate of experimental protein studies. SoluProt is freely available as a standalone program and a user-friendly webserver at https://loschmidt.chemi.muni.cz/soluprot/.

Availability and implementation: https://loschmidt.chemi.muni.cz/soluprot/.

Contact: jiri@chemi.muni.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Low protein solubility causes severe problems in protein science and industry; insufficient protein solubility is probably the most common cause of failure of protein production pipelines. The importance of solubility is underlined by the findings of the large-scale Protein Structure Initiative (PSI) project (Berman *et al.*, 2017), which sought to produce thousands of protein sequences from different organisms, crystallize them and resolve their tertiary structure. Unfortunately, in most cases it proved impossible to produce the target proteins in soluble form. The inherent low solubility of natural enzymes also limits the success of emerging high-throughput pipelines that explore protein databases to identify novel enzymes with diverse functions (Hon *et al.*, 2020; Vanacek *et al.*, 2018). Given the rapid growth of protein sequence databases driven by the capabilities of next-generation sequencing technologies, there is an urgent need to focus only on potentially soluble targets to avoid wasting resources on hard-to-produce orthologs. Solubility is thus a key attribute when choosing protein targets for experimental characterization (Vanacek *et al.*, 2018). Strictly speaking, solubility is a thermodynamic parameter defined as the protein's concentration in a saturated solution in equilibrium with a solid phase under specific conditions. However, it is challenging to quantitatively measure the solubility of large sets of proteins (Kramer *et al.*, 2012), so there is little quantitative experimental data on protein solubility. Moreover, this definition of solubility is too narrow to encompass many of the practical problems that may occur during protein production with common expression systems. Therefore, inspired by existing tools (Supplementary Table S1) (Agostini *et al.*, 2014; Khurana *et al.*, 2018; Raimondi *et al.*, 2020; Smialowski *et al.*, 2012), available data (Berman *et al.*, 2017) and laboratory practice, we use a slightly extended definition of protein solubility in this work. Specifically, by solubility, we mean the probability of soluble protein (over)expression in *Escherichia coli* cells. The difference from the classical thermodynamic solubility is in the perception of the insoluble class. We assume that insoluble proteins were either not expressed or were expressed in the insoluble form.

Solubility depends on many extrinsic and intrinsic factors. Extrinsic factors are dictated by the choice of expression system and the experimental conditions used in protein production. Expression systems may be either *in vivo* or *in vitro* (Carlson *et al.*, 2012; Rosano and Ceccarelli, 2014). *In vivo* protein expression is induced inside living cells of a host organism, whereas *in vitro* expression relies on the use of cell-free translational systems. Solubility can be increased by adjusting extrinsic solubility factors, especially by using different mutated host strains, codon optimization, coexpression of chaperones and foldases, lowering cultivation temperatures and adding suitable fusion partners (Costa *et al.*, 2014). However, tuning the expression system or experimental conditions is not always sufficient to confer solubility, and is not feasible in high-throughput protein production pipelines. If extrinsic factors cannot be varied, protein solubility will depend only on the intrinsic properties of the protein sequence. Unfortunately, the relationship between a protein's sequence and its solubility is poorly understood, mainly due to a lack of reproducible quantitative solubility measurements (Kramer *et al.*, 2012). Recent protein engineering studies suggest that charged amino acids on the protein surface are key intrinsic determinants of solubility (Carballo-Amador *et al.*, 2019; Chan *et al.*, 2013; Sankar *et al.*, 2018). However, this knowledge cannot be directly used for solubility prediction due to a lack of structural data. Despite the continuous growth of structural databases (Burley *et al.*, 2019), the structures of proteins of interest are generally unknown, and the limited availability of template structures prevents their accurate computational prediction.

The simultaneous effects of extrinsic and intrinsic factors make solubility prediction challenging. For example, the prediction of solubility from sequence data using machine learning is hindered by the high level of noise in typical training datasets due to the influence of diverse extrinsic variables. Because the molecular mechanisms governing protein solubility are poorly understood, recent solubility prediction tools rely heavily on statistical analysis and machine learning, using previously reported experimental data to train and validate model parameters. One of the most widely used data sources is the TargetTrack database (Berman *et al.*, 2017), formerly known as PepcDB or TargetDB, which integrates information from the Protein Structure Initiative projects. This database contains data from over 900 000 protein crystallization trials involving almost 300 000 unique protein sequences, which are referred to as targets. The database does not contain solubility data per se, but target proteins can be considered soluble if they were successfully purified in the experimental trials. A major limitation of this database is the low quality of its annotations. For example, reasons for failure are generally not provided for unsuccessful crystallization attempts. Therefore, it is impossible to distinguish failures due to insolubility from failures due to other problems later in the experimental pipeline. Second, the experimental protocols used for protein production and crystallization are described in free text with no internal structure, making it hard to automatically extract information about experimental conditions and expression systems for a given target. Filtering is therefore needed to reduce noise before using the TargetTrack data for model training. However, the application of stringent filtering rules to the target annotations can dramatically reduce the number of usable records.

eSOL is another well-known and commonly used solubility database (Niwa *et al.*, 2009, 2012) that contains experimentally measured solubilities for over 3 000 *E.coli* proteins produced in the PURE (Shimizu *et al.*, 2001) cell-free expression system. eSOL is an impressive collection of highly homogenous data but has its own limitations. First, it only contains data on proteins originating from *E.coli*. Second, it has relatively little negative data; adding the three main cytosolic *E.coli* chaperones (TF, DnaKJE and GroEL/GroES) to the PURE expression system reduced the number of insoluble proteins from 788 to 24 (Niwa *et al.*, 2012). eSOL is a valuable source of exact solubility data that were generated using a robust pipeline and provide a good quantitative measure of thermodynamic solubility. However, these data cannot be used to assess solubility according to our expanded definition, which also encompasses expressibility.

The relationship between protein sequence and solubility has been studied for over 30 years, leading to the development of several predictive models and software tools. There are 11 such models or tools that use definitions of solubility like that described above and take protein sequences as their sole input. These are the revised Wilkinson-Harrison model (rWH) (Davis *et al.*, 1999; Wilkinson and Harrison, 1991), SOLpro (Magnan *et al.*, 2009), RPSP (Diaz *et al.*, 2010), PROSO II (Smialowski *et al.*, 2012), ccSOL omics (Agostini *et al.*, 2012, 2014), ESPRESSO (Hirose and Noguchi, 2013), CamSol (Sormanni *et al.*, 2015), Protein-Sol (Hebditch *et al.*, 2017), DeepSol (Khurana *et al.*, 2018), SKADE (Raimondi *et al.*, 2020) and the Solubility-weighted index (SWI) (Bhandari *et al.*, 2020). However, the accuracy of these tools is limited, and there is clear room for improvement. Additionally, these tools exhibit poor generality when used to make predictions based on previously unseen data. A comprehensive review of advances in solubility prediction, including predictors that use protein structures as inputs, was published recently (Musil *et al.*, 2019). Here, we present a novel machine learning based tool, SoluProt, for predicting soluble expression from protein sequence data. SoluProt benefits from thorough dataset pre-processing and predicts soluble expression more accurately than previously reported methods.

2 SoluProt training and test set

We used the TargetTrack database to build the *SoluProt training set*. Since this database does not directly provide solubility information, we inferred solubility computationally, using an approach similar to those adopted previously (Magnan *et al.*, 2009; Smialowski *et al.*, 2012). A protein was considered *soluble* if it was recorded as having reached a soluble experimental state or any subsequent state requiring soluble expression (Supplementary Table S2). If failed expression or purification was mentioned in the experiment record's stop status, the protein was labeled *insoluble*. In contrast to a previous approach (Smialowski *et al.*, 2012), we required an explicit stop status relating to insolubility to reduce the frequency of incorrect classification of insoluble sequences. To improve the quality of the training set, we also performed several additional steps to clean the data.

Most importantly, we performed keyword matching combined with manual checking of TargetTrack annotations to extract only proteins expressed in the most common host organism, *E.coli*. This was necessary because a protein soluble in one organism might be insoluble in another. By focusing solely on the most common expression system, we reduced the noise in the training data. We also used specific keywords to search the unstructured descriptions of experimental protocols provided in the TargetTrack database (Supplementary Table S3). Generic search phrases like '*E.coli*' or '*Escherichia coli*' were used to identify potential *E.coli* related protocols. These protocols were then manually checked and confirmed (Supplementary Table S4). A full list of 248 TargetTrack protocols signifying expression in *E.coli* is available at the SoluProt website.

We next identified transmembrane proteins in the dataset based on direct annotations from the TargetTrack database and predictions generated using TOPCONS (Tsirigos *et al.*, 2015) with default settings. The transmembrane proteins were then removed, along with sequences shorter than 20 amino acids, and sequences with undefined residues. We also removed sequences that had been classified as insoluble but for which a protein structure was available in the Protein Data Bank (PDB) (Berman, 2000). To this end, we compiled an *E.coli* PDB subset containing sequences of proteins whose structures had been solved by NMR or X-ray crystallography and which had been expressed in *E.coli* according to the PDB annotations (64 416 sequences, downloaded April 4, 2018). Because both NMR and X-ray crystallography require soluble proteins, any protein in this PDB subset can be considered soluble in *E.coli*. This step reflects advances in molecular biology: methodological developments have made it possible to produce and crystallize some proteins that were previously considered insoluble.

Finally, we reduced the sequence redundancy in the training set by clustering to 25% identity using MMseqs2 (Steinegger and

Söding, 2017) and retaining only representative sequences from each cluster. This was done separately for positive and negative samples to avoid simplifying the prediction problem. We balanced the number of soluble and insoluble samples such that both classes were equally represented. Additionally, we balanced the sequence length distribution so that length alone would not play a dominant role in the predictions. Sequence length correlates with protein solubility—larger proteins are usually less soluble. However, we wanted to suppress its influence in the model because we anticipate that SoluProt would mainly be used to prioritize proteins of similar lengths, usually from a single protein family. A typical expected use case is that of the EnzymeMiner web server (Hon *et al.*, 2020) for automated mining of soluble enzymes. A prediction model relying heavily on sequence length would not perform well in this use case.

The *SoluProt test set* was built from a dataset generated by the North East Structural Consortium (NESG), which represents 9644 proteins expressed in *E.coli* using a unified production pipeline (Price *et al.*, 2011). The dataset contains two integer scores ranging from 0 to 5 for each target, indicating the protein's level of expression and the soluble fraction recovery. The reproducibility of the experimental results in the dataset was validated by performing repeat measurements for selected targets. The NESG dataset targets are included in the TargetTrack database because the NESG participated in the PSI project. However, the expression and solubility levels from the NESG dataset were not included in the TargetTrack database; instead, they were provided to us directly by the authors of the original study (W. Nicholson Price II, personal communication). The high consistency and quality of the NESG dataset make it suitable for benchmarking purposes. We processed the NESG dataset using the same procedure as the training set, although the computational solubility derivation and expression system filtration steps were omitted because they were pointless in this case. Instead, we transformed the solubility levels into binary classes: all proteins with a solubility level of 1 or above were considered soluble and all others insoluble.

Finally, we ensured that no pair consisting of a sequence from the test set and a sequence from the training set had a global sequence identity above 25% as calculated using the USEARCH software (Edgar, 2010). This made the test set more independent because it ensured that predictions were not validated against data similar to those used during training. In total, 11 436 protein sequences remained in the *SoluProt training set* and 3 100 in the independent *SoluProt test set*. Both datasets had equal numbers of soluble and insoluble samples with balanced sequence length distributions (Supplementary Fig. S1). The datasets are available at the SoluProt website. The dataset construction steps are summarized in Supplementary Table S5.

3 Prediction model

The SoluProt predictor is implemented in Python using scikit-learn (Pedregosa *et al.*, 2011), Biopython (Cock *et al.*, 2009) and pandas (McKinney, 2010) libraries. We used a gradient boosting machine (GBM) (Friedman, 2001) to generate the predictive model. Prediction features were selected from a set of 251 sequence characteristics that were divided into eight groups: (i) single amino acid content (20 features), (ii) amino acid dimer content (210 features), (iii), sequence physicochemical features (12 features, Supplementary Table S6), (iv) average flexibility as computed by DynaMine (Cilia *et al.*, 2014) (1 feature), (v) secondary structure content as predicted by FESS (Piovesan *et al.*, 2017) (3 features), (vi) average disorder as predicted by ESPRITZ (Walsh *et al.*, 2012) (1 feature), (vii) content of amino acids in transmembrane helices as predicted by TMHMM (Krogh *et al.*, 2001) (3 features) and (viii) maximum identity to the *E.coli* PDB subset as calculated using USEARCH (1 feature). All sequences equal to any sequence from the test set were excluded from the *E.coli* PDB subset for the calculation of maximum identity. The objective was to eliminate even the indirect presence of test set sequences from model training. We standardized all features by subtracting the mean and scaling to unit variance. The means and variances were calculated using the training set.

We removed correlated features in two steps. First, we fitted a GBM with default parameters using the full training set and all features. Second, we calculated Pearson's correlation coefficient for each pair of features. If the correlation between any two features exceeded 0.75, we removed the feature with the lesser importance in the fitted GBM model. We also removed irrelevant features using LASSO (Tibshirani, 1996). LASSO's alpha parameter was optimized to maximize the mean AUC of the GBM model with default parameters over 5-fold cross-validation. The alpha parameter was varied between 0.08 to 0 with a step size of 6.25×10^{-4}; its optimal value was 0.005. In total, 96 features were selected for inclusion in the predictive model (Supplementary Table S7). The DynaMine, FESS and ESPRITZ features were not included in the final feature set.

We next optimized the hyperparameters of the GBM model, using an iterative 7-stage strategy to maximize the mean AUC over 5-fold cross-validation using the training set (Supplementary Table S8). In each stage, one or two parameters were optimized using grid search; other parameters were left either at their final values from the previous stages or at the default value if the parameter had not yet been optimized. The best GBM model achieved mean AUC values of 0.85 ± 0.003 for the training part and 0.72 ± 0.02 for the validation part. Overall, the feature selection and hyperparameter optimization had little effect on the mean AUC: without these measures, the mean AUC values for the training and validation sets were 0.83 ± 0.003 and 0.72 ± 0.02, respectively. The main benefit of the feature selection and parameter tuning steps was that they reduced the number of features and thus made the feature calculation step roughly two times faster.

Finally, we used the best GBM hyperparameters to train the final SoluProt model using the full training set. The resulting model had an AUC of 0.84 and an accuracy of 76% for the full training set. The five most important features according to the GBM are: (i) maximum identity to the *E.coli* PDB subset (14.5%), (ii) isoelectric point (6.2%), (iii) predicted number of amino acids in transmembrane helices in the first sixty amino acids of the protein (4.2%), (iv) lysine content (4.0%) and (v) glutamine content (3.5%) (Supplementary Table S7).

4 Performance evaluation and comparison

We used the SoluProt test set to evaluate and compare SoluProt to 11 previously published tools. The evaluation relied on both threshold-independent (area under the ROC curve) and threshold-dependent metrics (accuracy, Matthew's correlation coefficient and confusion matrices). For the threshold-dependent metrics, we applied a threshold of 0.5 or the thresholds recommended by the authors of the corresponding method (Table 1). SoluProt achieved the highest accuracy (58.5%) and the greatest AUC (0.62) of the

Table 1. Performance of various solubility predictors using the balanced SoluProt test set of 3100 sequences

Method	AUC	T	ACC	MCC	TP	TN	FP	FN
SoluProt	0.62	0.50	58.5%	0.17	939	873	677	611
PROSO II	0.60	0.60	58.0%	0.17	630	1167	383	920
SWI	0.60	0.50	55.9%	0.13	1206	527	1023	344
CamSol	0.57	1.00	54.1%	0.08	676	1001	549	874
ESPRESSO	0.56	0.50	53.8%	0.08	1003	664	886	547
rWH	0.55	0.50	54.0%	0.08	670	1005	545	880
DeepSol	0.55	0.50	52.9%	0.09	230	1409	141	1320
Protein-Sol	0.54	0.45	51.6%	0.03	1056	544	1006	494
SOLpro	0.53	0.50	52.0%	0.04	654	959	591	896
SKADE	0.51	0.50	49.2%	−0.03	159	1366	184	1391
ccSOL omics	0.51	0.50	50.8%	0.02	884	690	860	666
RPSP	0.50	0.50	49.8%	0.00	501	1044	506	1049

Note: The different definitions of solubility and target expression system (Supplementary Table S1) should be considered when comparing the performance of individual tools.

AUC—area under the ROC curve, T—threshold for the soluble class, ACC—accuracy, MCC—Matthew's correlation coefficient, TP—true positives, TN—true negatives, FP—false positives, FN—false negatives.

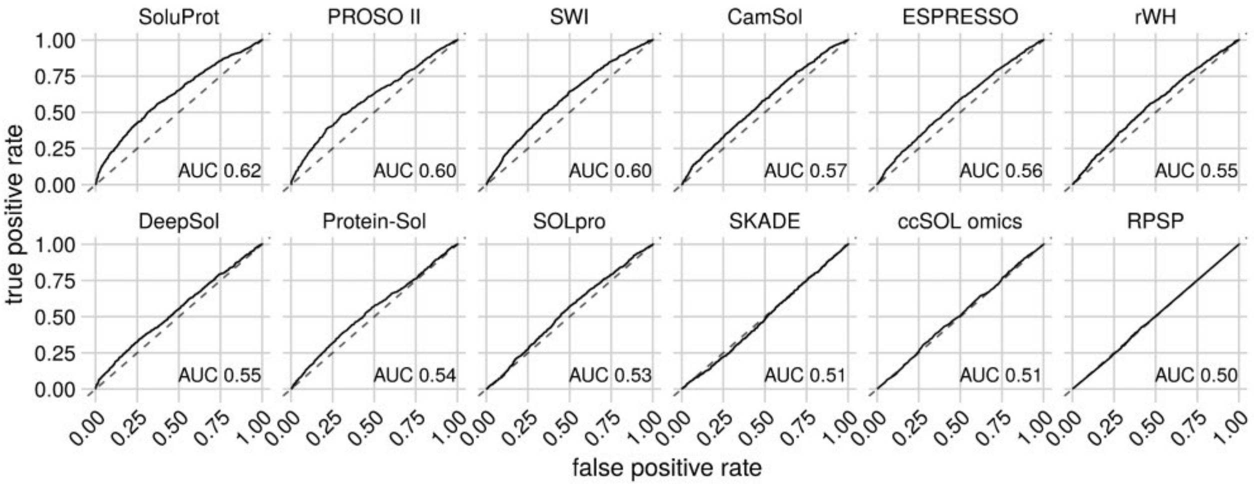


Fig. 1. Receiver operating curves (ROC) calculated for the balanced SoluProt test set of 3100 sequences. The predictors are ordered by the area under the receiver operating curve (AUC)

Table 2. Overlaps between the SoluProt test set and available training sets

Dataset	Size	Test set overlap	TP	TN	FP	FN
PROSO II initial	129643	2952 (95.2%)	951	1437	50	514
DeepSol/SKADE	69420	2294 (74.0%)	737	1130	67	360
SWI	12216	820 (26.5%)	537	210	53	20
SOLpro	17408	480 (15.5%)	178	120	39	143

Note: Two sequences were considered identical if their global sequence identity reported by USEARCH was 100%. Differences in solubility annotations for identical sequences were quantified using confusion matrix terms (TP, TN, FP and FN). The solubility annotations of the SoluProt test set are assumed to reflect the true solubilities of the proteins.

TP—true positives, TN—true negatives, FP—false positives, FN—false negatives. [a] DeepSol and SKADE share the same training set.

tested tools when evaluated against the SoluProt test set (Table 1 and Fig. 1),followed by PROSO II and SWI.

While the SoluProt test set is independent of the SoluProt training set, other tools' training sets might overlap with our test set. Therefore, we compared the SoluProt test set to the training sets of DeepSol, SKADE, SWI and SOLpro to quantify their overlaps (Table 2). DeepSol and SKADE have a common training set, which showed the largest overlap (74.0%), followed by the SWI training set (26.5%) and the SOLpro training set (15.5%). SWI benefits from the overlap; it was the third-best tool in our comparison. DeepSol and SKADE ranked 7th and 12th by accuracy with respect to the SoluProt test set despite having the greatest proportion of test sequences in their training set. This comparatively poor performance can be partly explained by differences in solubility annotations between the DeepSol training set and the SoluProt test set (Table 2): 360 (11.6% of the total) sequences annotated as insoluble in the DeepSol training set were annotated as soluble in the SoluProt test set. The total number of disagreements (the sum of false positives and false negatives) ranged from 336 to 551, depending on the binarization threshold applied to the SoluProt test set (Supplementary Table S9). No training set was published for PROSO II; only an initial set of soluble and insoluble sequences without pre-processing is available. However, the initial set exhibits 95.2% overlap with the SoluProt test set. Therefore, we expect the overlap of the PROSO II training set to also be very high, like the DeepSol training set. Unfortunately, the training sets of other previously developed tools have not been published, preventing a more comprehensive comparison.

The absolute accuracy of the available solubility prediction tools is low (below 60%), so there is clearly room for improvement. Nevertheless, SoluProt and other tools can be useful for protein sequence prioritization (Fig. 2), i.e. for selecting a small number of sequences for in-depth experimental characterization from a large database of several hundreds or thousands of sequences. Specifically, predicted solubility values can be used to select a limited number of high-scoring protein sequences. For example, if we use SoluProt predictions to order the SoluProt test set and remove all sequences bar the 10% with the highest scores, we get 232 true positives, i.e. 49.7% more true positives than would be expected with blind selection (155 true positives). This shows that despite their limited accuracy, current solubility predictors are valuable for protein sequence prioritization and can increase the success rate of experimental protein studies.

5 Conclusions

We have developed a novel method and software tool, SoluProt, for sequence-based prediction of soluble protein expression in *E.coli*. The tool simultaneously predicts the solubility and expressibility of the proteins under consideration. SoluProt achieved a higher accuracy (58.5%) and AUC (0.62) than a suite of alternative solubility prediction tools when evaluated using the balanced independent SoluProt test set of 3100 sequences. PROSO II, SWI and CamSol were the next best tools, achieving accuracies of 58.0%, 55.9% and 54.1%, respectively. SoluProt also performed well in protein prioritization. The main strengths of SoluProt are that it was trained using a dataset generated by thorough pre-processing of the noisy TargetTrack data, and was validated using a high-quality independent test set.

Surprisingly, the recently reported DeepSol (Khurana *et al.*, 2018) and SKADE (Raimondi *et al.*, 2020) tools, which are based on deep learning methods, performed worse than the simpler and mostly older methods PROSO II (Smialowski *et al.*, 2012), SWI (Bhandari *et al.*, 2020) and CamSol (Sormanni *et al.*, 2015) in our comparison. This may be partly due to the overlap of their training set with our test set and disagreements between these sets with respect to the solubility of certain sequences.

The SoluProt predictor is available via a user-friendly web server or as a standalone software package at https://loschmidt.chemi.muni.cz/soluprot/. The SoluProt web server has already predicted the solubility of over 4700 unique protein sequences in ten months since its launch in February 2020. It has also been integrated into the web server EnzymeMiner (Hon *et al.*, 2020) for automated

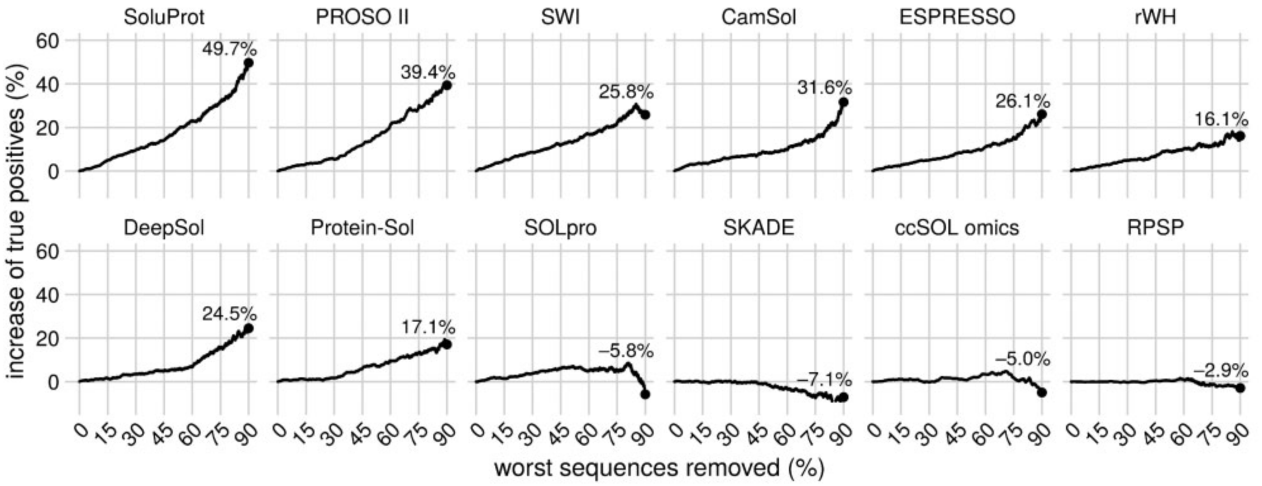


Fig. 2. Increases in the number of true positives resulting from sequence prioritization using the tested solubility prediction tools. The SoluProt test set sequences were ordered by predicted solubility based on each predictor's output, and a variable percentage of the sequences with the worst predicted solubility was then removed. The increase in the number of true positives was then calculated relative to a baseline random selection. For example, upon randomly removing 90% of the test set sequences (2790 samples), we would expect half of the remaining 310 sequences to be true positives

mining of novel soluble enzymes from protein databases (https://loschmidt.chemi.muni.cz/enzymeminer/).

Funding

This work was supported by Czech Ministry of Education [857560, 02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16_026/0008451, LQ1602]; Czech Grant Agency (20-15915Y); European Commission [857560, 720776, 814418]; and AI Methods for Cybersecurity and Control Systems project of the Brno University of Technology [FIT-S-20-6293]. Computational resources were supplied by the project 'e-Infrastruktura CZ' [e-INFRA LM2018140] and by the ELIXIR-CZ [LM2018131]. Funding for open access charge: Czech Ministry of Education.

Conflict of Interest: none declared.

References

Agostini,F. *et al.* (2014) ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics*, **30**, 2975–2977.

Agostini,F. *et al.* (2012) Sequence-based prediction of protein solubility. *J. Mol. Biol.*, **421**, 237–241.

Berman,H.M. *et al.* (2017) Protein Structure Initiative – TargetTrack 2000-2017 – all data files. *Zenodo*. doi:10.5281/zenodo.821654.

Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhandari,B.K. *et al.* (2020) Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics*, **36**, 4691–4698.

Burley,S.K. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464.

Carballo-Amador,M.A. *et al.* (2019) Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. *BMC Biotechnology*, **19**, 26.

Carlson,E.D. *et al.* (2012) Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.*, **30**, 1185–1194.

Chan,P. *et al.* (2013) Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci. Rep.*, **3**, 3333.

Cilia,E. *et al.* (2014) The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.*, **42**, W264–W270.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Costa,S. *et al.* (2014) Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system. *Front. Microbiol.*, **5**, 63.

Davis,G.D. *et al.* (1999) New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol. Bioeng.*, **65**, 382–388.

Diaz,A.A. *et al.* (2010) Prediction of protein solubility in *Escherichia coli* using logistic regression. *Biotechnol. Bioeng.*, **105**, 374–383.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Hebditch,M. *et al.* (2017) Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, **33**, 3098–3100.

Hirose,S. and Noguchi,T. (2013) ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics*, **13**, 1444–1456.

Hon,J. *et al.* (2020) EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res.*, **48**, W104–W109.

Khurana,S. *et al.* (2018) DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, **34**, 2605–2613.

Kramer,R.M. *et al.* (2012) Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys. J.*, **102**, 1907–1915.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Magnan,C.N. *et al.* (2009) SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, **25**, 2200–2207.

McKinney,W. (2010) Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. SciPy Organizers, Austin, Texas, pp. 56–61.

Musil,M. *et al.* (2019) Computational design of stable and soluble biocatalysts. *ACS Catal.*, **9**, 1033–1054.

Niwa,T. *et al.* (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. USA*, **106**, 4201–4206.

Niwa,T. *et al.* (2012) Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. USA*, **109**, 8937–8942.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Piovesan,D. *et al.* (2017) FELLS: fast estimator of latent local structure. *Bioinformatics*, **33**, 1889–1891.

Price,W.N. *et al.* (2011) Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microb. Inf. Exp.*, **1**, 6.

Raimondi,D. *et al.* (2020) Insight into the protein solubility driving forces with neural attention. *PLoS Comput. Biol.*, **16**, e1007722.

Rosano,G.L. and Ceccarelli,E.A. (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.*, **5**, 172.

Sankar,K. *et al.* (2018) AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins*, **86**, 1147–1156.

Shimizu,Y. *et al.* (2001) Cell-free translation reconstituted with purified components. *Nat. Biotechnol.*, **19**, 751–755.

Smialowski,P. *et al.* (2012) PROSO II - a new method for protein solubility prediction. *FEBS J.*, **279**, 2192–2200.

Sormanni,P. *et al.* (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.*, **427**, 478–490.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)*, **58**, 267–288.

Tsirigos,K.D. *et al.* (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–W407.

Vanacek,P. *et al.* (2018) Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization. *ACS Catal.*, **8**, 2402–2412.

Walsh,I. *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Wilkinson,D.L. and Harrison,R.G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N.Y.)*, **9**, 443–448

Supplementary Information

SoluProt: Prediction of Soluble Protein Expression in *Escherichia coli*

Jiri Hon[1,2,3], Martin Marusiak[3], Tomas Martinek[3], Antonin Kunka[1,2], Jaroslav Zendulka[3], David Bednar[1,2], Jiri Damborsky[1,2]

[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]International Clinical Research Center, St. Anne's University Hospital Brno, 656 91 Brno, Czech Republic; [3]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic

Table S1. The summary of solubility concepts used in existing tools. The key difference lies in the perception of the insoluble class. The tools predicting "soluble expression" assume that insoluble proteins were either not expressed or were expressed in the insoluble form. The tools predicting classical "solubility" assume that insoluble proteins were successfully expressed before the solubility was determined.

Method	**Predicted property**	**Expression system**	**Comment**
SoluProt	soluble expression	*E. coli*	Based on curated TargetTrack data.
PROSO II	soluble expression	mixed	Based on PepcDB data (predecessor of TargetTrack).
SWI	solubility	*E. coli*	Based on PSI:Biology data. All proteins in the dataset were successfully expressed.
CamSol	solubility	mixed	Based on literature data.
ESPRESSO	solubility	*E. coli*	Based on Hirose dataset (Hirose *et al.*, 2011).
rWH	solubility	*E. coli*	Based on literature data.
DeepSol	soluble expression	mixed	Based on TargetTrack data.
Protein-Sol	solubility	cell-free	Based on eSOL data.
SOLpro	solubility	mixed	Based on PDB, Swiss-Prot and TargetTrack data. Proteins marked as insoluble were required to reach at least "cloned" or "expressed" states.
SKADE	soluble expression	mixed	Based on TargetTrack data.
ccSOL omics	soluble expression	mixed	Based on TargetTrack data.
RPSP	solubility	*E. coli*	Based on literature data.

Table S2. TargetTrack experiment states signifying soluble expression. The list was compiled by the authors of PROSO II (Smialowski *et al.*, 2012).

Experiment states
soluble, purified, crystallized, hsqc, structure, in pdb, native diffraction-data, NMR assigned, phasing diffraction-data, diffraction, in bmrb, nmr structure, crystal structure, diffraction-quality crystals

Table S3. Specific keywords signifying expression in *E. coli*.

Specific keywords
BL21, DE3, rosetta, xl10, DH10B, CodonPlus, RIPL, RIL, DB3.1, DB3, arctic, origami

Table S4. Protocols identified by generic phrases and manually checked to signify expression in *E.coli*.

Protocol id
NYSGXRC-SGX_MOLBIO_TOPO_TRANSFORM JCSG-E_Ecoli_GNF_1 CSGID-NU_SelMet_expression CSGID-NU_native_expression MPP-LP.4341 MCSG-NU_default_expression NYSGXRC-SGX_FERM_ECOLI_LB MPP-LP.4813 SSGCID-33 NYSGXRC-SGX_FERM_ECOLI_M9 CSGID-NU_default_expression SSGCID-2 SSGCID-31 SSGCID-1 CESG-MAXWELL 16 EXPRESSION TESTING (R D) v.1.0.0 MPP-LP.4814 SSGCID-128 EFI-SeMET expression in HY Media-PSI2 SGX-SGX_FERM_ECOLI_LB_CFTR SGX-SGX_MOLBIO_EXPR_SOL_CFTR

Table S5. The number of sequences retained in each dataset construction step. The higher number of soluble sequences in comparison to insoluble sequences in the training set can be explained by the lack of stop status annotation in the TargetTrack database. Therefore, it is generally harder to reliably extract insoluble sequences from the TargetTrack database.

Construction step	**Training set**	**Soluble**	**Insoluble**	**Test set**	**Soluble**	**Insoluble**
Input	335,771[T]	-	-	9,703[R]	-	-
Pre-processing and solubility assignment	114,648[R]	-	-	-	-	-
Expression system detection	82,362[R]	-	-	-	-	-
Redundancy removal	54,969	40,905	14,064	9,423	5,718	3,705
Removal of short sequences and sequences with unknown residues	54,962	40,904	14,058	9,420	5,715	3,705
Removal of transmembrane proteins	51,380	38,633	12,747	8,769	5,421	3,348
Removal of insoluble sequences with available PDB structure	51,360	38,633	12,727	8,754	5,421	3,333
Overlap removal[a]	-	-	-	6,398	3.928	2,470
Clustering to 25% identity	22,169	16,422	5,747	3,545	1,990	1,555
Class and length balancing	**11,436**	5,718	5,718	**3,100**	1,550	1,550

[T] The number of targets in the TargetTrack database. [R] The number of extracted sequence records – possibly more than one record for a sequence. Without any superscript – the number of unique protein sequences.

[a] Test set sequences sharing >25% sequence identity to any training set sequence were removed. The input for this step was the final training set of 11,436 sequences to minimize the reduction of the test set.

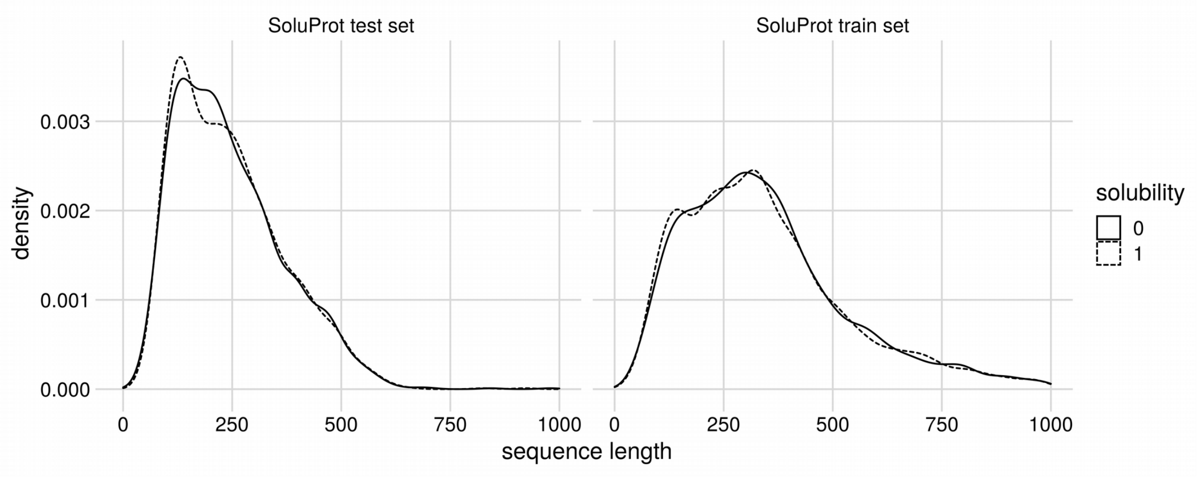


Figure S1. Sequence length distribution of soluble and insoluble proteins in the SoluProt datasets. The x-axis is limited to the range of 0–1,000 amino acids to improve readability. The longest sequences in the test and training sets have 979 and 2,825 amino acids, respectively.

Table S6. Sequence physicochemical features. Most of the features were extracted using the Biopython package (Cock *et al.*, 2009).

Feature	**Description**
physico_chemical_fracnumcharge	Fraction of charged amino acids (R, K, D, E).
physico_chemical_kr_ratio	Ratio of K and R content.
physico_chemical_aa_helix	Fraction of helix amino acids (V, I, Y, F, W, L).
physico_chemical_aa_sheet	Fraction of sheet amino acids (E, M, A, L).
physico_chemical_aa_turn	Fraction of turn amino acids (N, P, G, S).
physico_chemical_molecular_weight	Molecular weight.
physico_chemical_avg_molecular_weight	Molecular weight normalized by the sequence length.
physico_chemical_aromaticity	Fraction of aromatic amino acids (Y, W, F)
physico_chemical_flexibility	Flexibility according to (Vihinen *et al.*, 1994)
physico_chemical_gravy	Grand average of hydropathy according to (Kyte and Doolittle, 1982)
physico_chemical_isoelectric_point	Isoelectric point using methods of Bjellqvist (Bjellqvist *et al.*, 1993, 1994)
physico_chemical_instability_index	Instability index according to (Guruprasad *et al.*, 1990)

Table S7. Sequence features and their importance in the final SoluProt model.

#	Feature	Importance	#	Feature	Importance
1	ecoli_usearch_identity_identity	14.54%	26	dimers_comb_EM	0.96%
2	physico_chemical_isoelectric_point	6.23%	27	monomers_F	0.94%
3	tmhmm_first_60	4.18%	28	dimers_comb_EN	0.92%
4	monomers_K	3.95%	29	dimers_comb_AV	0.89%
5	monomers_Q	3.48%	30	dimers_comb_DL	0.89%
6	physico_chemical_aa_helix	1.91%	31	dimers_comb_IS	0.87%
7	monomers_E	1.84%	32	dimers_comb_EE	0.86%
8	physico_chemical_molecular_weight	1.77%	33	dimers_comb_CG	0.85%
9	monomers_M	1.70%	34	dimers_comb_PQ	0.85%
10	dimers_comb_DK	1.57%	35	dimers_comb_LQ	0.83%
11	dimers_comb_RR	1.55%	36	dimers_comb_EH	0.82%
12	dimers_comb_EK	1.49%	37	dimers_comb_AQ	0.82%
13	monomers_Y	1.39%	38	monomers_H	0.82%
14	dimers_comb_AA	1.35%	39	dimers_comb_CI	0.79%
15	monomers_C	1.28%	40	dimers_comb_EL	0.79%
16	dimers_comb_GK	1.13%	41	dimers_comb_HT	0.78%
17	dimers_comb_DT	1.09%	42	dimers_comb_EI	0.77%
18	dimers_comb_LN	1.09%	43	dimers_comb_QV	0.76%
19	dimers_comb_FT	1.08%	44	dimers_comb_DE	0.75%
20	dimers_comb_AI	1.05%	45	dimers_comb_DM	0.74%
21	dimers_comb_DI	1.02%	46	dimers_comb_MV	0.74%
22	dimers_comb_AG	1.01%	47	dimers_comb_GL	0.74%
23	dimers_comb_LT	1.00%	48	monomers_W	0.73%
24	dimers_comb_MN	0.98%	49	dimers_comb_TY	0.72%
25	dimers_comb_AN	0.98%	50	physico_chemical_fracnumcharge	0.72%

#	Feature	Importance
51	dimers_comb_EV	0.70%
52	dimers_comb_SV	0.65%
53	dimers_comb_RW	0.65%
54	dimers_comb_QT	0.64%
55	dimers_comb_KQ	0.61%
56	dimers_comb_GV	0.61%
57	dimers_comb_KV	0.60%
58	dimers_comb_HL	0.59%
59	dimers_comb_GN	0.58%
60	dimers_comb_RS	0.57%
61	dimers_comb_GG	0.57%
62	dimers_comb_AC	0.56%
63	dimers_comb_IL	0.55%
64	dimers_comb_FL	0.55%
65	dimers_comb_AM	0.54%
66	dimers_comb_LL	0.54%
67	dimers_comb_FI	0.52%
68	dimers_comb_MW	0.51%
69	dimers_comb_DR	0.51%
70	dimers_comb_EF	0.50%
71	dimers_comb_CY	0.50%
72	dimers_comb_GH	0.49%
73	dimers_comb_EP	0.48%

#	Feature	Importance
74	dimers_comb_CS	0.48%
75	dimers_comb_CP	0.47%
76	dimers_comb_AK	0.47%
77	dimers_comb_IY	0.46%
78	dimers_comb_PW	0.45%
79	dimers_comb_VY	0.45%
80	dimers_comb_NY	0.43%
81	dimers_comb_GM	0.42%
82	dimers_comb_IT	0.41%
83	dimers_comb_FP	0.40%
84	dimers_comb_HK	0.38%
85	dimers_comb_FM	0.38%
86	dimers_comb_GT	0.36%
87	dimers_comb_KR	0.34%
88	dimers_comb_FH	0.31%
89	dimers_comb_MM	0.31%
90	dimers_comb_KM	0.29%
91	dimers_comb_MY	0.28%
92	dimers_comb_WW	0.26%
93	dimers_comb_CC	0.21%
94	dimers_comb_DW	0.19%
95	dimers_comb_HW	0.17%
96	tmhmm_pred_hel	0.06%

Table S8. Optimized hyperparameters of the Gradient Boosting classifier. In each stage, one or two parameters were optimized while the other parameters were left either at their final values from previous stages or at their default values if they had not been optimized previously. The parameters were first optimized using a large step size. Smaller steps were then used for refinement. The learning rate was lowered from the default value of 0.1 to 0.01 before optimizing the number of estimators. Parameters not mentioned here were left at their default values. The procedure is based on the *Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python* by Aarshay Jain[1].

Stage	**Parameter**	**Range**	**Step**	**Final value**
1	n_estimators	20-100	10	-[a]
2	max_depth	3-17	2, 1	6
	min_samples_split	100-1400	100, 50	1250
3	min_samples_leaf	1-160	10, 5	6
4	max_features	5-96	5	40
5	subsample	0.5-1	1/40	0.525
6	learning_rate	-[b]	-[b]	0.01
7	n_estimators	200-1800	200, 50	1500

[a] The parameter was optimized again in the 7th stage, after which its final value was determined; [b] The learning rate was set to a fixed value; The final set of parameters was as follows: criterion='friedman_mse', init=None, learning_rate=0.01, loss='deviance', max_depth=6, max_features=40, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=6, min_samples_split=1250, min_weight_fraction_leaf=0.0, n_estimators=1500, n_iter_no_change=None, presort='auto', random_state=9, subsample=0.525, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False.

Table S9. Class disagreements between available training sets and the SoluProt test set when applying different binarization thresholds.

Dataset	**FP1**	**FP2**	**FP3**	**FP4**	**FP5**	**FN1**	**FN2**	**FN3**	**FN4**	**FN5**	**E1**	**E2**	**E3**	**E4**	**E5**
PROSO II initial	50	56	202	405	535	514	381	306	199	140	564	437	508	604	675
DeepSol/ SKADE	67	74	202	354	451	360	262	209	138	100	427	336	411	492	551
SWI	53	108	184	285	384	20	18	12	8	4	73	126	196	293	388
SOLpro	39	40	48	83	110	143	127	82	46	33	182	167	130	129	143
SoluProt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FP – false positives, FN – false negatives, E – total number of errors (FP + FN). The numerical suffix denotes the binarization threshold used for the SoluProt test set. For example, a binarization threshold of 2 means that all sequences with solubility scores of 2 or above are considered soluble, and all others are considered insoluble.

[1] https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

References

Bjellqvist,B. *et al.* (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, **15**, 529–539.

Bjellqvist,B. *et al.* (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, **14**, 1023–1031.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Guruprasad,K. *et al.* (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng Des Sel*, **4**, 155–161.

Hirose,S. *et al.* (2011) 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*, **150**, 73–81.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105–132.

Smialowski,P. *et al.* (2012) PROSO II - a new method for protein solubility prediction. *FEBS J*, **279**, 2192–2200.

Vihinen,M. *et al.* (1994) Accuracy of protein flexibility predictions. *Prot Struct Funct Bioinf*, **19**, 141–149.

Appendix C

Computational design of stable and soluble biocatalysts

ACS Catalysis

Perspective

Cite This: ACS Catal. 2019, 9, 1033−1054

pubs.acs.org/acscatalysis

Computational Design of Stable and Soluble Biocatalysts

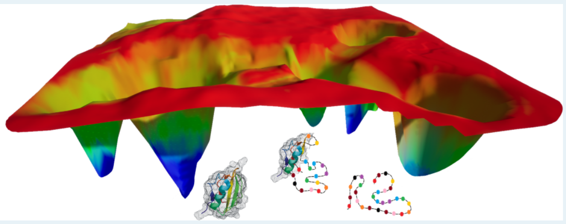
Milos Musil,[†,‡,§,∥] Hannes Konegger,[†,§,∥] Jiri Hon,[†,‡,§,∥] David Bednar,[†,§] and Jiri Damborsky*[,†,§]

[†] 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

[‡] IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic

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

Supporting Information



ABSTRACT: Natural enzymes are delicate biomolecules possessing only marginal thermodynamic stability. Poorly stable, misfolded, and aggregated proteins lead to huge economic losses in the biotechnology and biopharmaceutical industries. Consequently, there is a need to design optimized protein sequences that maximize stability, solubility, and activity over a wide range of temperatures and pH values in buffers of different composition and in the presence of organic cosolvents. This has created great interest in using computational methods to enhance biocatalysts' robustness and solubility. Suitable methods include (i) energy calculations, (ii) machine learning, (iii) phylogenetic analyses, and (iv) combinations of these approaches. We have witnessed impressive progress in the design of stable enzymes over the last two decades, but predictions of protein solubility and expressibility are scarce. Stabilizing mutations can be predicted accurately using available force fields, and the number of sequences available for phylogenetic analyses is growing. In addition, complex computational workflows are being implemented in intuitive web tools, enhancing the quality of protein stability predictions. Conversely, solubility predictors are limited by the lack of robust and balanced experimental data, an inadequate understanding of fundamental principles of protein aggregation, and a dearth of structural information on folding intermediates. Here we summarize recent progress in the development of computational tools for predicting protein stability and solubility, critically assess their strengths and weaknesses, and identify apparent gaps in data and knowledge. We also present perspectives on the computational design of stable and soluble biocatalysts.

KEYWORDS: *aggregation, computational design, force field, expressibility, machine learning, phylogenetic analysis, enzyme stability, enzyme solubility*

1. INTRODUCTION

Nature has developed a remarkable diversity of biochemical reactions that are vital to the continuing evolution of living organisms and the preservation of life. Enzymes are the most prominent catalytic entities in living cells and are collectively capable of catalyzing a vast range of biochemical reactions. The advent of next-generation sequencing together with recent advances in bioinformatics and molecular and structural biology have granted ready access to these rich genetic resources, facilitating the identification of efficient biocatalysts for diverse applications.[1−4] Moreover, the field of protein engineering has matured to a level that allows tailoring of native enzymes for specific practical applications.[5] However, the redesign of an enzyme sequence often imposes unintended secondary effects, frequently reducing the solubility and stability of the target enzyme.[6−9] Strategies for mitigating or eliminating these negative effects include chaperone buffering,[10] chemical modification of the protein structure,[11,12] protein immobilization,[13] medium engineering,[13] the addition of fusion proteins,[14,15] and the introduction of stabilizing or solubilizing mutations by protein engineering.[16−18]

Of particular interest for a mutational strategy is "directed evolution", which refers to experimental methods that emulate natural evolution by coupling molecular diversity generation to a selection or screening process. However, the immensity of an enzyme's sequence space prohibits global evaluation of all possible mutational combinations,[19] frequently causing optimization trajectories to become stuck in evolutionary dead ends.[20,21] This restricts the scope for creating stable and soluble biocatalysts by directed evolution alone and calls for knowledge-guided approaches to navigate the mutational space.[22] Rational protein design strategies can dramatically reduce the experimental effort required for successful directed evolution by consolidating pre-existing information.[23] Semirational strategies that combine directed evolution with structural and sequence data to help identify mutational hotspots amenable to focused screening efforts have been particularly popular recently.[24−26]

Received: September 7, 2018
Revised: December 15, 2018
Published: December 18, 2018

DOI: 10.1021/acscatal.8b03613

Table 1. Selected Experimentally Validated Cases of Successful Computational Redesigns of Stable and Soluble Biocatalysts

Stable Biocatalysts										
enzyme ∣ UniProt ID	substrate	method[j]	mutant code	mutations[a]	wild-type T_m [°C]	ΔT_m [°C][b]	$t_{1/2}$[c]	specific activity[c]	k_{cat}/K_m[c]	ref
cutinase ∣ P52956	4-nitrophenyl butyrate	force field	variant 10	7 of 197	62.3	5.7	12.9× (60 °C)	0.64× (25 °C)	n.d.[d]	41
keratinase ∣ Q1EM64	keratin	machine learning	quadruple mutant	4 of 379	n.d.	n.d.	8.6× (60 °C)	n.d.	4.11× (40 °C)	42
adenylate kinase ∣ P16304	Mg/ATP, AMP	phylogeny (ASR)	ANC1	66 of 218	53.6	35.4	n.d.	n.d.	1.79× (25 °C)	43
β-lactamase ∣ P62593	benzylpenicillin	phylogeny (CD)	ALL-CON	122 of 262	55.0	23.6	n.d.	n.d.	0.03× (25 °C)	44
kemp eliminase ∣ Q06121	5-nitrobenzisoxazole	phylogeny (CD)[e]	R2−4/3D	9 of 247	72.0	10.0	n.d.	n.d.	11.46× (25 °C)	31
haloalkane dehalogenase ∣ P59336	1-iodohexane	hybrid[f]	DhaA115	11 of 294	49.0	24.6	200× (60 °C)	0.31× (37 °C)	2.77× (37 °C)	45
halohydrin dehalogenase ∣ Q93D82	*rac*-*p*-nitro-2-bromo-1-phenylethanol	hybrid[g]	HheC-H12	13 of 253	57.0	25.5	n.d.	n.d.	0.88× (30 °C)	9

Soluble Biocatalysts										
enzyme ∣ UniProt ID	substrate	method[j]	mutant code	mutations[a]	wild-type T_m [°C]	ΔT_m [°C][b]	expr. yield[c]	specific activity[c]	expr. host	ref
haloalkane dehalogenase ∣ P59337	1,2-dibromoethane	phylogeny (ASR)	AncHLD2	69 of 317	53.6	21.9	4.8× (20 °C)	1.86× (37 °C)	*E. coli*	46
α-galactosidase ∣ P06280	*α*-D-galactose	hybrid[h]	A348R/A368P/S405L	3 of 397	n.d.	n.d.	1.4× (37 °C)	2.00× (37 °C)	*H. gartleri*	18
acetylcholinesterase ∣ P22303	acetylcholine	hybrid[i]	dAChE4	51 of 542	44.0	18.3	2000× (20 °C)	0.89× (25 °C)	*E. coli*	47

[a] Number of introduced mutations and total number of residues. [b] ΔT_m value of the mutant with respect to the wild-type enzyme. [c] Fold change in the specified property of the mutant relative to the wild-type enzyme. The temperature at which the given property was measured is given in parentheses. [d] n.d.: not determined. [e] Spiked Consensus Design, Directed Evolution. [f] FireProt: Rosetta, FoldX, Consensus Design. [g] FRESCO: Rosetta, FoldX, Disufide Bonds, MD. [h] SOLUBIS: TANGO, FoldX. [i] PROSS: Consensus Design, Rosetta. [j] CD - Consensus Design, ASR - Ancestral Sequence Reconstruction.

Table 2. Advantages and Disadvantages of Methods for the Computational Design of Stable and Soluble Biocatalysts

method	advantages	disadvantages
energy calculations	• granularity of predictions can be adjusted via different force fields	• high computational cost of accurate methods
	• web servers make predictions accessible to inexperienced users	• dependence on high-resolution structures
	• ever-growing structural databases together with advances in homology modeling and molecular threading	• trade-offs between stability and activity
	• high accuracy for the prediction of single-point mutations	• predicted stable mutants may not be expressible
		• epistatic effects are not well resolved
machine learning	• very rapid predictions	• lack of balanced high-quality experimental data
	• easy to implement and use	• limited accuracy of current models
	• wide applicability of features	• risk of overtraining
	• no need to understand all dependencies	
	• previously unknown patterns can be discovered	
phylogenetics[a]	• rich abundance of sequence data	• selection of relevant sequences is nontrivial
	• structures not needed for predictions	• profound understanding of the gene family is required
	• web servers available for certain tasks	• CD: epistatic effects are not considered
	• CD: simple and fast	• ASR: small data set size due to computational costs
	• CD: several filters are available to enhance prediction accuracies	• ASR: requires technical skills and experience
	• ASR: prediction of highly thermostable variants is achievable	
	• ASR: sequences of extremophilic proteins are not required	
	• ASR: sequence context and epistasis are maintained	

[a] CD, consensus design; ASR, ancestral sequence reconstruction.

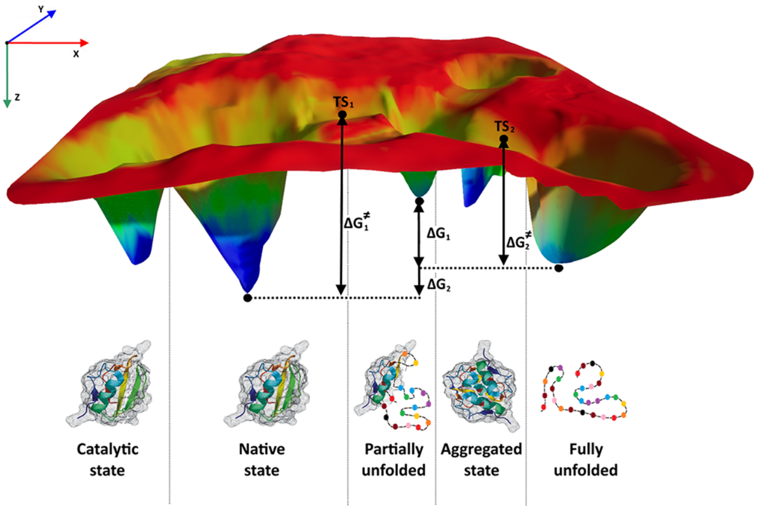


Figure 1. Simplified energy landscape with characteristic conformational states accessible from the native-state ensemble of a folded enzyme. Each point on the plane defined by the X axis and Y axis resembles a different conformation of the enzyme. The corresponding value on the Z axis is the free energy of folding, which has been color-coded to depict the spectrum from less probable high-energy states (red) to more probable low-energy states (blue). The catalytic state is readily accessible from the native-state ensemble but clearly separated by a free energy barrier. Catalysis based on a conformational selection model is assumed, which requires a distinct set of conformations prior to substrate binding and catalysis.[48] A reversible transition from the native state to a partially unfolded state via TS_1 is characterized by the free energy difference of folding ΔG_1 and its free energy barrier $\Delta G_1^{\ddagger}$. The partially unfolded state can also constitute the starting point for an irreversible unfolding transition via TS_2, leading to the fully unfolded state. Another irreversible pathway emanating from the partially unfolded state leads to an aggregated state, which is often characterized by the interactions of several biomolecules. ΔG_1 and ΔG_2 relate to thermodynamic stability, while $\Delta G_1^{\ddagger}$ and $\Delta G_2^{\ddagger}$ relate to kinetic stability.

This Perspective provides a thorough overview of contemporary data sets and computational protein redesign tools for enhancing enzyme stability or solubility. Preservation of enzymatic activity is of paramount importance in all protein engineering projects.[21,27] However, highly active and stable catalysts are evolutionarily disfavored because they could disrupt the host organism's homeostatic balance[28] or interfere with the cell's complicated metabolic regulatory networks.[29,30] Accordingly, several studies have indicated that most natural enzymes operate in a suboptimal regime,[21,28] leaving

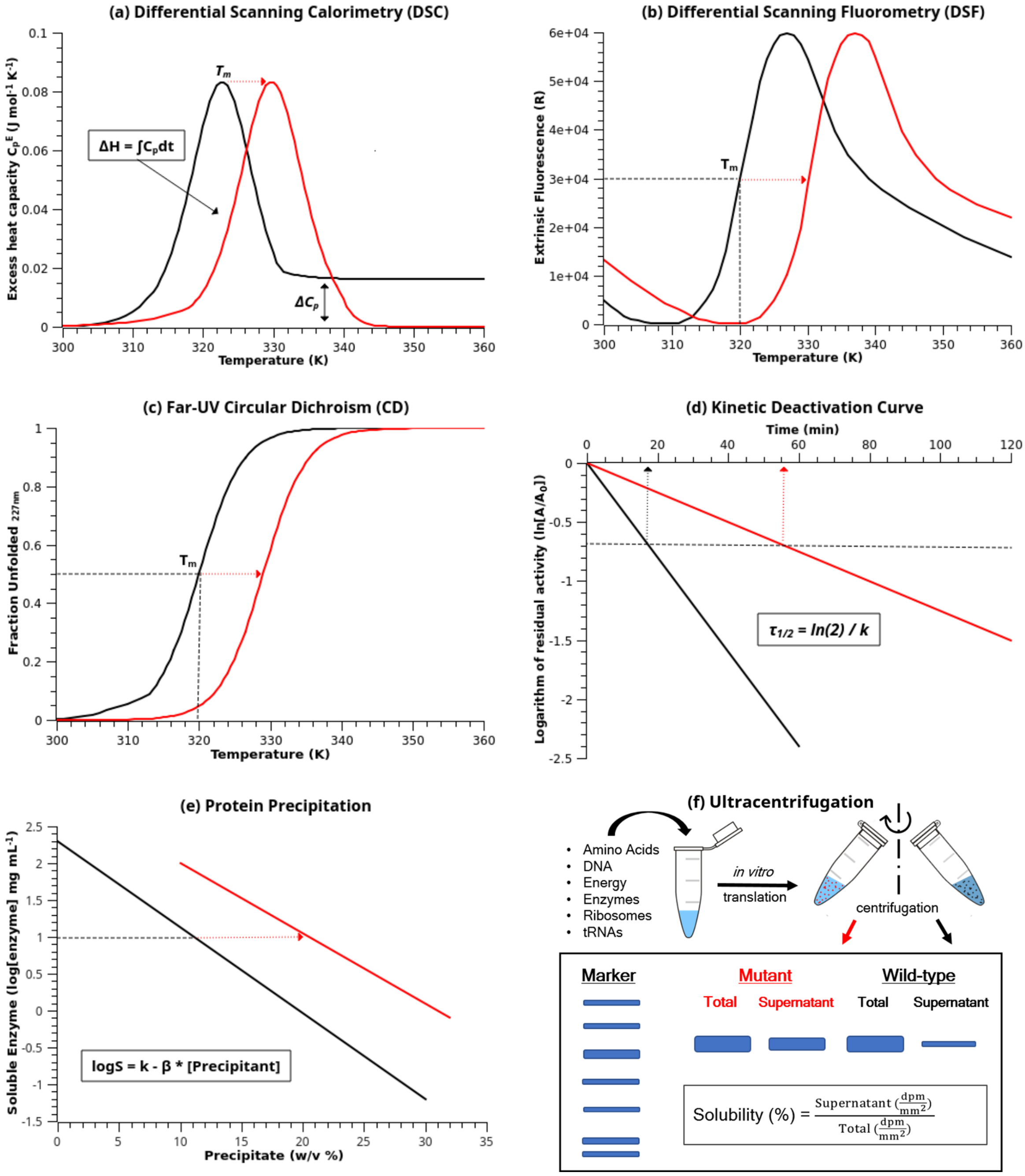


Figure 2. Representative experimental methods to quantify (a−d) protein stability and (e, f) solubility. Curves for a hypothetical wild-type enzyme (black) and an improved variant exhibiting higher stability or solubility (red) are shown. (a) Differential scanning calorimetry (DSC) curve. T_m is the midpoint of the transition, ΔC_p is the difference between the pre- and post-transition baselines, and ΔH is the area under the curve between the pre- and post-transition baselines. (b) Differential scanning fluorimetry (DSF) curve. Fluorescent dyes progressively bind to exposed hydrophobic regions of unfolding proteins, and the fluorescence signal is detected at different temperatures. T_m corresponds to the midpoint value of the stability curve. (c) Far-UV circular dichroism (CD) curve. Following the change of molar ellipticity at a specific wavelength over a wider temperature range monitors the change in secondary structure of an unfolding protein. The midpoint of the sigmoid curve is related to T_m of the protein. (d) Kinetic deactivation curve. For first-order deactivations, a plot of ln(activity) vs time yields a straight line with a slope of $-k$. The half-life can be calculated using the equation $\tau_{1/2} = \ln(2)/k$ and hence corresponds to the point ($\tau_{1/2}$, −0.69) on the fitted line. (e) Protein precipitation experiment. The addition of a precipitant is negatively correlated with the solubility of the folded protein. The parameter β is protein-specific and characterizes the dependence of the solubility on the precipitant concentration. (f) Record from ultracentrifugation. In vitro translation followed by ultracentrifugation allows quantification of protein solubility independent of the proteostatic network of a living cell (the PURE system). The solubility percentage is calculated as the ratio of protein in the supernatant to the total protein measured by autoradiography.[60] Adapted with permission from ref 37. Copyright 2007 Elsevier.

considerable room for further optimization (Table 1). Unfortunately, activity enhancements often come at the cost of reduced enzyme stability. The protein redesign tools presented here offer ways to avoid this trade-off and also to solubilize the polypeptides, facilitating the purposeful adaptation of natural enzymes.[31] Here we outline the theoretical frameworks of methods commonly used to analyze protein stability and solubility. We also critically review the data sets and software tools available for predictive purposes. This Perspective strives to evaluate the tools from the perspective of users, who are typically interested in accuracy, reliability, user-friendliness, and the strengths and weaknesses of the underlying methods (Table 2). We also present a personal perspective on existing gaps in knowledge and propose possible directions for future development.

2. EXPERIMENTAL FRAMEWORK TO DETERMINE PROTEIN STABILITY AND SOLUBILITY

2.1. Experimental Determination of Protein Stability. Globular proteins are known to be marginally stable, with free energy differences between the folded and unfolded states (Figure 1) being as low as 5 kcal/mol.[32] Two key concepts in the analysis of protein stability are thermodynamic and kinetic stability.[30,33−35] Thermodynamic stability can be defined on the basis of equilibrium thermodynamics as the Gibbs free energy difference of folding (ΔG). Exact quantification of absolute ΔG values is difficult,[36] so most stability predictors and experimental procedures determine the relative change in free energy ($\Delta\Delta G$) upon mutation. A commonly used experimental quantity related to $\Delta\Delta G$ is the change in melting temperature (ΔT_m). The melting temperature, T_m, is defined as the temperature at which half of the sample is in the unfolded state, and it can be determined using biophysical techniques (Figure 2) such as circular dichroism spectroscopy (CD), fluorescence spectroscopy (FS), dynamic light scattering (DLS), differential scanning microcalorimetry (DSC), or differential scanning fluorimetry (DSF).[37] The chemical equivalent of T_m is the half-concentration ($C_{1/2}$), i.e., the concentration of denaturant at which half the sample exists in the unfolded state. Kinetic stability, on the other hand, is a time-dependent property that is quantified by the height of the free energy barrier of unfolding ($\Delta G^\ddagger$) separating distinct folding states (Figure 1). Predicting kinetic stability is challenging,[38] and experimentally determined biological half-lives ($t_{1/2}$) are preferred to theoretical estimates (Figure 2). The kinetic stability is a key determinant of an enzyme's functional competence[30] because it is related to the rate at which the protein's structure is irreversibly altered by proteolysis or aggregation.[29,39,40]

2.2. Experimental Determination of Protein Solubility. Protein solubility is a thermodynamic parameter defined as the concentration of folded protein in a saturated solution that is in equilibrium with a crystalline or amorphous solid phase under given conditions.[49] Two methods can be used to estimate protein solubility in aqueous solutions in vitro: (i) adding lyophilized protein to the solvent and (ii) concentrating a protein solution by ultrafiltration and then estimating the protein fractions in the supernatant and the pellet. Both methods require that the concentration of protein in solution is increased until saturation is reached, which can be difficult to achieve.[49] The difficulties of measuring protein solubility can be alleviated by adding an agent—a precipitant—to reduce the protein's solubility. Precipitants may be salts, organic solvents, or long-chain polymers.

The term solubility can also be applied to the in vivo observable that describes protein expression quantitatively (expression yield) or qualitatively (soluble/insoluble). Besides the previously given definition of solubility, these two observables critically depend on the expressibility of a given enzyme inside the cell.[50,51] As a polypeptide is synthesized in the ribosome, the emerging chain enters the cell's highly regulated proteostasis network,[29,35,52] which assists the enzyme to attain its native-state structure. Protein folding does not rely on the random scanning of all accessible conformational states but follows a deterministic folding pathway[53,54] or multiple folding pathways.[55,56] Changes in the protein sequence can perturb such folding pathways, frequently diminishing the expressibility and solubility of an enzyme with a negative impact on its aggregation propensity or the formation of inclusion bodies.[8,9,57,58] One high-throughput in vivo experimental screening assay to test for properly folded enzyme variants is the Split-GFP system.[59] Besides the calculation of the expression yields via the Bradford method and the quantification of mRNA levels of the cells, the PURE system[60] might be a valuable experimental platform to investigate determinants of protein solubility and folding under in vitro conditions (Figure 2).

3. THEORETICAL FRAMEWORK FOR THE DESIGN OF ROBUST PROTEINS

3.1. Principles of Methods Based on Energy Calculations. In silico design of protein stability based on energy calculations has taken a long way from fairly simple[61,62] to more accurate and versatile methods, facilitating reliable high-throughput predictions of thermodynamically and kinetically stable enzymes.[41,63] A force field is a collection of bonded and nonbonded interaction terms[64,65] that are related by a set of equations that can be used to estimate the potential energy of a molecular system.[66] For stability predictions, such potential energy functions can be applied to a protein's structure to assess the energetic changes caused by the mutations. The most accurate but also the most computationally expensive methods are free energy methods, which rely on molecular dynamics (MD) or Metropolis Monte Carlo simulations. Free energy perturbation has proven to be a potent and rigorous alchemical approach that generates the most meaningful stability predictions, but only for a limited number of mutations.[67] Less accurate but considerably more performant are end-point methods such as molecular mechanics generalized Born[68] or linear interaction energy.[69] These free energy methods require a high level of technical expertise and access to supercomputing facilities, which can be challenging for experimental groups. Over the last 20 years, simpler and simulation-independent stability predictors have been developed. A subdivision into three categories has been proposed, namely, (i) statistical effective energy functions (SEEFs), (ii) empirical effective energy functions (EEEFs), and (iii) physical effective energy functions (PEEFs).[70,71]

SEEFs are fast and can predict changes in stability over the entire sequence space of an average-sized enzyme in a matter of seconds.[72,73] They are derived from curated data sets of folded protein structures, which are projected into a number of stability descriptors. An effective potential can be extracted for every descriptor distribution, and these can be combined to create an overall energy function.[72,74] SEEFs do not explicitly model physical molecular interactions, and the exact physical nature of statistical potentials remains obscure.[71] Consequently, overlapping and double counting of terms relating to the same causative interactions should be avoided.[70] EEEFs include both physical and statistical terms, which are carefully weighted and parametrized to match experimental data.[70,71] The thermodynamic data used in their derivation typically originate from mutational experiments conducted under standard conditions, which can be obtained from databases such as ProTherm.[75−77] EEEFs provide a reasonable compromise between computational cost and accuracy of the free energy function.[78] A major drawback of EEEFs and SEEFs is that their applicability is restricted to the environmental conditions under which the experimental data used for parametrization were acquired.[79,80] PEEFs are closely related to classical molecular mechanics force fields[81,82] and allow a fundamental analysis of molecular interactions.[66] PEEFs have more complex mathematical formalisms[71] and higher computational costs than EEEFs.[70] However, they are versatile, accurate, and capable of predicting behavior of the enzymes under nonstandard conditions, for instance at elevated temperature, nonphysiological pH, or nonstandard salinity.[83]

The accuracies of stability predictors based on such energy functions are still suboptimal[77,79,84−86] because of (i) imbalances in the force fields,[87,88] (ii) insufficient conformational sampling,[85,88] (iii) the occurrence of insoluble species,[8,9] and (iv) intrinsic problems with existing data sets (Table 2). The concept of free energy change upon mutation ($\Delta\Delta G$) was introduced for a fundamental analysis of the causative factors leading to these deficits. The computation of $\Delta\Delta G$ is based on a thermodynamic cycle (Figure 3), which requires modeling of the folded states of both the wild type and the mutant as well as their unfolded states.[36,67] Contemporary force fields describe enthalpic interactions reasonably well, although they are known to overestimate hydrophobicity and tend to favor nonpolar substitutions.[6,9,89] EEEFs and PEEFs generally underestimate the stability of buried polar residues because they overestimate the energetic cost of unsatisfied salt bridges and hydrogen bonds in the protein core.[58,90,91] The estimation of both conformational and solvent-related entropy is imprecise[9,92] because of the necessity of using computationally less expensive terms.[83] The inability of force field methods to account for entropy-driven contributions can be mitigated by using hybrid methods that incorporate complementary evolution-based approaches.[45,47,92,93] Moreover, most stability predictors have been parametrized using single-point-mutation data sets, resulting in higher prediction errors upon application to multiple-point mutants.[69,94] Whenever epistatic effects[20] are present between two or more individual mutations, force field predictions deviate from experimental results.

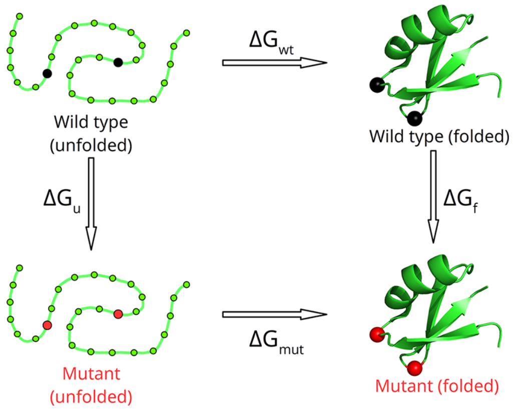


Figure 3. Thermodynamic cycle used to compute the free energy change upon mutation ($\Delta\Delta G$). $\Delta\Delta G$ is calculated according to the formula $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt} = \Delta G_f - \Delta G_u$. For better illustration, the hypothetical folded and unfolded states of the wild type and a two-point mutant are shown. The respective substitution sites have been color-coded in black (wild type) and red (mutant). Adapted with permission from ref 69. Copyright 2012 Wiley.

This shortcoming can be attributed to insufficient conformational sampling of the mutant's folded state, particularly when the introduced mutations induce large-scale backbone movements.[95] Tools based on EEEFs or PEEFs often apply rotamer libraries to fixed protein backbones, thereby reducing computational costs while providing comparable accuracies for the prediction of single-point mutations.[88] Multistate design[80,96] and flexible backbone sampling techniques[84,97−99] have partly alleviated the sampling problem for multiple-point substitutions by generating conformational ensembles and utilizing energetically more favorable conformations. Enzymes are intrinsically dynamic molecules and populate a high number of heterogeneous conformational substates[100] (Figure 1). Consequently, an adequate treatment of an enzyme's conformational plasticity[96,97] in the folded states of the wild type and mutant may be crucial for further advances of these methods. Structures obtained by X-ray crystallography do not essentially reflect the global energy minimum of the native state of an enzyme in its natural environment[101] and may therefore be nonideal starting points for stability predictions.[80,102] Besides the folded states, $\Delta\Delta G$ computations rely on sampling of the unfolded states of the wild type and the mutant. Simplifying and less realistic models (random coil or tetrapeptide) are frequently employed for explicit computations of the unfolded-state energies.[68,69] Generally, it is assumed that the free energy of the unfolded state does not change much upon mutation.[68,84]

The aforementioned explanations primarily relate to the prediction of thermodynamic stability. Not much work has been anticipated to predict kinetic stability, which can mostly be explained by the time-dependent nature[30] of this property and the time scales[103] assessable by energy-based methods. However, it is recognized that enhanced thermodynamic stability frequently goes hand in hand with enhanced kinetic stability.[41,45] One energy-based strategy to enhance the kinetic stability of an enzyme is to optimize solvent−solute interactions by introducing surface charges,[104] which can affect its expressibility.[105] The latter property may also be enhanced by computational linker design,[106] providing fusion enzymes with solubilizing protein tags.

3.2. Principles of Methods Based on Machine Learning.

Machine learning is a field of computer science that allows computational systems to be constructed without being explicitly programmed. Statistical techniques are used to analyze training data sets and recognize patterns that might be difficult to detect given the limitations of human knowledge and cognitive abilities. Machine learning systems can be trained with or without supervision. In supervised approaches, the system is given a set of example inputs and the corresponding desired outputs in the form of labels indicating the correct classification of each input. Supervised approaches are suitable for training predictive systems, while unsupervised approaches are more suitable for tasks involving data clustering. In recent years, machine learning has become one of the most common approaches for predicting the effects of mutations on protein stability[107−109] and solubility.[57,110] Machine learning does not require full understanding of the mechanistic principles underpinning the target function because they are modeled during the learning process. An important advantage of machine learning methods is that they are very flexible because any characteristic extracted from the data can be used as a feature if it improves the prediction accuracy, i.e., minimizes the prediction error (Table 2). Consequently, machine learning methods can reveal previously unrecognized patterns, relationships, and dependencies that are not considered in knowledge-based models. Moreover, machine learning is much less time-intensive than other methods because once a model has been constructed using the available data, predictions can be obtained almost instantaneously.

The reliability of machine learning approaches depends on the size and quality of the training data set. The weights representing the relative importance of the individual features and the relationships between them are based on experimental observations. Consequently, it is essential to use high-quality experimental data with high consistency when training and testing machine learning methods. The size and balance of the training data set must also be considered carefully. A modest data set with only a few hundred or a few thousand cases might be too small to identify useful descriptors during the learning process. Additionally, lower diversity of the training data set leads to a greater risk that the prediction tool will lose its ability to generalize. In such cases, the weights assigned to individual descriptors might be influenced by over-representation of some descriptors in the training data, while other descriptors that might be very important for general predictive ability could be omitted. Unbalanced training data sets with large differences in the numbers of cases representing individual categories could also lead to erroneous overestimations. For example, a training data set in which 80% of the mutations are destabilizing would allow the predictor to classify most mutations as destabilizing because of the prevalence of such mutations during the learning process. Methods like support vector machines and random forests are known to be more resistant to overfitting caused by unbalanced data sets,[111−113] while standard neural networks and decision trees are particularly sensitive to them. If the data set is too small to be balanced, the problem can be partially addressed by using cost-sensitive matrices,[114] which penalize the predictor more strictly for misclassifying mutations that are sparsely represented in the training data.

In parallel to the issue of the quality and availability of training data, one must address the problem of model validation. Ideally, the validation data set should be balanced and completely independent of the training set. In bioinformatics, it has become common to use *k*-fold cross-validation as a standard method for testing the performance of newly developed tools. This method entails randomly partitioning the original data set into *k* subsets. During the learning process, one of the *k* subsets is used for validation, while the remaining subsets are used as a training data set. This process is performed for each of the *k* subsets. The main reason for using cross-validation instead of splitting the data set into independent training and validation subsets is that the data set may be too small to support such splitting without harming the model's ability to learn the important predictive patterns. However, the combination of unbalanced data sets with the random aspect of *k*-fold cross-validation increases the risk of serious overestimation. Therefore, cross-validation is not a reliable method for measuring model accuracy when lower-quality data sets are used.[115] In conclusion, machine learning is a powerful approach that can reveal unknown interactions that are poorly defined in current force fields (Table 2). However, great care must be taken when constructing the training data set and during validation to avoid overfitting and overestimation of the results.

3.3. Principles of Methods Based on Phylogenetic Analysis.

The two most widely used phylogeny-based approaches for stability engineering are consensus design (CD) and ancestral sequence reconstruction (ASR). Continuous cycles of variation and selection have created an enormous diversity of modern-day enzyme sequences that can be processed using phylogenetic techniques (Table 2). Over the last two decades, the advent of next-generation sequencing methods has revolutionized life science but has also introduced new challenges arising from the vast amounts of sequence data that are now available.[116] When phylogenetic analyses are performed, this results in a selection problem: one must carefully decide which sequences to include in any analysis. Identifying suitable homologous sequences to a given target can be particularly challenging. Local alignment algorithms such as the Basic Local Alignment Search Tool (BLAST)[117] offer reasonable accuracy at minimal computational cost. More complex and computationally demanding signature-based and profile-based search algorithms[118−120] have further extended the boundaries of homology detection[121] beyond the twilight zone.[122] The twilight zone is an alignment-length-dependent pairwise sequence identity range above which homologous sequences can reliably be distinguished. When pairwise sequence identities fall within or below this specific range, a large number of false negative sequences will get incorporated into multiple sequence alignments (MSAs). Great care is needed in the construction of biologically relevant MSAs from distantly related homologues. The treatment of nontrivial evolutionary artifacts such as indels, translocations, and inversions within the coding sequence can profoundly affect the quality of an MSA.[123,124] Progressive, iterative, and consistency-based alignment algorithms[125] exclusively consider sequence data and often introduce topological inconsistencies that require manual correction.[126] These deficiencies have been alleviated by incorporating complementary structural or evolutionary information, but such approaches can be computationally demanding.[25,126,127]

CD starts from a set of homologous protein sequences. A genuine MSA is generated using a small number (between a dozen and a few hundred) of homologous sequences, which permits the computation of the frequency distribution of every amino acid position in the alignment.[128] A user-specified conservation threshold is then used to distinguish between ambiguous and conserved "consensus" positions. The core assumption of this method is that the most frequent amino acid at a given position is more likely to be stabilizing.[128−133] It has been noted that high levels of sequence diversity in the MSA can interfere with the preservation of catalytic activity in consensus enzymes; this problem can be particularly acute when the MSA incorporates both prokaryotic and eukaryotic sequences.[129,134] However, the assumption of statistical independence is central to CD. Excessively homogeneous

MSAs may violate this assumption, introducing phylogenetic bias that hinders the discovery of more thermostable proteins.[133] The proportions of neutral and destabilizing consensus mutations have been estimated to be 10 and 40%, respectively, among all characterized variants produced using consensus design to date, suggesting a need for a more focused selection of substitution sites.[128,132] To this end, Sullivan et al.[129] discarded mutations of residues with high statistical correlations to other positions in the MSA, thereby increasing the proportion of identified stabilizing mutations to 90%. Vazquez-Figueroa et al.[135] adopted a different approach, successfully using structural information (e.g., the distance between a possible mutation and the active site, secondary structure data, and the total number of intramolecular contacts) to complement traditional CD predictions. Another example of an effective structure-based CD approach involved the analysis of molecular fluctuations based on crystallographic B-factors.[136] Important drawbacks of CD are its inability to account for epistatic interactions[137,138] and an apparent phylogenetic bias in cases where the MSA is dominated by a few subfamilies.[130,139]

ASR is a probabilistic method for inferring primordial enzymes and ancestral mutations, which have proven to be very effective for thermostability engineering.[43,44,46,140] ASR explores the deep evolutionary history of homologous sequences to reassemble a gene's evolutionary trajectory.[138,141] As a starting point, a phylogenetic gene tree can be inferred from a manually curated MSA and a suitable evolutionary model using either the maximum-likelihood method[142,143] or Bayesian inference.[144] In the simplest case, such statistical inference methods derive parameters from the given MSA for the selected empirical evolutionary model, which defines the underlying amino acid substitution process. Once the gene phylogeny has been established, ancestral sequences corresponding to specific nodes of the tree can be computed, synthesized, overexpressed, and characterized in vitro. In addition to the difficulty of identifying and aligning legitimate sequences,[124] a major challenge encountered in ASR is the computation of a plausible phylogenetic tree that adequately explains the evolutionary relationships of the given sequences. Homogenous evolutionary models assume that amino acid substitutions are homogeneously distributed over time and among sites and are therefore heavily oversimplified models of evolution.[145] Maximum-likelihood methods have been shown to systematically overestimate the thermodynamic stability of deeper ancestors,[140,146] so Bayesian inference methods have been recommended as alternatives to account for this bias. However, Bayesian inference computes ancestral sequences with considerably lower posterior probabilities, sometimes leading to the loss of the biological function.[147] It is not entirely clear why ASR is successful at identifying sequences with improved thermostability.[141] One hypothesis states that its success is an artifact of the ancestral inference methods and resembles a possible bias toward stabilizing consensus sequences.[140,146] Another plausible explanation is based on the thermophilic origin of primordial life.[148,149] Regardless of the reasons for its effectiveness, ASR is clearly a very robust and efficient method for identifying enzyme sequences with high thermodynamic stability and elevated expression yields (Table 2). Furthermore, increases in kinetic stability resulting in higher $\tau_{1/2}$ have frequently been reported for ancestral enzymes in comparison with their extant forms.[140,150] The sequence context is maintained in the resurrected ancestral enzymes, enabling the conservation of historic mutations causing functionally important epistatic effects.[20,137,138] The fundamental drawbacks of ASR are that users must have considerable methodological skill and a good level of knowledge about the targeted gene family.

4. DATA SETS AND SOFTWARE TOOLS FOR DESIGNING STABLE PROTEINS

4.1. Data Sets for Protein Stability.

The accuracy and reliability of computational methods depends strongly on the size, structure, and quality of the chosen training and validation data sets. The primary source of validation data for protein stability is the ProTherm database.[75] ProTherm is the most extensive freely available database of thermodynamic parameters such as $\Delta\Delta G$, ΔT_m, and ΔC_p. It currently contains almost 26 000 entries representing both single- and multiple-point mutants of 740 unique proteins. Although ProTherm is the most common source of stability data, it suffers from high redundancy and serious inconsistencies. Particularly troubling are differences in the pH values at which the thermodynamic parameters were determined, missing values, redundancies, and strikingly even disagreements about the signs of $\Delta\Delta G$ values. ProTherm also neglects the existence of intermediate states.[57,107] To overcome the problems of the ProTherm database, the data must be filtered and manually repaired to construct a reliable data set.

Several subsets of the ProTherm database have been developed (Table S1) and used widely to train and validate new prediction tools. The most popular is the freely available PopMuSiC data set,[151] which contains 2648 mutations extracted from the ProTherm database. The data set is unbalanced because only 568 of its mutations are classified as stabilizing or neutral, while 2080 are classified as destabilizing. Furthermore, 755 of its 2648 mutations have reported $\Delta\Delta G$ values in the interval $\langle -0.5, 0.5\rangle$. Mutations with such $\Delta\Delta G$ values cannot be considered either stabilizing or destabilizing because the average experimental error in $\Delta\Delta G$ measurements is 0.48 kcal/mol.[152] Additionally, the data extracted from ProTherm are insufficiently diverse: around 20% of the PopMuSiC data set comes from a single protein, and 10 proteins (of 131 represented in the data set) account for half of the available data. Inspection of the data reveals that mutations to more hydrophobic residues located on the surface of the protein tend to be stabilizing, whereas mutations that increase the hydrophilicity in the protein core are usually destabilizing. Consequently, most computational tools are likely to identify mutations that increase surface hydrophobicity as stabilizing even though such designs often fail because of poor protein solubility.[58]

Some predictive tools use alternative data sets derived from ProTherm or PopMuSiC for training and validation. The most common benchmarking data set utilized for independent tests is S350,[151] which contains 90 stabilizing and 260 destabilizing mutations in 67 unique proteins. However, this data set is still small for comprehensive evaluation and unbalanced. The recently published PoPMuSiC^{sym} data set[153] tries to address these issues, containing 342 mutations inserted into 15 wild-type proteins and their inverse mutations inserted into the mutant proteins. A comparative study conducted using this data set showed a bias of the existing tools (Table S2) toward destabilizing mutations, as they performed significantly worse on the set of inverse mutations. Because of the overlaps of the mutations in training and validation data sets, the results of the individual tools can be overestimated. Even the new derivatives of the ProTherm database do not solve the problems arising from the size and structure of the available data. Therefore, there is an urgent need for new experimental data, particularly on the side of stabilizing mutations. Moreover, it would be of immense help for the future development of predictive tools to proceed with the standardization of the stability data, e.g., a unified definition of $\Delta\Delta G$ as a subtraction of the ΔG values for the mutant and the wild type. FireProt DB, a new publicly available database collecting carefully curated protein stability data, is being established at https://loschmidt.chemi.muni.cz/fireprotdb/.

Until the new unbiased data sets arise, a regular accuracy measure considering only the number of correctly predicted mutations from the testing set is not suitable for validation of the predictive tools. For binary classification, the Matthews correlation coefficient (MCC) can be utilized, as it was designed as a balanced measure that is usable even for data sets with a significant difference in the sizes of individual classes.[113] Similarly, when binary predictions are utilized as a filtration step in the hybrid approaches, metrics like sensitivity, specificity, and precision might be useful. When numerical measures are considered, the linear correlation between the predicted and experimental values can be estimated with the use of the Pearson correlation coefficient (PCC) and the average error established as the root-mean-square error (RMSE). Finally, the bias of the computational tools can be estimated as the sum of $\Delta\Delta G$ for the direct and inverse mutations according to Thiltgen and Goldstein.[94] Critical evaluation of the existing tools using the S350 data set revealed that the PCC ranges from 0.29 to 0.81 with an average RMSE of about 1.3 kcal/mol (Table S5).

4.2. Software Tools for Predicting Protein Stability Based on Energy Calculations. Software tools relying on force field calculations are based on either modeling the physical bonds between atoms (PEEFs) or utilizing methods of mathematical statistics (SEEFs). Rosetta[88] is one of the most versatile software suites for macromolecular modeling and consists of several modules. Rosetta Design is a generally applicable module for protein design experiments that evaluates mutations and assigns them scores (in physically detached Rosetta energy units) reflecting their predicted stability. In its newest version, the Rosetta force field converts Rosetta energy units into well-interpretable $\Delta\Delta G$ values.[83] Furthermore, the stand-alone ddg_monomer module was built on top of Rosetta Design and is parametrized specifically for predicting $\Delta\Delta G$ values and protein stability. The Rosetta suite is also supplemented by a wide variety of usable force fields and protocols. The Eris software[154] is based on the Medusa force field and incorporates a side-chain packing algorithm and backbone relaxation method. A similar physical approach is adopted in the Concoord/Poisson−Boltzmann surface area (CC/PBSA) method,[155] which uses the GROMACS force field[156] to evaluate an ensemble of structures initially generated by the Concoord program.[157]

Unlike the previously mentioned methods, in which the values of the individual terms in the force field equation are evaluated by performing calculations based on Newtonian physics, some tools simply fit equations using values derived from the available data. One of the main representatives of this approach is PopMuSiC,[73] whose force field equation includes 13 physical and biochemical terms with values derived from databases of known protein structures. Similar approaches are used by other statistical and empirical tools, including FoldX[78] and Dmutant.[158] Another tool in this class is HotMuSiC,[159] which is based on PopMuSiC and was parametrized specifically for estimating ΔT_m, since the correlation coefficient between $\Delta\Delta G$ and ΔT_m is −0.7.[159] HotMuSiC makes predictions using five temperature-dependent potentials based exclusively on data extracted from mesostable and thermostable proteins.

While PEEFs provide generally more accurate predictions of the effect of mutations on protein stability, there is an apparent trade-off between predictive power and computational demands. In the majority of cases, SEEFs still perform fairly well compared with most machine learning methods and are orders of magnitude faster than PEEFs. Therefore, SEEFs seem to be an acceptable compromise between accuracy and time demands, especially when utilized as filters for prioritization of the mutations in hybrid workflows.

4.3. Software Tools for Predicting Protein Stability Based on Machine Learning. Machine learning methods do not require comprehensive knowledge of the physical forces governing protein structure; their predictions are based exclusively on the available data. The most popular machine learning tools are based on the support vector machines (e.g., EASE-MM,[107] MuStab,[108] I-Mutant,[160] and MuPro[161]) and random forest (e.g., ProMaya[162] and PROTS-RF[163]) methods, which are known to be comparatively resistant to overtraining even when used with unbalanced training data sets (Table S2). Neural networks are rarely used for protein stability engineering because of their high sensitivity to the quality and size of the training data set.

In recent years, several new machine learning approaches have been applied to diverse problems in the field of bioinformatics. Deep learning is used to predict the effects of mutations on human health in DANN[164] and to predict protein secondary structure in SSREDNs.[165] Unfortunately, like regular neural networks, deep learning methods are prone to overfitting because adding extra layers of abstraction increases their ability to model rare dependencies, resulting in a loss of generality. This shortcoming can be addressed by using regularization methods such as Ivakhnenko's unit pruning.[166,167] However, this does not eliminate problems arising from inadequate training data sets because deep learning has very stringent data requirements. Consequently, deep-learning-based tools such as TopologyNet[168] still have very limited applicability in predicting protein stability.

The robustness and accuracy of computational tools can be increased by combining several machine learning approaches into a single multiagent system, as in the case of MAESTRO.[169] In MAESTRO, neural networks are combined with support vector machines, multiple linear regression, and statistical potentials. The outputs of the individual methods are then averaged to provide users with a single consensus prediction. In such tools, machine learning can be used to train the arbiter that decides how to combine the outputs of the individual methods and their weights, balancing the relative strengths of each method when applied to the type of mutation under consideration. This approach is widely used in metapredictors.[58]

It is difficult to compare individual tools on the basis of the results presented in the publications where they were first reported because most of them were validated using different data sets. This can bias a tool's performance toward particular proteins or mutation types, causing its general prediction accuracy to be overestimated. Therefore, independent comparative studies are needed. The critical evaluations reported by Kellogg et al.,[88] Potapov et al.,[77] and Khan and Vihinen[170] revealed that methods based on PEEF calculations systematically outperform tools relying only on machine learning techniques or statistical potentials in independent tests. Furthermore, machine learning methods tend to be more biased,[153,171] and their reported accuracies are overestimated as a result of overtraining. The PCC upper bound for the most commonly used stabilization data sets is about 0.8, and the lower bound of the RMSE is 1 kcal/mol.[172] The applicability of machine learning methods will increase with the size and diversity of the available data in the future.

4.4. Software Tools for Predicting Protein Stability Based on Phylogenetics. Phylogeny-based methods do not require knowledge of high-resolution protein structures; they can be applied to any protein with a known amino acid sequence and a sufficiently high number of sequence homologues. However, although phylogeny-based methods often improve some protein characteristics, the influence of individual mutations manifested during evolution is uncertain. About 50% of all mutations identified by CD are stabilizing, but some may affect protein solubility rather than stability.[131] CD-based methods are therefore frequently utilized as filters during core calculations of hybrid workflows or as components of predictive tools for hotspot identification.

CD is available in several bioinformatics suits (e.g., EMBOSS,[173] 3DM,[25] VectorNTI,[174] and HotSpot Wizard[175]). Although there are no stand-alone tools for CD, there are several for ASR, some using maximum-likelihood methods (e.g., RAxML,[176] FastML,[177] and Ancestors[178]) and others using Bayesian inference (e.g., HandAlign[179] and MrBayes[180]). A major limitation of these methods is that most of the tools require users to upload their own MSA and phylogenetic tree. Constructing these input data is the most important and demanding step of the entire process. To obtain reliable predictions, the initial set of homologue sequences must be filtered to identify a reasonably sized subset of biologically relevant sequences. At present, sets of homologous sequences obtained using BLAST,[117] profile-based methods such as position-specific iterated BLAST,[118] or hidden Markov models[120,181] must be manually curated to ensure reliable ancestral reconstructions.

4.5. Software Tools for Predicting Protein Stability Based on Hybrid Approaches. Hybrid methods make predictions by combining information from several fundamentally different approaches. They offer greater robustness and reliability than individual tools, allowing multiple-point mutants to be designed while reducing the risk of combining mutations with antagonistic effects. Consequently, several research groups are focusing on hybrid methods in their efforts to improve the rational design of thermostable proteins.

The Framework for Rapid Enzyme Stabilization by Computational Libraries (FRESCO)[93] is available as a set of individual tools and scripts, and its use requires a good knowledge of bioinformatics. FRESCO initially selects a pool of potentially stabilizing mutations (FoldX or Rosetta energy cutoff of −5 kJ/mol) and also filters out all residues in close proximity (<10 Å) to active sites. Disulfide bridges are designed by dynamic disulfide discovery using snapshots from MD simulations and subsequently evaluated using the set of geometric criteria. An energy criterion for the maximal molecular mechanics energy of the disulfide bond was also adopted. Furthermore, very short MD simulations predict changes in backbone flexibility upon mutation to remove designs with unreasonable features that are expected to destabilize the protein. About a hundred of the single-point mutants are then subjected to experimental validation to select mutations to be included in the combined multiple-point mutant. Experimental validation of individual mutations greatly reduces the risk of false positives and maximizes the stabilization effect but requires a substantial investment of time and effort.

FireProt[45,89] combines energy- and evolution-based approaches in a fully automated process for designing thermostable multiple-point mutants (Figure 4). FireProt integrates 16 computational tools, utilizing both sequence and structural information in the prediction process. When the energy-based approach is applied, information extracted from the protein sequences (e.g., lists of conserved and correlated residues) is used to exclude potentially deleterious mutations, while structural information is used to obtain estimated ΔΔG values with both FoldX and Rosetta. The second approach is based on back-to-consensus analysis followed by energy filtration using FoldX. Finally, a distance-based graph algorithm is used to create a multiple-point mutant by selecting the most favorable mutually nonconflicting mutations from the pool of all potentially stabilizing mutations. A stand-alone version of FireProt[45] has been implemented as an intuitive web-based application,[89] making this complex modeling workflow accessible to a wide user community. The automation of the whole procedure eliminates the need to select, install, and

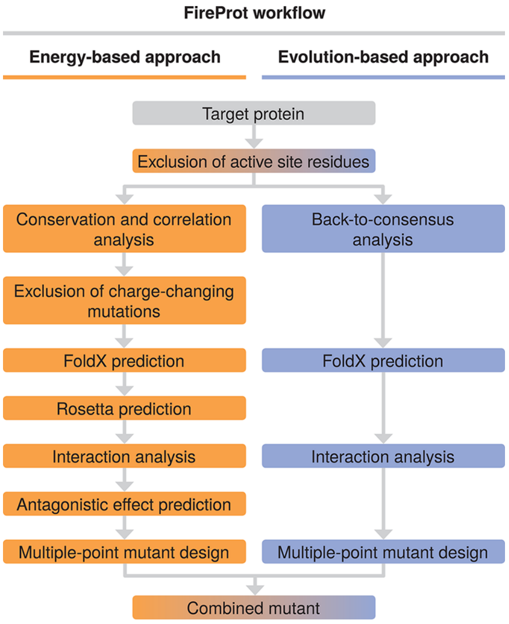

FireProt workflow
Energy-based approach
Evolution-based approach
Target protein
Exclusion of active site residues
Conservation and correlation analysis
Back-to-consensus analysis
Exclusion of charge-changing mutations
FoldX prediction
FoldX prediction
Rosetta prediction
Interaction analysis
Interaction analysis
Antagonistic effect prediction
Multiple-point mutant design
Multiple-point mutant design
Combined mutant

Figure 4. Workflow of the protein thermostabilization platform FireProt. The hybrid method combines evolutionary- and energy-based approaches and designs stable multiple-point mutants by fundamentally different methods.[45] The user is offered three different designs, two based solely on the energy- and evolution-based approaches and a third combining all of the identified mutations. FireProt has been made available as a fully automated and user-friendly web application[89] and is free of charge for academic users at http://loschmidt.chemi.muni.cz/fireprot.

evaluate tools, optimize their parameters, and interpret intermediate results.

Protein Repair One-Stop Shop (PROSS)[47] is another automated web-based protein stabilization platform. The PROSS workflow begins with a Rosetta design calculation in which the amino acids constituting the protein's active and binding sites are not eligible for mutation. A position-specific substitution matrix is analyzed to steer the design process away from amino acids that are rarely observed in the sequence homologues,[182] and Rosetta's computational mutation scanning tool[183] is used to scan the remaining pool of potential amino acid mutations. Finally, Rosetta's combinatorial sequence design tool is used to find an optimal combination of potentially stabilizing mutations, and an energy function is applied that favors amino acid identities on the basis of their frequency in the multiple-sequence alignment. This phylogeny-based biasing potential allows the designed variants to incorporate mutations found to be neutral or even slightly destabilizing in the Rosetta calculations,[35] which is desirable because some of these mutations might positively influence properties such as catalytic activity or protein solubility.

Hybrid methods represent a step forward in the prediction of protein stability because of their higher reliability at a decreased computational cost. These methods utilize evolution-based approaches as filters for removing potentially deleterious mutations in the conserved or correlated regions of the target protein. Furthermore, hybrid methods identify stabilizing mutations that would be missed by using only force field or phylogeny methods, as these two approaches are often complementary.[92] The increased robustness of the hybrid methods allows for a safer combination of single-point mutations into a multiple-point mutant. Hybrid methods can be further expanded by predictions of protein solubility or catalytic activity.

5. DATA SETS AND SOFTWARE TOOLS FOR THE DESIGN OF SOLUBLE PROTEINS

5.1. Protein Solubility Data Sets. Protein solubility, aggregation propensity, and expressibility are complex properties governed by several distinct biophysical and biological mechanisms. Progress in understanding these mechanisms depends on the availability of large, high-quality, diverse experimental data sets. In addition, the performance of prediction methods must be assessed with respect to the data used during their training. It is therefore important to recognize the strengths and limitations of the available experimental data sets on protein solubility and expressibility. To this end, this section presents a comprehensive review of the data sets available at the time of writing (Table S3).

5.1.1. Protein Solubility Data Sets Based on Full-Length Proteins. Data sources of this type contain information on the solubility of entire proteins produced in a specific expression system, either in vitro using a cell-free expression system or in vivo. Solubility can be determined by separating the liquid component of a sample by centrifugation or filtration and measuring the protein content in a solution, which is normalized by the protein content in the unseparated sample. The normalization removes the relationship between the solubility value and varying protein expression level. Alternatively, proteins may be simply classified as soluble or insoluble.

The Solubility Database of *E. coli* Proteins (eSOL)[60] contains experimentally measured solubilities for over 4000 *E. coli* proteins. The solubilities were determined by expressing the proteins using the PURE cell-free expression system[184] and using ultracentrifugation to measure their solubility as the ratio of the protein content in the supernatant to the total protein content of the sample. The limitations of eSOL are that only a moderate number of proteins are represented and that all of them originate from *E. coli*. In addition, in vitro cell-free expression systems cannot reproduce the post-transcriptional molecular processes that occur during protein expression in vivo. Interestingly, adding the three main cytosolic *E. coli* chaperones (TF, DnaKJE, and GroEL/GroES) to the in vitro cell-free expression system reduced the number of insoluble proteins from 788 to 24.[185]

TargetTrack,[186] formerly PepcDB or TargetDB, integrates vast amounts of information from the Protein Structure Initiative, a large-scale structure determination project. It contains data from over 900 000 protein crystallization trials using almost 300 000 unique protein sequences, which are termed targets. The database is not focused on solubility, but target proteins can be considered soluble if they reached a particular state in the experimental trial. We note that strictly speaking, this parameter reflects both the expressibility and the solubility of the target proteins. The major drawback of this database is the low quality of the annotations. No reason for failure is recorded for most of the unsuccessful crystallization attempts. Moreover, the experimental protocols are described in free text with no common structure. Therefore, it is difficult to automatically extract information about the expression systems. As a result, the application of strict rules to the target annotations dramatically reduces the number of usable records.

The Northeast Structural Consortium (NESG)[187] database is a subset of TargetTrack containing data on 9644 targets analyzed between 2001 and 2008. The NESG database contains explicit data on protein expression and solubility levels based on uniform protein production in *E. coli*. Two integer scores are recorded for each target, indicating the protein's level of expression and the recovery of the soluble fraction. The major drawback of this data set is that it was generated using outdated experimental methods; some of the targets could probably be solubilized using current techniques. Additionally, the database is too small to be used to train new machine learning algorithms. However, it can be used as a high-quality benchmark data set because its explicit experimental observations are more trustworthy than any other data in TargetTrack.

The Human Gene and Protein Database (HGPD)[188] contains expression and solubility measurements on over 9000 human proteins expressed in *E. coli*, a wheat-germ cell-free expression system, or *Brevibacillus*. The expression data were obtained using the Gateway system coupled with SDS-PAGE of C-terminal V5- or His-tagged proteins. Like the NESG data, these results originate from a uniform high-throughput protein production pipeline and thus constitute a consistent data set. Moreover, the HGPD provides information at the DNA level, so it includes codon composition data. Its major drawback is that it is focused exclusively on human proteins, so predictors constructed on the basis of its data will have an implicit bias toward human proteins.

AMYPdb[189] contains data on over 12 000 proteins belonging to amyloid precursor families as well as over 6000 generalized sequence patterns useful for assigning new sequences to poorly soluble amyloid precursor families. These data are derived from the literature and by UniProt and PROSITE mining, so they are useful only as training data and for concept verification; they are not suitable for performance validation. This database has not been updated since its release in 2008.

5.1.2. Protein Solubility Data Sets Based on Protein Fragments. Fragment databases often describe properties of short peptides and their tendency to aggregate when exposed to solvent. This tendency does not necessarily correlate with the peptide's behavior when it is incorporated into a larger globular protein, in which case it may be protected by the formation of a hydrophobic core. Therefore, great care is necessary when using these databases as a basis for solubility prediction.

AmylHex and AmylFrag[190] are literature-based collections of nearly 200 short peptide sequences known to form amyloid fibrils. The major flaws of this database are its strong over-representation (51%) of point variants of a single amyloidogenic hexapeptide (STVIIE) and its low content of data on longer protein fragments.

WALTZ-DB[191] integrates data obtained from the literature and by in-house experimental verification on over 1000 hexapeptides tested for amyloidogenicity. As such, it is a unique resource containing primary experimental data. Of the peptides represented in the data set, 22% were found to be amyloidogenic and 78% were found to be non-amyloidogenic.

AmyLoad[192] combines data collected from WALTZ-DB, AmylHex, AmylFrag, the AGGRESCAN and TANGO validation data sets, and manual reviews of over 90 publications. The data set contains information on almost 1500 amyloidogenic and non-amyloidogenic protein fragments that have been characterized experimentally or computationally. About 30% of the fragments are considered amyloidogenic.

The Human Protein Atlas (HPA)[193] contains data on over 16 000 protein epitope signature tags (PrESTs) that were produced using a uniform *E. coli* production pipeline. PrESTs are substantial fragments of human proteins ranging from 20 to 150 amino acids. Their expression and solubility were measured and are quantified using integer scores ranging from 0 to 5.

The Curated Protein Aggregation Database (CPAD)[194] is an integrated database that includes data on almost 1700 amyloidogenic protein fragments and aggregation changes upon mutation. The fragments represented in the database include peptides with known and unknown structures, almost 100 verified aggregation-prone regions, and over 2300 aggregation rate changes upon mutation. The database represents a unique resource for validating the effect of mutations on protein aggregation. Unfortunately, it is poorly structured, and the data are not easily downloadable in a machine-friendly format.

5.1.3. Protein Solubility Data Sets Based on Mutants. The existing data sets containing information on protein variants with measured effects on protein solubility are very small and were constructed ad hoc by the authors of prediction software on the basis of literature data. Three representatives of this small group of solubility data sources are OptSolMut,[195] CamSol,[17] and PON-Sol.[57] OptSolMut contains binary solubility data on 137 protein variants, and the amounts of positive and negative samples are nearly balanced. CamSol contains data on 56 protein variants, of which only three are classified as reducing solubility. The PON-Sol data set contains 443 protein variants, of which 222 reportedly have no effect on protein solubility.

5.2. Software Tools for Predicting Protein Solubility.

Unlike stability prediction tools, solubility prediction tools differ in their outputs rather than their fundamental operating principles. Almost all solubility prediction tools use some form of machine learning, ranging from simple statistical approaches to modern nonlinear methods such as support vector machines, random forests, or deep neural networks. The tools also use similar sets of features based on amino acid composition and physicochemical properties. Their outputs typically fall into one of three categories: (i) a single solubility score for the entire input sequence, (ii) a solubility profile with a unique score for each amino acid, or (iii) a score reflecting the effect of a specific mutation on solubility. All three outputs are expressed using arbitrary solubility scales with no physical meaning. The following section discusses the available predictive tools and their theoretical underpinnings and critically assesses their reliability (Table S4). Tools that predict single solubility scores for entire protein sequences are most useful for genomic projects because they can help prioritize protein sequences for laboratory production. Conversely, algorithms that provide quantitative scores over fixed-size sequence windows generate solubility profiles that can be used in the rational design of soluble proteins.

5.2.1. Software Tools for Protein Solubility Based on Primary Sequences. One of the first single-score solubility methods was the linear prediction model proposed by Wilkinson and Harrison,[196] which was later simplified by Davis and co-workers.[197] The revised model is surprisingly simple, using only two features (the approximate-charge average and turn-forming residue content) that both measure the relative abundance of specific amino acid types in the sequence. Despite its simplicity, the model can be useful for analyzing certain protein families. For example, it achieved a Spearman correlation coefficient of 0.54 and outperformed several newer tools in the same category (Table S4) when its predictions were compared to experimental data for 20 sequences closely related to a recently characterized haloalkane dehalogenase family.[4]

SOLpro,[198] PROSO II,[199] ccSOL omics,[200] and DeepSol[201] use the TargetTrack database as the source of training data. Consequently, although they use different features and machine learning models, they are quite similar to one another and have many shared strengths and weaknesses. Their most significant drawback is that they do not focus on any one expression system because it is hard to automatically extract expression system data from TargetTrack. Therefore, when validating these tools on a set of proteins expressed in a single expression system (e.g., *E. coli*), the observed prediction performance might differ significantly from that reported by the tools' creators. Published results suggest that DeepSol should have the highest prediction accuracy in general. However, this algorithm was created by using deep learning with a moderately sized training set and was validated against a data set representing protein families similar to those included in the training set. Moreover, although good performance is commonly claimed for tools based on TargetTrack, these claims have been strongly questioned.[199,201] In conclusion, the validation of these tools should be evaluated carefully, and further external validation using test sets independent of TargetTrack is needed. Unfortunately, the limitations of the TargetTrack database, from which solubility data can be extracted only via automated parsing, impose a strong performance limit on any tool that relies heavily on its data.

Periscope[202] attempts to predict soluble protein expression in the periplasm of *E. coli* rather than the cytosol. Although it was trained on a small data set, it was validated against an independent set of proteins and thus might be useful for predicting periplasmatic expression in *E. coli*.

ESPRESSO[203] estimates protein expression and solubility in both cell-free (wheat germ) and in vivo (*E. coli*) expression systems. The system has three unique aspects. First, it is based on measured expression and solubility levels of human proteins from the HGPD and thus may be useful for production of human proteins in either of the two relevant expression systems. Second, it offers two types of prediction: property-based and motif-based. The former type resembles the predictions offered by the other machine learning tools in this category. In contrast, motif-based predictions identify positive and negative solubility motifs extracted from the training data. For each negative motif, ESPRESSO suggests point mutations that should turn the negative motif into a positive one, so the tool can be used for the rational design of soluble proteins. Third, ESPRESSO also uses DNA-level features in its property-based method. However, direct verification of its reported performance is currently complicated because the original training and testing data are unavailable.

SoluProt[204] is one of the latest additions to the family of solubility predictors. Its training set is based on the TargetTrack database,[186] which was carefully filtered to keep only targets expressed in *E. coli*. The negative and positive samples were balanced and equalized with respect to protein length. The independent validation set was derived from the NESG data set.[187] The current version of the tool uses a predictor based on a random forest regression model that employs 36 sequence-based features, including amino acid content, predicted disorder, α-helix and β-sheet content, sequence identity to the Protein Data Bank (PDB), and several aggregated physicochemical properties. SoluProt currently achieves a prediction accuracy of 58.2%, which exceeds that of other currently available tools, and is under active development. An intuitive web interface to the tool will soon be made available to the community at https://loschmidt.chemi.muni.cz/soluprot/.

5.2.2. Software Tools for Predicting Protein Solubility Based on Sequence Profiles. A solubility profile is an abstract construct in which each residue of a given protein sequence is assigned a solubility score that contextually describes its relative contribution to the solubility of the protein as a whole. The solubility scores within a profile may represent aggregation rates or values on an arbitrary scale with no corresponding physical units. In either case, the highest scores represent solubility hotspots. Predictions based on such profiles must be interpreted with care because they rest on a hidden assumption: most profile-predicting methods are trained with data on short linear and unstructured peptides (Table S4), so there is an inherent assumption that the protein of interest is also at least partially unstructured. Therefore, these tools lack specificity when applied to natively folded globular proteins, in which many predicted low-solubility (or aggregation-prone) segments are stabilized by the interactions that maintain the protein's secondary and tertiary structure. If the protein's structure or a reasonable homology model is available, it is possible to compensate for these problems by applying structural corrections.

There are several profile-based tools, most of which share at least some concepts and/or training data sets. Zyggregator[205,206] uses a model fitted to the measured aggregation rates of nearly 100 variants of 15 proteins mined from the scientific literature. AGGRESCAN[207] is based on data from a single-codon saturation mutagenesis study of amyloid β 42 protein, in which aggregation rates were measured for 20 protein variants. Because both methods are based on very small data sets, the authors took care to bolster their credibility by applying the models in several case studies.

TANGO,[208] WALTZ,[209] and PASTA[210] predict amyloid plaque formation propensity on the basis of data for short experimentally characterized peptides (mostly hexapeptides). TANGO is the most famous of these tools and has been cited hundreds of times. However, the models used by the newer tools WALTZ and PASTA were inferred from larger experimental data sets, so they are claimed to outperform TANGO. A common concern is that the data sets of amyloidogenic peptides are unbalanced, containing too few non-amyloidogenic fragments (Table S3), which limits the generalizability of predictions obtained with these tools.

BETASCAN,[211] FoldAmyloid,[212] ZipperDB,[213] and ArchCandy[214] learn from experimentally determined structures of amyloidogenic proteins and apply the discovered general concepts at the sequence level. BETASCAN calculates likelihood scores for potential β-strands and strand pairs in sequences based on correlations observed in parallel β-sheets of experimental structures. FoldAmyloid uses the number of contacts per residue and statistics on hydrogen bonds in nearly 4000 PDB structures. In ZipperDB, the input protein is threaded onto a template cross-β spine structure, and the relative threading energy is used to predict amyloidogenicity. ArchCandy evaluates whether a protein segment can fold into β-arcade structures, which are often disease-related, and uses an empirical scoring function to evaluate interactions that disrupt β-arcade formation. These structure-based tools are expected to be inherently more specific than sensitive because structure-derived criteria tend to be relatively strict. When a high sensitivity is required and a structure is available, methods based on short peptides are expected to be more sensitive than structure-based alternatives. It is possible to compensate for false positives by checking the tool's output against known structures.

Because individual solubility prediction tools have different strengths and weaknesses, efforts have been made to create consensus-based methods that combine multiple tools to mitigate against the weaknesses of individual tools while preserving their strengths. The advantages of consensus methods have been proven both theoretically[215] and empirically.[216] Both AmylPred2[217] and MetAmyl[218] implement 11 individual methods, including AGGRESCAN, TANGO, and WALTZ. Although the primary publication on AmylPred2 claims superior performance to all of the individual methods, these results should be treated with care because the consensus threshold was validated using the entire data set chosen by the developers. Consequently, there was no independent validation set, and the claimed performance is very likely to be overestimated. MetAmyl uses a specially developed peptide set derived from the WALTZ data set to establish a logistic regression model that integrates the outputs of the individual tools. An evaluation using the AmylPred2 data set indicated that MetAmyl outperformed AmylPred2 despite having been optimized with a different data set.[218] This strongly suggests that MetAmyl performs better than AmylPred2 in general.

5.2.3. Software Tools for Protein Solubility Based on Mutations. While the profile-based tools discussed above can be used to design solubilizing mutations, the methods described in this section are tailored for this purpose and therefore are easier to use. Importantly, most of the methods discussed here require a protein structure as an additional input (Table S4).

OptSolMut[195] uses the concepts from computational geometry to define a scoring function reflecting the changes in solubility due to mutations. The scoring function was optimized using linear programming on the basis of a set of protein variants extracted from the literature. The reported 81% overall accuracy should be taken with care, as the training set was small and the model might not generalize well. In contrast to other tools in this section, OptSolMut is able to predict the effect of multiple-point mutations.

Several tools for predicting the effect of mutations on solubility have been developed from tools for predicting solubility profiles. For example, CamSol,[17] AGGRESCAN3D,[219] SolubiS,[220,221] and SODA[110] are based on the previously published profile-based methods Zyggregator, AGGRESCAN, TANGO, and PASTA, respectively. The workflows of these tools are all very similar: first a solubility profile is predicted, then a correction based on knowledge of the protein's structure is applied, and finally solubility hotspots are identified and specific mutations targeting low-solubility regions are suggested. CamSol, AGGRESCAN3D, and SODA use structural corrections to refine the predicted solubility profiles by averaging physicochemical properties over residues proximal in three-dimensional space or on the basis of solvent exposure of individual residues. SolubiS uses free energy calculations based on the FoldX force field to avoid potentially destabilizing mutations in aggregation-prone regions and can thus be classified as a hybrid method (Figure 5). CamSol and SODA can make predictions even without structural data. However, this necessarily eliminates the potential to exploit structure-based corrections and thus tends to reduce the prediction accuracy. The main issue with all of these tools is in the difficulty of validating their output. The data sets available for both training and testing are small, and they have only been validated using data for a small number of experimentally characterized protein variants.

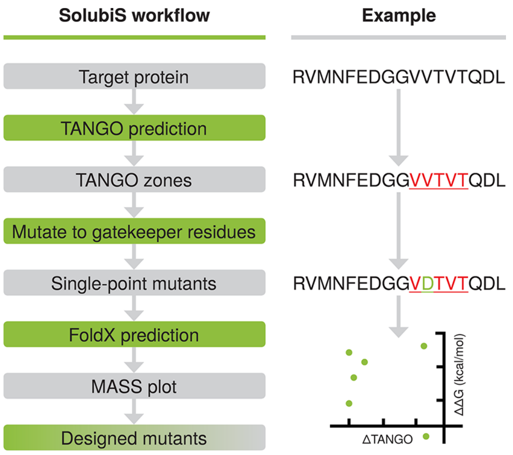


Figure 5. Workflow of the protein solubilization platform SolubiS. The platform uses free energy calculations performed with FoldX to avoid potentially destabilizing mutations in aggregation-prone regions identified by TANGO. The results are presented in form of a mutant aggregation and stability spectrum plot.[220] The web server is free of charge for academic users at http://solubis.switchlab.org/.

PON-Sol[57] uses a machine learning algorithm designed from scratch for solubility prediction of protein variants from protein sequences without structure-based corrections. The reported accuracy of this three-class classification method is 43%. The training data set was rather limited, representing a few tens of proteins.

6. PERSPECTIVES

Protein Structures from Cryoelectron Microscopy and Hardware-Accelerated Calculations. Access to large and diverse data sets is a key factor in the development of new predictive methods and tools. Therefore, the applicability of force field methods to stability prediction is limited by the availability of relevant tertiary structures. At present, the PDB contains over 77 000 unique protein structures, and around 10 000 new structures are added each year. Advances in structural genomics will provide access to an additional large pool of protein structures, including previously unattainable structures of membrane-bound proteins that will be solved by cryogenic electron microscopy. A tertiary structure of a biomolecule of interest is typically required for predictions employing energy calculations. The general applicability of these methods is also hindered by their computational cost, which imposes a trade-off between accuracy and throughput. The most precise alchemical free energy calculations rely on MD simulations in which both the solute and solvent are modeled atomistically. Such calculations are too costly to be used in systematic mutagenesis campaigns with currently available computational resources. However, they could be selectively used to design mutations whose effects are poorly predicted by otherwise reliable Rosetta or FoldX calculations (e.g., substitutions that change the charge at the protein surface). Their high computational cost could be alleviated by adopting computing employing graphics processing units (GPUs), which has not yet been implemented in a number of software tools. Wider use of GPUs will enable predictions of structures and complexes that are currently too large to process using computationally demanding physical force fields.

Consistent and Balanced Stability Data Sets Are Urgently Needed. Machine learning techniques are faster than force field methods and less dependent on the availability of tertiary structures because many features used in machine-learning-based predictors can be extracted from primary sequences. However, machine learning methods are very sensitive to the size and quality of the experimental data sets available for training and validation. At present, there is a serious lack of reliable experimental data suitable for use in protein stabilization efforts. The only available database—ProTherm—is burdened by errors and contains data on fewer than 2000 single-point mutations after rigorous filtering. This number is insufficient to train reliable machine learning systems without introducing a risk of overfitting. Moreover, the ProTherm database was most recently updated in February 2013, and several protein stabilization projects have been conducted since then. Systematic mining of the scientific literature to incorporate the stability data from these projects could provide valuable data resources for the training and validation of stability predictors. A new database, FireProt DB, is being established for this purpose at https://loschmidt.chemi.muni.cz/fireprotdb/. The research community should make an effort to establish validation procedures to assess the quality of predictions of protein stability and solubility. This could be done by releasing design challenges, but not experimental data, as in the well-known Critical Assessment of Protein Structure Prediction. Such a community-wide assessment is one of the most efficient ways to compare individual tools.

The Shift from Scores to Profiles and Specific Mutations in Solubility Predictions. The problem of unbalanced data sets also affects solubility predictors based on machine learning, especially those that use *k*-mer content and physicochemical properties as dominant features. The imbalance of the training data sets containing a larger number of negative samples and low diversity of protein structures limit the predictive performance and generalizability to unseen protein families. Over the short history of solubility prediction, there has been a significant and positive shift away from methods that provide single solubility scores toward alternatives that offer more detailed solubility profile predictions and even suggest mutations predicted to enhance protein solubility. However, this trend also poses problems because the quantity of relevant high-quality data decreases as the detail of the predictions increases. For single solubility score predictions, the TargetTrack database (which contains information on tens of thousands of samples) is large enough to support the development of machine learning models. For solubility profile predictions, the number of relevant samples decreases to hundreds or thousands, most of which are amyloidogenic peptides. Matters are worse still for attempts to predict the effect of mutations on protein solubility; in this case, the amount of relevant experimental data is arguably below the minimum needed to make adequate predictions. Therefore, mathematical models developed by machine learning frequently incorporate empirical components such as structure-based corrections. A mechanistic understanding of protein solubility justified by robust statistical analysis can only be expected once larger sets of experimental data become available.

High-Throughput Techniques for Highly Consistent Data Sets. We envisage that the lack of appropriate data for solubility prediction will be partially addressed by studies using novel high-throughput characterization techniques such as droplet microfluidics, fluorescence-activated cell sorting, fluorescence resonance energy transfer, deep sequencing, and deep mutational scanning. Experiments should be conducted under strictly controlled conditions to produce robust data and could employ one or more of the biomolecular and cellular systems that have recently been developed to monitor protein solubility and aggregation inside living cells. Additional high-quality data could be obtained from projects conducted by companies and other private organizations. The data generated under defined conditions need to be properly annotated, for example to report vectors, host organisms, buffers, laboratory conditions, and procedures used for protein expression, purification, and characterization. Proper controls should always be included and the statistics reported to allow a quantitative assessment of data variation. Collected data should be structured to allow processing using computers, which is for example not the case for the largest database of protein solubility data, TargetTrack. The data should be curated and stored in publicly accessible databases following the FAIR principles: Findable, Accessible, Interoperable, and Reusable. New data sets will enable the use of more sophisticated and data-intensive methods such as deep learning and allow proper external validation to be performed. Moreover, because solubility depends largely on the properties of the protein's surface, corrections based on protein structure and the inclusion of structural data in predictive tools could improve the prediction accuracy. Enhanced-sampling MD simulations of simplified molecular systems might reveal residue interactions that are important for protein folding, while advances in homology modeling and threading can complement sequence-based descriptors by providing structural information at a reasonable computational cost.

Robust Scaffolds for Directed Evolution by Phylogenetic Analyses. Whereas force field and machine learning methods are limited by a lack of data, the problem for phylogenetic approaches is different: high-throughput sequencing has made vast numbers of sequences available, allowing evolutionary analyses to be performed for the vast majority of protein families. The genomes of organisms living under extreme conditions are also becoming available, providing essential information for wider use of CD. This rapid expansion of the accessible sequence space has a downside for the ASR method, which can only use a limited number of homologous sequences for reconstruction. Therefore, large pools of potential homologues make sequence selection a challenging task. Homologue selection can be guided by annotation ontologies (e.g., molecular function, cellular component, and biological process) and other information from bioinformatics and biophysical databases. Furthermore, with increasing numbers of solved protein structures, structure-guided MSAs may displace sequence-based alternatives, and ASR may be more commonly used to generate robust scaffolds for directed evolution campaigns and de novo enzyme design. The degree of uncertainty in ASR increases the further back we go in evolutionary history. Therefore, the reliability of inference methods should be increased to more accurately predict folded, stable, and soluble ancestral proteins.

Addressing Stability−Activity Trade-Offs Using Metadata and Negative and Multistate Designs. The predictive power of computational methods has improved in recent years, with a positive impact mainly in the area of protein stabilization. A very challenging but important task is to predict thermodynamic as well as kinetic stability. There are several spectacular examples illustrating the improvement in kinetic stability by only a few mutations, but to the best of our knowledge, methods specifically targeting kinetic stability have not been developed. Connecting the design of kinetic stability with solubility within a single method could be particularly powerful. Stability−activity trade-offs are intrinsic to protein structures. Buried polar catalytic residues are suboptimal with respect to protein stability, and structural optimization of these functionally relevant regions is likely to also affect the biological activity. Mutations that stabilize regions whose conformational dynamics are important for enzyme activity can similarly be expected to negatively affect the catalytic performance. The incorporation of metadata and smart filters into engineering workflows will help preserve protein activity by enabling the identification of structurally and functionally important residues, which should be systematically excluded from mutagenesis. The incorporation of such negative designs will suppress misfolding and protein aggregation. Furthermore, prediction accuracy is sometimes compromised by using a single structure in calculations. Increasing computational power and the use of GPU hardware will allow the adoption of multistate designs. Extracting multiple representative conformations and averaging results over the ensemble will further improve the robustness and accuracy of predictions.

Enhancing Accuracy by Using Metapredictors, Consensual Force Fields, and Hybrid Methods. There is a clear trend toward combining multiple fundamentally different methods within single predictors, leading to the development of metapredictors, consensual force fields, and hybrid methods. Hybrid methods offer several advantages: (i) even a simple majority voting approach over several methods yields better results than any individual method, each of which has its own strengths and weaknesses; (ii) smart filtering out of "untouchable" residues reduces the time required for calculations to a degree that permits very thorough analysis of the designable residues; (iii) the phylogenetic components of hybrid methods can incorporate both positive and negative design elements; and (iv) the availability of reliable predictions will enable the combination of substitutions to create multiple-point mutants without risking the introduction of destabilizing or antagonistic effects. Hybrid methods represent a natural step forward in the rapidly evolving field of protein stability prediction because improvements in machine learning models are limited by the availability of adequate data sets, while the application of advanced force field methods is restrained by their computational cost. It was recently demonstrated that combining phylogenetic methods and atomistic force fields can effectively optimize stability−activity trade-offs. We also envisage the future enrichment of protein stabilization methods addressing both thermodynamic and kinetic stability with tools for predicting protein solubility, aggregation propensity, and expressibility, eventually yielding all-in-one software suites capable of designing "ideal" biocatalysts.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscatal.8b03613.

> Data sets for prediction of protein stability (Table S1); software tools for prediction of protein stability (Table S2); data sets for prediction of protein solubility (Table S3); software tools for prediction of protein solubility (Table S4); comparison of the existing tools with the S350 data set (Table S5) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jiri@chemi.muni.cz.

ORCID

Jiri Damborsky: 0000-0002-7848-8216

Author Contributions

[‖]M.M., H.K., and J.H. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Czech Ministry of Education (LM2015051, LM2015047, LM2015055, CZ.02.1.01/0.0/0.0/16_013/0001761, CZ.02.1.01/0.0/0.0/16_019/0000868, and CZ.02.1.01/0.0/0.0/16_026/0008451) and the European Commission (720776 and 722610) for financial support. H.K. is the MSCA ITN ES-Cat Research Fellow supported by the European Commission (722610). The work of M.M. and J.H. was supported by the ICT Tools, Methods and Technologies for Smart Cities Project of the Brno University of Technology (FIT-S-17-3964).

■ REFERENCES

(1) Choi, J.-M.; Han, S.-S.; Kim, H.-S. Industrial Applications of Enzyme Biocatalysis: Current Status and Future Aspects. *Biotechnol. Adv.* **2015**, *33*, 1443−1454.

(2) Mitchell, A. C.; Briquez, P. S.; Hubbell, J. A.; Cochran, J. R. Engineering Growth Factors for Regenerative Medicine Applications. *Acta Biomater.* **2016**, *30*, 1−12.

(3) Dvořák, P.; Nikel, P. I.; Damborský, J.; de Lorenzo, V. Bioremediation 3.0: Engineering Pollutant-Removing Bacteria in the Times of Systemic Biology. *Biotechnol. Adv.* **2017**, *35*, 845−866.

(4) Vanacek, P.; Sebestova, E.; Babkova, P.; Bidmanova, S.; Daniel, L.; Dvorak, P.; Stepankova, V.; Chaloupkova, R.; Brezovsky, J.; Prokop, Z.; Damborsky, J. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catal.* **2018**, *8*, 2402−2412.

(5) Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K. Engineering the Third Wave of Biocatalysis. *Nature* **2012**, *485*, 185−194.

(6) Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **2008**, *4*, e1000002.

(7) Dellus-Gur, E.; Toth-Petroczy, A.; Elias, M.; Tawfik, D. S. What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-Offs. *J. Mol. Biol.* **2013**, *425*, 2609−2621.

(8) Johansson, K. E.; Johansen, N. T.; Christensen, S.; Horowitz, S.; Bardwell, J. C. A.; Olsen, J. G.; Willemoës, M.; Lindorff-Larsen, K.; Ferkinghoff-Borg, J.; Hamelryck, T.; Winther, J. R. Computational Redesign of Thioredoxin Is Hypersensitive toward Minor Conformational Changes in the Backbone Template. *J. Mol. Biol.* **2016**, *428*, 4361−4377.

(9) Arabnejad, H.; Dal Lago, M.; Jekel, P. A.; Floor, R. J.; Thunnissen, A.-M. W. H.; Terwisscha van Scheltinga, A. C.; Wijma, H. J.; Janssen, D. B. A Robust Cosolvent-Compatible Halohydrin Dehalogenase by Computational Library Design. *Protein Eng., Des. Sel.* **2017**, *30*, 175−189.

(10) Wyganowski, K. T.; Kaltenbach, M.; Tokuriki, N. GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates. *J. Mol. Biol.* **2013**, *425*, 3403−3414.

(11) Lawrence, P. B.; Gavrilov, Y.; Matthews, S. S.; Langlois, M. I.; Shental-Bechor, D.; Greenblatt, H. M.; Pandey, B. K.; Smith, M. S.; Paxman, R.; Torgerson, C. D.; Merrell, J. P.; Ritz, C. C.; Prigozhin, M. B.; Levy, Y.; Price, J. L. Criteria for Selecting PEGylation Sites on Proteins for Higher Thermodynamic and Proteolytic Stability. *J. Am. Chem. Soc.* **2014**, *136*, 17547−17560.

(12) Rueda, N.; Dos Santos, J. C. S.; Ortiz, C.; Torres, R.; Barbosa, O.; Rodrigues, R. C.; Berenguer-Murcia, Á.; Fernandez-Lafuente, R. Chemical Modification in the Design of Immobilized Enzyme Biocatalysts: Drawbacks and Opportunities. *Chem. Rec.* **2016**, *16*, 1436−1455.

(13) Stepankova, V.; Bidmanova, S.; Koudelakova, T.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Strategies for Stabilization of Enzymes in Organic Solvents. *ACS Catal.* **2013**, *3*, 2823−2836.

(14) Butt, T. R.; Edavettal, S. C.; Hall, J. P.; Mattern, M. R. SUMO Fusion Technology for Difficult-to-Express Proteins. *Protein Expression Purif.* **2005**, *43*, 1−9.

(15) LaVallie, E. R.; DiBlasio, E. A.; Kovacic, S.; Grant, K. L.; Schendel, P. F.; McCoy, J. M. A Thioredoxin Gene Fusion Expression

System That Circumvents Inclusion Body Formation in the *E. coli* Cytoplasm. *Nat. Biotechnol.* **1993**, *11*, 187−193.

(16) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 5869−5874.

(17) 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.

(18) Ganesan, A.; Siekierska, A.; Beerten, J.; Brams, M.; Van Durme, J.; De Baets, G.; Van der Kant, R.; Gallardo, R.; Ramakers, M.; Langenberg, T.; Wilkinson, H.; De Smet, F.; Ulens, C.; Rousseau, F.; Schymkowitz, J. Structural Hot Spots for the Solubility of Globular Proteins. *Nat. Commun.* **2016**, *7*, 10816.

(19) Zeymer, C.; Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **2018**, *87*, 131−157.

(20) Starr, T. N.; Thornton, J. W. Epistasis in Protein Evolution. *Protein Sci.* **2016**, *25*, 1204−1218.

(21) Goldsmith, M.; Tawfik, D. S. Enzyme Engineering: Reaching the Maximal Catalytic Efficiency Peak. *Curr. Opin. Struct. Biol.* **2017**, *47*, 140−150.

(22) Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B. Synthetic Biology for the Directed Evolution of Protein Biocatalysts: Navigating Sequence Space Intelligently. *Chem. Soc. Rev.* **2015**, *44*, 1172−1239.

(23) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D. Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* **2017**, *357*, 168−175.

(24) 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.* **2018**, *46*, W356−W362.

(25) Kuipers, R. K.; Joosten, H.-J.; van Berkel, W. J. H.; Leferink, N. G. H.; Rooijen, E.; Ittmann, E.; van Zimmeren, F.; Jochens, H.; Bornscheuer, U.; Vriend, G.; Martins dos Santos, V. A. P.; Schaap, P. J. 3DM: Systematic Analysis of Heterogeneous Superfamily Data to Discover Protein Functionalities. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2101−2113.

(26) Reetz, M. T.; Carballeira, J. D. Iterative Saturation Mutagenesis (ISM) for Rapid Directed Evolution of Functional Enzymes. *Nat. Protoc.* **2007**, *2*, 891−903.

(27) Liskova, V.; Stepankova, V.; Bednar, D.; Brezovsky, J.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Different Structural Origins of the Enantioselectivity of Haloalkane Dehalogenases toward Linear *β*-Haloalkanes: Open-Solvated versus Occluded-Desolvated Active Sites. *Angew. Chem., Int. Ed.* **2017**, *56*, 4719−4723.

(28) Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **2011**, *50*, 4402−4410.

(29) Balchin, D.; Hayer-Hartl, M.; Hartl, F. U. In Vivo Aspects of Protein Folding and Quality Control. *Science* **2016**, *353*, aac4354.

(30) Colón, W.; Church, J.; Sen, J.; Thibeault, J.; Trasatti, H.; Xia, K. Biological Roles of Protein Kinetic Stability. *Biochemistry* **2017**, *56*, 6179−6186.

(31) Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S. Bridging the Gaps in Design Methodologies by Evolutionary Optimization of the Stability and Proficiency of Designed Kemp Eliminase KE59. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10358−10363.

(32) Taverna, D. M.; Goldstein, R. A. Why Are Proteins Marginally Stable? *Proteins: Struct., Funct., Genet.* **2002**, *46*, 105−109.

(33) Sanchez-Ruiz, J. M. Protein Kinetic Stability. *Biophys. Chem.* **2010**, *148*, 1−15.

(34) Bommarius, A. S.; Paye, M. F. Stabilizing Biocatalysts. *Chem. Soc. Rev.* **2013**, *42*, 6534−6565.

(35) Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **2018**, *87*, 105−129.

(36) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2632−2647.

(37) Polizzi, K. M.; Bommarius, A. S.; Broering, J. M.; Chaparro-Riggers, J. F. Stability of Biocatalysts. *Curr. Opin. Chem. Biol.* **2007**, *11*, 220−225.

(38) Buck, P. M.; Kumar, S.; Wang, X.; Agrawal, N. J.; Trout, B. L.; Singh, S. K. Computational Methods To Predict Therapeutic Protein Aggregation. *Methods Mol. Biol.* **2012**, *899*, 425−451.

(39) Jaswal, S. S.; Sohl, J. L.; Davis, J. H.; Agard, D. A. Energetic Landscape of *α*-Lytic Protease Optimizes Longevity through Kinetic Stability. *Nature* **2002**, *415*, 343−346.

(40) Young, T. A.; Skordalakes, E.; Marqusee, S. Comparison of Proteolytic Susceptibility in Phosphoglycerate Kinases from Yeast and *E. coli*: Modulation of Conformational Ensembles Without Altering Structure or Stability. *J. Mol. Biol.* **2007**, *368*, 1438−1447.

(41) Shirke, A. N.; Basore, D.; Butterfoss, G. L.; Bonneau, R.; Bystroff, C.; Gross, R. A. Toward Rational Thermostabilization of Aspergillus Oryzae Cutinase: Insights into Catalytic and Structural Stability. *Proteins: Struct., Funct., Genet.* **2016**, *84*, 60−72.

(42) Liu, B.; Zhang, J.; Li, B.; Liao, X.; Du, G.; Chen, J. Expression and Characterization of Extreme Alkaline, Oxidation-Resistant Keratinase from Bacillus Licheniformis in Recombinant Bacillus Subtilis WB600 Expression System and Its Application in Wool Fiber Processing. *World J. Microbiol. Biotechnol.* **2013**, *29*, 825−832.

(43) Nguyen, V.; Wilson, C.; Hoemberger, M.; Stiller, J. B.; Agafonov, R. V.; Kutter, S.; English, J.; Theobald, D. L.; Kern, D. Evolutionary Drivers of Thermoadaptation in Enzyme Catalysis. *Science* **2017**, *355*, 289−294.

(44) Risso, V. A.; Gavira, J. A.; Gaucher, E. A.; Sanchez-Ruiz, J. M. Phenotypic Comparisons of Consensus Variants versus Laboratory Resurrections of Precambrian Proteins. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 887−896.

(45) Bednar, D.; Beerens, K.; Sebestova, E.; Bendl, J.; Khare, S.; Chaloupkova, R.; Prokop, Z.; Brezovsky, J.; Baker, D.; Damborsky, J. FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants. *PLoS Comput. Biol.* **2015**, *11*, e1004556.

(46) Babkova, P.; Sebestova, E.; Brezovsky, J.; Chaloupkova, R.; Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem* **2017**, *18*, 1448−1456.

(47) 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.

(48) Hammes, G. G.; Chang, Y.-C.; Oas, T. G. Conformational Selection or Induced Fit: A Flux Description of Reaction Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 13737.

(49) Kramer, R. M.; Shende, V. R.; Motl, N.; Pace, C. N.; Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* **2012**, *102*, 1907−1915.

(50) Khow, O.; Suntrarachun, S. Strategies for Production of Active Eukaryotic Proteins in Bacterial Expression System. *Asian Pac. J. Trop. Biomed.* **2012**, *2*, 159−162.

(51) Sørensen, H. P.; Mortensen, K. K. Soluble Expression of Recombinant Proteins in the Cytoplasm of *Escherichia coli*. *Microb. Cell Fact.* **2005**, *4*, 1.

(52) Hartl, F. U.; Bracher, A.; Hayer-Hartl, M. Molecular Chaperones in Protein Folding and Proteostasis. *Nature* **2011**, *475*, 324−332.

(53) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

(54) Englander, S. W.; Mayne, L. The Case for Defined Protein Folding Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 8253−8258.

(55) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.

(56) Eaton, W. A.; Wolynes, P. G. Theory, Simulations, and Experiments Show That Proteins Fold by Multiple Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E9759−E9760.

(57) 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.

(58) Broom, A.; Jacobi, Z.; Trainor, K.; Meiering, E. M. Computational Tools Help Improve Protein Stability but with a Solubility Tradeoff. *J. Biol. Chem.* **2017**, *292*, 14349−14361.

(59) Cabantous, S.; Waldo, G. S. *In Vivo* and *in Vitro* Protein Solubility Assays Using Split GFP. *Nat. Methods* **2006**, *3*, 845−854.

(60) 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. U. S. A.* **2009**, *106*, 4201−4206.

(61) Eijsink, V. G.; Vriend, G.; van den Burg, B.; van der Zee, J. R.; Veltman, O. R.; Stulp, B. K.; Venema, G. Introduction of a Stabilizing 10 Residue Beta-Hairpin in Bacillus Subtilis Neutral Protease. *Protein Eng., Des. Sel.* **1992**, *5*, 157−163.

(62) Lee, C.; Levitt, M. Accurate Prediction of the Stability and Activity Effects of Site-Directed Mutagenesis on a Protein Core. *Nature* **1991**, *352*, 448−451.

(63) Buß, O.; Muller, D.; Jager, S.; Rudat, J.; Rabe, K. S. Improvement in the Thermostability of a β-Amino Acid Converting ω-Transaminase by Using FoldX. *ChemBioChem* **2018**, *19*, 379−387.

(64) Modarres, H. P.; Mofrad, M. R.; Sanati-Nezhad, A. Protein Thermostability Engineering. *RSC Adv.* **2016**, *6*, 115252−115270.

(65) Pace, C. N.; Scholtz, J. M.; Grimsley, G. R. Forces Stabilizing Proteins. *FEBS Lett.* **2014**, *588*, 2177−2184.

(66) Lazaridis, T.; Karplus, M. Effective Energy Functions for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.

(67) Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98*, 2309−2316.

(68) Zhang, Z.; Wang, L.; Gao, Y.; Zhang, J.; Zhenirovskyy, M.; Alexov, E. Predicting Folding Free Energy Changes upon Single Point Mutations. *Bioinformatics* **2012**, *28*, 664−671.

(69) Wickstrom, L.; Gallicchio, E.; Levy, R. M. The Linear Interaction Energy Method for the Prediction of Protein Stability Changes Upon Mutation. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 111−125.

(70) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More than 1000 Mutations. *J. Mol. Biol.* **2002**, *320*, 369−387.

(71) Mendes, J.; Guerois, R.; Serrano, L. Energy Estimation in Protein Design. *Curr. Opin. Struct. Biol.* **2002**, *12*, 441−446.

(72) Dehouck, Y.; Gilis, D.; Rooman, M. A New Generation of Statistical Potentials for Proteins. *Biophys. J.* **2006**, *90*, 4010−4017.

(73) 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 Bioinf.* **2011**, *12*, 151.

(74) Liu, H. On Statistical Energy Functions for Biomolecular Modeling and Design. *Quant. Biol.* **2015**, *3*, 157−167.

(75) Kumar, M. D. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: Thermodynamic Databases for Proteins and Protein−Nucleic Acid Interactions. *Nucleic Acids Res.* **2006**, *34*, D204−206.

(76) Pucci, F.; Bourgeas, R.; Rooman, M. High-Quality Thermodynamic Data on the Stability Changes of Proteins Upon Single-Site Mutations. *J. Phys. Chem. Ref. Data* **2016**, *45*, 023104.

(77) Potapov, V.; Cohen, M.; Schreiber, G. Assessing Computational Methods for Predicting Protein Stability upon Mutation: Good on Average but Not in the Details. *Protein Eng., Des. Sel.* **2009**, *22*, 553−560.

(78) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33*, W382−388.

(79) Kepp, K. P. Towards a "Golden Standard" for Computing Globin Stability: Stability and Structure Sensitivity of Myoglobin Mutants. *Biochim. Biophys. Acta, Proteins Proteomics* **2015**, *1854*, 1239−1248.

(80) Christensen, N. J.; Kepp, K. P. Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol. *J. Chem. Inf. Model.* **2012**, *52*, 3028−3042.

(81) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(82) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656−1676.

(83) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031−3048.

(84) Davey, J. A.; Damry, A. M.; Euler, C. K.; Goto, N. K.; Chica, R. A. Prediction of Stable Globular Proteins Using Negative Design with Non-Native Backbone Ensembles. *Structure* **2015**, *23*, 2011−2021.

(85) Ó Conchúir, S.; Barlow, K. A.; Pache, R. A.; Ollikainen, N.; Kundert, K.; O'Meara, M. J.; Smith, C. A.; Kortemme, T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLoS One* **2015**, *10*, e0130433.

(86) Trainor, K.; Broom, A.; Meiering, E. M. Exploring the Relationships between Protein Sequence, Structure and Solubility. *Curr. Opin. Struct. Biol.* **2017**, *42*, 136−146.

(87) Das, R. Four Small Puzzles That Rosetta Doesn't Solve. *PLoS One* **2011**, *6*, e20044.

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

(89) 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.

(90) Bush, J.; Makhatadze, G. I. Statistical Analysis of Protein Structures Suggests That Buried Ionizable Residues in Proteins Are Hydrogen Bonded or Form Salt Bridges. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 2027−2032.

(91) Stranges, P. B.; Kuhlman, B. A Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds. *Protein Sci.* **2013**, *22*, 74−82.

(92) Beerens, K.; Mazurenko, S.; Kunka, A.; Marques, S. M.; Hansen, N.; Musil, M.; Chaloupkova, R.; Waterman, J.; Brezovsky, J.; Bednar, D.; Prokop, Z.; Damborsky, J. Evolutionary Analysis Is a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catal.* **2018**, *8*, 9420−9428.

(93) Wijma, H. J.; Floor, R. J.; Jekel, P. A.; Baker, D.; Marrink, S. J.; Janssen, D. B. Computationally Designed Libraries for Rapid Enzyme Stabilization. *Protein Eng., Des. Sel.* **2014**, *27*, 49−58.

(94) Thiltgen, G.; Goldstein, R. A. Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. *PLoS One* **2012**, *7*, e46084.

(95) Buß, O.; Rudat, J.; Ochsenreither, K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25−33.

(96) Allen, B. D.; Nisthal, A.; Mayo, S. L. Experimental Library Screening Demonstrates the Successful Application of Computational Protein Design to Large Structural Ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19838−19843.

(97) Barlow, K. A.; Ó Conchúir, S.; Thompson, S.; Suresh, P.; Lucas, J. E.; Heinonen, M.; Kortemme, T. Flex DdG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **2018**, *122*, 5389−5399.

(98) Ludwiczak, J.; Jarmula, A.; Dunin-Horkawicz, S. Combining Rosetta with Molecular Dynamics (MD): A Benchmark of the MD-Based Ensemble Protein Design. *J. Struct. Biol.* **2018**, *203*, 54−61.

(99) Davis, I. W.; Arendall, W. B.; Richardson, D. C.; Richardson, J. S. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **2006**, *14*, 265−274.

(100) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116*, 6516−6551.

(101) Fan, H.; Mark, A. E. Relative Stability of Protein Structures Determined by X-Ray Crystallography or NMR Spectroscopy: A Molecular Dynamics Simulation Study. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 111−120.

(102) Kuzmanic, A.; Pannu, N. S.; Zagrovic, B. X-Ray Refinement Significantly Underestimates the Level of Microscopic Heterogeneity in Biomolecular Crystals. *Nat. Commun.* **2014**, *5*, 3220.

(103) Karshikoff, A.; Nilsson, L.; Ladenstein, R. Rigidity versus Flexibility: The Dilemma of Understanding Protein Thermal Stability. *FEBS J.* **2015**, *282*, 3899−3917.

(104) Der, B. S.; Kluwe, C.; Miklos, A. E.; Jacak, R.; Lyskov, S.; Gray, J. J.; Georgiou, G.; Ellington, A. D.; Kuhlman, B. Alternative Computational Protocols for Supercharging Protein Surfaces for Reversible Unfolding and Retention of Stability. *PLoS One* **2013**, *8*, e64363.

(105) Chan, P.; Curtis, R. A.; Warwicker, J. Soluble Expression of Proteins Correlates with a Lack of Positively-Charged Surface. *Sci. Rep.* **2013**, *3*, 3333.

(106) Rezaie, E.; Mohammadi, M.; Sakhteman, A.; Bemani, P.; Ahrari, S. Application of Molecular Dynamics Simulations To Design a Dual-Purpose Oligopeptide Linker Sequence for Fusion Proteins. *J. Mol. Model.* **2018**, *24*, 313.

(107) 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.

(108) 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.

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

(110) 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.

(111) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18−22.

(112) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(113) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS One* **2017**, *12*, e0177678.

(114) Ling, C. X.; Sheng, V. S. Cost-Sensitive Learning and the Class Imbalance Problem. In *Encyclopedia of Machine Learning*; Sammut, C., Ed.; Springer: New York, 2007.

(115) Rao, R.; Fung, G.; Rosales, R. On the Dangers of Cross-Validation. An Experimental Evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 2008; pp 588−596.

(116) Stephens, Z. D.; Lee, S. Y.; Faghri, F.; Campbell, R. H.; Zhai, C.; Efron, M. J.; Iyer, R.; Schatz, M. C.; Sinha, S.; Robinson, G. E. Big Data: Astronomical or Genomical? *PLoS Biol.* **2015**, *13*, e1002195.

(117) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(118) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(119) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14*, 755−763.

(120) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM−HMM Alignment. *Nat. Methods* **2012**, *9*, 173−175.

(121) Pearson, W. R. An Introduction to Sequence Similarity ("Homology") Searching. *Curr. Protoc. Bioinf.* **2013**, *42*, 3.1.1−3.1.8.

(122) Rost, B. Twilight Zone of Protein Sequence Alignments. *Protein Eng., Des. Sel.* **1999**, *12*, 85−94.

(123) Fletcher, W.; Yang, Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Mol. Biol. Evol.* **2010**, *27*, 2257−2267.

(124) Vialle, R. A.; Tamuri, A. U.; Goldman, N. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol. Biol. Evol.* **2018**, *35*, 1783−1797.

(125) Chowdhury, B.; Garai, G. A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. *Genomics* **2017**, *109*, 419−431.

(126) Taly, J.-F.; Magis, C.; Bussotti, G.; Chang, J.-M.; Di Tommaso, P.; Erb, I.; Espinosa-Carrasco, J.; Kemena, C.; Notredame, C. Using the T-Coffee Package to Build Multiple Sequence Alignments of Protein, RNA, DNA Sequences and 3D Structures. *Nat. Protoc.* **2011**, *6*, 1669−1682.

(127) Pei, J.; Grishin, N. V. PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information. *Methods Mol. Biol.* **2014**, *1079*, 263−271.

(128) Steipe, B.; Schiller, B.; Plückthun, A.; Steinbacher, S. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.* **1994**, *240*, 188−192.

(129) Sullivan, B. J.; Nguyen, T.; Durani, V.; Mathur, D.; Rojas, S.; Thomas, M.; Syu, T.; Magliery, T. J. Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *J. Mol. Biol.* **2012**, *420*, 384−399.

(130) Lehmann, M.; Kostrewa, D.; Wyss, M.; Brugger, R.; D'Arcy, A.; Pasamontes, L.; van Loon, A. P. From DNA Sequence to Improved Functionality: Using Protein Sequence Comparisons to Rapidly Design a Thermostable Consensus Phytase. *Protein Eng., Des. Sel.* **2000**, *13*, 49−57.

(131) Magliery, T. J. Protein Stability: Computation, Sequence Statistics, and New Experimental Methods. *Curr. Opin. Struct. Biol.* **2015**, *33*, 161−168.

(132) Porebski, B. T.; Buckle, A. M. Consensus Protein Design. *Protein Eng., Des. Sel.* **2016**, *29*, 245−251.

(133) Jäckel, C.; Bloom, J. D.; Kast, P.; Arnold, F. H.; Hilvert, D. Consensus Protein Design without Phylogenetic Bias. *J. Mol. Biol.* **2010**, *399*, 541−546.

(134) Goyal, V. D.; Magliery, T. J. Phylogenetic Spread of Sequence Data Affects Fitness of SOD1 Consensus Enzymes: Insights from Sequence Statistics and Structural Analyses. *Proteins: Struct., Funct., Genet.* **2018**, *86*, 609−620.

(135) Vázquez-Figueroa, E.; Chaparro-Riggers, J.; Bommarius, A. S. Development of a Thermostable Glucose Dehydrogenase by a

Structure-Guided Consensus Concept. *ChemBioChem* **2007**, *8*, 2295−2301.

(136) Parthasarathy, S.; Murthy, M. R. Protein Thermal Stability: Insights from Atomic Displacement Parameters (B Values). *Protein Eng., Des. Sel.* **2000**, *13*, 9−13.

(137) Cole, M. F.; Gaucher, E. A. Exploiting Models of Molecular Evolution to Efficiently Direct Protein Engineering. *J. Mol. Evol.* **2011**, *72*, 193−203.

(138) Hochberg, G. K. A.; Thornton, J. W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biophys.* **2017**, *46*, 247−269.

(139) Aerts, D.; Verhaeghe, T.; Joosten, H.-J.; Vriend, G.; Soetaert, W.; Desmet, T. Consensus Engineering of Sucrose Phosphorylase: The Outcome Reflects the Sequence Input. *Biotechnol. Bioeng.* **2013**, *110*, 2563−2572.

(140) Trudeau, D. L.; Kaltenbach, M.; Tawfik, D. S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol.* **2016**, *33*, 2633−2641.

(141) Wheeler, L. C.; Lim, S. A.; Marqusee, S.; Harms, M. J. The Thermostability and Specificity of Ancient Proteins. *Curr. Opin. Struct. Biol.* **2016**, *38*, 37−43.

(142) Yang, Z. PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood. *Bioinformatics* **1997**, *13*, 555−556.

(143) Stamatakis, A. RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* **2006**, *22*, 2688−2690.

(144) Huelsenbeck, J. P.; Ronquist, F.; Nielsen, R.; Bollback, J. P. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* **2001**, *294*, 2310−2314.

(145) Goldstein, R. A.; Pollard, S. T.; Shah, S. D.; Pollock, D. D. Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Mol. Biol. Evol.* **2015**, *32*, 1373−1381.

(146) Williams, P. D.; Pollock, D. D.; Blackburne, B. P.; Goldstein, R. A. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. *PLoS Comput. Biol.* **2006**, *2*, e69.

(147) Eick, G. N.; Bridgham, J. T.; Anderson, D. P.; Harms, M. J.; Thornton, J. W. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Mol. Biol. Evol.* **2016**, *34*, 247−261.

(148) Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K. Palaeotemperature Trend for Precambrian Life Inferred from Resurrected Proteins. *Nature* **2008**, *451*, 704−707.

(149) Akanuma, S. Characterization of Reconstructed Ancestral Proteins Suggests a Change in Temperature of the Ancient Biosphere. *Life (Basel, Switz.)* **2017**, *7*, 33.

(150) Gumulya, Y.; Baek, J.-M.; Wun, S.-J.; Thomson, R. E. S.; Harris, K. L.; Hunter, D. J. B.; Behrendorff, J. B. Y. H.; Kulig, J.; Zheng, S.; Wu, X.; Wu, B.; Stok, J. E.; De Voss, J. J.; Schenk, G.; Jurva, U.; Andersson, S.; Isin, E. M.; Bodén, M.; Guddat, L.; Gillam, E. M. J. Engineering Highly Functional Thermostable Proteins Using Ancestral Sequence Reconstruction. *Nat. Catal.* **2018**, *1*, 878.

(151) Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M. Fast and Accurate Predictions of Protein Stability Changes upon Mutations Using Statistical Potentials and Neural Networks: PoPMuSiC-2.0. *Bioinformatics* **2009**, *25*, 2537−2543.

(152) Khatun, J.; Khare, S. D.; Dokholyan, N. V. Can Contact Potentials Reliably Predict Stability of Proteins? *J. Mol. Biol.* **2004**, *336*, 1223−1238.

(153) Pucci, F.; Bernaerts, K. V.; Kwasigroch, J. M.; Rooman, M. Quantification of Biases in Predictions of Protein Stability Changes upon Mutations. *Bioinformatics* **2018**, *34*, 3659−3665.

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

(155) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Böckmann, R. A. Predicting Free Energy Changes Using Structural Ensembles. *Nat. Methods* **2009**, *6*, 3−4.

(156) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(157) de Groot, B. L.; van Aalten, D. M.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. C. Prediction of Protein Conformational Freedom from Distance Constraints. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 240−251.

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

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

(160) 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−W310.

(161) Cheng, J.; Randall, A.; Baldi, P. Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1125−1132.

(162) 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. *Bioinformatics* **2011**, *27*, 3286−3292.

(163) Li, Y.; Fang, J. PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. *PLoS One* **2012**, *7*, e47247.

(164) Quang, D.; Chen, Y.; Xie, X. DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants. *Bioinformatics* **2015**, *31*, 761−763.

(165) Wang, Y.; Mao, H.; Yi, Z. Protein Secondary Structure Prediction by Using Deep Learning Method. *Knowl.-Based Syst.* **2017**, *118*, 115−123.

(166) Ivakhnenko, A. G. Polynomial Theory of Complex Systems. *IEEE Trans. Syst., Man, Cybern.* **1971**, *SMC-1*, 364−378.

(167) Bengio, Y.; Boulanger-Lewandowski, N.; Pascanu, R. Advances in Optimizing Recurrent Networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, 2013; pp 8624−8628.

(168) 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.

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

(170) Khan, S.; Vihinen, M. Performance of Protein Stability Predictors. *Hum. Mutat.* **2010**, *31*, 675−684.

(171) Usmanova, D. R.; Bogatyreva, N. S.; Ariño Bernad, J.; Eremina, A. A.; Gorshkova, A. A.; Kanevskiy, G. M.; Lonishin, L. R.; Meister, A. V.; Yakupova, A. G.; Kondrashov, F. A.; Ivankov, D. N. Self-Consistency Test Reveals Systematic Bias in Programs for Prediction Change of Stability upon Mutation. *Bioinformatics* **2018**, *34*, 3653−3658.

(172) Montanucci, L.; Martelli, P. L.; Ben-Tal, N.; Fariselli, P. A Natural Upper Bound to the Accuracy of Predicting Protein Stability Changes upon Mutations. 2018, arXiv:1809.10389 [q-bio.BM]. arXiv.org e-Print archive. https://arxiv.org/abs/1809.10389.

(173) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276−277.

(174) Lu, G.; Moriyama, E. N. Vector NTI, a Balanced All-in-One Sequence Analysis Suite. *Briefings Bioinf.* **2004**, *5*, 378−388.

(175) 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.* **2016**, *44*, W479−487.

(176) Stamatakis, A. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **2014**, *30*, 1312−1313.

(177) Ashkenazy, H.; Penn, O.; Doron-Faigenboim, A.; Cohen, O.; Cannarozzi, G.; Zomer, O.; Pupko, T. FastML: A Web Server for

Probabilistic Reconstruction of Ancestral Sequences. *Nucleic Acids Res.* **2012**, *40*, W580−584.

(178) Diallo, A. B.; Makarenkov, V.; Blanchette, M. Ancestors 1.0: A Web Server for Ancestral Sequence Reconstruction. *Bioinformatics* **2010**, *26*, 130−131.

(179) Westesson, O.; Barquist, L.; Holmes, I. HandAlign: Bayesian Multiple Sequence Alignment, Phylogeny and Ancestral Reconstruction. *Bioinformatics* **2012**, *28*, 1170−1171.

(180) Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D. L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M. A.; Huelsenbeck, J. P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst. Biol.* **2012**, *61*, 539−542.

(181) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* **2011**, *39*, W29−37.

(182) Altschul, S. F.; Gertz, E. M.; Agarwala, R.; Schäffer, A. A.; Yu, Y.-K. PSI-BLAST Pseudocounts and the Minimum Description Length Principle. *Nucleic Acids Res.* **2009**, *37*, 815−824.

(183) Whitehead, T. A.; Chevalier, A.; Song, Y.; Dreyfus, C.; Fleishman, S. J.; De Mattos, C.; Myers, C. A.; Kamisetty, H.; Blair, P.; Wilson, I. A.; Baker, D. Optimization of Affinity, Specificity and Function of Designed Influenza Inhibitors Using Deep Sequencing. *Nat. Biotechnol.* **2012**, *30*, 543−548.

(184) Shimizu, Y.; Inoue, A.; Tomari, Y.; Suzuki, T.; Yokogawa, T.; Nishikawa, K.; Ueda, T. Cell-Free Translation Reconstituted with Purified Components. *Nat. Biotechnol.* **2001**, *19*, 751−755.

(185) Niwa, T.; Kanamori, T.; Ueda, T.; Taguchi, H. Global Analysis of Chaperone Effects Using a Reconstituted Cell-Free Translation System. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 8937−8942.

(186) Berman, H. M.; Gabanyi, M. J.; Kouranov, A.; Micallef, D. I.; Westbrook, J. Protein Structure Initiative - TargetTrack 2000−2017 - All Data Files. DOI: 10.5281/zenodo.821654.

(187) 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. Inf. Exp.* **2011**, *1*, 6.

(188) 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.* **2011**, *150*, 73−81.

(189) Pawlicki, S.; Le Béchec, A.; Delamarche, C. AMYPdb: A Database Dedicated to Amyloid Precursor Proteins. *BMC Bioinf.* **2008**, *9*, 273.

(190) 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. U. S. A.* **2006**, *103*, 4074−4078.

(191) 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.

(192) Wozniak, P. P.; Kotulska, M. AmyLoad: Website Dedicated to Amyloidogenic Protein Fragments. *Bioinformatics* **2015**, *31*, 3395−3397.

(193) 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.

(194) 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.

(195) Tian, Y.; Deutsch, C.; Krishnamoorthy, B. Scoring Function To Predict Solubility Mutagenesis. *Algorithms Mol. Biol.* **2010**, *5*, 33.

(196) Wilkinson, D. L.; Harrison, R. G. Predicting the Solubility of Recombinant Proteins in *Escherichia coli*. *Nat. Biotechnol.* **1991**, *9*, 443−448.

(197) Davis, G. D.; Elisee, C.; Newham, D. M.; Harrison, R. G. New Fusion Protein Systems Designed to Give Soluble Expression in *Escherichia coli*. *Biotechnol. Bioeng.* **1999**, *65*, 382−388.

(198) Magnan, C. N.; Randall, A.; Baldi, P. SOLpro: Accurate Sequence-Based Prediction of Protein Solubility. *Bioinformatics* **2009**, *25*, 2200−2207.

(199) Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; Frishman, D. PROSO II—A New Method for Protein Solubility Prediction. *FEBS J.* **2012**, *279*, 2192−2200.

(200) 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.

(201) 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.

(202) 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.

(203) Hirose, S.; Noguchi, T. ESPRESSO: A System for Estimating Protein Expression and Solubility in Protein Expression Systems. *Proteomics* **2013**, *13*, 1444−1456.

(204) Hon, J.; Marusiak, M.; Martinek, T.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: Prediction of Protein Solubility. *Nucleic Acids Res.* **2018**, in preparation.

(205) 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.

(206) 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.

(207) 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 Bioinf.* **2007**, *8*, 65.

(208) 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.

(209) Maurer-Stroh, S.; Debulpaep, M.; Kuemmerer, N.; Lopez de la Paz, M.; 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.

(210) 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*, W301−307.

(211) Bryan, A. W.; Menke, M.; Cowen, L. J.; Lindquist, S. L.; Berger, B. BETASCAN: Probable Beta-Amyloids Identified by Pairwise Probabilistic Analysis. *PLoS Comput. Biol.* **2009**, *5*, e1000333.

(212) 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.

(213) Goldschmidt, L.; Teng, P. K.; Riek, R.; Eisenberg, D. Identifying the Amylome, Proteins Capable of Forming Amyloid-like Fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 3487−3492.

(214) Ahmed, A. B.; Znassi, N.; Château, M.-T.; Kajava, A. V. A Structure-Based Approach to Predict Predisposition to Amyloidosis. *Alzheimer's Dementia* **2015**, *11*, 681−690.

(215) Krogh, A.; Vedelsby, J. Neural Network Ensembles, Cross Validation and Active Learning. In *Proceedings of the 7th International*

Conference on Neural Information Processing Systems (NIPS'94); MIT Press: Cambridge, MA, 1994; pp 231−238.

(216) Maclin, R.; Opitz, D. Popular Ensemble Methods: An Empirical Study. *J. Artif. Intell. Res.* **1999**, *11*, 169−198.

(217) 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.

(218) Emily, M.; Talvas, A.; Delamarche, C. MetAmyl: A METa-Predictor for AMYLoid Proteins. *PLoS One* **2013**, *8*, e79722.

(219) 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*, W306−313.

(220) De Baets, G.; Van Durme, J.; van der Kant, R.; Schymkowitz, J.; Rousseau, F. Solubis: Optimize Your Protein. *Bioinformatics* **2015**, *31*, 2580−2582.

(221) 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.

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.

Appendix D

Contents of CD

The attached CD contains all the presented manuscripts and their supplementary materials.

```
/
├── computational_design_manuscript.pdf
├── computational_design_supplement.pdf
├── enzymeminer_manuscript.pdf
├── functional_annotation_manuscript.pdf
├── pqsfinder_manuscript.pdf
├── pqsfinder_supplement.zip
├── pqsfinder_web_manuscript.pdf
├── pqsfinder_web_supplement.pdf
├── soluprot_manuscript.pdf
└── soluprot_supplement.pdf
```