



BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

FACULTY OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SYSTEMS

ÚSTAV INFORMAČNÍCH SYSTÉMŮ

COMPUTATIONAL DESIGN OF STABLE PROTEINS

AUTOMATIZOVANÝ NÁVRH STABILNÍCH PROTEINŮ

DOCTORAL THESIS

DIZERTAČNÍ PRÁCE

AUTHOR

AUTOR PRÁCE

Ing. Miloš Musil

ADVISOR

VEDOUCÍ PRÁCE

Doc. Ing. Jaroslav Zendulka, CSc.

BRNO 2020

Abstract

Stable proteins are utilized in a vast number of medical and biotechnological applications. However, the native proteins have mostly evolved to function under mild conditions inside the living cells. As a result, there is a great interest in increasing protein stability to enhance their utility in the harsh industrial conditions. In recent years, the field of protein engineering has matured to the point that enables tailoring of native proteins for specific practical applications. However, the identification of stable mutations is still burdened by costly and laborious experimental work. Computational methods offer attractive alternatives that allow a rapid search of the pool of potentially stabilizing mutations to prioritize them for further experimental validation. A plethora of the computational strategies was developed: i) force-field-based energy calculations, ii) evolution-based techniques, iii) machine learning, or iv) the combination of several approaches. Those strategies are usually limited in their predictions to less impactful single-point mutations, while some more sophisticated methods for prediction of multiple-point mutations require more complex inputs from the side of the user.

The main aim of this Thesis is to provide users with a fully automated workflow that would allow for the prediction of the highly stable multiple-point mutants without the requirement of the extensive knowledge of the bioinformatics tools and the protein of interest.

FireProt is a fully automated workflow for the design of the highly stable multiple-point mutants. It is a hybrid method that combines both energy- and evolution-based approaches in its calculation core, utilizing sequence information as a filter for robust force-field calculations. FireProt workflow not only detects a pool of potentially stabilizing mutations but also tries to combine them together while reducing the risk of antagonistic effects.

FireProt^{ASR} is a fully automated workflow for ancestral sequence reconstruction, allowing users to utilize this protein engineering strategy without the need for the laborious manual work and the

knowledge of the system of interest. It resolves all the steps required during the process of ancestral sequence reconstruction, including the collection of the biologically relevant homologs, construction of the rooted tree, and the reconstruction of the ancestral sequences and ancestral gaps.

HotSpotWizard is a workflow for the automated design of mutations and smart libraries for the engineering of protein function and stability. It allows for a wider analysis of the protein of interest by utilizing four different protein engineering strategies: i) identification of the highly mutable residues located in the catalytic pockets and tunnels, ii) identification of the flexible regions, iii) calculation of the sequence consensus, and iv) identification of the correlated residues.

FireProt^{DB} is a database of the known experimental data quantifying a protein stability. The main aim of this database is to standardize protein stability data, provide users with well-manageable storage, and allow them to construct protein stability datasets to use them as training sets for various machine learning applications.

Introduction

Proteins are the building blocks of every living organism, where they perform a wide variety of functions, including DNA replication, catalysis of metabolic reactions, responding to the stimuli, and transporting molecules between different parts of the living structures^[1]. They consist of one or more long chains of amino acid residues connected by peptide bonds. The sequence of the amino acids in the protein determines its structure and function. Therefore, mutations leading to amino acid alteration are the driving force of evolution at the molecular level.

Over time, Nature has developed a remarkable diversity of biochemical reactions vital to the continuing evolution of living organisms and the preservation of life. These biochemical reactions scale from the simple one-step degradation processes to more complex pathways employing several different proteins. The recent advances of the next-generation sequencing, together with the steady growth of the computational resources and advances in bioinformatics have allowed wider access to these naturally evolved processes and their utilization in various medical, industrial and biotechnological applications. Furthermore, protein engineering has matured to the point that enables tailoring of native proteins for specific practical applications, thus overcoming the limitations of the native variants that have evolved to function in mild conditions^[2].

As a result, the ability to understand what drives the protein folding, its function, and other characteristics is crucial for further advances in the field of protein engineering as the mutations introduced into a modified protein can affect it in many different ways. Only a small portion of the mutations will have a beneficial impact on the protein characteristics, considering its intended purpose in the specific medical or industrial applications. Some of the mutations can influence protein stability, while others will affect its solubility, activity, expression yields, or ability to fold into the 3D structure and create more complex quaternary structures by interacting with other molecules. Both positive and harmful effects can be observed by

introducing mutations into the sequence of the protein of interest, and in many cases, there is an apparent trade-off between some of the characteristics of the proteins^[3-5]. As a result, mutation improving protein stability can harm its function and vice versa. Thus, it is necessary to analyze a large number of mutations to obtain the variant most suitable for its intended use.

This Thesis focuses mainly on the aspect of protein stability as one of the main characteristics that determine the usability of the natural biochemical reactions in the harsh environment of the medical and industrial applications. Stable proteins are able to withstand extreme temperatures, acidic or basic pH, or an unfavorable effect of organic solvents and proteases^[6]. Furthermore, stable proteins are often distinguished by higher half-life, making them easier to transport and store for later use^[7]. As a result, there is a high interest in increasing protein stability, and many different methods were designed over the years to accomplish such a task.

In the ideal case, the saturation mutagenesis would be applied to evaluate every possible mutation on every position of the engineered protein. However, such search space would be enormous, and the experimental evaluation laborious and costly. Therefore, there rises a need for effective and precise computational methods to predict protein stability. To satisfy this goal, a number of in silico tools have been developed recently. Unfortunately, due to the limited reliability and potential antagonistic effect between individual mutations, only single-point mutations with an almost negligible effect on protein stability are usually predicted in the existing tools. Such mutations typically enhance the stability of the target proteins only mildly, while higher stabilization can be achieved by engineering multiple-point mutants^[8].

Objectives of the Thesis

The main aim of this Thesis is to develop new methods that would allow for the design of highly stable multiple-point mutants, and it

presents several possible solutions. FireProt is a hybrid method that combines several different computational approaches into a single workflow, allowing for a more robust and reliable construction of the stable multiple-point mutants. The second solution, FireProt^{ASR}, is based on natural evolution and the observation that the ancestral proteins were significantly more stable than their extant counterparts. Finally, HotSpotWizard is presented as a tool that can be utilized to highlight potentially interesting residues in the protein, where mutations could have a positive impact not only on the stability but also on other protein characteristics. The new database FireProt^{DB} is introduced as a possible solution for a current troubling situation surrounding the storage and management of the existing data obtained from the laboratory measurements of the protein stability. Such a compilation of manually curated data is very much needed for future development of reliable predictive tools based on machine learning.

The main goals of this Thesis are:

- to analyze the physico-chemical forces that participate in the increase of protein stability
- to construct a reliable protein stability dataset that could be used for the validation of the existing tools and force-fields and for the training of the methods based on machine learning
- to develop, integrate and thoroughly validate a hybrid workflow for an automated design of the stable multiple-point mutants
- to resolve the algorithmic and technical problems connected with the automatization of the ancestral sequence reconstruction with the primary focus on improvement of the proteins' thermal stability
- to develop, integrate and validate a fully automated workflow for ancestral sequence reconstruction

Computational approaches for prediction of protein stability

In the ideal case, saturation mutagenesis of each possible mutation would be carried by the rigorous experimental validation. However, in most projects, such validation is close to impossible due to the costly and laborious nature of those experiments. Considering a standard protein consisting of approximately 300 amino acids, this leaves us with over 5,000 single-point mutations. Furthermore, single-point mutations often provide an almost negligible effect on protein stability (< 2 kcal/mol)^[9,10], and therefore combining several stabilizing mutations is typically required to procure a significant improvement of protein stability^[8]. Unfortunately, the additive effect of stabilizing mutations is not guaranteed as synergistic or antagonistic effects can occur between any subset of stabilizing single-point mutations. Mutations are considered synergistic if their combined effect on protein stability is notably higher than the sum of the individual mutations, while the antagonistic effect means the exact opposite. The synergistic effect usually appears due to the creation of a new physico-chemical interaction such as a salt bridge between anionic carboxylate and cationic ammonium or a disulphide bridge between two cysteine residues. On the other hand, the antagonistic effect disturbs some of the newly introduced interactions or creates clashes between the side chains of the mutated or original residues. This, for example, can be easily observed when several mutations are designed to fill the same space in the structure of the protein, filling the void each by itself, however being unable to fit in if combined. This could either damage protein stability or even completely prevent it from a successful folding.

In most cases, antagonistic effects are not easily detectable, and therefore further experimental validation is needed. With only 100 potentially stabilizing mutations, close to 5,000 experiments would have to be performed to evaluate all possible double-point mutants, and this number is exponentially increasing with each added mutation. As a result, there is an ever-growing need for fast and accurate

computational methods that would allow for rapid evaluation of the potentially stabilizing mutations, and serve as a reliable tool for the prioritization of mutations for the rigorous laboratory experiments.

In general, the computational methods for the prediction of the effect of mutations on protein stability can be divided into four categories^[11]:

Force-field methods relying on the calculation of the $\Delta\Delta G$ based on the models of molecular mechanics.

Phylogenetic analysis utilizing the evolutionary information contained in the set of homolog sequences.

Machine learning methods constructing a computational model based on the stability data provided by previous experimental validation.

Hybrid methods and meta-predictors combining several of the previous approaches or several different methods of a single approach together to obtain more robust and reliable results.

Principles of methods based on force-field calculations

In silico design of the stable proteins based on the calculation of the energy force-fields is deeply rooted in our current state of knowledge of the physico-chemical properties of the individual amino acids and their description by molecular mechanic force-fields. Therefore, these calculations do not rely on the availability of the diverse, high-quality experimental data. In general terms, a force-field is a description of all bonded and non-bonded interactions in the protein of interest^[2,12]. These interactions are captured in the energy-field equation used to estimate the potential energy of a molecular system^[13]. The most accurate methods in this category are the free energy methods, relying on molecular dynamics (MD), or Metropolis Monte Carlo simulations. Unfortunately, those methods require a tremendous amount of computational power and are viable only for a limited number of mutations or smaller, less expensive systems of interest^[14]. A number

of heuristic approaches were created over the last decades to overcome this bottleneck, however huge analysis is still viable only with the use of simulation-independent stability predictors that can be divided into three categories^[15,16]:

Physical effective energy functions (PEEFs) are closely related to classical molecular mechanic force-fields, which allow for a fundamental analysis of the molecular interactions^[13]. The individual terms of the energy-field equations are calculated via the simplification of the known physical laws and are still burdened by high computational demands reaching from hours up to several days for a single mutation. However, similarly to the molecular dynamics methods, they are versatile, accurate, and capable of predicting the behaviour of the enzymes under non-standard conditions such as non-physiological pH, non-standard salinity, or elevated temperature^[17].

Statistical effective energy functions (SEEFs) are viable for rapid analysis as they can predict changes in stability over the entire sequence space of an average-sized enzyme in a matter of minutes^[18,19]. Compared to PEEFs, terms used in the SEEFs energy-field equations are derived from curated data sets of available experimental protein structures projected into several stability descriptors. An effective potential can be then extrapolated for every descriptor distribution and utilized as a part of the overall energy function^[18]. SEEFs do not explicitly model physical molecular interactions and are strongly dependent on the folded protein structures' availability and diversity^[16].

Empirical effective energy functions (EEEFs) represent a bridge between PEEFs and SEEFs as they include both physical and statistical terms in their energy-field equations, which are weighted and parametrized to match experimental data^[15,16]. The thermodynamic data used in the derivation of terms typically originate from mutational experiments conducted under standard conditions. As a result, EEEFs provide a reasonable compromise between computational demands and the accuracy of the free energy function^[20]. A major drawback of EEEFs is that their applicability is

restricted to the environmental conditions under which the experimental data used for the parametrization were acquired^[21,22].

Even though force-field-based calculations are currently considered the most powerful tool for predicting the effect of mutations on protein stability, the accuracy of the energy functions is still suboptimal due to insufficient conformational sampling, imbalances in force-fields, and the problems connected with the existing data sets^[21,23]. The computation of $\Delta\Delta G$ is based on the thermodynamic cycle, and therefore it requires modelling the folded and unfolded states of both wild-type and mutant protein^[14,24]. The value of $\Delta\Delta G$ is then established as the difference between both folded states with several issues reported for various energy functions. All energy functions are known to overestimate hydrophobicity and tend to favour nonpolar mutations as the stabilizing ones^[25-27]. PEEFs often underestimate the stabilization provided by the buried polar residues as they overestimate the energetic cost of unsatisfied salt bridges and hydrogen bonds in the protein core^[28-30]. The estimation of the conformational and solvent-related entropy is also imprecise. The inability of force-field methods to account for entropy-driven contributions can be partially resolved by utilizing evolutionary-based approaches inside the more robust hybrid workflows^[2,12]. Another shortcoming comes with the prediction of the multiple-point mutants as most stability predictors have been parametrized using only a single-point mutant datasets. As a result, the predictive power for the multiple-point mutants is limited for most of the existing force-fields^[31,32]. This shortcoming can also be attributed to the insufficient conformational sampling of the folded state, especially in the case of mutations introducing large-scale backbone movements into the mutant protein structure^[33]. In PEEFs and EEEFs, such movements are simulated by the utilization of the rotamer libraries to the fixed protein backbones, thereby reducing computational demands while providing comparable precision for the predictions of the single-point mutations^[34]. However, this approach does not stand in the case of the multiple-point mutants and multistate designs. Therefore, flexible backbone sampling techniques^[23,35], generating conformational ensembles and utilizing energetically more

favourable conformations, are required. Finally, the accuracy of the force-field methods is strongly dependent on the quality of the available tertiary structure. Their applicability for the proteins without resolved tertiary structure is given by the reliability of the structure modelling tools and the similarity of the closest sequence homology with a known tertiary structure. Furthermore, structures obtained by X-ray crystallography (>90% proteins in PDB database^[36]) do not necessarily reflect the global energy minimum of the native state of the protein in its natural environment^[37] and may, in some cases, be misleading starting point for a comprehensive prediction of protein stability^[22,38].

Principles of methods based on phylogenetic analysis

A phylogenetic or evolutionary analysis are methods that take advantage of the information hidden in the set of homolog sequences. The evolutionary approach's main advantage is that those methods do not require tertiary structure and are therefore viable for the majority of known protein sequences (about 200 million of sequences in UniProt compared to 100 thousand structures in the PDB database). The only limitation in its applicability occurs in the families with the low representation of sequences in the database. However, with the rise of the next-generation sequencing methods, this limitation slowly mitigates as the number of sequences in the databases almost doubles every three years. The two most widely used phylogeny-based methods are consensus design and ancestral sequence reconstruction, both built on top of the reasonably-sized set of homolog sequences.

Consensus design (CD) starts by building a compact multiple-sequence alignment (MSA) using a small number of homolog sequences ranging between a dozen and a few hundred. This MSA allows for a computation of every amino acid's frequency distribution in each position in the sequence alignment. Positions, where one or just a few amino acids are significantly more prevalent than others, are conserved as those residues changed only sparsely during evolution.

CD's core assumption is that conserved positions are somehow crucial for the function of the protein (stability, activity, protein folding, etc.), and the most frequent amino acid at the given position is more likely to be stabilizing^[39]. CD can be utilized when amino acid in the designed sequence differs from the most dominant ones in those conserved regions. This residue's mutation to the dominant amino acid suggested by evolution often leads to a non-negligible improvement of protein's thermal stability. It has been observed that high levels of sequence diversity in the MSA can interfere with the preservation of catalytic activity in the designed proteins, particularly if the MSA contains both prokaryotic and eukaryotic sequences. On the other hand, including only closely related homologs might introduce an evolutionary bias that prohibits CD from discovering more thermostable variants^[40]. In recent studies, the proportions of neutral and destabilizing CD mutations have been estimated to be 10 and 40%, respectively^[41]. In 2012, Sullivan was able to increase the proportion of correctly identified stabilizing mutations to 90% by discarding mutations of the residues with high statistical correlations to other positions in the MSA^[39]. This would suggest an inability of the CD analysis to account for any synergic or antagonistic effects. The second possible weakness comes from an apparent phylogenetic bias when the MSA is dominated by a small number of highly similar subfamilies^[42]. If tertiary structure for the protein of interest is available, the CD can be further refined by utilizing information about an active site, secondary structures, and intramolecular contacts or by analyzing molecular fluctuations based on crystallographic B-factors or MD simulations^[43].

Ancestral sequence reconstruction (ASR) is a probabilistic method that explores the deep evolutionary history of homolog sequences to reassemble protein's evolutionary trajectory^[44]. The method was initially developed to study molecular evolution. ASR is able to unearth sequences of the long-extinct genes and organisms from which the current ones evolved and is, therefore, an invaluable tool in the field of evolutionary biology. ASR has also been shown to be a very effective strategy for thermostability engineering^[45] and for improving

other protein's characteristics such as specificity, activity, or expression rates. Similarly to CD, ASR starts with the MSA's construction from the set of relevant homolog sequences. However, while CD relies on the simple analysis of the conservation of amino acids on the individual positions in the sequence alignment, ASR goes much further by considering evolutionary information depicted by the phylogenetic tree. Two main algorithms, maximum-likelihood^[46] (ML) and Bayesian inference^[47] (BI) were designed to interfere with ancestral sequences from MSA and phylogenetic tree. Over the years, many tools were built to make those algorithms accessible to the broad scientific community. However, several crucial steps in the calculation of ASR were not yet resolved in a satisfactory way that would allow for a fully automatized inference of the ancestral proteins, i.e., selection of the biologically relevant subset of homolog sequences, rooting of the phylogenetic tree and the reconstruction of the ancestral gaps. This limits the ASR's applicability as the method requires an in-depth knowledge of the biological system of interest and necessary bioinformatics tools together with the abysmal amount of manual work.

Principles of the methods based on machine learning

In recent years, machine learning has become one of the most dominant approaches in predicting protein stability^[48-50] and many other fields reaching far above the limited scope of protein engineering applications. The popularity of machine learning methods comes mostly from their ability to construct computational systems 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. The system based on the machine learning approach can be trained either with or without supervision. Both find their utilization in the field of protein engineering. In the supervised approaches, the system is given a set of training inputs and the expected outputs in the form of labels indicating each input's correct classification. Those

methods are well-suitable for training predictive systems. On the other hand, unsupervised approaches are mostly implemented for tasks involving data clustering.

As the machine learning systems are constructed during the learning process, they do not require a full understanding of the mechanistic principles underpinning the target function. This advantage shines, especially in situations where there is a severe gap in human knowledge-base, and therefore expert construction of the predictive systems is not entirely possible. Machine learning can also expand existing systems by discovering previously unrecognized features, patterns, and relationships hidden in the training dataset. Furthermore, machine learning methods 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. Moreover, machine learning is also much less time demanding than other methods because once the model has been constructed using the training data, predictions can be obtained at an almost instant rate.

However, the reliability of the machine learning approaches strongly depends on the quality and size of the training data set. The weights representing the relative importance of the individual features and the relationships between them are based on the provided experimental observations. Consequently, it is crucial to use high-quality experimental data with high consistency of experimental measurements and wide diversity when training and testing machine learning methods. The size and balance of the training dataset must also be considered. A modest dataset with only a few hundreds of cases might be too small to establish useful descriptors during learning. Additionally, lower diversity of the training data usually leads to a higher risk of overtraining and, therefore, losing its ability to generalize on a new, previously unknown data. In such cases, the weights assigned to the individual descriptors tend to be influenced by over-representing some of the descriptors in the training data, while other features with high informational value are under-estimated or omitted entirely. Unbalanced training datasets with substantial differences in the individual prediction categories' size could also lead

to erroneous predictions. For example, a training dataset in which more than two-thirds of the mutations are stated as deleterious would mislead the predictor to classify most mutations as deleterious because of the prevalence of such mutations during the learning. Some methods, namely support vector machines and random forests, are known to be more resistant to overfitting caused by unbalanced datasets^[51], while decision trees and standard neural networks are particularly sensitive. If the dataset is not sufficiently sized for the manual balancing by cutting part of the mutations out of the training set, this problem can be partially addressed using cost-sensitive matrices^[52], which penalize the system more strictly for misclassifying mutations that are sparsely represented in the training set. Some oversampling techniques such as SMOTE^[53] or ADASYN^[54] can be also utilized.

In parallel to the issue of the construction of the high-quality training data set, there arises the problem of model validation. In the ideal scenario, the validation data should also be balanced and utterly independent of the data used for training. However, due to the limited amount of experimental data, this scenario is often hard to reach. In bioinformatics, especially in the prediction of the effect of mutations on protein stability, it has become a common practise to use k-fold cross-validation as a standard method to validate the performance of the newly developed tools. This method entails randomly partitioning the original dataset into k subsets, using k - 1 subsets for the system's training, and the last random subset is left for the following validation. This process is then performed for each of the k subsets. The main argument of the utilization of cross-validation instead of splitting the data into independent training and testing datasets is that the available set of experimental data is often too small to support such a division without compromising the model's ability to identify the essential patterns and relationships. However, combining unbalanced datasets with the random aspect of k-fold cross-validation further increases the risk of overestimating the system's accuracy on the general data^[55]. Therefore, cross-validation is often no longer accepted as a means of validation of the bioinformatics tools. This is particularly problematic

in protein stability, where the construction of the sizeable, high-quality training dataset is impossible due to the lack of experimental data.

In summary, machine learning is a powerful approach that allows for detecting the previously unknown dependencies and interactions in the protein molecules. However, the utilization of the machine learning approaches in the predictions of the protein stability currently suffers from the overestimation of the accuracy of the existing machine learning-based tools due to the usage of the k-fold cross-validation as the method for their validation. This disadvantage is partially mitigated by using less vulnerable methods, such as random forests, and the cost-sensitive matrices.

Meta-predictors and principles of the methods based on hybrid approach

Methods based on the hybrid approaches cannot be considered a singular tool but more as a combination of several different methods, tools, and computational strategies. Those methods are usually more robust and provide users with mostly reliable results as the hybrid methods usually incorporate both energy- and evolution-based approaches into their workflows, utilizing their strengths and mitigating their shortcomings.

The analysis of the highly conserved regions and the residues that show a high correlation with one or more other residues in the MSA (correlated residues are usually changing together during evolution) is a starting point for most of the hybrid methods^[27,56]. This comes from the presumption that the conserved or highly correlated residues are somehow crucial for the correct function of the target protein, and therefore mutations designed on those positions would be at high risk of damaging some of the characteristics of the proteins. Conserved regions are often clustered around active sites, while the evolutionary correlation of two or more residues suggests an important intramolecular interaction. For this reason, hybrid approaches often

exclude those positions from further calculation, making the mutational space safer and, at the same time, reducing the computational demands. Furthermore, it was previously proven that evolution-based and force-field methods are complementary in many proteins as there is only a partial overlap of the stabilizing mutations designed by force-fields and evolution^[24]. This complementarity might be in part caused by the inability of the energy-based methods to correctly classify the charge changing mutations due to their weak implementation in the current force-fields and by the inability to estimate the effect of mutation on the unfolded state of the protein. As a result, hybrid methods are able to identify potentially stabilizing mutations that would be omitted by using only energy- or evolution-based approaches.

Due to the higher complexity and robustness of the hybrid methods, these methods are often viable not only for predicting the effect of single-point mutations but also for significantly more stable multiple-point mutants. In general, multiple-point mutants are unattainable by the tools based on a singular approach, as there is a high risk of undesired antagonistic effects. However, this issue is tackled in hybrid methods such as PROSS^[56] and our novel FireProt strategy^[27].

Finally, meta-predictors are the special subset of the hybrid methods that combine the results of several different tools into one consensual prediction using the simple majority voting or utilizing some form of weights. Those predictors are usually more accurate than their components. However, they lack the complexity and robustness of the real hybrid workflows.

Research summary

This part summarizes the research that was conducted in connection with the main topic of this thesis, i.e. the development of the *in silico* tools that can be employed to design stable protein structures. Four original publications describing three tools and one database: FireProt, FireProt^{ASR}, FireProt^{DB}, and HotSpotWizard 2.0 are included. A brief list of the research published by the author that is not mentioned in this thesis is attached at the end of this section.

Original publications

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 Research*. 2017, 45(W1), W393-W399.

Author participation: 60%

Journal impact factor: 11.501 (Q1)

MUSIL M, KHAN RT, BEIER A, STOURAC J, KONEGGER H, DAMBORSKY J, BEDNAR D. FireProt-ASR: web server for fully automated ancestral sequence reconstruction. *Briefings in bioinformatics*. 2020. (accepted for publication)

Author participation: 60%

Journal impact factor: 8.990 (Q1)

STOURAC J, DUBRAVA J, MUSIL M, HORACKOVA J, DAMBORSKY J, MAZURENKO S, BEDNAR D. FireProt-DB: Database of Manually Curated Protein Stability Data. *Nucleic Acids Research*. 2020. (accepted for publication)

Author participation: 20\%

Journal impact factor: 11.501 (Q1)

BENDL J, STOURAC J, SEBESTOVA E, VAVRA O, MUSIL M, BREZOVSKY J, DAMBORSKY J. HotSpotWizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Research*. 2016, 44(W1), W479-W487.

Author participation: 15\%

Journal impact factor: 11.501 (Q1)

BENDL J, MUSIL M, ZENDULKA J, DAMBORSKY J, BREZOVSKY J. PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Computational Biology*. 2016, 12, e1004962.

Author participation: 35\%

Journal impact factor: 4.428 (Q1)

MUSIL M, KONEGGER H, HON J, BEDNAR D, DAMBORSKY J. Computational design of stable and soluble biocatalysts. *ACS Catalysis*. 2018, 9, 1033-1054.

Author participation: 35\%

Journal impact factor: 12.350 (Q1)

BEERENS K, MAZURENKO S, KUNKA A, MARQUES S, HANSEN N, MUSIL M, CHALOUPKOVA R, WATERMAN J, BREZOVSKY J, BEDNAR D, PROKOP Z, DAMBORSKY J. Evolutionary analysis as a powerful complement to energy

calculations for protein stabilization. *ACS Catalysis*. 2018, 8, 9420-9428.

Author participation: 15%

Journal impact factor: 12.350 (Q1)

KHAN RT, MUSIL M, STOURAC J, DAMBORSKY J, BEDNAR D. Fully automated ancestral sequence reconstruction using FireProt-ASR. *Current protocols in bioinformatics*. 2020. (*under review*).

Author participation: 40%

Journal impact factor: 9.630 (Q1)

PLANAS-IGLESIAS J, MARQUES S, PINTO G, MUSIL M, STOURAC J, BEDNAR D, DAMBORSKY J. Computational design of enzymes for biotechnological applications. *Biotechnology advances*. 2020. (*under review*)

Author participation: 20%

Journal impact factor: 12.831 (Q1)

FireProt

There is a continuous interest in increasing proteins stability to enhance their usability in numerous biomedical and biotechnological applications. A number of in silico tools for the prediction of the effect of mutations on protein stability have been developed recently. However, only single-point mutations with a small effect on protein stability are typically predicted with the existing tools and have to be followed by laborious protein expression, purification, and characterization. Here, we present FireProt, a web server for the automated design of multiple-point thermostable mutant proteins that combines structural and evolutionary information in its calculation core. FireProt utilizes sixteen tools and three protein engineering strategies for making reliable protein designs. The server is complemented with interactive, easy-to-use interface that allows users to directly analyze and optionally modify designed thermostable mutants. FireProt is freely available at <http://loschmidt.chemi.muni.cz/fireprot>.

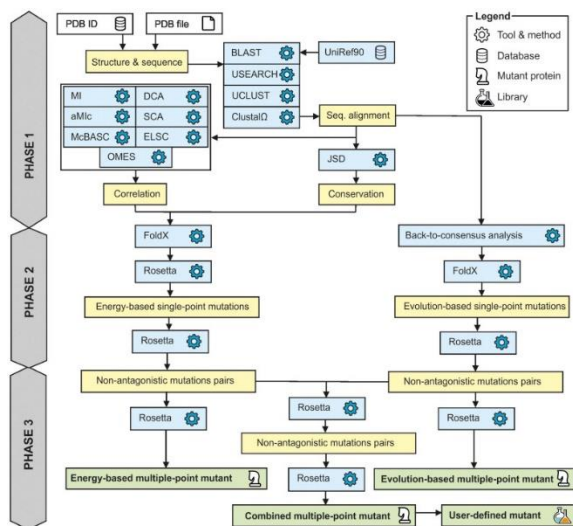


Figure 1: Workflow of the FireProt method.

FireProt^{ASR}

There is a great interest in increasing proteins' stability to widen their usability in numerous biomedical and biotechnological applications. However, native proteins cannot usually withstand the harsh industrial environment, since they are evolved to function under mild conditions. Ancestral sequence reconstruction is a well-established method for deducing the evolutionary history of genes. Besides its applicability to discover the most probable evolutionary ancestors of the modern proteins, ancestral sequence reconstruction has proven to be a useful approach for the design of highly stable proteins. Recently, several computational tools were developed, that make the ancestral reconstruction algorithms accessible to the community, while leaving the most crucial steps of the preparation of the input data on users' side. FireProt^{ASR} aims to overcome this obstacle by constructing a fully automated workflow, allowing even the unexperienced users to obtain ancestral sequences based on a sequence query as the only input. FireProt^{ASR} is complemented with an interactive, easy-to-use web interface and is freely available at <https://loschmidt.chemi.muni.cz/fireprotasr/>.

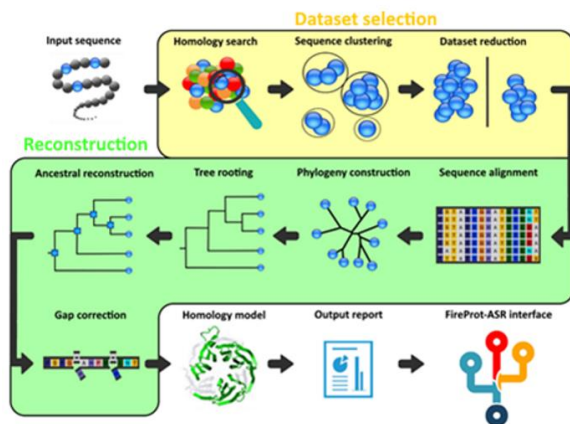


Figure 2: Workflow of the FireProt^{ASR} calculation.

FireProt^{DB}

The majority of naturally occurring proteins have evolved to function under mild conditions inside the living organisms. One of the critical obstacles for the use of proteins in biotechnological applications is their insufficient stability at elevated temperatures or in the presence of salts. Since experimental screening for stabilizing mutations is typically laborious and expensive, *in silico* predictors are often used for narrowing down the mutational landscape. The recent advances in machine learning and artificial intelligence further facilitate the development of such computational tools. However, the accuracy of these predictors strongly depends on the quality and amount of data used for training and testing, which have often been reported as the current bottleneck of the approach. To address this problem, we present a novel database of experimental thermostability data for single-point mutants FireProt^{DB}. The database combines the published datasets, data extracted manually from the recent literature, and the data collected in our laboratory. Its user interface is designed to facilitate both types of the expected use: (i) the interactive explorations of individual entries on the level of a protein or mutation and (ii) the construction of highly customized and machine learning-friendly datasets using advanced searching and filtering. The database is freely available at <https://loschmidt.chemi.muni.cz/fireprotdb>.

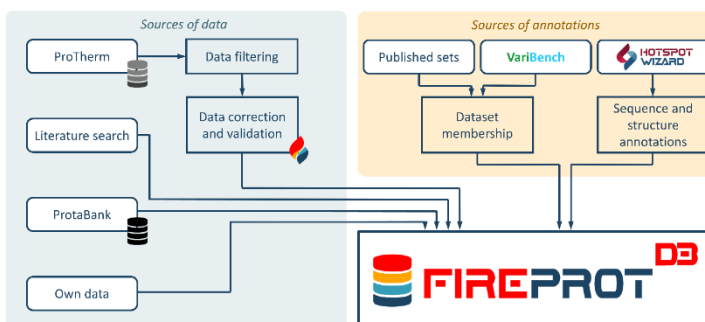


Figure 3: Schematic representation of the data stored in the FireProt^{DB} database.

HotSpotWizard

HotSpot Wizard 2.0 is a web server for automated identification of hot spots and design of smart libraries for engineering proteins' stability, catalytic activity, substrate specificity and enantioselectivity. The server integrates sequence, structural and evolutionary information obtained from 3 databases and 20 computational tools. Users are guided through the processes of selecting hot spots using four different protein engineering strategies and optimizing the resulting library's size by narrowing down a set of substitutions at individual randomized positions. The only required input is a query protein structure. The results of the calculations are mapped onto the protein's structure and visualized with a JSmol applet. HotSpot Wizard lists annotated residues suitable for mutagenesis and can automatically design appropriate codons for each implemented strategy. Overall, HotSpot Wizard provides comprehensive annotations of protein structures and assists protein engineers with the rational design of site-specific mutations and focused libraries. It is freely available at <http://loschmidt.chemi.muni.cz/hotspotwizard>.

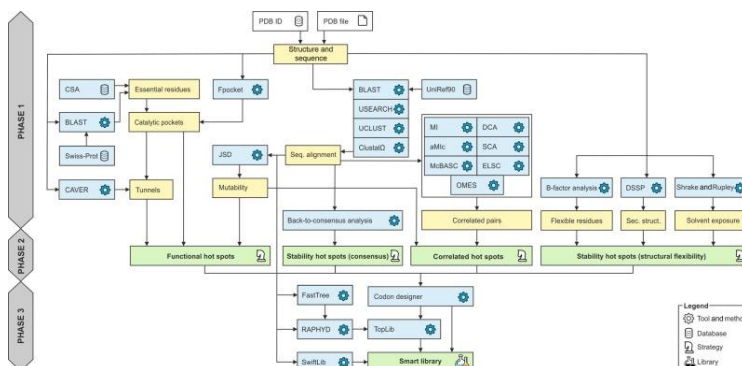


Figure 4: Workflow of the HotSpotWizard service.

Concluding remarks

Stable proteins are utilized in various medical and biotechnological applications. However, native proteins have evolved to function in very mild conditions. Therefore, there is an increasing interest in improving protein stability by introducing mutations into the sequences of modern proteins. However, the saturation mutagenesis of all possible mutations is still far out of reach for many academic laboratories, creating the need for fast and reliable computational approaches. In the recent years, a plethora of computational tools was designed to deal with such a task, falling into one of the three main categories: i) tools based on force-field calculations, ii) tools utilizing the evolutionary information extracted from the set of homolog sequences, and iii) models built on top of the existing experimental data with the use of the modern machine learning methods.

The steady growth of the computational resources allowed for a comprehensive analysis of the mutational space, while the accuracy of stability-predicting methods is currently well-sufficient for the prioritization of experimentally validated mutations. Thus, *in silico* approaches are reducing the need for expensive and laborious laboratory experiments. However, most of the existing methods are viable only for predicting the single-point mutations with only a negligible effect on protein stability, while the construction of the multiple-point mutants is more complicated due to the possible occurrence of the antagonistic effects.

In this Thesis, several computational tools were presented to deal with designing stable multiple-point mutants. FireProt is a fully automated hybrid workflow that combines both energy- and evolution-based approaches in its calculation core. The tool utilizes sequence information, such as conservation and correlation of the amino acids in the MSA, as an initial filter to exclude those risky regions from the further calculation. Force-field approaches are then employed to select a pool of the potentially stable single-point mutations, which are then combined while eliminating most of the antagonistic effects by evaluating all the mutations' pairs. The second approach, FireProt^{ASR},

is based on the idea that the ancestral proteins were significantly more stable than their extant counterparts. It is a fully automated workflow that allows users to utilize ancestral sequence reconstruction for their proteins without the deep knowledge of the essential bioinformatics tools and the biological system. FireProt^{ASR} deals with all steps of the ancestral reconstruction, including the search for the biologically relevant homolog sequences, construction of the MSA and phylogenetic tree, rooting of the tree without the need to specify its outgroup and finally the reconstruction of the ancestral sequences together with the identification of the ancestral gaps.

As the introduction of the stabilizing mutations into the protein structure often causes deterioration of other protein properties, the protein engineering tool HotSpotWizard was designed to add another level of abstraction. HotSpotWizard allows observing the protein by many different criteria, including its conservation and flexibility. Moreover, it provides the visualization of the sites and tunnels that are crucial for the function of the protein of interest. Stabilizing mutations designed by other methods can be analyzed in the HotSpotWizard tool to consider their position within a tertiary structure and the distance of those mutations from the sites essential for protein function. Such an analysis can unearth mutations that could (while stabilizing) compromise proteins activity and other properties, and therefore removing such a mutation could lead to the safer design of the engineered variant.

Finally, the work presented in this Thesis takes a stance on the current unsatisfactory situation surrounding the storage and management of the experimental data that are crucial for the training and validation of the computational tools based on the machine learning approaches. FireProt^{DB} is a comprehensive database of a protein stability data, supplemented with a sophisticated search engine and expanded by various annotations from the sequence and structural databases.

In conclusion, this Thesis presents a set of methods that aim to ease the engineering of highly stable multiple-point mutants, while providing users with a further analysis of the designed protein by

considering other factors such as protein flexibility and location of the functional sites. Furthermore, it aims to simulate further improvement of the protein stability predictors by providing the research community with easy access to reliable experimental data.

In the future, the plan is to utilize the new high-quality dataset that was compiled for FireProt^{DB} to train a novel machine learning-based predictor of the effect of mutations on protein stability. This novel predictor would not be just a simple implementation of some of the standard machine learning techniques (e.g., SVM, RF), but rather a more complex multi-agent system that would focus more deeply on the mutations that are hard to predict by the existing predictors such as charge changing mutations located on the protein surface.

References

- [1] "Proteins: Structure and Function | Wiley,"
- [2] H. P. Modarres, M. R. Mofrad, A. Sanati-Nezhad, *RSC Adv.* **2016**, *6*, 115252.
- [3] R. Kurahashi, S. Tanaka, K. Takano, *J. Biosci. Bioeng.* **2019**, *128*, 405.
- [4] "Defying the activity–stability trade-off in enzymes: taking advantage of entropy to enhance activity and thermostability: Critical Reviews in Biotechnology: Vol 37, No 3,"
- [5] H. Yu, P. A. Dalby, *Proc. Natl. Acad. Sci.* **2018**, *115*, E11043.
- [6] K. M. Polizzi, A. S. Bommarius, J. M. Broering, J. F. Chaparro-Riggers, *Curr. Opin. Chem. Biol.* **2007**, *11*, 220.
- [7] "Thermostable Variants of Cocaine Esterase for Long-Time Protection against Cocaine Toxicity | Molecular Pharmacology,"
- [8] A. S. Bommarius, M. F. Paye, *Chem. Soc. Rev.* **2013**, *42*, 6534.
- [9] H. J. Wijma, R. J. Floor, D. B. Janssen, *Curr. Opin. Struct. Biol.* **2013**, *23*, 588.
- [10] "Enhancing the Thermal Robustness of an Enzyme by Directed Evolution: Least Favorable Starting Points and Inferior Mutants Can Map Superior Evolutionary Pathways - Gumulya - 2011 - ChemBioChem - Wiley Online Library,"
- [11] M. Musil, H. Konegger, J. Hon, D. Bednar, J. Damborsky, *ACS Catal.* **2019**, *9*, 1033.
- [12] C. Nick Pace, J. M. Scholtz, G. R. Grimsley, *FEBS Lett.* **2014**, *588*, 2177.
- [13] T. Lazaridis, M. Karplus, *Curr. Opin. Struct. Biol.* **2000**, *10*, 139.
- [14] D. Seeliger, B. L. de Groot, *Biophys. J.* **2010**, *98*, 2309.
- [15] R. Guerois, J. E. Nielsen, L. Serrano, *J. Mol. Biol.* **2002**, *320*, 369.
- [16] J. Mendes, R. Guerois, L. Serrano, *Curr. Opin. Struct. Biol.* **2002**, *12*, 441.
- [17] "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design | Journal of Chemical Theory and Computation,"
- [18] Y. Dehouck, J. M. Kwasigroch, D. Gilis, *BMC Bioinformatics* **2011**, *12*, 151.
- [19] Y. Dehouck, D. Gilis, M. Rooman, *Biophys. J.* **2006**, *90*, 4010.
- [20] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, *Nucleic Acids Res.* **2005**, *33*, W382.
- [21] K. P. Kepp, **2015**.
- [22] N. J. Christensen, K. P. Kepp, *J. Chem. Inf. Model.* **2012**, *52*, 3028.
- [23] J. A. Davey, A. M. Damry, C. K. Euler, N. K. Goto, R. A. Chica, *Structure* **2015**, *23*, 2011.
- [24] K. Beerens, S. Mazurenko, A. Kunka, S. M. Marques, N. Hansen, M. Musil, R. Chaloupkova, J. Waterman, J. Brezovsky, D. Bednar, Z. Prokop, J. Damborsky, *ACS Catal.* **2018**, *8*, 9420.
- [25] N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, *PLOS Comput. Biol.* **2008**, *4*, e1000002.
- [26] H. Arabnejad, M. Dal Lago, P. A. Jekel, R. J. Floor, A.-M. W. H. Thunnissen, A. C. Terwisscha van Scheltinga, H. J. Wijma, *Prot. Eng. Des. Sel.* **2017**, *30*, 175.
- [27] M. Musil, J. Stourac, J. Bendl, J. Brezovsky, Z. Prokop, J. Zendulka, T. Martinek, D. Bednar, J. Damborsky, *Nucleic Acids Res.* **2017**, *45*, W393.

- [28] A. Broom, Z. Jacobi, K. Trainor, *J. Biol. Chem.* **2017**, *292*, 14349.
- [29] J. Bush, G. I. Makhataдзе, *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 2027.
- [30] P. B. Stranges, B. Kuhlman, *Protein Sci.* **2013**, *22*, 74.
- [31] L. Wickstrom, E. Gallicchio, R. M. Levy, *Proteins Struct. Funct. Bioinforma.* **2012**, *80*, 111.
- [32] G. Thiltgen, R. A. Goldstein, *PLOS ONE* **2012**, *7*, e46084.
- [33] O. Buß, J. Rudat, K. Ochsenreither, *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25.
- [34] E. H. Kellogg, A. Leaver-Fay, D. Baker, *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 830.
- [35] K. A. Barlow, S. Ó Conchúir, S. Thompson, P. Suresh, J. E. Lucas, M. Heinonen, T. Kortemme, *J. Phys. Chem. B* **2018**, *122*, 5389.
- [36] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, E. E. Abola, *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 1078.
- [37] H. Fan, A. E. Mark, *Proteins Struct. Funct. Bioinforma.* **2003**, *53*, 111.
- [38] A. Kuzmanic, N. S. Pannu, B. Zagrovic, *Nat. Commun.* **2014**, *5*, 3220.
- [39] B. J. Sullivan, T. Nguyen, V. Durani, D. Mathur, S. Rojas, M. Thomas, T. Syu, T. J. Magliery, *J. Mol. Biol.* **2012**, *420*, 384.
- [40] C. Jäckel, J. D. Bloom, P. Kast, F. H. Arnold, *J. Mol. Biol.* **2010**, *399*, 541.
- [41] "Consensus protein design | Protein Engineering, Design and Selection | Oxford Academic,"
- [42] M. Lehmann, D. Kostrewa, M. Wyss, R. Brugger, A. D'Arcy, L. Pasamontes, A. P. G. M. van Loon, *Protein Eng. Des. Sel.* **2000**, *13*, 49.
- [43] E. Vázquez-Figueroa, J. Chaparro-Riggers, A. S. Bommarium, *Chembiochem Eur. J. Chem. Biol.* **2007**, *8*, 2295.
- [44] G. K. A. Hochberg, J. W. Thornton, *Annu. Rev. Biophys.* **2017**, *46*, 247.
- [45] V. A. Risso, J. A. Gavira, E. A. Gaucher, J. M. Sanchez-Ruiz, *Proteins Struct. Funct. Bioinforma.* **2014**, *82*, 887.
- [46] Z. Yang, *PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood*, **1997**.
- [47] "Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology | Science,"
- [48] L. Folkman, B. Stantic, A. Sattar, Y. Zhou, *J. Mol. Biol.* **2016**, *428*, 1394.
- [49] S. Teng, A. K. Srivastava, L. Wang, *BMC Genomics* **2010**, *11*, S5.
- [50] L.-T. Huang, M. M. Gromiha, S.-Y. Ho, *Bioinformatics* **2007**, *23*, 1292.
- [51] "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric,"
- [52] C. X. Ling, V. S. Sheng, **n.d.**, 8.
- [53] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.* **2002**, *16*, 321.
- [54] Haibo He, Yang Bai, E. A. Garcia, Shutao Li, *2008 IEEE Int. Jt. Conf. Neural Netw. IEEE World Congr. Comput. Intell.*, IEEE, China, **2008**, pp. 1322–1328.
- [55] R. B. Rao, G. Fung, R. Rosales, in *Proc. 2008 SIAM Int. Conf. Data Min.*, Society For Industrial And Applied Mathematics, **2008**, pp. 588–596.
- [56] A. Goldenzweig, M. Goldsmith, S. E. Hill, O. Gertman, P. Laurino, Y. Ashani, O. Dym, T. Unger, S. Albeck, J. Prilusky, R. L. Lieberman, A. Aharoni, I. Silman, J. L. Sussman, D. S. Tawfik, S. J. Fleishman, *Mol. Cell* **2016**, *63*, 337.

Personal information

Name, Title	Miloš Musil, Ing.
Address	Starohorská 37, Oslavany, 664 12
Telephone	+420 736 164 330
E-mail	imusilm@fit.vutbr.cz
Nationality	Czech
Date of birth	24.6.1991

Education

2015 – present	<i>PhD's degree study</i> University: Faculty of Information Technology, Brno University of Technology Study programme: Computer Science and Engineering Date of acceptance: 26.6.2015 Date of enrollment to the first semester: 3.9.2015 Doctoral thesis: Advanced bioinformatic tools for identification, annotation and engineering of proteins Supervisors: Doc. Ing. Jaroslav Zendulka, Csc. And Prof. Mgr. Jiří Damborský, Dr.
2013 – 2015	<i>Master's degree study</i> University: Faculty of Information Technology, Brno University of Technology Study programme: Bioinformatics and Biocomputing Master thesis: Prediction of the effect of amino acid substitutions on protein function
2010 – 2013	<i>Bachelor's degree study</i> University: Faculty of Information Technology, Brno University of Technology Study programme: Information technology Bachelor thesis: Interactive application in Unreal Script

Internships & Academic stays

2017	Westfälische Wilhelms-Universität Münster – Institute for Evolution and Biodiversity (Germany) Topic: Ancestral sequence reconstruction as a protein stabilization method
------	---

Honours & awards

2013	Dean's award for very good study results during bachelor studies and very good knowledge during state final examination
------	--

2015	Dean's award for very good study results during master studies and very good knowledge during state final examination
2015	Dean's award for very good master's thesis
2018	Young Scientist Award , ProtStab2018
2018	Joseph Fourier Award (2 nd place + IT4Innovation award)
2020	PredictSNP2 : 10% of most cited articles published in 2016

Professional experience

2015 – present	Research Assistant <i>International Clinical Research Centre</i> St. Anne's University Hospital Brno, Brno, Czech Republic
2014 – present	Research Assistant <i>Loschmidt Laboratories</i> Department of Experimental Biology, Research Centre for Toxic Compounds in the Environment, Faculty of Science, Masaryk University, Brno, Czech Republic
2014 – present	PhD student Faculty of Information Technology, Brno University of Technology, Czech Republic
2011 – 2013	Research Assistant <i>Natural Language Processing group</i> Department of Computer Graphics and Multimedia, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

Teaching activities

2014 – present	Bioinformatics practices (VUT & MU)
2014 – present	Supervision of the bachelor and diploma theses 1x project praxes 2x bachelor thesis (1 in progress) 4x diploma thesis (3 in progress)
2020	Lecturer at 1st hands-on computational enzyme design course

Conferences & Proceedings

2015	Student conference EXCEL@FIT , Brno, Czech Republic <i>RAPHYD: Predictor of the effect of amino acid substitutions on protein function</i>
2016	ENBIK2016 , Loučeň, Czech Republic <i>FireProt: výpočetní platforma pro návrh termostabilních vícebodových mutantů</i>
2016	CSMB2016 , Prague, Czech Republic

	<i>PredictSNP: A family of tools for the prediction of the effect of mutations on human health</i>
2017	ISMB/ECCB2017 , Prague, Czech republic <i>FireProt: Computational design of thermostable multiple-point mutants</i>
2018	ProtStab2018 , Vilnius, Lithuania <i>FireProt: a web server for automated design of thermostable multiple-point mutants</i>
2018	ENBIK2018 , Bystřice nad Pernštejnem, Czech Republic <i>An automated design of thermostable multiple-point mutants</i>
2019	ISMB/ECCB2019 , Basilej, Switzerland <i>FireProt-ASR: automatized workflow for ancestral sequence reconstruction</i>

Software & databases

Software	Development and maintenance of PredictSNP2 Development of HotSpotWizard 2.0 Development and maintenance of FireProt Development and maintenance of FireProt-ASR
Databases	Development of FireProt-DB

Publications

Published	BENDL J, MUSIL M, ZENDULKA J, DAMBORSKY J, BREZOVSKY J. PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. <i>PLoS Computational Biology</i> . 2016, 12, e1004962. BENDL J, STOURAC J, SEBESTOVA E, VAVRA O, MUSIL M, BREZOVSKY J, DAMBORSKY J. HotSpotWizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. <i>Nucleic Acids Research</i> . 2016, 44(W1), W479-W487. 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. <i>Nucleic Acids Research</i> . 2017, 45(W1), W393-W399. BEERENS K, MAZURENKO S, KUNKA A, MARQUES S, HANSEN N, MUSIL M, CHALOUPKOVA R, WATERMAN J, BREZOVSKY
-----------	--

Accepted	<p>J, BEDNAR D, PROKOP Z, DAMBORSKY J. Evolutionary analysis as a powerful complement to energy calculations for protein stabilization. <i>ACS Catalysis</i>. 2018, 8, 9420-9428.</p> <p>MUSIL M, KONEGGER H, HON J, BEDNAR D, DAMBORSKY J. Computational design of stable and soluble biocatalysts. <i>ACS Catalysis</i>. 2018, 9, 1033-1054.</p> <p>MUSIL M, KHAN RT, BEIER A, STOURAC J, KONEGGER H, DAMBORSKY J, BEDNAR D. FireProt-ASR: web server for fully automated ancestral sequence reconstruction. <i>Briefings in bioinformatics</i>. 2020.</p> <p>STOURAC J, DUBRAVA J, MUSIL M, HORACKOVA J, DAMBORSKY J, MAZURENKO S, BEDNAR D. FireProt-DB: Database of Manually Curated Protein Stability Data. <i>Nucleic Acids Research</i>. 2020.</p>
Under review	<p>KHAN RT, MUSIL M, STOURAC J, DAMBORSKY J, BEDNAR D. Fully automated ancestral sequence reconstruction using FireProt-ASR. <i>Current protocols in bioinformatics</i>. 2020.</p> <p>PLANAS-IGLESIAS J, MARQUES S, PINTO G, MUSIL M, STOURAC J, BEDNAR D, DAMBORSKY J. Computational design of enzymes for biotechnological applications. <i>Biotechnology advances</i>. 2020.</p>

International courses

2016	Summer school of protein engineering, Brno, Czech Republic
2017	Machine learning for biologists, San Michele All'Adige, Italy