MASARYK UNIVERSITY

FACULTY OF SCIENCE

RECETOX

Advanced Data Modelling for Protein Engineering

Habilitation Thesis

Stanislav Mazurenko

Brno, 2025

Acknowledgements

I would like to express my sincere gratitude to my former supervisors for shaping my scientific outlook and providing invaluable mentorship throughout my research journey. I am especially indebted to Prof. Jiří Damborský for teaching me not only how to be a scientist but also how to lead a scientific team and guide others in exploring new and exciting directions in science.

I would also like to extend my heartfelt thanks to all my colleagues at the Loschmidt Laboratories and to my co-authors from around the world for our fruitful collaborations and inspiring interactions. Their tremendous support, illuminating insights, and passionate enthusiasm have greatly contributed to the success of my research.

Finally, I wish to thank my parents and my sister for their unconditional love and unwavering support.

Contents

Commentary

This habilitation thesis is a compilation of publications from the domain of protein engineering authored or co-authored by Stanislav Mazurenko. The methods introduced in the publications were developed between the years 2015 and 2025 mainly at Masaryk University, Brno. The main motivation for the methods was to design computational tools for the analysis and modelling of complex biological data. Such a toolset enables protein engineers to select protein targets, plan experiments, analyse the collected data, and formulate a hypothesis about biological phenomena of interest in a more informed way, drastically enlarging the space of possibilities while reducing the experimental effort required to explore this space. The developed methods focus on two main approaches: (1) low-parameter modelling of the data, primarily based on physical principles, and (2) machine learning-based modelling, which leverages existing data sets to identify useful patterns in the data.

The thesis is divided into two parts. The first part provides a commentary on the contributions the thesis is based on. In Chapter 1, we motivate the need for data modelling and computational tools to study proteins. Then Chapter 2 discusses the bottom-up low-parameter modelling, which leverages biophysical principles. The alternative, top-down modelling that is based on machine learning methods, is presented in Chapter 3. Finally, Chapter 4 summarises the presented methods and proposes directions for future research. The second part consists of 11 publications in which these contributions were introduced. In the first two papers, the author was responsible for the design of the study, developing the mathematical framework, and programming the algorithms. In the remaining articles, the author was designing and overseeing data modelling and machine learning contributions to the corresponding studies.

The list of publications is as follows (the asterisk indicates corresponding authors):

1. **Mazurenko, S.**, Kunka, A., Beerens, K., Johnson, C. M., Damborsky, J., & Prokop, Z. (2017). Exploration of protein unfolding by modelling calorimetry data from reheating. *Scientific Reports*, 7(1), 16321.
2. **Mazurenko, S.**, Stourac, J., Kunka, A., Nedeljković, S., Bednar, D., Prokop, Z., & Damborsky, J. (2018). CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Research*, *46*(W1), W344-W349.
3. Stourac, J., Dubrava, J., Musil, M., Horackova, J., Damborsky, J., **Mazurenko, S.***, & Bednar, D.* (2021). FireProtDB: database of manually curated protein stability data. *Nucleic Acids Research*, *49*(D1), D319-D324.
4. Kunka, A., Lacko, D., Stourac, J., Damborsky, J., Prokop, Z.*, & **Mazurenko, S.*** (2022). CalFitter 2.0: Leveraging the power of singular value decomposition to analyse protein thermostability. *Nucleic Acids Research*, *50*(W1), W145-W151.
5. Velecký, J., Hamsikova, M., Stourac, J., Musil, M., Damborsky, J., Bednar, D.*, & **Mazurenko, S.*** (2022). SoluProtMutDB: A manually curated database of protein solubility changes upon mutations. *Computational and Structural Biotechnology Journal*, *20*, 6339-6347.
6. Khan, R. T., Pokorna, P., Stourac, J., Borko, S., Arefiev, I., Planas-Iglesias, J., Dobias, A., Pinto, G., Szotkowska, V., Sterba, J., Slaby, O., Damborsky, J., **Mazurenko, S.***, & Bednar, D.*

(2024). A computational workflow for analysis of missense mutations in precision oncology. *Journal of Cheminformatics*, *16*(1), 86.
7. Marques, S. M., Kouba, P., Legrand, A., Sedlar, J., Disson, L., Planas-Iglesias, J., Sanusi, Z., Kunka, A., Damborsky, J., Pajdla, T., Prokop, Z., **Mazurenko, S.***, Sivic, J.*, & Bednar, D.* (2024). CoVAMPnet: comparative Markov state analysis for studying effects of drug candidates on disordered biomolecules. *JACS Au*, *4*(6), 2228-2245.
8. Velecký, J., Berezný, M., Musil, M., Damborsky, J., Bednar, D.*, & **Mazurenko, S.*** (2024). BenchStab: a tool for automated querying of web-based stability predictors. *Bioinformatics*, *40*(9), btae553.
9. Vavra, O., Tyzack, J., Haddadi, F., Stourac, J., Damborsky, J., **Mazurenko, S.***, Thornton J.*, Bednar, D.* (2024). Large-scale annotation of biochemically relevant pockets and tunnels in cognate enzyme–ligand complexes. *Journal of Cheminformatics*, *16*(1), 114.
10. Kohout, P., Vasina, M., Majerova, M., Novakova, V., Damborsky, J., Bednar, D., Marek, M., Prokop, Z.*& **Mazurenko, S.*** (2025). Engineering dehalogenase enzymes using variational autoencoder-generated latent spaces and microfluidics. *JACS Au*, *5*(2), 838-850.
11. Khan, R. T., Kohout, P., Musil, M., Rosinska, M., Damborsky, J., **Mazurenko, S.***, & Bednar, D.* (2025). Anticipating protein evolution with successor sequence predictor. *Journal of Cheminformatics*, *17*(1), 34.

1. Introduction

1.1. Protein engineering and its impact

Proteins are fundamental biomolecules performing a vast array of essential functions in living organisms, from acting as the building blocks of cells and tissues to catalysing chemical reactions and regulating physiological processes (1). Understanding protein structure, dynamics, and functions thus provides crucial insights into biological mechanisms, making the study of proteins a cornerstone of life sciences and biomedicine. Existing protein databases, such as UniProt and Protein Data Bank, reveal unparalleled diversity of proteins by cataloguing hundreds of millions of protein sequences and hundreds of thousands of protein structures from a broad range of organisms, representing the vast evolutionary landscape (2, 3). Tapping into this protein diversity fuels the development of several domains, including biotechnology, medicine, and synthetic biology. For instance, enzymes from extremophiles, microorganisms thriving in extreme environments, are highly stable and function efficiently under harsh conditions, such as extreme temperatures, pH, or salinity, making them ideal for industrial applications, including biofuel production, pharmaceutical synthesis, and bioremediation (4). Thermostable DNA polymerases have revolutionised PCR, a crucial technique in molecular biology and forensic science (5). Constructing genetic, metabolic, or signalling networks with predictable and controllable properties has made it possible to reprogram cells to produce drugs, biofuels, biomaterials, and fine chemicals (6).

Apart from understanding the function of a particular protein, the knowledge of how proteins change upon mutations in their primary sequence may lead to actionable insights for developing drugs against inherited diseases (7, 8) and biotechnologically relevant protein modifications (9). Protein engineering refers to the process of designing and modifying proteins to alter their functions in the desired direction. By using techniques such as directed evolution, site-directed mutagenesis, and computational modelling, protein engineers can change the amino acid sequence of a given protein, for example, to enhance its stability, activity, or specificity, enabling the development of more effective therapeutic enzymes, environmentally friendly biocatalysts for manufacturing, and novel biomaterials (10). Overall, protein engineering is a key technology driving innovation in biotechnology and synthetic biology.

In our research, we have studied and engineered a range of proteins with potential for applications in biotechnology and medicine, including haloalkane dehalogenases, fibroblast growth factors, amyloid-beta peptides, apolipoprotein E, and staphylokinase. Haloalkane dehalogenases are model enzymes for understanding catalytic mechanisms and engineering bioremediation tools for toxic halogenated compounds (11–13). Fibroblast growth factors are key signalling proteins with crucial roles in development, tissue repair, and cancer research (14). Amyloid-beta peptides are central to Alzheimer’s disease research because of the association between their aggregation into plaques and disruption of neural function (15). The genetic variants of Apolipoprotein E, a protein involved in the metabolism of fats, have a profound impact on several neurological diseases, particularly in modulating the risk of Alzheimer's disease (16). Staphylokinase is a bacterial protein promising for its potent fibrinolytic activity, making it an important candidate for developing thrombolytic therapies (17, 18).

Despite numerous successful protein engineering cases, protein engineering remains a challenging task. Its complexity stems from multiple sources, from the lack of annotated data in protein databases to difficulties associated with experimental data collection and interpretation. Our next section elaborates on those challenges.

1.2. Challenges in engineering proteins

Several groups of challenges can be identified depending on the target area of protein science: protein discovery, protein characterisation, protein engineering, and molecular simulation.

Protein discovery is significantly hindered by the paucity of annotations in databases. Based on the UniProtKB/Swiss-Prot protein knowledgebase statistics 2025_04, fewer than 180,000 protein sequences have evidence of protein existence at the protein or transcript level, which is less than 0.1% of the total number of sequences available in UniProtKB. Fewer than 1.2 million gene ontology annotations have been inferred from experiments[1]. This lack creates an enormous gap between the number of available protein sequences and the information available for search. Therefore, the known protein universe remains largely experimentally unannotated, motivating the use of i*n silico* tools for predicting protein properties (19, 20).

Protein characterisation requires the availability of sufficient quantities of the sample, access to appropriate experimental assays, sufficiently high throughput, and, more importantly, proper data analysis methods to extract the desired property from raw data (21). Assays that enable the measurement of the desired protein property directly are quite rare, and in many cases, one must apply data analysis methods to model the observed experimental signal and determine the underlying parameters by fitting into experimental signals (22). However, a typical profile of a wet lab researcher features limited programming and data analysis skills. Moreover, some parameters are not identifiable from the available measurements, even in the idealised noiseless case, due to the mathematical properties of the governing equations, often leading to data misinterpretations and parameter misestimations (23–25). This highlights the need for the development of more robust parameter estimation protocols and user-friendly tools for data modelling and analysis.

Protein engineering must efficiently explore the vast space of possible amino acid substitutions, an arduous task given that a single mutation can already lead to a misfolded or nonfunctional protein (26, 27). Even if we had a method capable of accurately predicting the effect of a single set of amino acid substitutions on a property of interest in 1 μs, the evaluation of all possible five-point mutants of an “average” 300-amino-acid-long protein would take almost 2000 years[2]. This is primarily the reason why the majority of mutational data comes from single-point mutants (26, 28), and larger steps in a protein sequence require advanced tools for smart navigation in the sequence space.

Finally, molecular simulations are a powerful group of methods to provide insights into protein function (29, 30). However, their capabilities are limited by the complex protein mechanisms,

[1] https://release.geneontology.org/2025-10-10/release_stats/index.html

[2] The total number of possible 5-point mutants for a 300-amno-acid-long protein is 19582837560 position combinations x 20^5 amino acid identities.

intricate interactions with the environment, and the sizes of the systems one is capable of modelling (31). Simulating several microseconds of a molecular dynamics trajectory for a single protein can still take weeks, even on a high-performance cluster, significantly limiting the accessibility of such simulations for protein engineering campaigns. Moreover, the vast amounts of data generated by simulations still require data processing to provide interpretable and actionable insights for protein engineering.

It is therefore clear that a large subset of challenges in protein engineering refers to data generation, analysis, and modelling, in particular, methods for extracting patterns from available data. Those methods can be roughly subdivided into two classes: low-parameter modelling and machine learning. The next section provides a brief introduction to these two different paradigms of modern data modelling approaches used for proteins.

1.3. Two paradigms of data modelling

Historically, modelling of protein-related data, e.g., measurements from enzyme kinetics, protein denaturation or aggregation, or protein-ligand interactions, was performed using equations derived from physics (24). Such equations will describe the time course of the reaction, the properties of interactions, and the outcomes of the process typically by means of differential operators and parameters of the system under study. They are estimated from the data using scientific software, such as KinTek Global Kinetic Explorer or Amylofit (32, 33). Their numbers are typically limited to a few dozen at most, with higher numbers requiring significant simplifications of the models, e.g., linearisation to enable feasible parameter estimation (34, 35). Otherwise, the parameters stop being constrained by available data. In what follows, we will refer to such an approach as **low-parameter modelling**.

An alternative approach, **machine learning**, pursues the opposite: start with equations featuring a large number of parameters without any particular physical interpretation, sometimes in billions, and find any set of their values that will explain the patterns in the data. The important distinction from the previous approach, making such vast fitting useful, is the protocol whereby a part of the data is initially put aside and used only to evaluate the generalisability of the fitted model. In other words, the data model is assessed based on how well it performs with new data. Since the patterns are extracted from the data directly and only very limited prior knowledge of the system is used for modelling, this approach critically depends on the amount and quality of the available data. Thus, although the first attempts at applying machine learning to proteins date back several decades ago (36, 37), they have become increasingly widespread more recently, primarily due to the growing availability of data and computational resources.

Both these approaches, low parameter modelling and machine learning, have their advantages and disadvantages. In the low-data regime, when only a limited set of measurements is available, the data set size may not be enough to learn generalisable patterns via machine learning. Moreover, low parameter modelling is inherently interpretable as the parameters used have physical meaning. However, it struggles in the regimes where the observed property is too complex to be modelled or requires too many equations to be fitted into the measurements reliably. In such cases,

machine learning offers a powerful alternative provided the available dataset is large enough to discern the required relationships.

This work explores both approaches. Our contribution to low-parameter modelling primarily focuses on protein stability and analysis of protein thermal denaturation. Machine learning approaches address a wide range of topics, from predicting the effects of mutations on protein properties to the analysis of molecular dynamics data.

2. Low-Parameter Modelling in Protein Engineering

2.1. General overview

Biological data is notoriously complex and difficult to analyse. Even when experimental conditions are carefully designed and controlled for, protein datasets collected in those experiments almost always require subsequent data analysis to extract useful information. This primarily stems from the inability to directly observe a protein property of interest, such as protein folding and unfolding energy differences, the rate of conversion of a substrate into product, or protein aggregation rates. Most modern experimental techniques rely on measuring proxy signals that are affected by desired quantities, such as folding energies or kinetic rates, and come from fluorescence, absorbance, bioluminescence, electric resistance, or heat flow, to name a few. Hence, one requires data analysis to deconvolute these signals into interpretable insights and parameters.

A typical data analysis workflow for extracting the required quantities from experimental signals unfolds as follows (Fig. 1). First, one suggests a mathematical model that is based on the physical nature of the experiment and consists of several parameters, including the desired quantity. Then, for a given set of initial parameters, the model is used to simulate the signal one should observe in experiments. This simulated signal is compared with the experimentally obtained measurements, and the model parameters are adjusted to account for any discrepancy, usually in an iterative fashion until they converge to a specific set of values. Finally, the desired quantities of interest are calculated based on the fitted parameters.

Such a workflow usually requires specialised knowledge of mathematics, programming, and biophysics, which is typically beyond the skill set of an average wet lab researcher. While sometimes instruments already include software for data processing, its functionality is often limited to only a few simple mathematical models, which are usually insufficient for analysing signals beyond simple standard cases.

One such example is the analysis of protein unfolding data, where software tools can only account for a simple one-step reversible protein unfolding from native to unfolded states (38). Such a simple unfolding model is usually not enough in many protein stabilisation campaigns, in which the shift of unfolding to higher temperatures usually entails irreversible transitions (11). Moreover, multidomain proteins rarely unfold in one step. Thus, in the absence of an off-the-shelf tool to account for more complex behaviour or programming expertise to create a custom-made script, protein engineers are forced to apply incorrect models, compromising the data quality.

In Chapter 2, we will consider the problem of analysing protein thermal unfolding data in more detail. We will first introduce the basic workflow, main formulas and notations, and the challenges in data processing. We will then describe the methods we developed to tackle these challenges. We will conclude this chapter with the outlook.

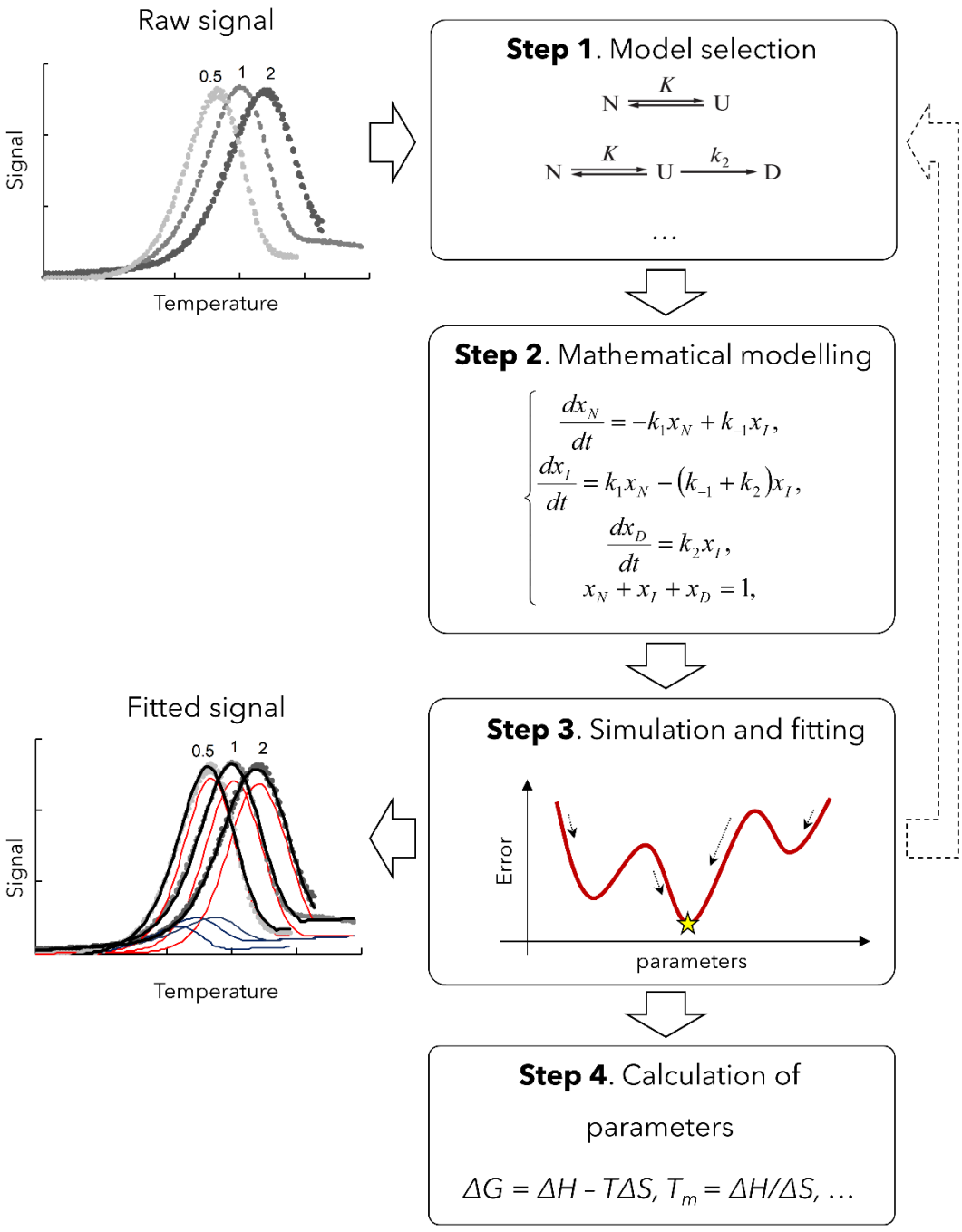


Figure 1. A typical workflow for processing experimental data when studying proteins. An example of the raw data is heat capacity change upon protein unfolding, generated by differential scanning calorimetry at the scan rates of 0.5°C/min, 1°C/min , and 2°C/min. First, a candidate model of unfolding is selected (e.g., one- or two-step unfolding). Second, the corresponding parametrised equations are modelled and programmed (e.g., fractions of states in time). Third, their parameters are optimised to minimise the error of the fit (the yellow star corresponds to the global minimum). Optionally, a different candidate model is selected and fitted to the raw signal. Finally, the quantities of interest (e.g., Gibbs free energy difference ΔG or melting temperature T_m) are derived from the parameters.

2.2. Protein thermostability and unfolding

Naturally occurring proteins have primarily evolved to function in mild conditions of a living cell, limiting their applications for biotechnology (39). Protein engineers generally aim to improve protein stability, and thermostability is their primary target as it is correlated with half-life, expression yield, and activity in the presence of denaturants.

Several techniques are most commonly used to measure protein thermostability. In differential scanning calorimetry (DSC), the native state of the protein is slowly perturbed by gradually increasing temperature, and the difference in the heat capacity between the sample and a reference cell with buffer is recorded (40, 41). This technique is one of the most powerful methods as it records the energetic footprint of unfolding directly, in terms of the amount of heat necessary to unfold a protein. The changes in proteins upon thermal unfolding can also be detected using fluorescence/absorbance spectroscopy, light scattering, and circular dichroism (CD) (42). Such measurements do not give a full energy profile of unfolding, e.g., CD is sensitive to changes in the secondary structure only, but they are still powerful in providing structural insights. Finally, the protein sample can also be perturbed by a rapid increase in temperature, e.g., in temperature jump experiments, and then the kinetic trace of protein transitioning from native to denatured states can be recorded by spectroscopic reading. Such experiments help evaluate the kinetic stability of the protein (43), i.e., the energy barrier separating the native and denatured states, rather than the equilibrium distribution of protein in those states.

All three types of thermostability measurements are important as they provide insights from different perspectives. A common feature of these measurements is the need to apply advanced data analysis techniques to extract meaningful quantities. Indeed, protein engineers are usually interested in identifying the protein melting temperature (T_m, the temperature at which half of the protein in the sample is in the denatured state in equilibrium), unfolding intermediates, and the energies separating those intermediates (ΔG, the Gibbs free energy change). All these quantities are not observed directly and typically must be derived from modelling the signal and curve-fitting (Fig. 1). The next subsection introduces the basics of such data modelling.

2.3. Workflow for data analysis of protein thermostability

The fundamental mathematical framework to deconvolute the signal is the Lumry-Eyring model, given by the following scheme:

[1] $$N \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} I \xrightarrow{k_2} D,$$

in which protein sample undergoes the first reversible transition from the native (N) to intermediate (I) states, characterised by the rate constants k_1 for forward and k_{-1} for reverse reactions, respectively, followed by the second irreversible step at the rate k_2 to the denatured state (D). This model then leads to the following system of ordinary differential equations describing the fraction of the protein in each state:

[2]
$$\begin{cases} \dfrac{dx_N}{dt} = -k_1 x_N + k_{-1} x_I, \\ \dfrac{dx_I}{dt} = k_1 x_N - (k_{-1} + k_2) x_I, \\ \dfrac{dx_D}{dt} = k_2 x_I, \\ x_N + x_I + x_D = 1, \end{cases} \qquad \begin{cases} v\dfrac{dx_N}{dT} = -k_1 x_N + k_{-1} x_I, \\ v\dfrac{dx_I}{dT} = k_1 x_N - (k_{-1} + k_2) x_I, \\ v\dfrac{dx_D}{dT} = k_2 x_I, \\ x_N + x_I + x_D = 1. \end{cases}$$

where x denotes protein fractions in corresponding states. The right-hand side version exploits the fact that in DSC experiments, the temperature (T) changes with a preset scan rate *v:* dT/dt=v. In fact, this dependence of the temperature on time is the major challenge in simulating system (2), as it translates to the time-dependence of the rate constants, typically modelled according to Arrhenius or, equivalently, Eyring laws (44, 45):

[3]
$$k = A\exp\left\{-\frac{E_a}{RT}\right\} = \exp\left\{\frac{E_a}{R}\left(\frac{1}{T_f} - \frac{1}{T}\right)\right\} \qquad \text{or} \qquad k = \frac{k_B T}{h}\exp\left\{-\frac{\Delta G^{\ddagger}}{RT}\right\}.$$

Here, R is the universal gas constant, k_B is Boltzmann's constant, h is Planck's constant. In the Arrhenius model, A (or T_f) and E_a are protein-specific parameters, obtained from fitting the experimental data. The Eyring law model of kinetic rates usually provides a more familiar representation from the transitional state theory by using the Gibbs energy of activation $\Delta G^{\ddagger}$. Since the Gibbs energy consists of enthalpy and entropy components (protein-specific parameters in the Eyring law) according to the equation $\Delta G^{\ddagger}=\Delta H^{\ddagger} - T\Delta S^{\ddagger}$, the connection between the two formalisms becomes clear when the first-order approximation in T is taken:

[4]
$$E_a = \Delta H^{\ddagger} + RT, \qquad A = \frac{eTk_B}{h}\exp\left\{\frac{\Delta S^{\ddagger}}{R}\right\}.$$

In the simplest form, the enthalpy $\Delta H^{\ddagger}$ and entropy $\Delta S^{\ddagger}$ changes are assumed to be temperature-independent constants, derived from the fitting of the data. A more general approach assumes a linear dependence of the enthalpy on temperature with the coefficient ΔC_p. However, for the sake of not overwhelming the readers, we will not go into these details here.

Knowing the fraction of protein in different states allows modelling the observed signal in experiments. For instance, in many kinetic experiments based on fluorescence, one will model the signal as a weighted sum of contributions from individual states:

[5]
$$Signal(T) = f_N x_N(T) + f_I x_I(T) + f_D x_D(T),$$

where each coefficient f is another parameter to be determined from the fitting. Calorimetry modelling is more challenging because the measured signal is the amount of heat absorbed to transition between the states, and in its simplest form for system (1), it is derived as follows:

[6]
$$Signal(T) = f_{baseline} + \frac{dx_I}{dT}\Delta H_I + \frac{dx_D}{dT}\Delta H_D.$$

Substituting equations for the temperature derivatives from system [2] then allows simulating the signal for a given set of protein-specific parameters.

To sum up, the standard protocol of solving the unfolding mechanism consists of the following steps: (i) collect the data, (ii) select a model of unfolding similar to scheme [1]; (iii) set initial parameters (e.g., $\Delta H^{\ddagger}$, $\Delta S^{\ddagger}$ for each rate constant and *f* for each fraction) to reasonable values; (iv) fit the data by iteratively updating the parameters until convergence; and (v) check the fitting and change the model or adjust the starting parameters if the outcome is unsatisfactory. The following subsection elaborates on the main challenges associated with this protocol.

2.4. Gaps in the state of the art and our contribution

Several challenges to proper data analysis for protein thermal denaturation needed to be addressed before our contributions. First, due to the temperature dependence of system [2], no tool was capable of processing signals from DCS, fluorescence, and temperature jump experiments simultaneously. General-purpose but programming-intensive tools for data analysis, such as MATLAB, Origin, or Igor Pro, required programming skills typically beyond those that protein engineers possess. A few software packages could handle different types of thermal denaturation experiments, e.g., KinTek Explorer allowed fitting kinetic traces in temperature jump experiments, MicroCal DSC Origin could process DSC data, and CDpal enabled fitting of CD signals. However, such tools were unable to fit global data from different sources, e.g. equilibrium and kinetic data simultaneously. And separate data analyses may eventually lead to conflicting models of protein unfolding. Moreover, consecutive unfolding steps sometimes overlap significantly and produce an apparent single transition which cannot be resolved by fitting into just one data type. Finally, many independent variables must be introduced in separate data analyses, increasing the uncertainty and risks of overinterpreting the data. These issues could be avoided in the global data fitting, but no tool was available for that purpose.

The second challenge concerned the models. The unfolding of large proteins may follow more complex schemes with multiple intermediate states and/or their conformations than scheme [1]. In such cases, each step of unfolding, either reversible or irreversible, must be modelled using rate constants according to the same principles as in the Lumry-Eyring model. Yet, the software on DSC or CD devices, for example, could only handle a very basic fully reversible model, in which k_2 was assumed zero, and the scan rate *v* was slow enough to ensure equilibrium for each temperature, with the equilibrium unfolding constant K:

[7] $$K(T) = \frac{k_1}{k_{-1}} = \exp\left\{-\frac{\Delta G_1^{\ddagger} - \Delta G_{-1}^{\ddagger}}{RT}\right\} = \exp\left\{-\frac{\Delta H}{RT}\left(1 - \frac{T}{T_m}\right)\right\}.$$

In such a simplified setting, only two parameters suffice to model the DSC signal: the enthalpy change upon unfolding $\Delta H = \Delta H_1^{\ddagger} - \Delta H_{-1}^{\ddagger}$, and the melting temperature T_m at which K=1, i.e., half of the protein in the sample is the denatured state. This simplification allows replacing the integration of system [2] by modelling each fraction as a simple equation: $x_N = 1/(1+K)$ and $x_D = K/(1+K)$. However, many proteins do not produce unfolding curves that can be adequately described by such a simplified approach, e.g., due to the presence of unfolding intermediates. Therefore, simple models

provided by the manufacturers of devices were not enough and one had to search the literature for more complex models and program them from scratch.

The third challenge was the inability to distinguish between fully and partially reversible models of unfolding. Protein stabilisation campaigns aim to shift unfolding to high temperature ranges, where the transition often becomes irreversible. Therefore, one must know if the reversibility assumption can be made in data processing and if the rates are fast enough to enable the equilibrium approximation. Protein engineers often apply cooling and reheating of the sample to gather some insights. However, no data model was available to incorporate such measurements into the global data fitting.

Finally, the fourth challenge was specific to the spectroscopic signal. Quite often, multiple wavelengths are recorded during the data collection. Understanding which wavelengths are most informative is critical for further data analysis. A naïve approach is to select a single wavelength to record the signal, risking losing much information contained in other wavelengths. A more advanced approach applies the singular value decomposition to the matrix of measurements at different time points and wavelengths and selects the linear combinations of different wavelengths corresponding to the largest singular values. However, such decomposition was also beyond reach for protein engineers without programming skills, as no suitable analytical toolbox was available for researchers without an advanced data analysis background.

Our main contributions address these four limitations in a series of papers. First, we developed the workflow and the web server CalFitter, capable of globally fitting the data from different types of protein thermal denaturation experiments (38). Second, we collected from the literature and implemented 14 different models of protein unfolding known to date. These are now available for users dealing with complex unfolding profiles. Third, we derived a mathematical framework for modelling the data from reheating experiments and incorporated it in the CalFitter web server (46). Finally, with our second version of CalFitter, we introduced the singular value decomposition analysis as an optional first step in data processing (47). The overall pipeline of the tool is shown in Figure 2, and all these developments are accessible at https://loschmidt.chemi.muni.cz/calfitter/.

2.5. Outlooks

Despite the aforementioned developments in the data modelling for protein thermal denaturation in our studies, several directions can be pursued to advance the field forward. One of the most significant obstacles is the need to supply reasonable starting parameters in any data-fitting campaign. While several heuristics exist to help set up parameters for simple mechanisms, more complex unfolding mechanisms require hours of trial and error in selecting starting parameters, fitting the models, simulating the signal, and adjusting the starting parameters. Moreover, protein denaturation often incorporates aggregation, and adding steps for protein aggregation might better elucidate the unfolding mechanism. However, aggregation is notoriously difficult to model mathematically. Finally, the application of recent advances in machine learning to the analysis of protein denaturation data has largely been unexplored and may substantially simplify the workflow.

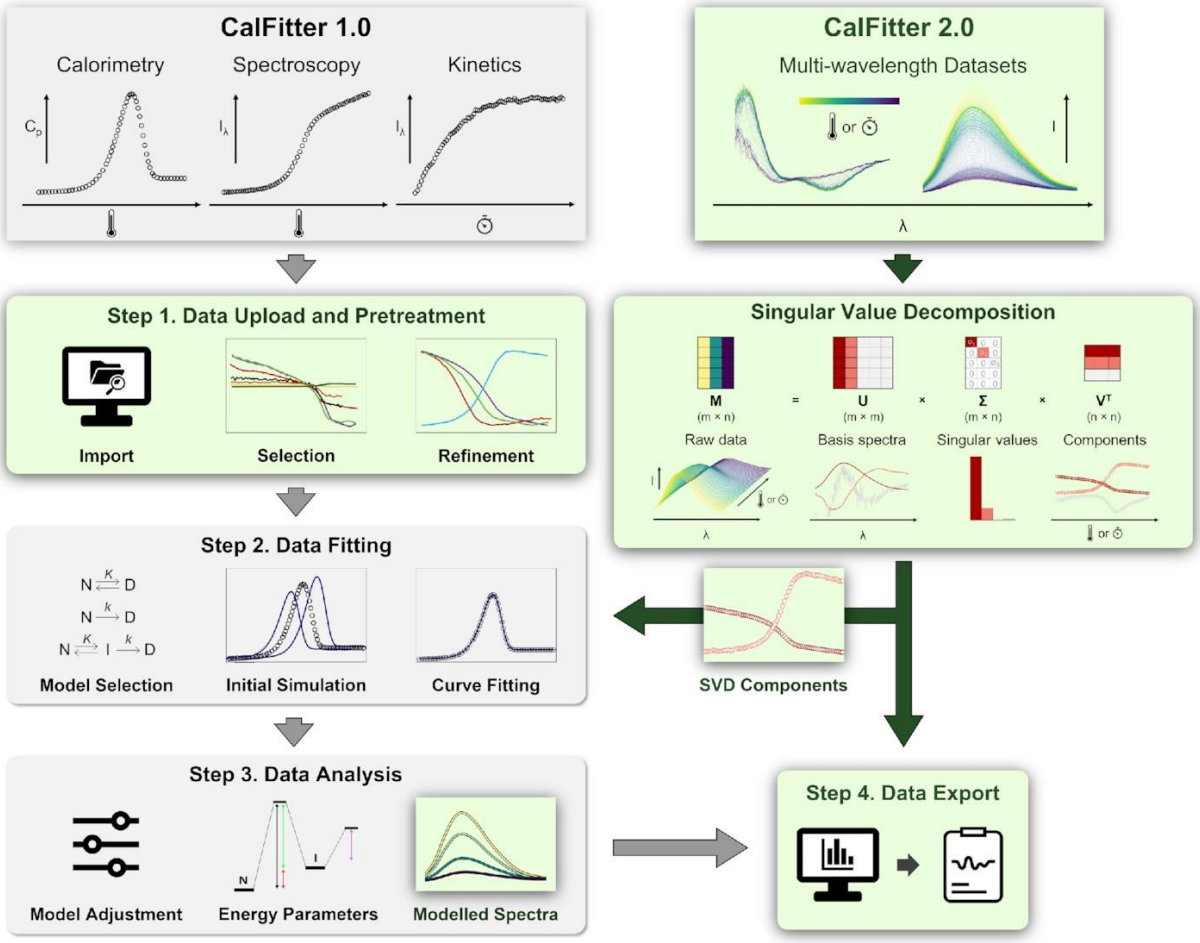


Figure 2. The overall pipeline of CalFitter. The webserver was a first-of-the-kind tool to globally fit the data coming from different types of protein thermal denaturation experiments: calorimetry, spectroscopy, and kinetics measurements. Once the data are uploaded, the interactive user interface offers several unfolding mechanisms to select from and fit into the data. The final step provides the calculation of the main characteristics of unfolding with their confidence intervals and different visualisations. The second release (green boxes) additionally introduced a simpler data upload module and, more importantly, an extra step enabling the singular value decomposition. The entire web server thus allows scientist without a data analysis background to process their data in a user-friendly graphical interface. The figure is adopted from (47).

3. Machine Learning in Protein Engineering

3.1. General overview

The previous chapter introduced data modelling primarily based on biophysical principles. In those approaches, one assembles equations to model the data, solves them, simulates an expected signal, compares it with the observed signal step by step, and changes the parameters of the models to match the data better. The equations used are derived from basic laws of physics, such as protein thermodynamics and kinetics, Arrhenius formalism, and the Eyring law. These methods are powerful in providing mechanistic insights into the underlying physical processes, such as protein unfolding pathways, reaction mechanisms, or aggregation pathways. Yet, they are limited in their ability to grasp complex patterns in the data, since every unexplained variability in data must be carefully described mathematically and properly modelled alongside the underlying model. In contrast, in many protein engineering tasks, the patterns in the data can be too subtle, intricate, or time-consuming to be modelled explicitly with a set of well-defined equations. For instance, while in theory we can simulate the effect of individual amino acid substitutions on protein function (to a certain extent) based on physical principles, scaling up such modelling to large datasets is prohibitively expensive. Moreover, accurately modelling certain effects in simulations, such as enthalpic contributions to free energies or protein interactions with the environment, is currently beyond the reach of existing approaches (31).

More recently, an alternative group of methods for modelling biological data has emerged whereby the equations are generated at scale, within a specific general mathematical framework, with little or no use of laws of physics (19). Such equations typically have simple structures but consist of many more parameters (often millions, in contrast to several dozens, as is the case, for example, of CalFitter models) whose values are determined by fitting much larger sets of data, often thousands of instances. This parameter determination, often referred to as training, is implemented according to a set of strict rules to ensure that the final equations capture generalisable patterns in the data instead of simple memorisation of the data. The domain that develops such methods is called machine learning. This chapter introduces the main principles of machine learning and discusses our contribution to advancing machine learning methods for various protein engineering tasks.

3.2. Machine learning methodology

Machine learning (ML) primarily aims to learn patterns directly from available data and leverage this knowledge to make predictions or explanations for new data. It uses mathematical functions of a general form that depend on many parameters. The values of these parameters are then learned using the available data, often through iterative minimisation of the error, similar to the data-fitting protocols covered in Chapter 2. The main difference of this approach compared to low-parameter modelling discussed in Chapter 2 is a much larger search space of possible functions explored to describe the data, as the models are no longer constrained by leveraging specific laws of physics. To make sure that only meaningful patterns are captured, only part of the data is typically used for fitting the parameters, and the remaining part is used for evaluation of generalisability, in contrast to low parameter modelling, where the entire dataset is used for fitting the model. This

pipeline is quite universal and can thus be applied to tasks and datasets for which we might not have any explicit mechanistic description of the process, e.g., for predicting melting temperatures of proteins from protein sequences, effects of mutations on protein stability or solubility, or types of protein binding pockets. The only major requirement is the availability of a data set of instances based on which the patterns could be identified.

If such a dataset is available, every data point there needs to be represented as a vector of numbers, commonly referred to as features. Features may be obtained through a wide variety of means (48). For example, protein sequences can be turned into feature vectors by assigning each position in a sequence a vector of 19 zeros and a single “1” corresponding to the specific amino acid in this position. Features may also be derived, examples being simple amino acid counts, propensities of different residues to form secondary structures, conservation scores, or various physicochemical characteristics, to name a few.

Once every data point in the data set is represented as a set of feature vectors, ML can be applied. Several different categories of ML problems exist. Supervised learning deals with methods for the task of predicting a particular property known as a label for each data point (Fig. 3A) (49). For example, a mutation in human protein can be labelled as "oncogenic” or “benign” or a protein sequence can be labelled as “soluble” or “insoluble”. Labels can form a set of classes (e.g., class 1: "oncogenic”; class 2: “benign”) or fall within a range of numerical values (e.g., protein yield in mg/mL), giving rise to two subtypes of supervised learning problems: classification problems involving labels with no inherent order (e.g., “oncogenic” or “benign”) and regression problems involving labels corresponding to numerical values (e.g., protein yields). The best practice in supervised learning is to split the available data into three disjoint subsets: a training set (used for fitting the model and determining the model parameters), a validation set (used for comparing different models and optimising model architectures), and a test set (only used for final evaluation as a representation of the future, “unseen” data).

In the absence of labels, unsupervised learning can be applied, whose goal is usually to identify patterns in unlabelled data, e.g., by using clustering algorithms and data compression or projection methods (Fig. 3B) (49). More recently, the boundary between supervised and unsupervised machine learning has been blurred by the emergence of methods that can create labels synthetically – self-supervised learning (Fig. 3D). For example, in a data compression method, the label might be the input itself, and the algorithm may impose constraints (e.g., a bottleneck in the architecture) that force the model to learn a more compact data representation, as is the case of variational autoencoders (50).

Finally, the algorithms that aim to learn the data distribution to generate new samples belong to a class of ML models called generative models (51). The most recent examples of this class include diffusion models (Fig. 3C), which create artificial labels by altering the input data, e.g. masking amino acids in protein sequences or adding noise to atom coordinates in protein structures, and learn to remove the alteration. Such a self-supervision approach turns out to be capable of learning useful characteristics of the data. In natural language processing, for instance, language models trained by predicting masked words manage to learn grammar and semantics to

generate sentences, and similar algorithms can thus be leveraged to generate new protein sequences, structures, or even protein ensembles (52).

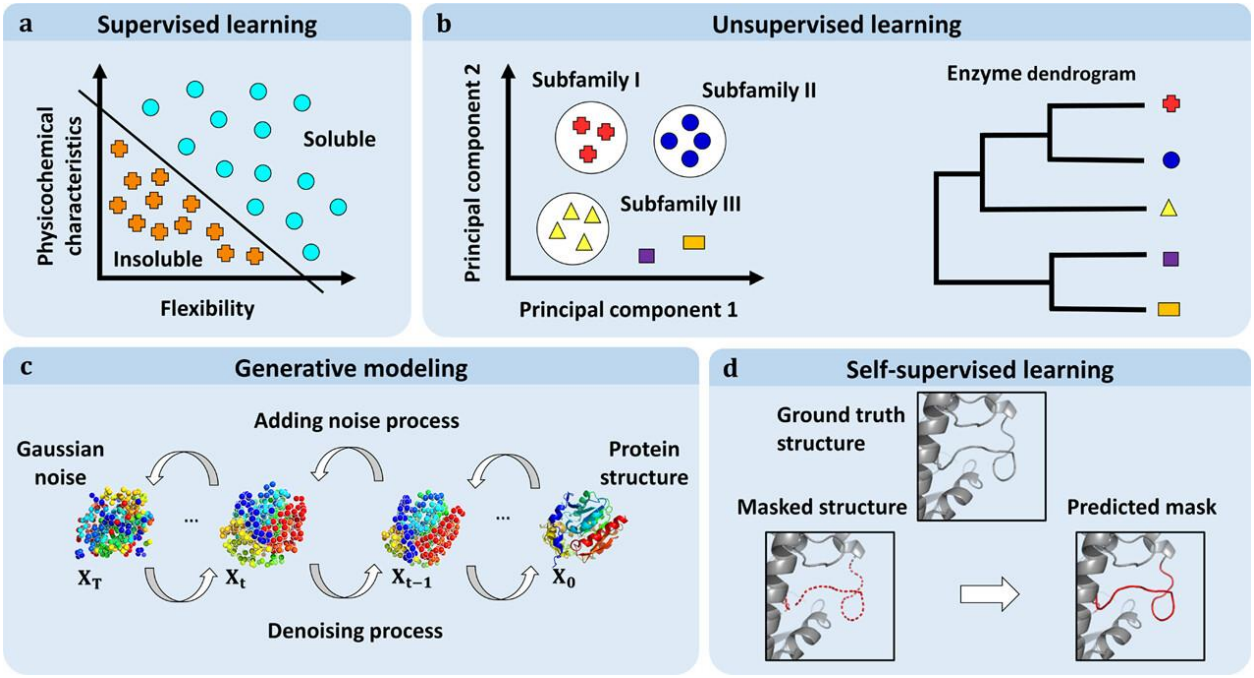


Figure 3. An overview of different machine learning approaches. Machine learning methods can be categorised into the following groups: (a) supervised learning methods aim to predict a specific label, (b) unsupervised learning methods typically find clusters in unlabelled data, (g) generative models learn the distribution of the training data to generate new instances corresponding to that distribution, and (d) self-supervised learning methods transform an unsupervised problem into a supervised problem by creating synthetic labels, e.g., masking part of the input. The figure is adopted from (20).

The following subsections will cover different tasks in protein engineering that we have tackled by machine learning. In our studies, we have explored a range of ML algorithms, from classical methods, such as decision trees, random forests, extreme gradient boosting, K-nearest neighbours, to methods from deep learning, i.e., based on artificial neural networks. Both supervised and self-supervised tasks were explored.

3.3. Supervised learning for protein engineering

In this subsection, we will cover two examples of supervised learning tasks we explored in our research: classifying pockets in enzymes into two types, buried or surface, and classifying mutations in human proteins related to oncology as oncogenic or benign.

Classification of pockets in enzyme structures. Critical structural elements of enzymes are active sites, which are often located in surface clefts or internal cavities, facilitating chemical reactions. Various computational tools, such as Fpocket (53), CASTp (54), and P2Rank (55), may help identify and rank potential binding pockets, but their accuracy is limited by sparse structural annotations. When buried, these sites connect to the surface via access tunnels that regulate ligand movement and influence enzyme activity and specificity. Tunnels in enzymes with buried active sites

allow the entry of substrates and the release of products, thus contributing to the catalytic efficiency. Targeting the bottlenecks of protein tunnels thus represents a powerful protein engineering strategy (56). Our colleagues recently developed a set of software tools CAVER for constructing and analysing the tunnels connecting a buried pocket to the surface and intended to apply this tool at scale, to thousands of proteins, to better understand structural determinants of enzyme activities (57). To this end, they required an efficient algorithm for discriminating between buried and surface pockets, as CAVER is sensitive to the starting point, and calculating tunnels is not applicable to surface pockets. However, none of the pocket features produced by Fpocket was sufficient to classify the pocket type, and no other tool existed to tackle this task.

Therefore, the main goal of our contribution was to create a predictor capable of assessing and differentiating between buried and surface-exposed protein pockets based on the features produced by Fpocket. In total, 20 features readily available from Fpocket were used, e.g., total surface area, volume, mean local hydrophobic density and others. For the training of the predictor, we manually labelled 200 pockets by an expert. Pockets were categorised into three classes: buried, surface, and borderline cases based on visual inspection of protein structures. We further analysed the distribution of the Enzyme Commission (EC) classes in the dataset to negate any bias in the EC class distribution. Given the small dataset size and heterogeneous features, we tested a range of ML-based algorithms, including the Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Shallow Neural Network (ANN), Gaussian Naive Bayes, and Random Forest. In each case, we tuned hyperparameters by a grid search with five-fold cross-validation. We also explored whether Kolmogorov–Smirnov feature filtering could improve the accuracy of predictors. For final validation, we employed an independent test set of 100 manually labelled additional samples, mirroring the class distribution of the training set. For the three-class problem, the ANN achieved the highest accuracy (54%) and F1 score (50%), and the second-highest 1-FPR score (67%) on the test set. ANN was also among the top-performing models for the two-class prediction, with all three metrics of 70% on the test set. The Python code for the pocket discrimination predictor is available at https://github.com/Faranehhad/large-scale-pocket-tunnel-annotation.git.

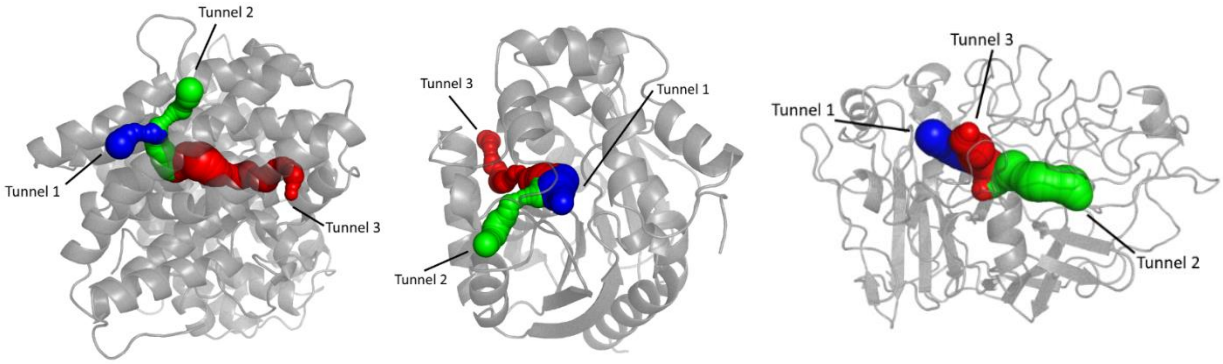


Figure 4. Examples of tunnels in proteins leading to the active site identified with the automated pipeline. The sample proteins are Pyrroloquinoline-quinone synthase (1OTW, left); Haloalkane dehalogenase LinB (2BFN, centre); and Cellobiohydrolase (2RFY, right).

The optimised and validated pipeline was then applied to annotate more than 17,000 cognate enzyme–ligand complexes, creating the first-of-a-kind dataset of this scale (Fig. 4**)**. Further analysis of ligand un-/binding energies revealed that the top priority tunnel had the most favourable energies in 75% of cases, and a simple geometry analysis could correctly identify tunnel bottlenecks only in 50% of cases. Thus, the pipeline gave essential information for the interpretation of results from tunnel calculation and energy profiling in mechanistic enzymology and protein engineering.

Classifying mutations in human proteins. Over 19 million cancer cases are diagnosed annually, and this number continues to rise[3]. As standard treatments vary in effectiveness across cancer types, understanding tumour biology is critical, especially for hard-to-treat cases. Personalised high-throughput profiling using next-generation sequencing enables comprehensive analysis of biopsy samples and has generated vast data on cancer-specific gene alterations. In many cases, after a cancer diagnosis, treatment is a race against time, and with the variable success rates of conventional "one size fits all" therapies, fast and accurate interpretation of molecular findings and assessment of their actionability are of vital importance, especially in difficult-to-treat cases. This is where an automated precision oncology approach can be most useful as it can optimise treatment strategies, improve outcomes, and increase the quality of life for many patients. However, a major gap remained between these alterations and their proven effects on protein function.

To fill this gap, we developed a bioinformatics pipeline for rapid analysis of missense mutations in oncogenic proteins (Fig. 5), assessing their impact on stability and function (58). The predictive part of the pipeline is a machine-learning-based tool trained on 1073 single-point mutants in 42 proteins. For a set of known cancer-related human proteins, we assigned the labels "Oncogenic" or "Benign" based on the annotations from a range of publicly available databases, including gnomAD, ClinVar, OncoKB, The JAX Clinical Knowledgebase, Personalized Cancer Therapy Knowledge Base, cBioPortal, and the DoCM database (59–62). The entire dataset of proteins and mutations was then annotated by the computational biology-based pipeline of PredictONCO to produce a set of sequence-based (essentiality of, conservation, domain, the PredictSNP score (63), the number of essential residues) and structure-based (FoldX and Rosetta ddg_monomer scores, ligand-binding pocket, and the pKa changes of essential residues obtained from PROPKA3) features. The final training dataset with features and labels is available at https://zenodo.org/records/10013764.

For the training of a predictor, 20% of the data was kept aside for testing, chosen randomly but grouped by positions to ensure that no specific position in a protein from the test set appears in the training set. We explored a range of methods, with the extreme gradient boosting algorithm producing the best results (AUC ROC of 0.97 and 0.94, and the average precision of 0.99 and 0.94 for structure-based and sequence-based predictions for the test set, respectively, Fig. 6). We demonstrated the applicability of the tool by presenting its usage for variants in two cancer-associated proteins, cellular tumour antigen p53 and fibroblast growth factor receptor FGFR1 (64).

[3] https://www.wcrf.org/preventing-cancer/cancer-statistics/worldwide-cancer-data/

To facilitate access and analysis of cancer-related mutations, we also implemented the pipeline as a web server (65), which is freely available at https://loschmidt.chemi.muni.cz/predictonco/.

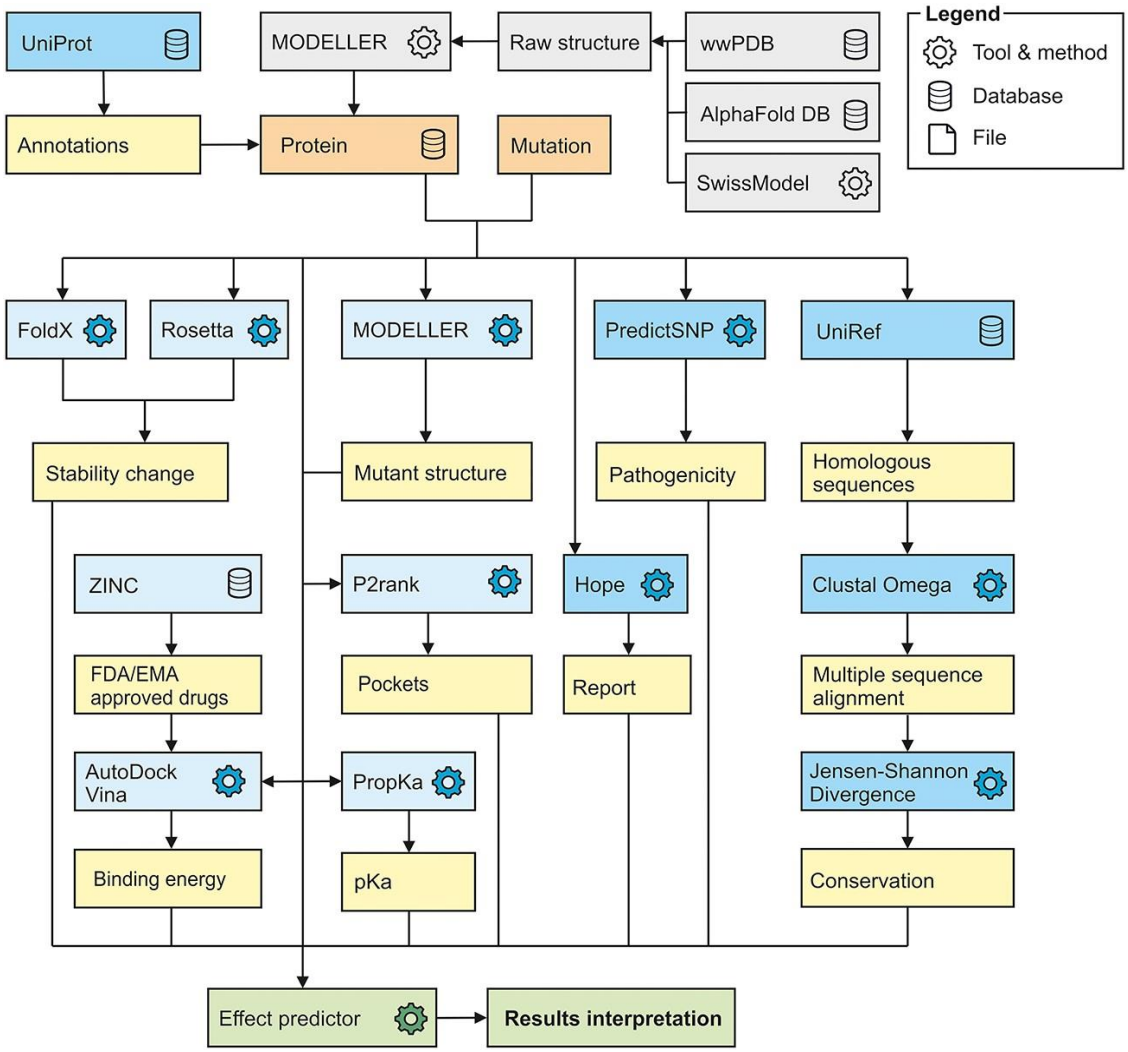


Figure 5. The entire workflow of the PredictONCO tool. The only required input is a protein and a mutation (orange boxes). The grey boxes show pre-treatment steps done manually in advance to prepare high-quality protein structures as a reliable starting point for the calculation. Once the calculation is submitted, multiple analyses are executed. The sequence-based analyses are performed for all mutations (dark blue boxes), such as fetching annotations from public databases, pathogenicity prediction, conservation prediction and HOPE. For mutations in the catalytic domains with available 3D structure, structure-based analyses are additionally performed (light blue boxes), such as stability and pKa prediction, pocket detection, and virtual screening. Once all the features are collected, the effect of the mutation is predicted using an XGBoost-based effect predictor (green box). The yellow boxes briefly describe outputs collected from each analysis. The figure is adopted from (65).

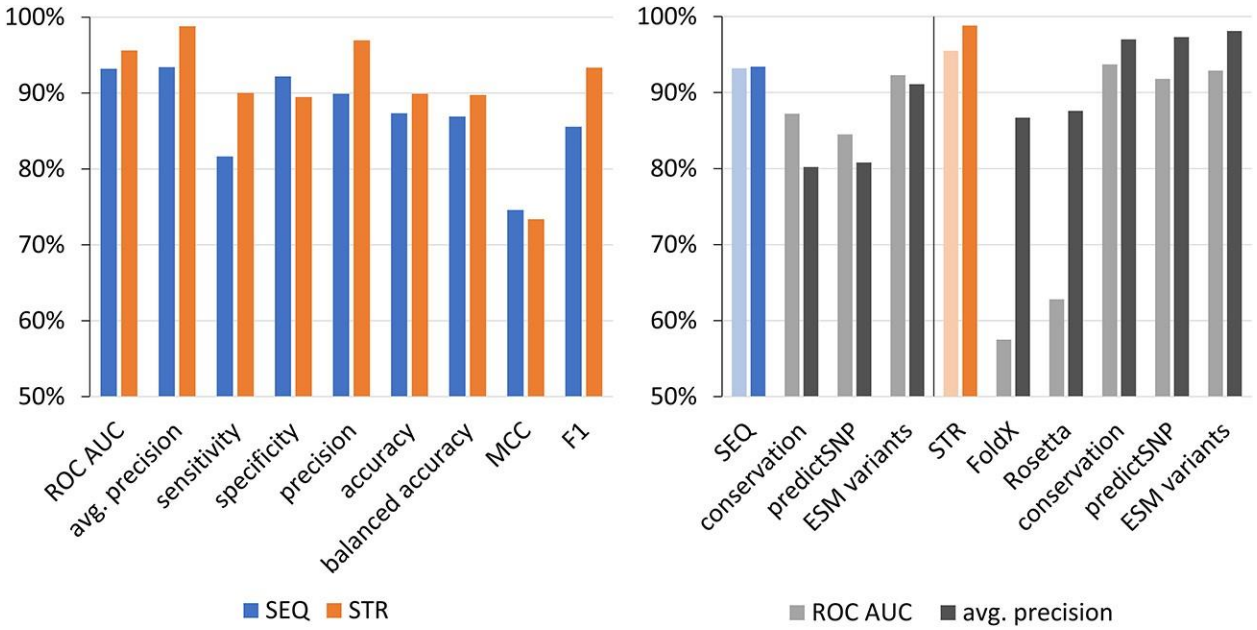


Figure 6. The performance of the structure-based (STR) and sequence-based (SEQ) PredictONCO models on the held-out test set of 213 and 89 mutations, respectively. Left: The area under the receiver operating characteristic curve (ROC AUC) and average precision values show strong performance for the probability of the oncogenic effect of a mutation returned by the predictors. The remaining values were calculated for the cutoff of 0.50 applied to this probability. **Right**: The comparison to the individual tools and the state-of-the-art method ESM variants according to ROC AUC and average precision metrics shows overall better performance in both SEQ and STR evaluations. The figure is adopted from (65).

3.4. Self-supervised learning for protein engineering

Despite the significant potential of machine learning in predicting specific labels, unlabelled protein datasets possess incredible potential in guiding protein engineering. In our research, we explored several tasks related to such methods: identifying evolutionary trends in protein phylogenetic trees to suggest promising mutations; leveraging the geometry of the latent space of a neural network to suggest ancestor-like protein sequences; and comparing protein dynamics in the presence and absence of a potential drug candidate for challenging disordered peptides.

Protein successor prediction. Protein evolution can be distilled into two key steps: the occurrence of amino acid mutations, e.g., from errors in DNA replication during cell division, exposure to mutagens, or viral infections, and the subsequent fixation of these mutated proteins within a population based on their impact on fitness. While this two-step model provides a useful framework, it remains primarily descriptive rather than predictive as it cannot be used to accurately forecast future mutations or their likelihood of fixation. Consequently, evolutionary predictions generally focus on projecting adaptive processes, rather than mutations at the amino acid level and concentrate on the evolution of infectious diseases, cancer, or other somatic processes at the phenotypic level.

In the protein engineering domain, ancestral sequence reconstruction (ASR) was shown to be a powerful method for suggesting protein variants by leveraging phylogenetic trees and sequence

alignments to trace evolutionary changes and infer ancestral protein sequences. Through the reconstruction of evolutionary histories, ASR can identify specific positions in protein sequences where selective pressures have driven adaptation. This approach is a very effective strategy not only for thermostability engineering (66), but also for improving other protein characteristics such as specificity, activity, or expression (67).

However, ASR is inherently limited to exploring the evolutionary past of a sequence. In this study, we hypothesised that integrating evolutionary insights with physicochemical properties, such as descriptors in the AAindex database (68), holds significant promise for predicting evolutionary successors that conform to physical evolutionary pressures, possibly revealing promising protein variants. To explore this hypothesis, we proposed a novel approach called the Successor Sequence Predictor (SSP), designed to find trends in phylogenetic trees based on AAindices and propagate those trends for protein design (69). SSP reconstructs the evolutionary history of a given protein sequence and suggests amino acid substitutions by projecting observed evolutionary trends through a range of carefully selected physicochemical descriptors (Fig. 7). Introducing these predicted mutations is expected to enhance specific protein properties.

We tested SSP using various published datasets and observed intriguing results for various properties. In the case of thermostability, SSP made 14 predictions for the cold shock protein CspB, eight of which had a stabilising effect on the protein ($\Delta\Delta G < 0$), while the remaining six were neutral with $\Delta\Delta G$ values between 0 kcal/mol and 1 kcal/mol, including the highest increase in melting temperature of + 16.6°C. We also applied SSP to make predictions for aminoglycoside 3'-phosphotransferase, which confers resistance to aminoglycosides with antibiotic properties. SSP made 221 predictions with a significant improvement in the average enrichment value across a set of antibiotics compared to the baseline random selection, thus demonstrating predictive prowess in the context of enhancing enzymatic activity. Another validation of mutations for the solubility dataset of the levoglucosan kinase showed a higher likelihood of a positive or neutral effect on the solubility of the protein. Thus, our study showed that the SSP approach could enhance specialised proteins by predicting mutations that may improve desired properties, such as thermostability, activity, and solubility.

Leveraging latent spaces to produce ancestor-like proteins. Recent advances in analysing multiple sequence alignments (MSAs) of homologous proteins leverage deep learning models, such as diffusion models, GANs, and variational autoencoders (VAEs), to extract meaningful patterns and generate novel protein variants. Among these, VAEs are particularly intriguing due to their ability to model latent space representations that capture biophysical and evolutionary properties (50). VAEs have already proved useful in several supervised learning applications, including predicting protein structures, discovering novel drugs, and predicting protein functions. More recently, the latent space of the variational autoencoders was shown to capture the biophysical properties of protein variants and the phylogenetic relationships within protein families (70). In particular, the authors observed that ancestors tend to be positioned close to the origin of the latent space. However, the study did not go further to offer any strategy that would allow exploiting these relationships to generate new proteins from the latent space.

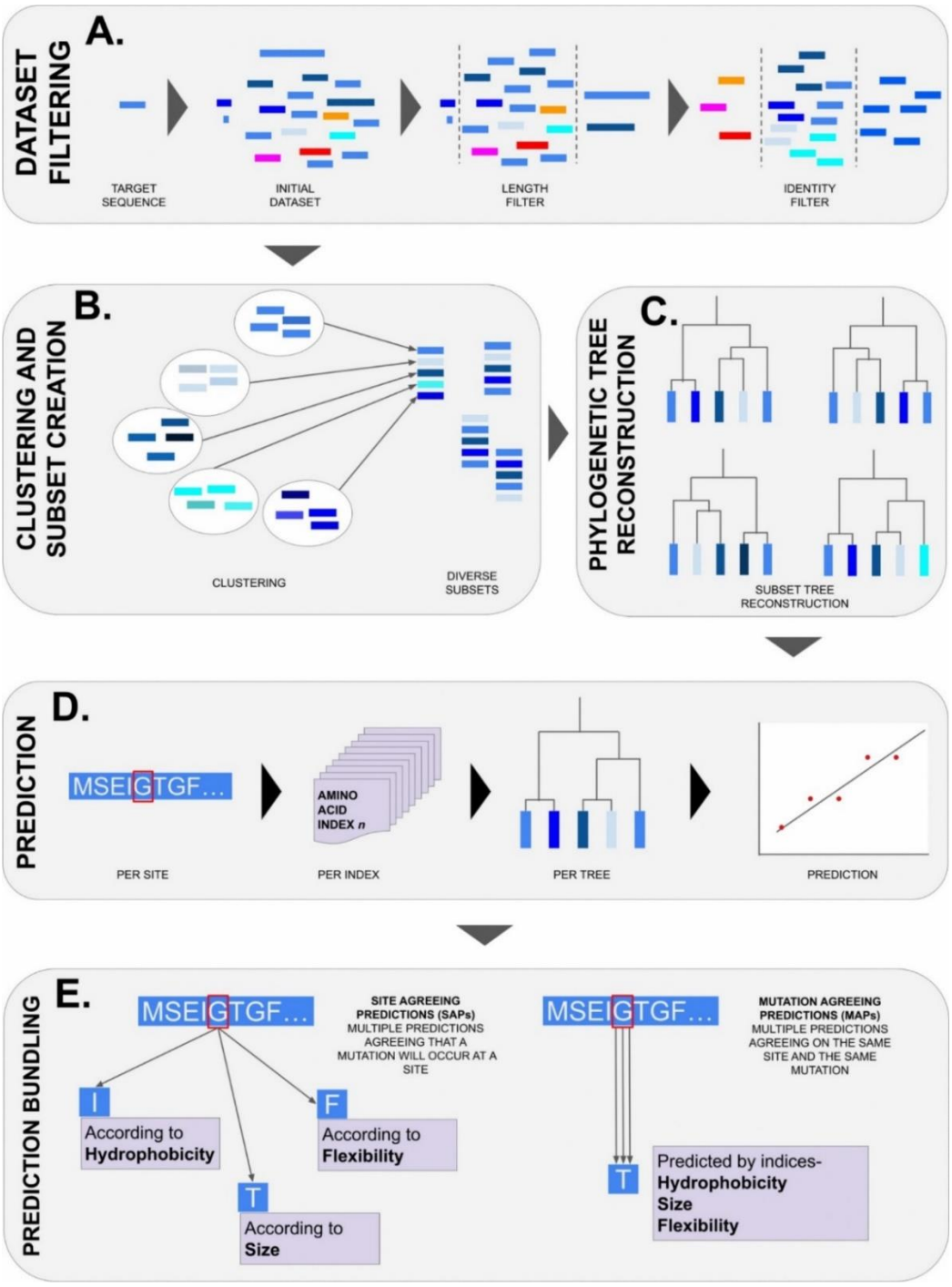


Figure 7. A generalised overview of the Successor Sequence Predictor. Initial curation and filtering of the target protein dataset include mining public databases and filtering out the obtained multiple sequence alignment (**A**). The sequences are then clustered by sequence identity (**B),** and multiple phylogenetic trees are constructed by sampling sequences from different clusters, along with ancestral sequence reconstruction for the nodes on the trees (**C**). For each tree and each position in the reference sequence, linear regressions for ten pre-selected AA indices are calculated along the evolutionary trajectory (**D**). If the directional trends in the data were detected, the closest matching amino acids that would continue the trend are assigned and reported (**E**). When several AA indices show trends, the mutations are marked as Site Agreeing Predictions (if different amino acids are suggested) or Mutation Agreeing Predictions (the same amino acid is suggested). The figure is adopted from (69).

Building on this observation, we hypothesised that a simple straight line connecting a query protein sequence and the origin of the latent space may suggest protein variants with ancestral-like properties, e.g., improved protein stability and preserved function. To test this hypothesis, we adopted the approach suggested previously (70) and augmented it with the straight-line evolutionary strategy (13). In addition to suggesting the strategy to generate protein sequences, we also implemented several modifications. First, we introduced the step of mined sequences with preserved catalytic residues using EnzymeMiner to obtain an MSA of functionally related proteins. Second, we introduced a reference sequence that is used as a template to narrow down the MSA to filter out distantly related sequences and reduce the number of insertions and deletions in the MSA. Third, we trained a VAE and explored several metrics to measure its capacity to generate protein sequences and capture the phylogeny in the constructed latent representations.

As a result of three rounds of experimental validation and protocol optimisation, we applied the straight-line strategy, reconstructed protein sequences with the VAE-decoder part, and generated 20 new ancestral-like designs of haloalkane dehalogenases, sharing as little as 67% sequence similarity to known sequences (Fig. 8). Using state-of-the-art microfluidics devices developed earlier (71), we subjected the designs to a thorough experimental characterisation, including the determination of their biophysical properties and substrate specificity profiles (Fig. 9). Obtained enzymes showed up to a 9 °C increase in melting temperatures and an average improvement of 3 °C across all soluble variants. We also observed a boost in activity, up to 3.5-fold for the most stable variant, whereas most of the other expressed variants showed activity levels comparable to benchmark enzymes. Our study demonstrated that the structure of the latent space and the generative potential of VAEs can guide the sequence search and designing novel soluble and functional proteins with enhanced stability. Moreover, to facilitate access to our methods for a wider audience of protein engineers, we integrated both the SSP and VAE design modules in our easy-to-use web server FireProtASR (https://loschmidt.chemi.muni.cz/fireprotasr/).

Analysis of complex protein dynamics with CoVAMPNets. Computational study of the effect of drug candidates on protein dynamics is quite challenging due to the data complexity. A popular approach to studying such effects is running molecular dynamics (MD) simulations and identifying notable conformational states by building so-called Markov state models (MSMs) (72). Under the assumption of the dynamics being Markovian (memoryless), these models cluster the conformational space into states, preserving the Markovianity of the transitions and estimate the equilibrium distribution and transition rates between the states. In general, selecting the variables derived from MDs for clustering is of critical importance for successful creation of an MSM and was often performed manually. Recent progress in variational approaches for conformational dynamics allowed scoring different MSMs, e.g., based on their ability to approximate the slowest modes of the dynamics, thus facilitating the development of automatic frameworks for the identification of Markov states (73). A powerful framework based on deep learning is VAMPnet, a neural network that learns a probabilistic assignment of each simulation frame to individual states in an unsupervised manner by maximising a variational score representing the quality of the model approximation of the slowest modes (74).

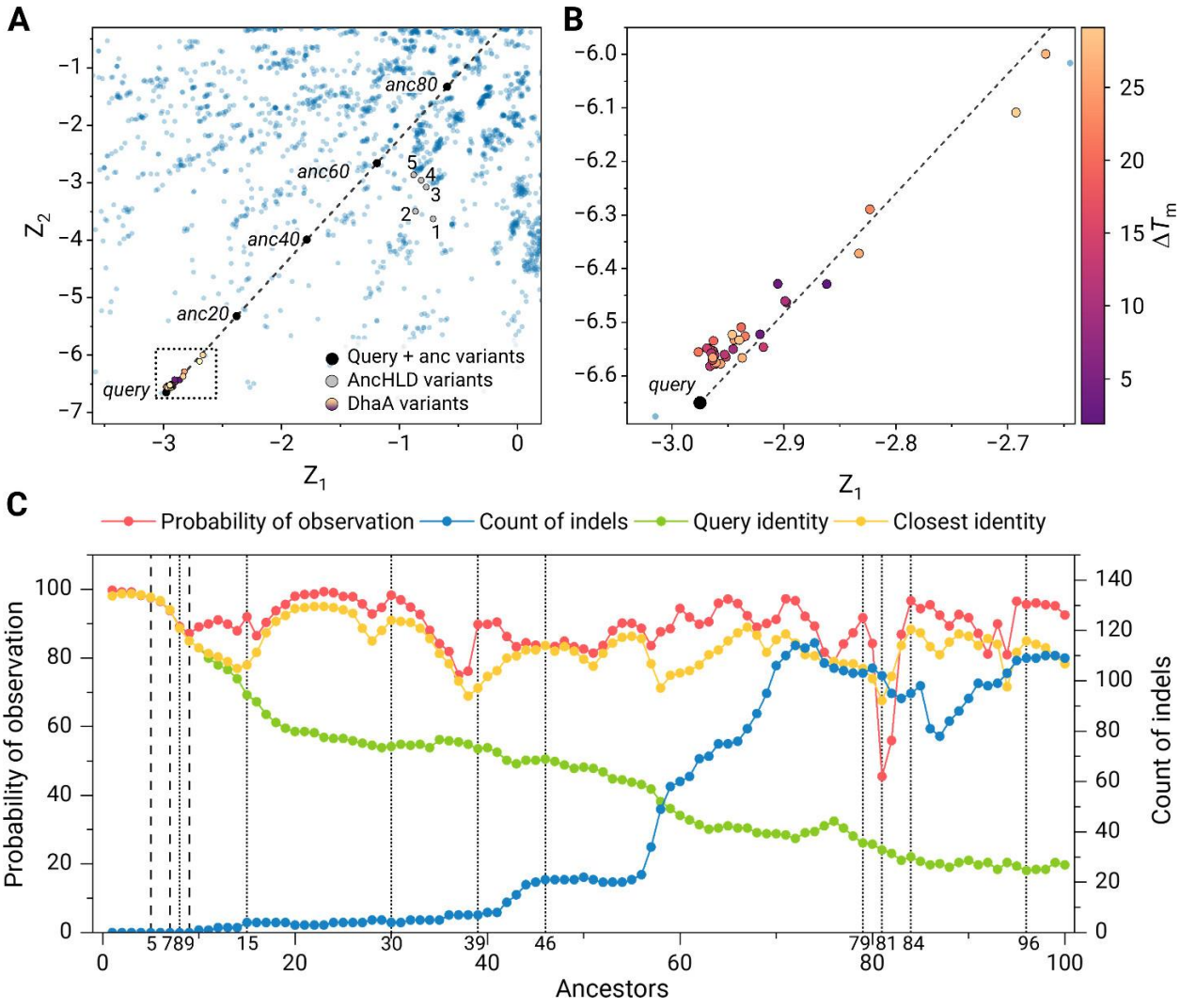


Figure 8. The summary of the straight-line evolutionary strategy for the Haloalkane Dehalogenase engineering case study. (**A**) Straight-line evolutionary strategy reconstructed 100 sequences along the trajectory from query embedding to the latent space origin (black dashed line), based on training on the aligned sequences of functionally related proteins (blue dots). The embeddings of previously characterised ancestors (grey points 1–5 denoting AncHLD variants) and engineered DhaA variants (magenta spectrum points) are mapped closer to the latent space origin, supporting the idea behind our ancestral generation strategy. (**B**) A detailed view of the previously engineered DhaA variants. While there is no strong correlation between the positions in the latent space and the stability gain (melting temperature difference, ΔTm) of variants up to 28 °C, some of the most stable points are situated closer to the origin. (**C**) The statistical profile of 100 sequences from the straight-line evolutionary strategy in rounds 1-2 of the selection. The ancestors are numbered 1 to 100 based on their order in the VAE-generated latent space, with lower numbers being closer to the starting sequence and higher numbers representing more divergent designs closer to the latent space origin. Number 0 corresponds to the reconstruction of the original embedding of the query sequence. The vertical lines represent sequences selected for experimental characterisation: dashed line variants were successfully expressed, dotted line variants were not soluble. The figure is adopted from (13).

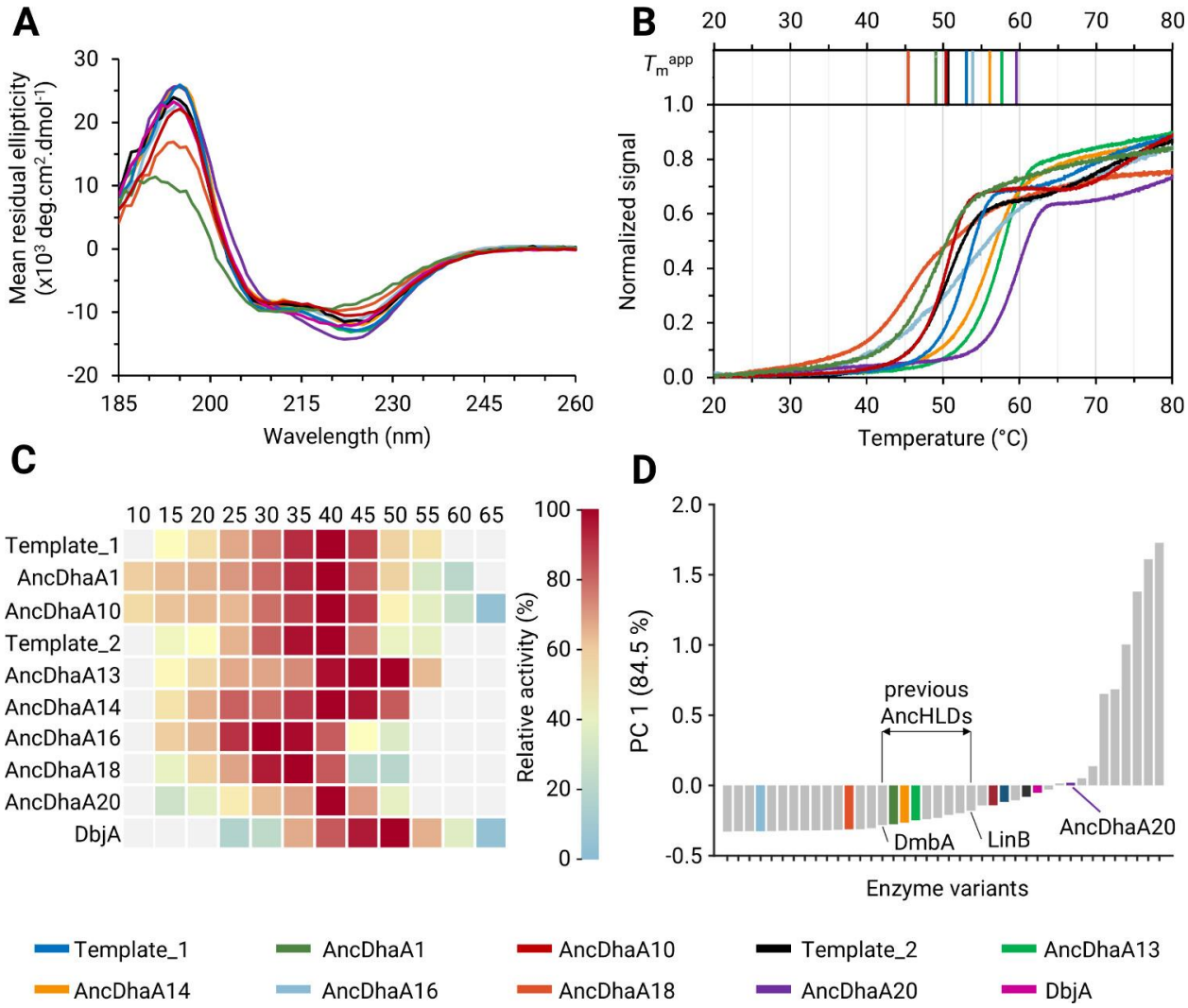


Figure 9. Experimental characterisation of selected variants from the straight-line evolutionary strategy for Haloalkane Dehalogenase engineering case study. (**A**) Far-UV circular dichroism spectra probing the correct folding and secondary structure of the variants. (**B**) Normalised thermal denaturation curves from nanoDSF spectroscopy with apparent melting temperatures (T_m^{app}) are shown above the curves. (**C**) The dependence of specific activity on temperature. The heatmap represents the relative activity of individual variants. (**D**) The score plot shows the first principal component PC 1 explaining 84.9% of the data variance in substrate specificity profiles, which compares VAE-based designs (in colour) with previously characterised wild-type haloalkane dehalogenases (grey) in terms of their activity with 27 substrates being determined by the MicroPEX method (71). The highlighted range between DmbA and LinB corresponds to the ranges of values observed for previously characterised AncHLD variants. The values of PC1 above this range imply that the overall activity of the corresponding designs was higher than that of previous AncHLD variants. The figure is adopted from (13).

In one of our research projects, we were interested in assessing the effects of the ongoing phase 3 therapeutics tramiprosate (TMP) and its metabolite 3-sulfopropanoic acid (SPA) on the disordered Aβ42 peptide involved in Alzheimer's disease. Alzheimer's disease is the fifth leading cause of death globally and the fourth cause of disability in people over 75 years (75). Aβ peptides play a major role in the development of the disease, although the mechanism behind their toxicity is still debated. To make matters worse, Aβ peptides are intrinsically disordered and difficult to study

both experimentally and computationally. Intrinsically disordered proteins do not adopt a single well-defined structure, but rather exist as ensembles of conformations with similar energies, complicating the construction of a good MSM.

The application of VAMPnets to the analysis of Aβ42 trajectories already showed great potential in producing robust MSMs for quantification of the Aβ42 kinetics and equilibrium properties previously (76). Thus, we embarked on the development of a comparative Markov state analysis (CoVAMPnet) framework to quantify changes in the conformational distribution and dynamics of a disordered biomolecule in the presence and absence of small organic drug candidate molecules (15). To this end, we generated molecular dynamic trajectories using enhanced sampling and computed an ensemble of soft MSMs for each system by training VAMPnet neural networks (Fig. 10). Then, using our novel alignment method, these ensembles were aligned to identify similar conformational states across the different systems based on a solution to an optimal transport problem. Finally, we applied explainable AI, in particular a discriminative analysis of aggregated neural network gradients, to assess the directional importance of inter-residue distances for the assignment to different conformational states (Fig. 11).

In the case of Aβ42 trajectories, our CoVAMPnet analysis revealed that both TMP and SPA preserved more structured conformations of Aβ42 by interacting nonspecifically with charged residues. SPA impacted Aβ42 more than TMP, protecting α-helices and suppressing the formation of aggregation-prone β-strands (77). While our experimental data suggested that TMP/SPA might also target biomolecules other than Aβ peptides, the CoVAMPnet approach can be applied to study and compare any related molecular systems. It can be especially useful to study the impact of small molecules on intrinsically disordered proteins and peptides, whose quantitative analysis can be extremely difficult.

3.5. Outlooks

Our studies leveraging supervised and self-supervised approaches have reconfirmed a tremendous potential for the use of machine learning methods in protein engineering. However, they have also identified major limitations hindering the progress in the domain, most pressing of which are as follows: (i) ML heavily relies on data availability and quality, and protein data are notoriously difficult to work with; (ii) constantly expanding datasets and changing data splits make benchmarking extremely time-consuming; and (iii) reporting standards for ML studies in biology are yet to be improved, further complicating reproducibility and comparison. In what follows, we will provide more details on each of those limitations and our contributions to overcoming them.

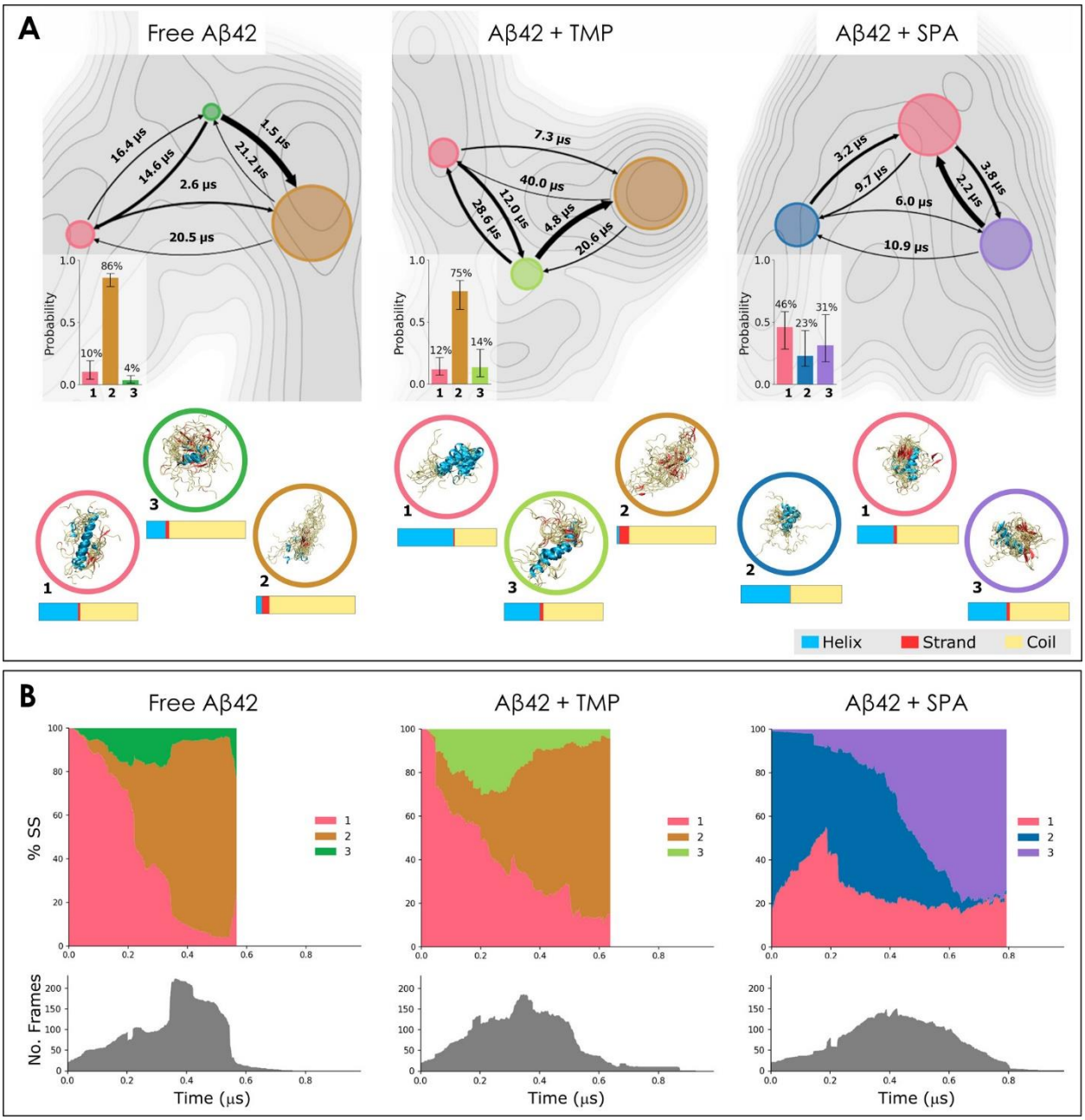


Figure 10. Analysis of conformational states learned using VAMPnets on the adaptive simulations and their evolution in time. (A) Properties of the states. For each system, we report: the free energy surface projected on the first two tICA dimensions (grey maps); flux diagrams projected on the same tICA space, where each state is represented by a coloured circle with the area proportional to the state probability, and the arrows indicate the mean first-passage times TM between the states, with the thickness proportional to the transition probability; equilibrium distribution of the states (the bars represent the 95th percentile of values from the ensemble of 20 learned models); superimposition of 20 representative structures from each state; and global mean secondary structure content of each state. (**B**) Distribution of the CoVAMPNet learned states in time (top) and the number of frames available at each time point (bottom). The adaptive sampling trajectories were aligned in time and concatenated. The state probability at a given time point was computed as the average soft assignment of all available frames at this time point. From left to right, the state assignments evolve from the beginning to the end of the simulation time. The states are numbered and colour-coded consistently across the entire panel. The figure is adopted from (15).

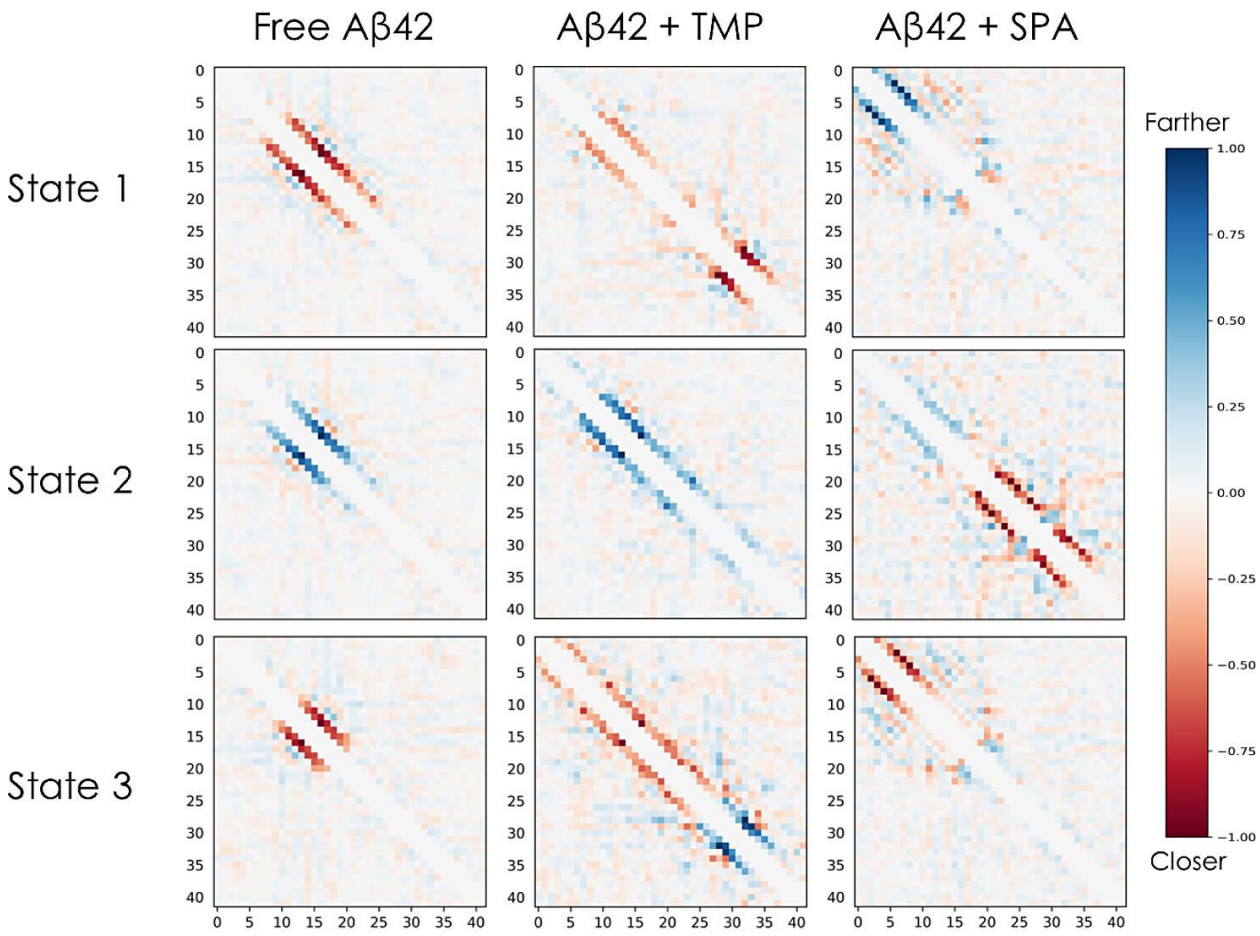


Figure 11. Gradients of the state assignment probabilities of the learned variational Markov state models using VAMPnets. Each 42 × 42 heatmap shows the ensemble-averaged gradients of the model probabilities for the corresponding system and state with respect to the input inter-residue Cα distances. The colour indicates how the probability of the particular state would change for an input frame if the distance between the particular pair of residues increased (blue: the probability of the state assignment would increase; red: decrease). The presented visualisations correspond to ensemble-averaged gradients evaluated and aggregated over 10,000 randomly selected simulation frames. The figure is adopted from (15).

Data availability. Protein-based datasets are challenging to work with for several reasons. First and foremost, protein engineering deals with enormous sequence spaces growing exponentially with the number of mutations in a protein to consider. But typical datasets used for training have only hundreds to thousands of labelled sequences. This limited data is sparse in exploring the mutational landscape and biased toward a few overrepresented proteins, greatly restricting ML model generalisation and extrapolation capabilities. Second, data often come from multiple sources with varying experimental biases, differing data normalisation practices, and inconsistent definitions of key protein properties, such as stability, solubility, or activity (78). These inconsistencies complicate dataset construction and can lead to contradictory labels for the same protein, thereby affecting ML reliability. Third, certain mutational types, such as alanine scanning, are overrepresented in datasets, introducing biases that impact ML predictions. Finally, many datasets are proprietary or

published in unstructured formats, limiting availability for ML training. We embarked on addressing those limitations in two related domains: protein stability and protein solubility. Our main contribution was two novel databases: FireProtDB and SoluProtMutDB for protein stability and solubility changes upon mutation, respectively (26, 27) .

Naturally occurring proteins face in maintaining stability under biotechnologically relevant conditions, such as elevated temperatures or high salt concentrations. Experimental screening of stabilising mutations is often labour-intensive and costly, which makes the use of computational predictors highly desirable to narrow down potentially beneficial mutations. Yet, existing databases contained outdated, inaccurate data and lacked advanced usability features. To address these drawbacks, we created FireProtDB, a manually curated source for experimental thermostability data for single-point mutants collected from ProTherm (79), ProtaBank (80), recent literature, and our own laboratory data (26). The key contribution of FireProtDB was a comprehensive, high-quality, and systematically curated dataset of over 15,000 protein stability changes, accessible through a user-friendly web interface (Fig. 12). This interface supports interactive exploration of individual mutations at the protein or mutation level and the construction of customised, ML-friendly datasets utilising advanced search, filtering, and export. All entries are carefully annotated to indicate their origin from existing datasets, allowing developers to create distinct training and testing datasets. FireProtDB thus filled a crucial gap by offering a freely available resource that facilitates data-driven approaches in protein engineering, particularly by enabling the creation of reliable datasets for the validation and benchmarking of stability prediction tools. The database is available at https://loschmidt.chemi.muni.cz/fireprotdb/.

In addition to protein stability, protein solubility is another key factor in protein research and applications. It is also connected to protein aggregation, which is linked to serious human diseases. Structural determinants governing protein solubility changes upon mutations are poorly understood, and the available data on this topic are scattered across the literature. To address this gap, we created SoluProtMutDB as the first manually curated database that compiles protein solubility change data upon mutations from various published sources (27). This extensive collection aims to facilitate better understanding and prediction of mutational effects on solubility, benefiting researchers in protein engineering and machine learning tool development. Our main contribution was assembling a large, high-quality dataset containing approximately 33,000 measurements covering 17,000 protein variants across 103 different proteins. The database integrated previously published solubility datasets along with thousands of new data points from recent studies, including deep mutational scanning experiments (27). It also incorporated detailed experimental conditions that affect protein solubility and underwent extensive manual curation to improve data quality for machine learning applications. This curated database is now available online (https://loschmidt.chemi.muni.cz/soluprotmutdb/) and serves as a valuable resource for designing improved protein variants and developing computational predictors for mutation-induced solubility changes, filling an important need for structured solubility data in protein science.

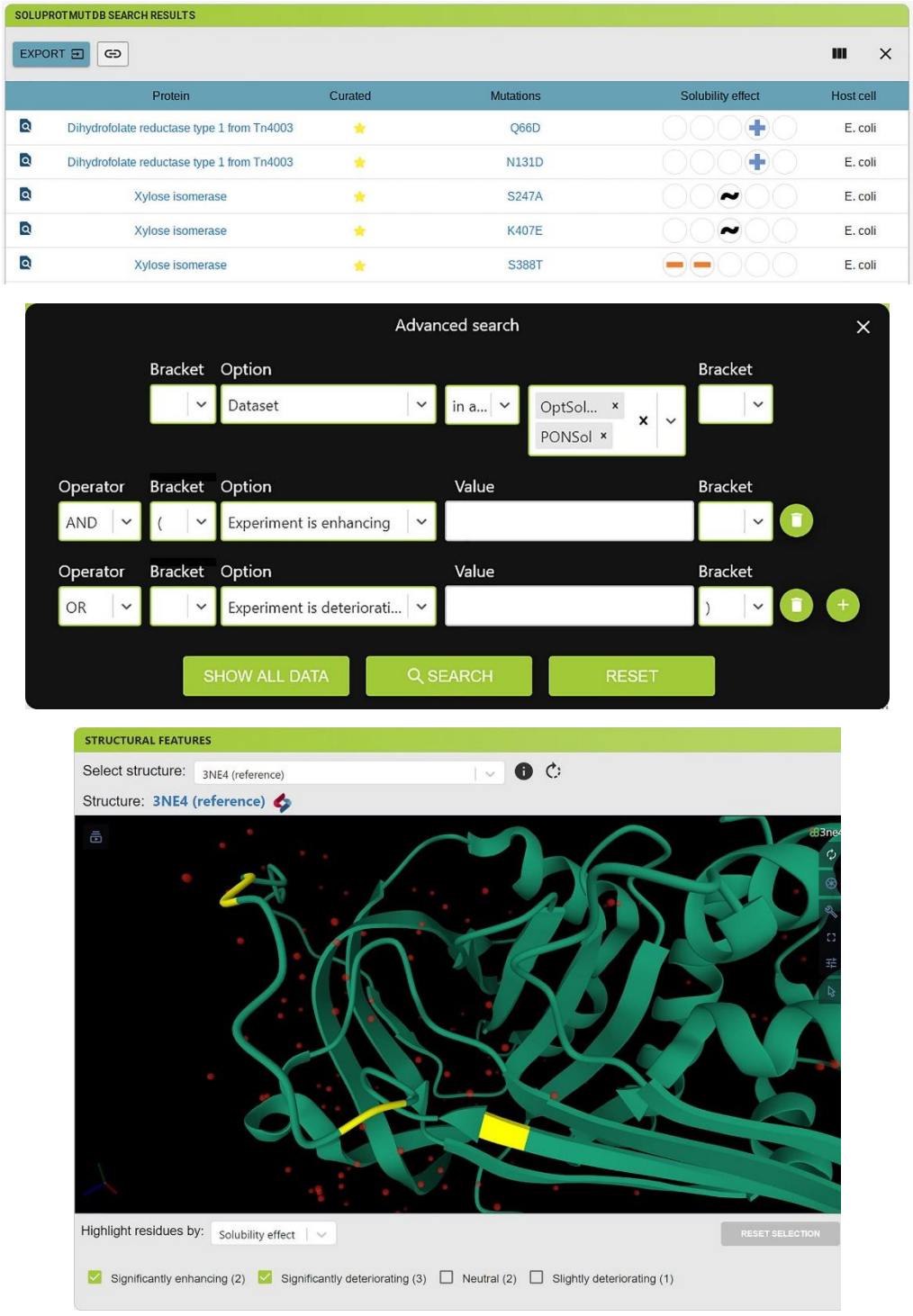


Figure 12. Examples of the user interface of SoluProtMutDB. **Top**: a table with search results. For clarity, only the most important columns are displayed by default: protein names, curation flags, mutations, solubility effects, and host cells. **Middle**: The advanced search with an example of a filtering protocol. In this example, the database will find measurements from OptSolMut and PON-Sol datasets with enhancing or deteriorating solubility effect. **Bottom**: visualisation of mutations in a protein with a known 3D structure. User-selected mutations can be highlighted in the structure. In this example, the mutated positions resulting in a significant change in solubility are highlighted in yellow. The database is available at https://loschmidt.chemi.muni.cz/soluprotmutdb/.

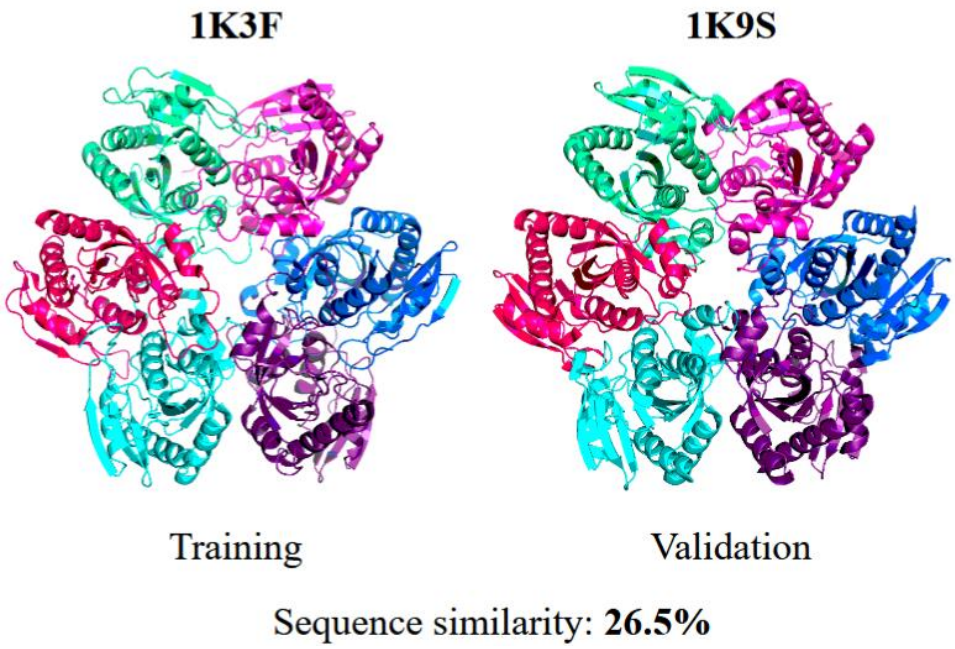


Figure 13. Examples of data leakage in a common benchmark for protein-protein interaction engineering. The two phosphorylase homooligomers, taken from DIPS, a standard dataset for training and validating machine learning models for protein-protein interactions. Both complexes are composed of five identical proteins (highlighted with colours) and have very low sequence similarity (26.5%). Despite the sequences in the complexes being different, the secondary structure of the chains, the topology of the interactions, as well as the 3D structure and the amino acids at the interfaces are highly similar across the entries (iDist score below 0.04, the near-duplicate threshold; iAlign p-value < 10^{-6}). Therefore, recent machine learning research for protein docking and interface prediction employed data splitting, resulting in data leakage. Figure is taken from (81).

Benchmarking. A fair comparison of the performance among state-of-the-art tools is critical in any machine learning workflow. With the growing number of computational web-based tools available for predicting the effects of mutations, e.g., on protein stability, benchmarking against these predictors becomes a major bottleneck. Researchers face difficulties conducting large-scale evaluations due to diverse input formats, overlapping training and test datasets, limited availability of some predictors as web services, input size restrictions, variable response times, and occasional downtimes. To address these issues, we developed BenchStab, an open-source Python package and command-line tool that automates the querying of multiple online stability predictors and collects their results efficiently, enabling straightforward benchmarking on user-defined lists of mutants (82). Our core contribution is providing a unified, modular platform that currently integrates 19 web-based protein stability prediction tools, facilitating automated, fast evaluation and comparison of different methods. BenchStab is extensible for integration of new predictors and promotes ongoing development in mutation stability prediction through open-source community contributions (https://github.com/loschmidt/BenchStab).

We also curated an independent test dataset derived from FireProtDB, carefully filtered to avoid overlap with predictor training data, comprising 289 mutation records across 36 proteins with diverse structural folds and reported the performance evaluation of the web-based protein stability prediction tools integrated in BenchStab (https://zenodo.org/records/10637728). This evaluation

reconfirmed limitations of existing predictors, such as bias towards destabilising mutations and the lack of a clear advantage for structure-based tools over sequence-only ones.

In addition to creating a tool for benchmarking mutational predictors, we also investigated problems with common benchmarks used to evaluate mutational predictors for protein-protein interactions. When working on such a predictor ourselves (18), we observed that common data splitting strategies based on protein sequence or metadata similarity introduced substantial data leakage (Fig. 13). This leakage caused overly optimistic assessments of model generalisation and unfair benchmarking because test interactions often had near-duplicates in the training data, sometimes leading to leakage rates as high as 80% and compromising the evaluation of predictive models. To address this, we proposed an improved approach to dataset splitting based on 3D structural similarity of protein-protein interfaces using the iDist algorithm, which significantly reduced leakage and led to more realistic evaluations (81).

Reporting standards. As the use of ML rapidly expands in genomics, proteomics, and other life sciences, the transparency and reproducibility of reported results are often limited due to insufficient details about dataset origin, optimisation strategies, model architecture, or evaluation protocols. The DOME recommendations (Data, Optimisation, Model, Evaluation) emerged as a structured checklist to guide authors toward comprehensive reporting, ensuring that critical aspects of data handling, algorithm design, and evaluation are explicitly described, thereby fostering reproducibility, accountability, and trust in ML-driven biological research (83). Building on these guidelines, we developed the DOME Registry, a web-based platform that enables researchers to curate, annotate, and access ML studies in a standardised manner (84). The registry was integrated with various tools, such as ORCID for researcher identity, APICURON for recognising curation efforts, and the Data Stewardship Wizard for a guided annotation workflow, offering a user-friendly interface. Each entry received a unique identifier and a DOME score, calculated as the proportion of recommendation items adequately addressed, fostering consistent evaluation standards across studies.

4. Discussion and Future Directions

With the recent progress in computation and the growing availability of high-throughput experiments, data processing is becoming a major bottleneck in protein science. We are observing an unprecedented penetration of data modelling and analysis methods for protein engineering. Two major paradigms: (i) bottom-up low-parameter modelling and (ii) top-down machine learning methods, are increasingly enabling us to extract meaningful biological insights from the available and newly collected data.

Low-parameter modelling allows well-controlled creation of readily interpretable models, capable of producing mechanistic insights into experimental signals, such as protein unfolding pathways, intermediate states, and Gibbs free energy barriers separating those states. Such insights can then guide protein engineers in the deep exploration of the mutational landscape. Machine learning methods approach the problem from a different perspective by building models from the available data. This perspective allows tackling much more complex and intricate phenomena but comes at a cost of lower interpretability and higher dependence on the available high-quality data.

Our research focused on both approaches to modelling protein-related data. In the case of low-parameter modelling, we investigated the current challenges in analysing protein stability and thermal denaturation signals. We developed the workflow implementing fourteen different models of protein unfolding and implemented it as a user-friendly web server CalFitter (38, 47). The workflow is capable of globally fitting the data from different types of protein thermal denaturation experiments. It also included the new mathematical framework we designed for modelling the data from reheating experiments and singular value decomposition analysis as an optional first step in data processing (46). As far as machine learning methods are concerned, we explored a wide range of models, from classical small-scale to deep learning-based, and protein engineering tasks. We ventured into both supervised and unsupervised learning, learning from protein sequences (13), structures (18, 85), molecular dynamics trajectories (15), and mutational data (64, 65). We have also created new benchmarking tools for the community of machine learning developers (82), assembled two new databases (26, 27), suggested more robust dataset splits (81), and contributed to improving reporting standards for ML in biology (84, 86).

Based on the current state of the art in computational protein engineering, it is difficult to foresee which one of those paradigms will dominate. In fact, the most promising direction might be a smart combination of the two. Such hybridisation can be implemented via several routes.

First, physics-based constraints can be integrated into the design of an ML predictor. For example, employing SE(3)-invariant models when learning on protein structures is gradually becoming a standard to ensure that the final prediction will not depend on the shifts and rotations of the protein 3D model (87). In the tasks of predicting mutational effects, the anti-symmetry is often enforced at the level of the ML model architecture to guarantee that a reverse mutation will lead to the reverse prediction of its effect on protein properties such as stability, solubility or activity (88). In our CoVAMPNet study, the architecture of the artificial neural network ensured that the learned Markov state model is reversible and that the elements of the matrix representing the governing

Koopman operator (a linear operator propagating the state probabilities in time) are non-negative (15, 89).

Second, ML predictors can be used as part of physics-based models to account for the effects that are hard to model. For instance, enzyme kinetics models can include terms calculated by a neural network, leading to so-called neural ordinary differential equations (ODEs). Unlike traditional kinetic models, which require prior knowledge of all reaction mechanisms, neural ODEs learn correction terms from experimental time-series data, allowing them to capture hidden interactions and nonlinearities not accounted for in the theoretical models (90). This leads to improved fitting of enzyme kinetic data by adapting the system dynamics during training. As a result, the method offers computational biologists a flexible tool to reconcile discrepancies between mechanistic models and experimental observations and allowing inference of unknown pathways and better modelling of complex enzyme-catalysed reaction networks.

Third, these two paradigms can be combined on an equal basis. One such example is the recently suggested kinetics-informed neural networks (91). These are specialised feed-forward neural networks designed to solve ODEs constrained by kinetic models, often microkinetic models describing biochemical or chemical reaction networks. They integrate knowledge of reaction kinetics directly into the neural network training, allowing the network to fit kinetics data and estimate kinetic parameters simultaneously. This approach improves noise tolerance and performance compared to traditional optimisation methods and enables interpolation and prediction of unseen reaction behaviours. Another example is the possibility of combining the two paradigms on the level of the data. For instance, one can simulate protein ensembles using demanding physics-based molecular dynamics tools and then use such a dataset to train an ML-based tool (52, 92). Such a tool eventually provides a faster yet less exhaustive generation of protein conformations.

Finally, the two paradigms can be applied side-by-side to provide two different perspectives on the task at hand. For example, predicting effects of mutations on protein stability can be done via force fields, such as FoldX (93) or Rosetta (94), or conservation scores from protein evolutionary data (95). Such predictions are not perfect and can thus be augmented with the predictions obtained by machine learning (96, 39, 82). While some evidence is starting to appear that combined these approaches may be able to compensate for each other's weaknesses (97), this hypothesis is yet to be validated in large-scale experiments.

In summary, advanced methods of data analysis are transforming protein research by enabling researchers to interpret vast and complex biological datasets with unprecedented depth and speed. They enable protein engineers to select protein targets, plan experiments, analyse the collected data, and formulate a hypothesis about biological phenomena of interest in a more informed way. These insights thus help accelerate the design of novel proteins with tailored properties and unveil the effects of mutations on health, leading to next-generation solutions across a range of domains, from medicine and diagnostics to bioengineering, synthetic biology, and materials science. Overall, advanced data analysis unlocks deeper insights into protein behaviour, bridging fundamental science with practical applications.

References

1. Petsko,G.A. and Ringe,D. (2004) Protein structure and function New Science Press, London, U.K.

2. The UniProt Consortium, Bateman,A., Martin,M.-J., Orchard,S., Magrane,M., Adesina,A., Ahmad,S., Bowler-Barnett,E.H., Bye-A-Jee,H., Carpentier,D., *et al.* (2025) UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res*, **53**, D609–D617.

3. Berman,H.M. and Burley,S.K. (2025) Protein Data Bank (PDB): Fifty-three years young and having a transformative impact on science and society. *Quart Rev Biophys*, **58**, e9.

4. Raddadi,N., Cherif,A., Daffonchio,D., Neifar,M. and Fava,F. (2015) Biotechnological applications of extremophiles, extremozymes and extremolytes. *Appl Microbiol Biotechnol*, **99**, 7907–7913.

5. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science*, **239**, 487–491.

6. Ausländer,S., Ausländer,D. and Fussenegger,M. (2017) Synthetic Biology—The Synthesis of Biology. *Angew Chem Int Ed*, **56**, 6396–6419.

7. Stefl,S., Nishi,H., Petukh,M., Panchenko,A.R. and Alexov,E. (2013) Molecular Mechanisms of Disease-Causing Missense Mutations. *J Mol Biol*, **425**, 3919–3936.

8. Katsonis,P., Wilhelm,K., Williams,A. and Lichtarge,O. (2022) Genome interpretation using in silico predictors of variant impact. *Hum Genet*, **141**, 1549–1577.

9. Bayer,T., Wu,S., Snajdrova,R., Baldenius,K. and Bornscheuer,U.T. (2025) An Update: Enzymatic Synthesis for Industrial Applications. *Angew Chem Int Ed*, **64**, e202505976.

10. Lutz,S., Lutz,S. and Bornscheuer,U.T. eds (2012) Protein Engineering Handbook 1. Auflage. Wiley-VCH, Weinheim.

11. Beerens,K., Mazurenko,S., Kunka,A., Marques,S.M., Hansen,N., Musil,M., Chaloupkova,R., Waterman,J., Brezovsky,J., Bednar,D., *et al.* (2018) Evolutionary analysis as a powerful complement to energy calculations for protein stabilization. *ACS Catal*, **8**, 9420–9428.

12. Kokkonen,P., Beier,A., Mazurenko,S., Damborsky,J., Bednar,D. and Prokop,Z. (2021) Substrate inhibition by the blockage of product release and its control by tunnel engineering. *RSC Chem Biol*, **2**, 645–655.

13. Kohout,P., Vasina,M., Majerova,M., Novakova,V., Damborsky,J., Bednar,D., Marek,M., Prokop,Z. and Mazurenko,S. (2025) Engineering dehalogenase enzymes using variational autoencoder-generated latent spaces and microfluidics. *JACS Au*, **5**, 838–850.

14. Dvorak,P., Bednar,D., Vanacek,P., Balek,L., Eiselleova,L., Stepankova,V., Sebestova,E., Kunova Bosakova,M., Konecna,Z., Mazurenko,S., *et al.* (2018) Computer-assisted engineering of hyperstable fibroblast growth factor 2. *Biotechnol Bioeng*, **115**, 850–862.

15. Marques,S.M., Kouba,P., Legrand,A., Sedlar,J., Disson,L., Planas-Iglesias,J., Sanusi,Z., Kunka,A., Damborsky,J., Pajdla,T., *et al.* (2024) CoVAMPnet: comparative markov state analysis for studying effects of drug candidates on disordered biomolecules. *JACS Au*, **4**, 2228–2245.

16. Legrand,A., Cerna,K.A., Marques,S.M., Verma,N., Kopko,J., Vanova,T., Subramanian,M., Bendl,J., Henek,T., Vanacek,P., *et al.* (2025) Taurine Inhibits Apolipoprotein E4 Aggregation. 10.1101/2025.08.13.669519.

17. Štulajterová,M., Ambro,Ľ., Sedláková,D., Nemergut,M., Kohout,P., Mazurenko,S., Varhač,R., Strunga,A., Toul,M., Prokop,Z., *et al.* (2025) Assessing the impact of His-tags on activity and stability of staphylokinase variants. *Int J Biol Macromol*, **328**, 147655.

18. Bushuiev,A., Bushuiev,R., Kouba,P., Filkin,A., Gabrielova,M., Gabriel,M., Sedlar,J., Pluskal,T., Damborsky,J., Mazurenko,S., *et al.* (2024) Learning to design protein-protein interactions with enhanced generalization. In *ICLR Proceedings*. Vienna.

19. Mazurenko,S., Prokop,Z. and Damborsky,J. (2020) Machine learning in enzyme engineering. *ACS Catal*, **10**, 1210–1223.

20. Kouba,P., Kohout,P., Haddadi,F., Bushuiev,A., Samusevich,R., Sedlar,J., Damborsky,J., Pluskal,T., Sivic,J. and Mazurenko,S. (2023) Machine learning-guided protein engineering. *ACS Catal*, **13**, 13863–13895.

21. Vasina,M., Kovar,D., Damborsky,J., Ding,Y., Yang,T., deMello,A., Mazurenko,S., Stavrakis,S. and Prokop,Z. (2023) In-depth analysis of biocatalysts by microfluidics: An emerging source of data for machine learning. *Biotechnol Adv*, **66**, 108171.

22. Bagshaw,C.R. (2017) Biomolecular kinetics: a step-by-step guide CRC Press, Boca Raton London New York.

23. Chis,O.-T., Banga,J.R. and Balsa-Canto,E. (2011) Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods. *PLoS ONE*, **6**, e27755.

24. Johnson,K.A. (2013) A century of enzyme kinetic analysis, 1913 to 2013. *FEBS Letters*, **587**, 2753–2766.

25. Aledo,J.C. (2021) Enzyme kinetic parameters estimation: A tricky task? *Biochem Molecular Bio Educ*, **49**, 633–638.

26. Stourac,J., Dubrava,J., Musil,M., Horackova,J., Damborsky,J., Mazurenko,S. and Bednar,D. (2021) FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res*, **49**, D319–D324.

27. Velecký,J., Hamsikova,M., Stourac,J., Musil,M., Damborsky,J., Bednar,D. and Mazurenko,S. (2022) SoluProtMutDB: a manually curated database of protein solubility changes upon mutations. *Comput Struct Biotechnol J*, **20**, 6339–6347.

28. Notin,P., Kollasch,A.W., Ritter,D., Van Niekerk,L., Paul,S., Spinner,H., Rollins,N., Shaw,A., Weitzman,R., Frazer,J., *et al.* (2023) ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. In *NeurIPS Proceedings*.

29. Roy,A., Ward,E., Choi,I., Cosi,M., Edgin,T., Hughes,T.S., Islam,M.S., Khan,A.M., Kolekar,A., Rayl,M., *et al.* (2025) MDRepo—an open data warehouse for community-contributed molecular dynamics simulations of proteins. *Nucleic Acids Res*, **53**, D477–D486.

30. Souza,P.C.T., Alessandri,R., Barnoud,J., Thallmair,S., Faustino,I., Grünewald,F., Patmanidis,I., Abdizadeh,H., Bruininks,B.M.H., Wassenaar,T.A., *et al.* (2021) Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat Methods*, **18**, 382–388.

31. Mathur,A., Ghosh,R. and Nunes-Alves,A. (2024) Recent progress in modeling and simulation of biomolecular crowding and condensation inside cells. *J Chem Inf Model*, **64**, 9063–9081.

32. Johnson,K.A., Simpson,Z.B. and Blom,T. (2009) Global Kinetic Explorer: A new computer program for dynamic simulation and fitting of kinetic data. *Anal Biochem*, **387**, 20–29.

33. Meisl,G., Kirkegaard,J.B., Arosio,P., Michaels,T.C.T., Vendruscolo,M., Dobson,C.M., Linse,S. and Knowles,T.P.J. (2016) Molecular mechanisms of protein aggregation from global fitting of kinetic models. *Nat Protoc*, **11**, 252–272.

34. Klamt,S., Stelling,J., Ginkel,M. and Gilles,E.D. (2003) FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinform*, **19**, 261–269.

35. Antoniewicz,M.R. (2021) A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications. *Metab Eng*, **63**, 2–12.

36. Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol*, **202**, 865–884.

37. Casadio,R., Compiani,M., Fariselli,P. and Vivarelli,F. (1995) Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. *Proc Int Conf Intell Syst Mol Biol*, **3**, 81–88.

38. Mazurenko,S., Stourac,J., Kunka,A., Nedeljković,S., Bednar,D., Prokop,Z. and Damborsky,J. (2018) CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res*, **46**, W344–W349.

39. Musil,M., Konegger,H., Hon,J., Bednar,D. and Damborsky,J. (2019) Computational design of stable and soluble bocatalysts. *ACS Catal*, **9**, 1033–1054.

40. Freire,E. (1995) Differential scanning calorimetry. In *Protein Stability and Folding*. Humana Press, New Jersey, Vol. 40, pp. 191–218.

41. Johnson,C.M. (2013) Differential scanning calorimetry as a tool for protein folding and stability. *Arch Biochem Biophys*, **531**, 100–109.

42. Kelly,S.M. and Price,N.C. (1997) The application of circular dichroism to studies of protein folding and unfolding. *BBA - Protein Struct Molec Enzymol*, **1338**, 161–185.

43. Sanchez-Ruiz,J.M. (2010) Protein kinetic stability. *Biophys Chem*, **148**, 1–15.

44. Arrhenius,S. (1889) Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Zeitschrift für Physikalische Chemie*, **4U**, 96–116.

45. Evans,M.G. and Polanyi,M. (1935) Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans Faraday Soc*, **31**, 875.

46. Mazurenko,S., Kunka,A., Beerens,K., Johnson,C.M., Damborsky,J. and Prokop,Z. (2017) Exploration of protein unfolding by modelling calorimetry data from reheating. *Sci Rep*, **7**, 16321.

47. Kunka,A., Lacko,D., Stourac,J., Damborsky,J., Prokop,Z. and Mazurenko,S. (2022) CalFitter 2.0: leveraging the power of singular value decomposition to analyse protein thermostability. *Nucleic Acids Res*, **50**, W145–W151.

48. Harding-Larsen,D., Funk,J., Madsen,N.G., Gharabli,H., Acevedo-Rocha,C.G., Mazurenko,S. and Welner,D.H. (2024) Protein representations: encoding biological information for machine learning in biocatalysis. *Biotechnol Adv*, **77**, 108459.

49. Marsland,S. (2015) Machine learning: an algorithmic perspective 2nd ed. CRC press, Boca Raton.

50. Kingma,D.P. and Welling,M. (2014) Auto-encoding variational bayes. In *ICLR Proceedings*.

51. Strokach,A. and Kim,P.M. (2022) Deep generative modeling for protein design. *Curr Opin Struct Biol*, **72**, 226–236.

52. Lewis,S., Hempel,T., Jiménez-Luna,J., Gastegger,M., Xie,Y., Foong,A.Y.K., Satorras,V.G., Abdin,O., Veeling,B.S., Zaporozhets,I., *et al.* (2025) Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, **389**, eadv9817.

53. Le Guilloux,V., Schmidtke,P. and Tuffery,P. (2009) Fpocket: An open source platform for ligand pocket detection. *BMC Bioinform*, **10**, 168.

54. Binkowski,T.A. (2003) CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res*, **31**, 3352–3355.

55. Krivák,R. and Hoksza,D. (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform*, **10**, 39.

56. Kaushik,S., Marques,S.M., Khirsariya,P., Paruch,K., Libichova,L., Brezovsky,J., Prokop,Z., Chaloupkova,R. and Damborsky,J. (2018) Impact of the access tunnel engineering on catalysis is strictly ligand-specific. *FEBS J*, **285**, 1456–1476.

57. Stourac,J., Vavra,O., Kokkonen,P., Filipovic,J., Pinto,G., Brezovsky,J., Damborsky,J. and Bednar,D. (2019) Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Res*, **47**, W414–W422.

58. Khan,R.T., Pokorna,P., Stourac,J., Borko,S., Arefiev,I., Planas-Iglesias,J., Dobias,A., Pinto,G., Szotkowska,V., Sterba,J., *et al.* (2024) A computational workflow for analysis of missense mutations in precision oncology. *J Cheminform*, **16**, 86.

59. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, **42**, D980–D985.

60. Chakravarty,D., Gao,J., Phillips,S., Kundra,R., Zhang,H., Wang,J., Rudolph,J.E., Yaeger,R., Soumerai,T., Nissan,M.H., *et al.* (2017) OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*, 10.1200/PO.17.00011.

61. Patterson,S.E., Liu,R., Statz,C.M., Durkin,D., Lakshminarayana,A. and Mockus,S.M. (2016) The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum Genomics*, **10**, 4.

62. Scheps,K.G., Hasenahuer,M.A., Parisi,G., Targovnik,H.M. and Fornasari,M.S. (2020) Curating the gnomAD database: Report of novel variants in the globin-coding genes and bioinformatics analysis. *Hum Mutat*, **41**, 81–102.

63. Bendl,J., Stourac,J., Salanda,O., Pavelka,A., Wieben,E.D., Zendulka,J., Brezovsky,J. and Damborsky,J. (2014) PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol*, **10**, e1003440.

64. Khan,R.T., Pokorna,P., Stourac,J., Borko,S., Dobias,A., Planas-Iglesias,J., Mazurenko,S., Arefiev,I., Pinto,G., Szotkowska,V., *et al.* (2024) Analysis of mutations in precision oncology using the automated, accurate, and user-friendly web tool PredictONCO. *Comput Struct Biotechnol J*, **24**, 734–738.

65. Stourac,J., Borko,S., Khan,R.T., Pokorna,P., Dobias,A., Planas-Iglesias,J., Mazurenko,S., Pinto,G., Szotkowska,V., Sterba,J., *et al.* (2023) PredictONCO: a web tool supporting decision-making in precision oncology by extending the bioinformatics predictions with advanced computing and machine learning. *Brief Bioinform*, **25**, bbad441.

66. Livada,J., Vargas,A.M., Martinez,C.A. and Lewis,R.D. (2023) Ancestral sequence reconstruction enhances gene mining efforts for industrial ene reductases by expanding enzyme panels with thermostable catalysts. *ACS Catal*, **13**, 2576–2585.

67. Zakas,P.M., Brown,H.C., Knight,K., Meeks,S.L., Spencer,H.T., Gaucher,E.A. and Doering,C.B. (2017) Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat Biotechnol*, **35**, 35–37.

68. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, **36**, D202-205.

69. Khan,R.T., Kohout,P., Musil,M., Rosinska,M., Damborsky,J., Mazurenko,S. and Bednar,D. (2025) Anticipating protein evolution with successor sequence predictor. *J Cheminform*, **17**, 34.

70. Ding,X., Zou,Z. and Brooks Iii,C.L. (2019) Deciphering protein evolution and fitness landscapes with latent space models. *Nat Commun*, **10**, 5644.

71. Vasina,M., Vanacek,P., Hon,J., Kovar,D., Faldynova,H., Kunka,A., Buryska,T., Badenhorst,C.P.S., Mazurenko,S., Bednar,D., *et al.* (2022) Advanced database mining of efficient haloalkane dehalogenases by sequence and structure bioinformatics and microfluidics. *Chem Catalysis*, **2**, 2704–2725.

72. Noé,F. (2020) Machine learning for molecular dynamics on long timescales. In Schütt,K.T., Chmiela,S., Von Lilienfeld,O.A., Tkatchenko,A., Tsuda,K., Müller,K.-R. (eds), *Machine Learning Meets Quantum Physics*, Lecture Notes in Physics. Springer International Publishing, Cham, Vol. 968, pp. 331–372.

73. Wu,H. and Noé,F. (2020) Variational approach for learning markov processes from time series data. *J Nonlinear Sci*, **30**, 23–66.

74. Mardt,A., Pasquali,L., Wu,H. and Noé,F. (2018) VAMPnets for deep learning of molecular kinetics. *Nat Commun*, **9**, 5.

75. 2024 Alzheimer's disease facts and figures (2024) *Alzheimers Dement*, **20**, 3708–3821.

76. Löhr,T., Kohlhoff,K., Heller,G.T., Camilloni,C. and Vendruscolo,M. (2021) A kinetic ensemble of the Alzheimer's Aβ peptide. *Nat Comput Sci*, **1**, 71–78.

77. Hampel,H., Hardy,J., Blennow,K., Chen,C., Perry,G., Kim,S.H., Villemagne,V.L., Aisen,P., Vendruscolo,M., Iwatsubo,T., *et al.* (2021) The Amyloid-β Pathway in Alzheimer's Disease. *Mol Psychiatry*, **26**, 5481–5503.

78. Mazurenko,S. (2020) Predicting protein stability and solubility changes upon mutations: data perspective. *ChemCatChem*, 10.1002/cctc.202000933.

79. Kumar,M.D.S., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Res*, **34**, D204–D206.

80. Wang,C.Y., Chang,P.M., Ary,M.L., Allen,B.D., Chica,R.A., Mayo,S.L. and Olafson,B.D. (2018) ProtaBank: A repository for protein design and engineering data. *Protein Sci*, **27**, 1113–1124.

81. Bushuiev,A., Bushuiev,R., Sedlar,J., Pluskal,T., Damborsky,J., Mazurenko,S. and Sivic,J. (2024) Revealing data leakage in protein interaction benchmarks. In *ICLR Workshop on Generative and Experimental Perspectives for Biomolecular Design*. Vienna.

82. Velecký,J., Berezný,M., Musil,M., Damborsky,J., Bednar,D. and Mazurenko,S. (2024) BenchStab: a tool for automated querying of web-based stability predictors. *Bioinform*, **40**, btae553.

83. Walsh,I., Fishman,D., Garcia-Gasulla,D., Titma,T., Pollastri,G., ELIXIR Machine Learning Focus Group, Capriotti,E., Casadio,R., Capella-Gutierrez,S., Cirillo,D., *et al.* (2021) DOME: recommendations for supervised machine learning validation in biology. *Nat Methods*, **18**, 1122–1127.

84. Attafi,O.A., Clementel,D., Kyritsis,K., Capriotti,E., Farrell,G., Fragkouli,S.-C., Castro,L.J., Hatos,A., Lenaerts,T., Mazurenko,S., *et al.* (2024) DOME Registry: implementing community-wide recommendations for reporting supervised machine learning in biology. *GigaScience*, **13**, giae094.

85. Vavra,O., Tyzack,J., Haddadi,F., Stourac,J., Damborsky,J., Mazurenko,S., Thornton,J.M. and Bednar,D. (2024) Large-scale annotation of biochemically relevant pockets and tunnels in cognate enzyme–ligand complexes. *J Cheminform*, **16**, 114.

86. Domínguez-Romero,E., Mazurenko,S., Scheringer,M., Martins Dos Santos,V.A.P., Evelo,C.T., Anton,M., Hancock,J.M., Županič,A. and Suarez-Diez,M. (2024) Making PBPK models more reproducible in practice. *Brief Bioinform*, **25**, bbae569.

87. Yim,J., Trippe,B.L., De Bortoli,V., Mathieu,E., Doucet,A., Barzilay,R. and Jaakkola,T. (2023) SE(3) diffusion model with application to protein backbone generation. In *Proceedings of the 40th International Conference on Machine Learning*. arXiv.

88. Benevenuta,S., Pancotti,C., Fariselli,P., Birolo,G. and Sanavia,T. (2021) An antisymmetric neural network to predict free energy changes in protein variants. *J Phys D: Appl Phys*, **54**, 245403.

89. Mardt,A., Pasquali,L., Noé,F. and Wu,H. (2020) Deep learning Markov and Koopman models with physical constraints. In *Proceedings of Machine Learning Research*.Vol. 107, pp. 451–475.

90. Giampiccolo,S., Reali,F., Fochesato,A., Iacca,G. and Marchetti,L. (2024) Robust parameter estimation and identifiability analysis with hybrid neural ordinary differential equations in computational biology. *npj Syst Biol Appl*, **10**, 139.

91. Gusmão,G.S., Retnanto,A.P., Cunha,S.C.D. and Medford,A.J. (2023) Kinetics-informed neural networks. *Catal Today*, **417**, 113701.

92. Janson,G., Jussupow,A. and Feig,M. (2025) Deep generative modeling of temperature-dependent structural ensembles of proteins. *bioRxiv*, 10.1101/2025.03.09.642148.

93. Tiberti,M., Terkelsen,T., Degn,K., Beltrame,L., Cremers,T.C., Da Piedade,I., Di Marco,M., Maiani,E. and Papaleo,E. (2022) MutateX: an automated pipeline for *in silico* saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform*, **23**, bbac074.

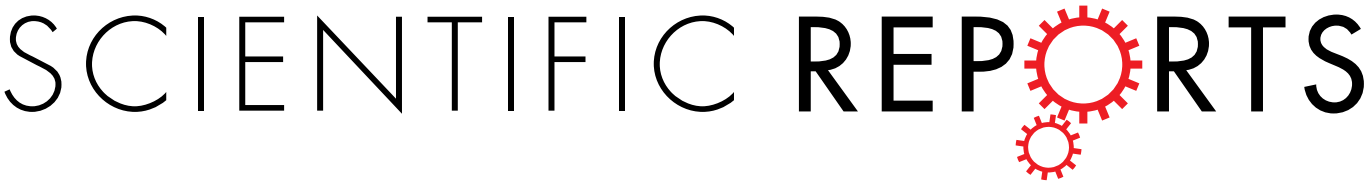
94. Barlow,K.A., Ó Conchúir,S., Thompson,S., Suresh,P., Lucas,J.E., Heinonen,M. and Kortemme,T. (2018) Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *J Phys Chem B*, **122**, 5389–5399.

95. Hopf,T.A., Ingraham,J.B., Poelwijk,F.J., Schärfe,C.P.I., Springer,M., Sander,C. and Marks,D.S. (2017) Mutation effects predicted from sequence co-variation. *Nat Biotechnol*, **35**, 128–135.

96. Musil,M., Stourac,J., Bendl,J., Brezovsky,J., Prokop,Z., Zendulka,J., Martinek,T., Bednar,D. and Damborsky,J. (2017) FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res*, **45**, W393–W399.

97. Ertelt,M., Meiler,J. and Schoeder,C.T. (2024) Combining Rosetta Sequence Design with Protein Language Model Predictions Using Evolutionary Scale Modeling (ESM) as Restraint. *ACS Synth Biol*, **13**, 1085–1092.

Selected Publications



OPEN

Exploration of Protein Unfolding by Modelling Calorimetry Data from Reheating

Received: 3 August 2017
Accepted: 10 November 2017
Published online: 24 November 2017

Stanislav Mazurenko[1], Antonin Kunka[1,2], Koen Beerens[1], Christopher M. Johnson[3], Jiri Damborsky[1,2] & Zbynek Prokop[1,2]

Studies of protein unfolding mechanisms are critical for understanding protein functions inside cells, de novo protein design as well as defining the role of protein misfolding in neurodegenerative disorders. Calorimetry has proven indispensable in this regard for recording full energetic profiles of protein unfolding and permitting data fitting based on unfolding pathway models. While both kinetic and thermodynamic protein stability are analysed by varying scan rates and reheating, the latter is rarely used in curve-fitting, leading to a significant loss of information from experiments. To extract this information, we propose fitting both first and second scans simultaneously. Four most common single-peak transition models are considered: (i) fully reversible, (ii) fully irreversible, (iii) partially reversible transitions, and (iv) general three-state models. The method is validated using calorimetry data for chicken egg lysozyme, mutated Protein A, three wild-types of haloalkane dehalogenases, and a mutant stabilized by protein engineering. We show that modelling of reheating increases the precision of determination of unfolding mechanisms, free energies, temperatures, and heat capacity differences. Moreover, this modelling indicates whether alternative refolding pathways might occur upon cooling. The Matlab-based data fitting software tool and its user guide are provided as a supplement.

Understanding the mechanisms of protein folding and unfolding is of particular importance to identifying relationships between amino acid sequences and protein function and stability. These mechanisms are crucial for comprehensive protein engineering, ranging from the *de novo* design of proteins[1,2] to analysis of different variants of existing proteins and their biological function[3]. It has also been reported that protein misfolding and aggregation are primary causes of many human diseases[4], and therefore knowledge of protein folding mechanisms may help to develop effective treatments. Although there have been significant advances in protein unfolding simulations *in silico* recently[5], their experimental validation and measurement of protein stability usually have to be performed indirectly. Several experimental techniques can be used either separately or in combination, e.g. high resolution hydrogen-deuterium exchange methods[6], nuclear magnetic resonance spectroscopy coupled with mass spectroscopy[7] as well as less expensive methods such as differential scanning calorimetry (DSC)[8], circular dichroism, and fluorescence spectroscopy[5,9,10]. In this paper, our main interest lies in DSC.

In DSC, the native state of the protein is perturbed by increasing the temperature and the difference in the heat capacity between sample and reference cells is recorded. This technique is one of the most powerful methods of protein folding analysis as it records the energetic profile of unfolding directly in terms of the amount of heat necessary to unfold a protein. As summarized in a number of reviews[8,11–13], DSC studies have already had a great impact on the current understanding of protein stability and its energetic profiles. In particular, DSC has contributed towards (1) the currently accepted framework of temperature dependence in studies of protein stability, heat and cold denaturation[14,15]; (2) quantification of the interplay between equilibrium thermodynamics and kinetics[13,16]; (3) our understanding of structure-energy relationships in proteins, bridging the gap between experimental folding/unfolding data and *in silico* protein models and energy landscapes[12,17,18]; and (4) insights into aggregation mechanisms and the unfolding intermediates involved[19].

[1]Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, Kamenice 5/A13, 625 00, Brno, Czech Republic. [2]International Clinical Research Center, St. Anne's University Hospital, Pekarska 53, 656 91, Brno, Czech Republic. [3]Biophysics Facilities, MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, UK. Correspondence and requests for materials should be addressed to J.D. (email: jiri@chemi.muni.cz) or Z.P. (email: zbynek@chemi.muni.cz)

Since modern instruments now provide a high precision of DSC measurements, proper data analysis is crucial for understanding the data collected. The popularity of DSC measurements stems from the fact that complete energy profiles of unfolding can be analysed to quantify unfolding pathways by mathematical modelling and curve fitting. The modelling is usually based on the premise that protein stability comes in two different forms: thermodynamic stability in terms of the low fraction of unfolded protein *versus* folded protein in equilibrium and kinetic stability in terms of energy barriers separating the native and unfolded states[16,20]. A set of parameters for both types of stability can be obtained from analysis of DSC data since the thermogram data can be curve-fitted to analytically or numerically derived solutions for a given unfolding mechanism[21–23].

It has also often been reported that proteins undergoing heating in DSC show unmistakable signs of irreversible transition[24]. Protein engineering of more stable variants usually involves an increase in melting temperatures, shifting the transition to the denatured state to higher temperatures where irreversibility more commonly occurs. Multi-domain proteins sometimes exhibit irreversible denaturation due to domain interactions upon unfolding and/or irreversible changes to secondary structures, e.g. decreased fractions of α-sheet and β-turn conformations and increased fraction of α-helix upon thermal unfolding of mouse monoclonal immunoglobulin[25,26]. To perform proper modelling of such irreversible denaturation, two techniques are commonly used, i.e. using different scan rates and reheated runs[23]. The latter is mostly used to draw inferences about reversibility in general by repeated unfolding/refolding experiments to high temperatures. Only a few articles have dealt with reheating in a more sophisticated way, e.g. for decomposition of peaks[27], calculation of the proportion of irreversibly denatured protein at different temperatures[28], and analysis of the DSC profiles of irreversibly denaturing multidomain proteins[29]. While the abovementioned studies have provided valuable insights into the process of unfolding, only a limited amount of information from reheated runs has been captured for data analysis. However, curves obtained from reheated runs are usually recorded at the same number of temperature points as first runs, and thus can also be used for curve fitting. Their information content goes arguably far beyond that of the first run, and global fitting both runs can substantially enhance the modelling. Indeed, apart from the shape of the curve, reheating curves contain data on the change in the native state of a protein as a function of temperature.

This paper aims to demonstrate how data from reheating runs can help determine protein unfolding mechanisms, such as the number of intermediate states, reversibility of each transition and alternative refolding pathways. We also give explicit equations for fitting curves from reheated runs and subsequent quantification of states in terms of activation energies, enthalpies, entropies, Gibbs energies, critical temperatures, and heat capacity changes. While the techniques presented in this paper are general and can be applied to various models, this paper only covers the four most common fitting models for apparent single peak transitions, namely a (A) fully reversible transition, (B) fully irreversible transition, (C) partially reversible transition with equilibrium at the first step, and (D) general three-state model. Fully reversible transitions are of little interest in the current framework because their reheated runs are expected to almost precisely follow the first runs. Consequently, they do not contribute any new information apart from evidence of full reversibility. Conversely, as far as irreversible transitions are concerned, there seems to be no upper limit on the possible complexity of models describing protein denaturation. Hence, we limited ourselves to basic models demonstrating major derivation principles, according to which more complicated models may be extended to include reheating. It should be noted that proteins demonstrating complex, e.g. multi-peak, DSC profiles of unfolding must be modelled with extra care since their dynamics may rarely be described by a precise kinetic model and may include aggregation with considerable complexity[30]. Moreover, the methodology used in this paper is based on discreet macrostates of unfolding pathways, such as native, intermediate, denatured states, etc., and statistical free energy surface models of microstates[11] were beyond the scope of this study.

The suggested method was tested on DSC thermograms of wild type chicken egg lysozyme, wild type haloalkane dehalogenases LinB, DbjA and DhaA, the mutant DhaA115 thermostabilized by protein engineering and the mutant of Protein A from *Staphylococcus aureus* SpA. The proposed methodology was implemented as a graphic user interface for fitting based on MATLAB 2016a (MathWorks, United States). A link to a computer program calculating modelled heat capacities for the four basic mechanisms of unfolding as well as some more complex models is provided in DSC data analysis section of Materials and Methods. It can be used for global curve fitting for different scan rates and reheating.

Materials and Methods

Theoretical Basis. Modelling of the cooling and reheating processes was similar to existing models for the first scan based on explicit formulas used to fit apparent heat capacity data. As far as experiments are concerned, it is often expedient to conduct cooling and reheating at rates similar to that of the first run to ensure that the mechanism of folding/unfolding is not disrupted by a change in scan rate. On the other hand, if no such disruption is expected, e.g. unfolding is fully irreversible, the cooling rate might not necessarily be the same as the scan rate. It is also advisable to verify that the temperature profile of the heating, cooling, and reheating is linear in time and does not have any artefacts, especially at high temperatures. An example of such a profile obtained for the analyses in this paper is given in Supplement 2.

The assumptions used for mathematical modelling were as follows:

- The process of unfolding was represented as a sequence of steps, e.g. the following equation

$$N \overset{K}{\longleftrightarrow} I \overset{k}{\longrightarrow} D \tag{1}$$

  stands for a three-state unfolding reaction, in which the first step (from native to intermediate states) is reversible and characterized by an equilibrium constant K, whereas the second step (from intermediate to denatured states) is irreversible with rate constant k.

- At the beginning of the DSC scan, the fraction of protein in states other than N was assumed to be negligibly small.
- Each equilibrium constant K as a function of temperature was parameterized as follows:

$$K(T) = \exp\left\{-\frac{\Delta G(T)}{RT}\right\}, \tag{2}$$

where R is the gas constant and ΔG is the differences in Gibbs energies of the respective states:

$$\Delta G = \Delta H(T) - T\Delta S(T). \tag{3}$$

Here ΔH stands for the enthalpy change and ΔS is the change in entropy.

- Each rate constant k for an irreversible step was assumed to satisfy the Arrhenius equation:

$$k(T) = \exp\left\{-\frac{E}{R}\left(\frac{1}{T} - \frac{1}{T_f}\right)\right\}, \tag{4}$$

where T_f is the temperature at which $k = 1$ and E is the energy of activation for the respective step.

- The difference in heat capacities ΔC_p between different states was assumed to be independent of T. In the case ΔC_p does depend on the temperature, the modeling will estimate ΔC_p value that will correspond to the average ΔC_p over the temperature range of transition[31]. Hence, ΔH and ΔS were functions of temperature and the ground level had to be defined. In line with previous studies, we selected T_m and T_f as reference points of the ground state for reversible and irreversible unfolding, respectively.

We will now briefly summarize the mathematical treatment of reheating for the four simple models of unfolding. Further details and final equations used for fitting can be found in Supplement 1.

(A) ***Reversible two-state denaturation***

$$N \overset{K}{\longleftrightarrow} D \tag{5}$$

In this case, there is an explicit equation for the heat capacity as a function of T_m, the melting temperature, i.e. the temperature at which half of the protein is denatured, ΔC_p, the constant change in heat capacity between the folded and denatured states, and ΔH, the enthalpy change at T_m. For totally reversible protein unfolding, the reheated run should match the first run. It should be noted that the modelled reheated run should follow the first run in any equilibrium fully reversible model of unfolding, e.g. multi-step model based on calculation of van't Hoff's enthalpy[9], given that the cooling scan is performed at the same scan rate as the first run. This follows from the fact that the rate of approaching a new equilibrium is the sum of the rates of folding and unfolding. Thus, if a fully reversible model is valid, and equilibrium is assumed to take place during heating, the time needed for a protein to refold is exactly the same as the time of unfolding. Hence, there should be no change to the thermogram during reheating as compared to the first run.

(B) ***Irreversible two-state denaturation***

$$N \overset{k}{\longrightarrow} D \tag{6}$$

This model is often considered as a simplification of the more general Lurmy–Eyring model (see models C and D) when the intermediate state I is barely populated due to faster transition to state D during the scan. If we define the relative concentrations of the states as x_n and $x_d = 1 - x_n$ respectively, the equation for the heat capacity is as follows:

$$C_p(T) = B_0 + B_1 T + (1 - x_n(T))\Delta C_p + \frac{k(T)}{\nu} x_n(T)\Delta H(T), \tag{7}$$

where

$$x_n(T) = X(T, T_0, \nu)x_n(T_0). \tag{8}$$

Here T_0 is the initial temperature (low enough to ensure that $x_n = 1$, i.e. all the protein is in the native state) and $X(T_2, T_1, \nu)$ represents the decay factor of the native state relative concentration from temperature T_1 to T_2 given the scan rate ν. In other words, it shows the ratio of the protein concentration in the native state at temperature T_2 to that at temperature T_1 after changing the temperature at a constant rate of ν. If the first run is stopped at temperature T', the terminal amount of protein in the native state will be $x_n(T') = X(T', T_0, \nu)x_n(T_0)$, or if we assume $x_n(T_0) = 1$, it is $x_n(T') = X(T', T_0, \nu)$. Hence, after cooling to temperature T_0 at a rate ν, this amount is reduced to

$$x_n(T; T') = X(T_0, T', -\nu)x_n(T') = X(T', T_0, \nu)x_n(T') = X(T', T_0, \nu)^2. \tag{9}$$

Subsequent reheating results in a fraction of the protein in its native state equal to

$$x_n^R(T;\ T') = X(T,\ T_0,\ \nu)x_n(T_0;\ T') = X(T,\ T_0,\ \nu)X(T',\ T_0,\ \nu)^2. \tag{10}$$

The heat capacity for the reheated run is then as follows:

$$C_p^R(T;\ T') = B_0 + B_1T + (1 - x_n^R(T;\ T'))\Delta C_p + \frac{k(T)}{\nu}x_n^R(T;\ T')\Delta H(T). \tag{11}$$

(**C**) ***Partially reversible three-state denaturation with equilibrium***

$$N \overset{K}{\longleftrightarrow} I \overset{k}{\longrightarrow} D \tag{12}$$

This is a more general model in which an irreversible step follows reversible unfolding. It is assumed that the rates of the reaction at the first step allow approximation of the step with equilibrium constant K. As in (B), we define the relative concentrations of the states as x_n, x_i and $x_d = 1 - x_n - x_i$, respectively. Then, according to already published results[23,32]:

$$x_i = Kx_n,\ \ x_n = \frac{1 - x_d}{1 + K},\ \ \frac{dx_n}{dT} = -\frac{K}{1 + K}x_n\left(\frac{k}{\nu} + \frac{\Delta H_R(T)}{RT^2}\right). \tag{13}$$

There is one differential equation for x_n left; thus one decay factor for the native state from T_1 to T_2 given the scan rate ν as $X_I(T_2, T_1, \nu)$. Following the same logic as for model B, the terminal amount of protein in the native state after the first run up to temperature T' will be $x_n(T') = X_I(T', T_0, \nu)x_n(T_0)$. After cooling to temperature T_0 at rate ν and reheating, the following equation applies:

$$x_n^R(T;\ T') = X_I(T,\ T_0,\ \nu)X_I(T_0,\ T',\ -\nu)X_I(T',\ T_0,\ \nu). \tag{14}$$

Here, we again assumed $x_n(T_0) = 1$. Thus, the formula for the heat capacity of the reheated run is the same as that for the first run but with x_n^R substituted for x_n. Direct numerical integration was used to calculate the decay factor X_I as there is no explicit solution currently available.

(**D**) ***General partially reversible three-state denaturation***

$$N \overset{k_1,k_{-1}}{\longleftrightarrow} I \overset{k_2}{\longrightarrow} D \tag{15}$$

This is a classical Lumry-Eyring model, in which the first step is not approximated by an equilibrium constant as in (C), rather it is parameterized by two rate constants: k_1 for the forward reaction and k_{-1} for the reverse one. In this case, there are two differential equations governing the temperature changes in protein fractions that have to be integrated numerically[21], and consequently, two decay factors that have to be found for the first, cooling and reheated scans.

More complicated models of unfolding can be supplemented with formulas for reheating according to principles similar to those in the above four models. The computer software detailed in the supplementary material includes several more complex models apart from the four presented here. Nonetheless, difficult cases that require additional steps should be treated with caution since the model of unfolding may be exceedingly complex, e.g. include protein-protein interactions.

Protein sample preparation. Chicken egg white lysozyme (lot BCBM6718V) was purchased from Sigma-Aldrich (USA). The His6-tagged haloalkane dehalogenases DbjA, LinB, DhaA and DhaA115 variant were overexpressed in *Escherichia coli* BL21 (DE3) cells as previously described[33]. Proteins were purified using Ni-NTA Superflow Cartridges (Qiagen) and a previously described method[34]. Protein samples were dialyzed to 50 mM potassium phosphate buffer, pH 7.5 and their concentration was determined by the Bradford assay from the calibration curve of bovine serum albumin. Lysozyme concentration was determined spectroscopically by absorption measurement at 280 nm and using the calculated extinction coefficient of 37,970 M^{-1} cm^{-1}. The purity of purified proteins was checked via densitometric analysis using a GS-800 Calibrated Densitometer (Bio-Rad, USA) after sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) followed by Coomassie Brilliant Blue R-250 staining. The experimental data for the mutant of SpA (L20A + Y15W) were collected as given in the study of Sato and coworkers[35].

DSC experiments. DSC measurements were performed using MicroCal VP-Capillary DSC system (GE Healthcare, Sweden). Prior to scanning, samples were degassed under vacuum for 15 min using MicroCal ThermoVac (GE Healthcare, Sweden). DSC thermograms were determined by monitoring the difference in heat capacity in solution upon increasing temperature at a scan rate of 1 °C min^{-1}, followed by cooling and subsequent re-heating of the sample at the same scan rate to the same final temperature as in the first scan. While this final temperature is a limitation of MicroCal VP-Capillary instrument, the method and software provided in the Supplement do not have this limitation and are able to model reheated runs for any temperature ranges. The time delay between the end of heating and start of cooling was set at zero. Moreover, temperature profiles of the DSC instrument were collected for different scan rates to ensure that the temperature changed linearly in time and no artefact took place at high temperatures upon the start of the cooling (see Supplement 2). Scans were performed under increased pressure (3 atm) and varying terminal temperatures for consecutive scans were determined from the initial thermogram obtained by heating the sample from 20 °C to 100 °C. All proteins used in this study

(A) Reversible two-state denaturation $N \overset{K}{\longleftrightarrow} D$	(B) Irreversible two-state denaturation $N \overset{k}{\longrightarrow} D$
(C) Partially reversible three-state denaturation with equilibrium $N \overset{K}{\longleftrightarrow} I \overset{k}{\longrightarrow} D$	(D) General partially reversible three-state model without equilibrium $N \overset{k_1,k_{-1}}{\longleftrightarrow} I \overset{k_2}{\longrightarrow} D$

Table 1. Definition of the models and respective schemes for protein unfolding.

were extensively dialyzed against 50 mM potassium phosphate buffer, pH 7.5, and dialysis buffers were used for instrumental baseline scans and as reference samples. Protein concentrations used were typically between 1.0–1.5 mg mL^{-1}, corresponding to 70–105 μM and 30–45 μM for egg white lysozyme and the dehalogenases, respectively.

DSC data analysis. After each data set was collected, the buffer-buffer baselines were subtracted and then concentration normalization was performed to obtain apparent heat capacity per mole of protein. The data were then exported into Excel and fed to a computer program written in MATLAB (MathWorks) for curve fitting (CalFitter). The software, source code, and binaries are freely available for download at http://loschmidt.chemi.muni.cz/peg/software/calfitter. The program uses standard built-in functions from Optimization and Statistics toolboxes and allows its users to simultaneously fit data with reheating and different scan rates according to specified models, most basic of which were discussed in detail above. The developed software tool is freely available, and the link can be found in the Supplement.

The number of steps for unfolding mechanisms in each one of our examples was selected according to the following conservative rule: the first apparent peak can be modelled by at most two steps (models C and D), with the first step being reversible and the second step corresponding to the loss of reversibility, and each subsequent apparent peak in the thermogram is modelled by a single step to avoid over-fitting.

Results and Discussion

Here, we focus our attention on the four most commonly used models of protein unfolding (Table 1). For derivation of the formula for reheating as well as a detailed explanation of parameters, please see the Materials and Methods section. The key challenge in distinguishing between the four models is that models A, B, and C are in fact limiting cases of model D. Model A represents the case when the rate constant k_2 is negligibly close to zero. Model B is an approximation of model D when $k_2 \gg k_1$ and k_{-1}, in which case the apparent rate k corresponds to either k_1 or k_2K depending on whether k_{-1} is small or large compared to k_1, respectively[36]. Finally, model C is the limiting case of model D when $k_1 + k_{-1} \gg k_2$. Hence, the first step equilibrates at a much higher rate than the second step proceeds. Therefore, fitting the data from only one scan is usually insufficient for proper model selection and additional techniques have to be used to discriminate between the models, such as varying the scan rate. However, in some cases, even using different scan rates may still not be enough and reheating may be the only solution. The modelling of reheating also provides other advantages, such as a better estimation of the heat capacity change, but comes at a cost – alternative refolding pathways must be discarded first before a final decision about the models is made. In what follows, we will elaborate on the above mentioned points, discuss the possible procedure for final temperature selection and present results of data analysis for various proteins using the software provided in the Supplement.

Enrichment of scan rate dependence with reheating. Initially, we compared the proposed method with the well-established technique of changing scan rates during DSC experiments and demonstrated how considering reheating may improve analysis of protein unfolding in both qualitative and quantitative ways.

One of the most commonly used approaches for the study of irreversible protein denaturation and verification of the selected model is to vary the scan rate. Several important equations and analysis in this respect can be found in literature[21,23,24,28,32]. The basis of this approach is to change the scan rates in DSC experiments and then compare the apparent shifts in DSC curves/peak temperatures with those predicted by unfolding models being tested. However, this method has several drawbacks.

Although the scan rate dependence of the thermogram may indicate that model A is not valid, the main weakness of the method of varying scan rates is that it poorly discriminates between models B, C, and D. Indeed, consider the following example for the DhaA115 mutant (Fig. 1). Model B gives a reasonably good fit for different scan rates (Fig. 1A). However, this simple two-step transition model fails to explain the reheated runs (Fig. 1B). On the contrary, if reheating is taken into consideration and a global fit performed, model D turns out to be the simplest model that accounts for both the scan-rate dependence and reheated runs (Fig. 1C,D). The main reason for such behaviour is that there has to be a reverse component of unfolding at the first step to account for the reheating data. Hence, model B is not applicable. Due to a very fast drop in the reheated peaks, the first step cannot be approximated by an equilibrium. Consequently, model C can also be eliminated. As a result, the simplest model that can explain the data with great approximation is model D, in which the first step is described by two rate constants.

Another potential problem of using different scan rates stems from the different heating rates, which might shift the model from C, in which equilibrium at the first step can be assumed, to D, in which no such simplification can be made, due to different values of the rate constants adjusted by the scan rate k/v. For instance, equilibrium at the first step may no longer be attained if k/v of the second irreversible step is significant, thus driving the protein from the intermediate to final state more rapidly. Conversely, in the analysis of the reheated runs, all the

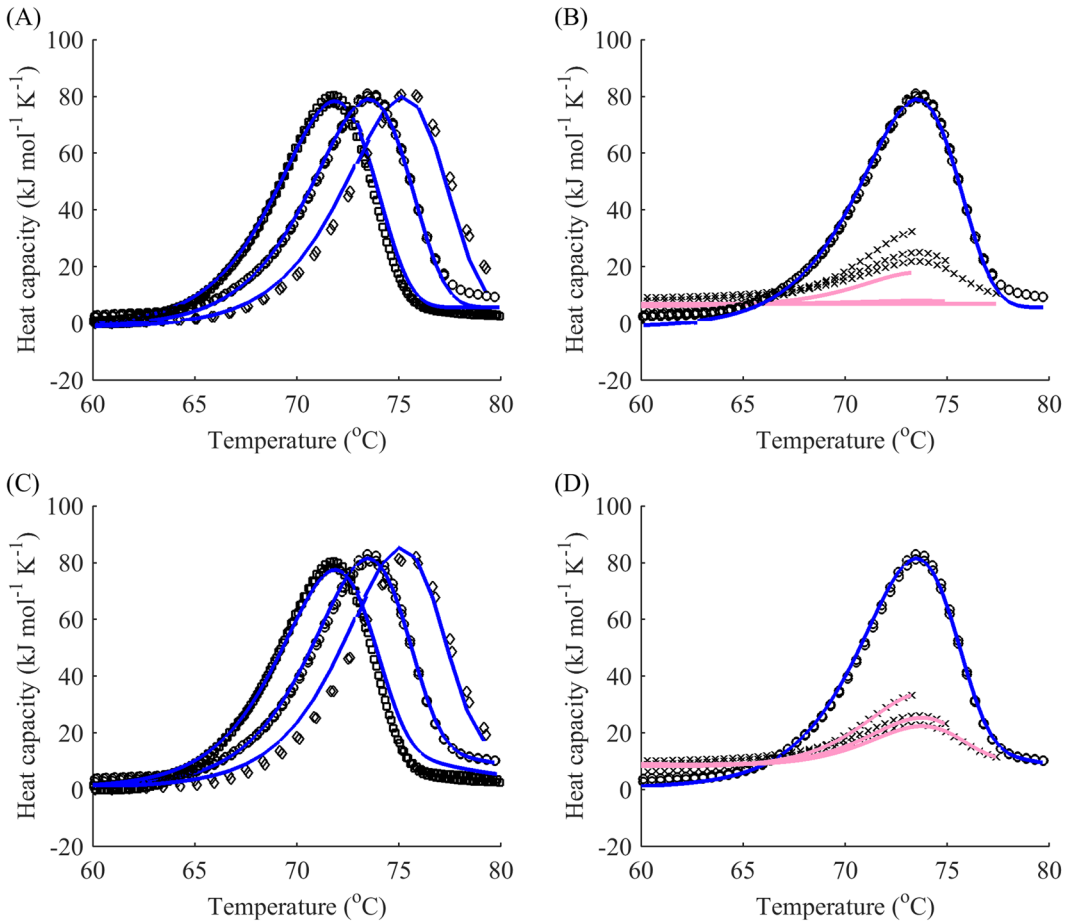


Figure 1. Fitting to different scan rates alone fails to account for reheating. (**A**) Global fitting of model B (blue) to DSC data (black) for denaturation of the DhaA115 mutant while disregarding reheated runs. The scan rates were 0.5 (□), 1 (○) and 2 °C min^{-1} (◇). (**B**) Resulting reheated runs (pink) and actual reheated runs (black crosses) for a scan rate of 1 °C min^{-1}. (**C**) Global fitting of model D (blue) to DSC data (black) for denaturation of the DhaA mutant with reheated runs. The scan rates were 0.5 (□), 1 (○) and 2 (◇)°C min^{-1}. (**D**) Resulting reheated runs (pink) and actual reheated runs (black crosses) for a scan rate of 1 °C min^{-1}.

parameters of the process except for the direction of heating/cooling remain the same, providing a better tool for model verification.

Finally, some difficulty in performing a global fit with different scan rates may occur as the corresponding DSC curves are usually shifted vertically and sometimes have different base slopes (notice the differences in heights of the peaks for modelled and actual data in Fig. 1 for different scan rates). Hence, to carry out a global fit, one should either shift them manually to some predefined starting point or introduce additional individual parameters for different baselines, which would further complicate the calculations. This is usually one of the main reasons for using limited information, e.g. dependence of the peak temperature on scan rate alone, rather than fitting the whole curves simultaneously. However, such problems do not seem to arise when analyzing reheated runs as they behave in a similar manner to first runs. Therefore, data gathered from different experiments with independent protein batches can be superimposed according to the first runs (Fig. 2).

Study of the heat capacity difference (ΔC_p) effect. It is often observed that the pre-transitional baseline in a DSC thermogram is lower than the post-transitional baseline, indicating that there is a positive heat capacity difference between the denatured and native states. Such a phenomenon is usually attributed to hydration of the protein residues that are exposed to water upon protein unfolding[37]. Reheating runs provide additional information about the heat capacity difference between the native and denatured states. Indeed, different starting points of the reheated run can be indicative of ΔC_p accumulation during unfolding. This improves the precision of the estimate from fitting because no manual baseline subtraction is needed. On the contrary, this subtraction may decrease the information content of the data and distort the outcome of the fitting. The gradual shift of reheating data with respect to the first run with increasing terminal temperature is shown in Fig. 3. Since after the first run, some portion of the protein is irreversibly denatured, its heat capacity differs from that of the native state by exactly ΔC_p. The change in the slope of the reheating runs with temperature indicates that ΔC_p is temperature-dependent and this dependence can also be included in the modelling and estimated with high accuracy in global fitting to first and reheated runs.

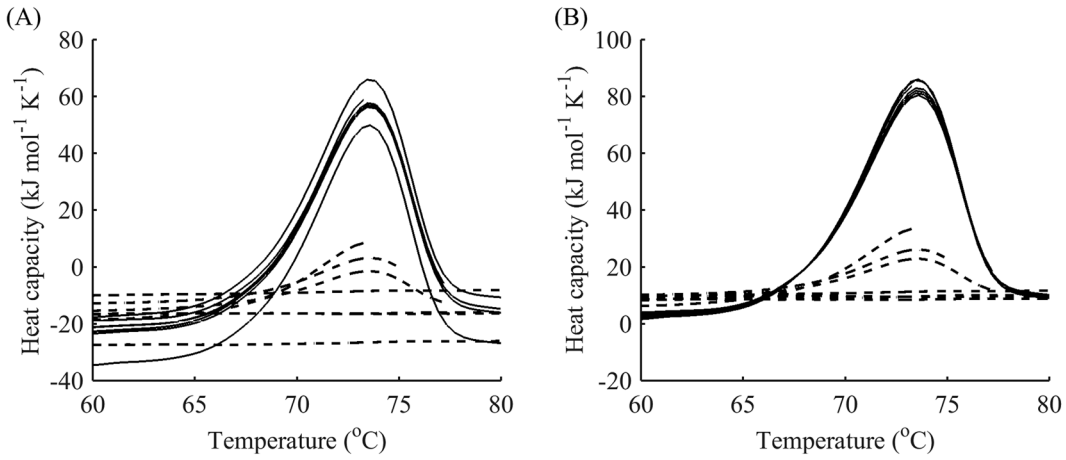


Figure 2. Superimposition of data from different experiments according to the first run. DSC data of denaturation of the DhaA115 mutant: raw data (**A**) and superimposed data (**B**) allows global fitting without any additional parameters.

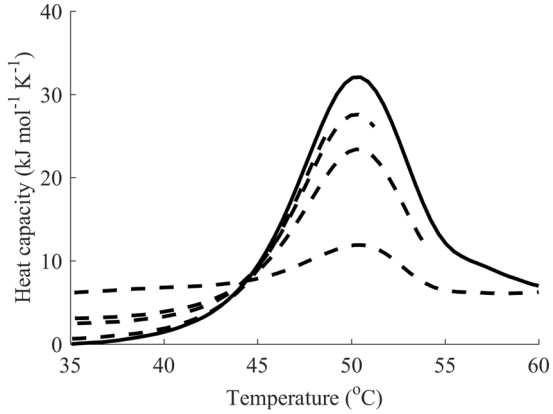


Figure 3. Vertical shift of data from reheating accounts for a non-zero ΔC_p. DSC data (black) for the first peak of denaturation of DhaA: first runs (solid line), reheated runs for terminal temperatures 51, 54, 61 and 69 °C (dashed line). The reheating data gradually increases with increasing final temperature, indicating that the total ΔC_p for the first peak is around 7 kJ mol^{-1} K^{-1}, which is further supported by the fitting (see Table 2).

Nonetheless, the proposed method revealed several drawbacks with regard to ΔC_p estimations. First, some vertical drift of the data between the two runs may occur due to errors in measurement rather than unfolding, which is why it is expedient to perform replicates of the runs for a higher precision of ΔC_p estimation. Moreover, unfortunately, the method usually fails to distinguish between ΔC_p of the two consecutive steps in models C and D if they contribute to the same apparent peak. This is because reheating usually highlights the difference in the native fraction of the heat capacity and irreversibly denatured one, which includes both ΔC_{pR} and ΔC_{pI} (or ΔC_{p1} and ΔC_{p2} for model D). It should be possible to separate those two values if at least some fraction of the intermediate state can be preserved at the beginning of reheating. However, in this study, only a combined estimate of ΔC_p was achieved.

In a similar way, aggregation and refolding might result in the same apparent ΔC_p changes upon reheating if the former takes place at unfolding temperatures. However, if two transitions are separated from each other in the thermogram, the reheating analysis may be conducted for each transition separately (for details, see the section "Selection of optimal points for reheating" below), which helps to quantify ΔC_p contributions by different steps during unfolding. For instance, ΔC_p of aggregation is likely to manifest during reheating from high temperatures, whereas ΔC_p of unfolding should appear during reheating from the peak temperature, similar to the case in Fig. 3.

Analysis of alternative refolding. It has been reported in the literature that some proteins exhibit a DSC profile of reheating that does not correspond well to the first run, whereby more complex schemes of unfolding have to be applied[38]. Simultaneous modelling of the first and reheated runs may provide additional information regarding the extent to which the simple models agree with the data.

Protein	Model*	1st step	2nd step
Lysozyme wild type	Partially reversible three-state equilibrium (C)	$T_m = 345.03 \pm 0.05$ K $\Delta H = 516 \pm 3$ kJ mol^{-1} $\Delta C_p = 7.5 \pm 0.7$ kJ mol^{-1}K^{-1}**	$E = 18 \pm 1$ kJ mol^{-1} $T_f = 3.2 \pm 1.4$ kK $\Delta H = -37 \pm 10$ kJ mol^{-1}
SpA mutant	Reversible two-state (A)‡‡	$T_m = 328.44 \pm 0.07$ K $\Delta H = 102.4 \pm 0.7$ kJ mol^{-1} $\Delta H_{vh} = 169 \pm 1$ kJ mol^{-1} $\Delta C_p = 1.38 \pm 0.04$ kJ mol^{-1}K^{-1}	N/A
DbjA wild type	Irreversible two-state (B)	$E = 418 \pm 3$ kJ mol^{-1} $T_f = 337.6 \pm 0.1$ K $\Delta H = 337 \pm 3$ kJ mol^{-1} $\Delta C_p = 5.4 \pm 0.1$ kJ mol^{-1}K^{-1}	N/A
LinB wild type	Irreversible two-state (B)	$E = 294 \pm 3$ kJ mol^{-1} $T_f = 338.6 \pm 0.2$ K $\Delta H = 397 \pm 5$ kJ mol^{-1} $\Delta C_p = 8.0 \pm 0.1$ kJ mol^{-1}K^{-1}	N/A
DhaA wild type	Partially reversible three-state equilibrium (C)	$T_m = 323.8 \pm 0.1$ K $\Delta H = 338 \pm 2$ kJ mol^{-1}	$E = 75 \pm 13$ kJ mol^{-1} $T_f = 436 \pm 27$ K $\Delta H = 70 \pm 19$ kJ mol^{-1}*** $\Delta C_p = 10.2 \pm 0.5$ kJ mol^{-1}K^{-1}**
DhaA115 mutant	General three-state (D)	$E_1 = 436.5 \pm 0.1$ kJ mol^{-1} $T_{f1} = 358.0 \pm 0.01$ K $E_{-1} = 46.7 \pm 0.1$ kJ mol^{-1} $T_{f-1} = 689.4 \pm 0.1$ K $\Delta H = 596 \pm 5$ kJ mol^{-1}	$E = 109.1 \pm 0.1$ kJ mol^{-1} $T_f = 431.6 \pm 0.1$ K $\Delta H = -160 \pm 40$ kJ mol^{-1}‡ $\Delta C_p = 6.1 \pm 0.4$ kJ mol^{-1}K^{-1}**

Table 2. Results of the global fitting of DSC thermograms for various proteins. *The model is defined for the main peak; **ΔC_p is given as a combined value for two steps; N/A – not applicable. The values are given with 95% confidence intervals from the fitting; ‡ΔH values calculated at the apparent peak temperature and, therefore, resembling the area under the peak; the values used for modelling of $\Delta H(T)$, i.e. values of $\Delta H(T_f)$, were 1126 ± 301 and -566 ± 40 kJ mol^{-1} for DhaA wild type and 115, respectively; ‡‡The reversible model was augmented by the van't Hoff enthalpy.

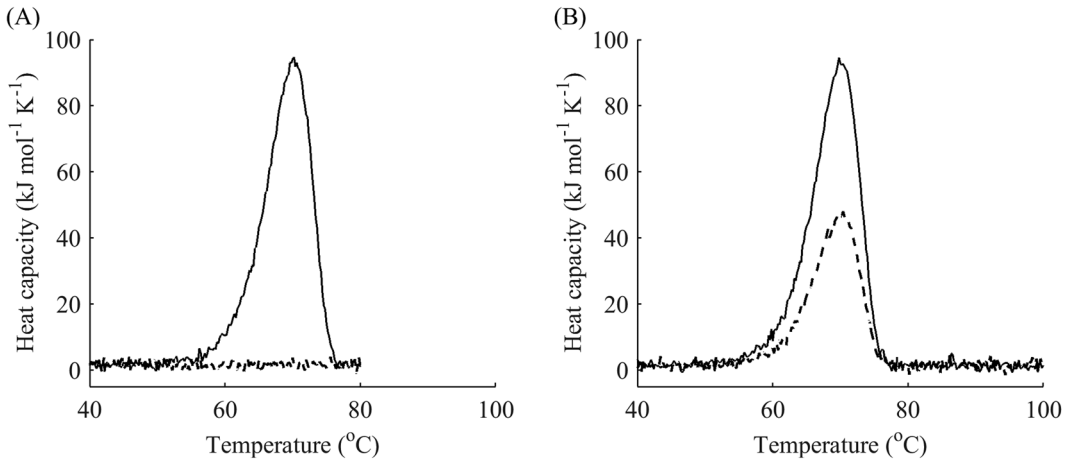


Figure 4. Simulated DSC data suggesting an alternative refolding conformation: (**A**) the first dataset at a terminal temperature of 80 °C (reheating represented by dashed line) showing no signal during reheating, whereas (**B**) the second dataset at a terminal temperature of 100 °C shows a reheating peak with area of 50%.

We calculated residuals from global fitting and separated them into two groups: those based on data from the first run and those based on reheating data. Next, average distances and standard deviations were calculated for the respective groups. We assumed that if the average distances and standard deviations were of the same magnitude, the data did not suggest that there was an alternative conformation upon refolding. We selected a practical threshold of 5% in the signal units based on the precision of the concentration measurements. Thus, if the threshold is not surpassed, the conformation of the refolded protein resembles that of the native state. Moreover, due to the smaller magnitude of the signal, the calculated distance and standard deviation of the reheating run might be lower than those of the first run. Hence, we considered the extreme case where after fitting, the calculated average distance and standard deviation of the reheating case were significantly higher, indicating that the model predicted reheating significantly less precisely than the first run. We tested the methodology on a simulated case that produced datasets similar to the denaturation of lipase from *Thermomyces lanuginosa*[38]. In those experiments, reheating from a temperature immediately after the main transition (80 °C) did not result in any signal during reheating, whereas reheating from 100 °C showed a signal of almost 50% of that during the first run. We simulated the first scan as a one-step irreversible model ($E_a = 300$ kJ mol^{-1}, $T_f = 360$ K, $\Delta H = 800$ kJ mol^{-1}) with

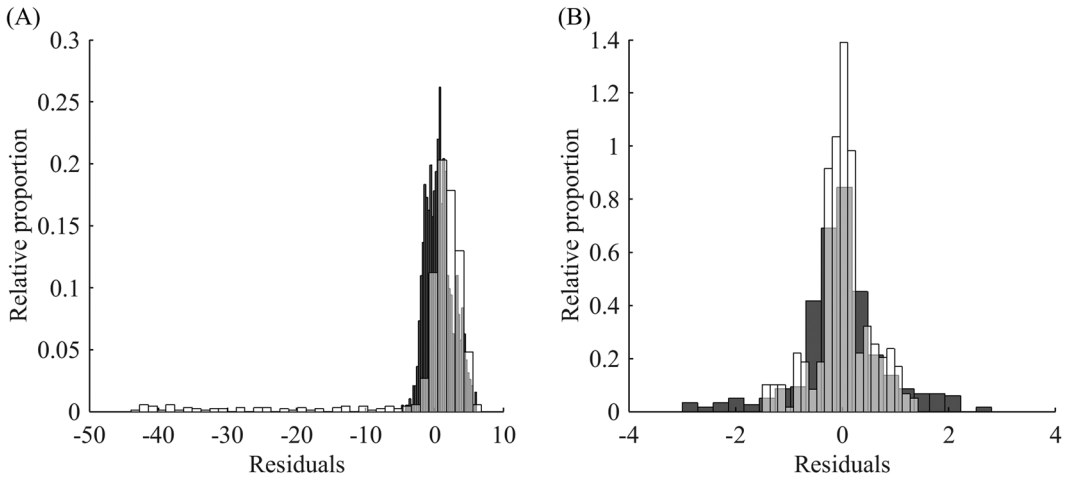


Figure 5. Analysis of alternative refolding for the simulated case *versus* DbjA. (**A**) Histograms of residuals from the first run (black) and reheating (white) of the simulated dataset. (**B**) DbjA with clear one-step irreversible unfolding: the residuals from the first run (black) and reheating run (white) are in good agreement regarding their means and standard deviations.

added normally distributed noise, and the second scan followed the same model but with a reduced ΔH and from an alternative native state (Fig. 4). If the second peak had an area of 50% of the first peak, the calculated standard deviation of the residuals for the second run was 4-times higher than that of the first run and the value of noise used in the simulations. Even when the area under the peak for reheating was lowered to as little as 10% of the first run, the calculated standard deviation of the residuals from the second run was still 2-times higher, which should raise concerns. Hence, the alternative refolding defined by the authors of the article cited above may also arise as a result of fitting using the methodology proposed in this article.

To demonstrate the difference in analysis of the datasets with possible alternative refolding and ones with no signs of alternative refolding, we compared the residuals and calculated statistics of the simulated datasets described above with fitted calorimetry data for DbjA, whose DSC thermogram was found to be in perfect agreement with a one-step irreversible model (Fig. 5). The residuals of the simulated case (Fig. 5A) showed significant bias of the reheating as well as substantial deviation of its residuals from the average. The residuals from the first run had a calculated standard deviation of 2.1 kJ mol^{-1} K^{-1}, whereas residuals for the reheating run exhibited a calculated standard deviation of 8.8 kJ mol^{-1} K^{-1}. In the case of DbjA, the residuals from the first and reheating runs were in good agreement in terms of their means and standard deviations, the latter being 0.85 kJ mol^{-1} K^{-1}and 0.61 kJ mol^{-1} K^{-1} for the first and second runs, respectively (Fig. 5B).

Selection of optimal points for reheating. Next, we decided to tackle the sensitivity of modelling and analysis of data from DSC experiments with reheating with respect to the choice of terminal temperatures for the first run. If the assumed model is correct for heating, it should also be valid for cooling and reheating subject to the selection of different end points for the first run. First, one should verify reversibility of each peak. The only reliable way to do this is by reheating from a point at the foot of the peak immediately after the end of the transition. If reheating produces the same peak as the one in the first run, model A and the classical analysis of reversible denaturation should be considered (Fig. 6).

If irreversible denaturation is observed, one more point is required, as in model B, C, and D. In these cases, reheating from the end point of the peak usually results in no peak during reheating. Since in the irreversible case, we are interested in measuring the speed of the peak reduction, at least one more point for reheating should be added. We studied the dependence of reheating runs on the final temperature of the first run of DhaA wild type denaturation. Reheating showed that the protein unfolds in a partially reversible manner. Fitting revealed that the thermal unfolding was in relatively good agreement with a two intermediate model, i.e. model C plus one negative peak at high temperatures (Fig. 7). Since the protein exhibited a rather complex unfolding pathway, we limited our analysis to the first peak of the DSC thermogram.

For a better understanding of the sensitivity of the reheated run to the experimental setup, we investigated different terminal points for the first run. As can be seen from the graph in Fig. 8, the reheating data was far more sensitive to the final temperature immediately after the peak temperature (points III – V) than before; the lack of a significant portion of irreversibly denatured protein and almost complete refolding during cooling for the temperature range before the peak (points I – II) drastically reduced the new information obtainable from reheating. Thus, only the final temperatures on the downward slope of the DSC curve were used for further analysis.

To allow discrimination between models B and C, an additional point for the reheating run at the summit of the peak (point III) seems to suffice. Indeed, as demonstrated earlier, model A exhibits almost no change in the height of the peak during reheating (Fig. 6). In contrast, model B results in a dramatic reduction of the native state after reheating, as can be seen in the examples of other haloalkane dehalogenases DbjA (Fig. 9) and LinB (Fig. 10). The DSC thermograms of these proteins were almost perfectly fitted by model B, although we also captured a

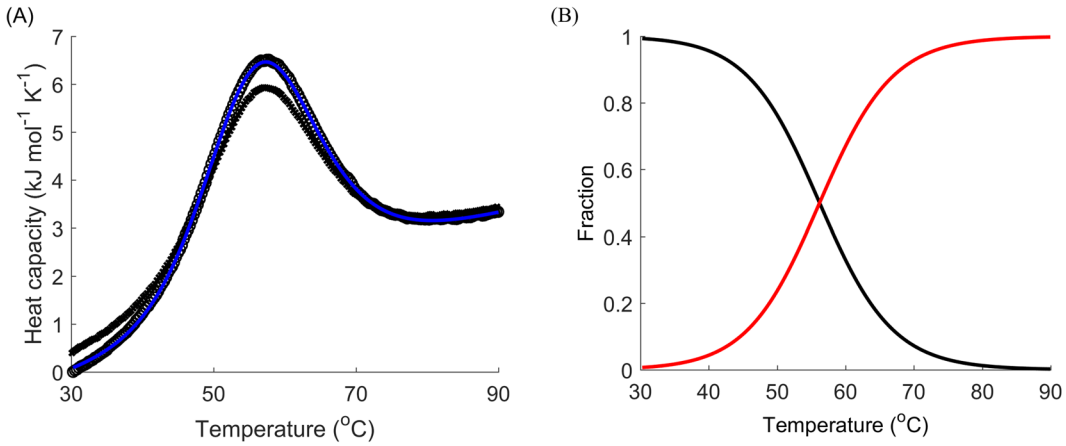


Figure 6. Reversible unfolding of the mutant SpA. (**A**) DSC data (black) for the denaturation of SpA (L20A + Y15W): first runs (circles), reheated runs (crosses), fitted curves for the first run (blue) for model A (with van't Hoff enthalpy). (**B**) Respective modelled fractions of states for a given temperature: native folded (black) and denatured (red) states. The scan rate was 4 °C min^{-1}.

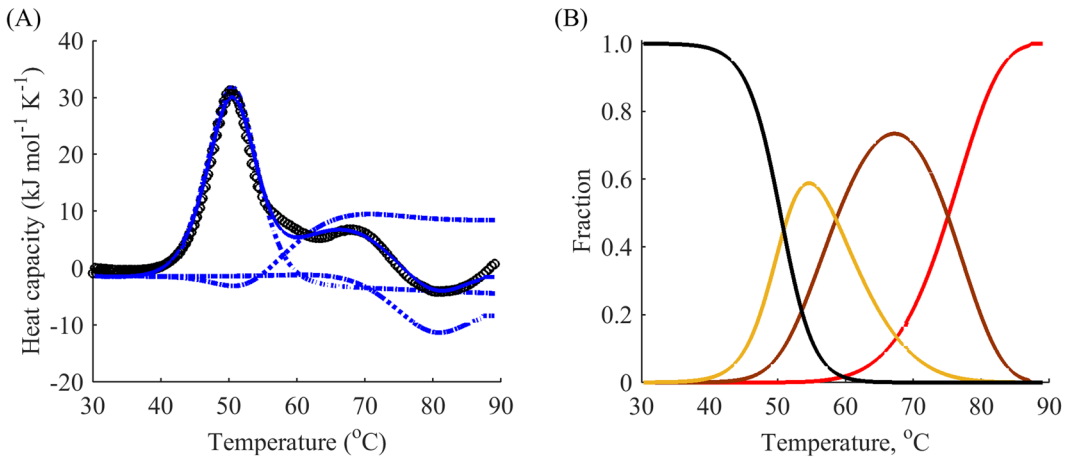


Figure 7. The complex three-step modelling of thermal unfolding of wild type DhaA. (**A**) DSC data (black) for the denaturation of DhaA wild type: first runs (circles), reheated runs for terminal temperatures 49, 51, 54, 61, and 69 °C (not shown), fitted curves for the first run (blue) with decomposition by peaks (dotted) from model C plus one negative peak at high temperatures. (**B**) Respective modelled fractions of states for a given temperature: native folded (black), first intermediate (yellow), second intermediate (brown) and denatured (red) states. The scan rate was 1 °C min^{-1}.

second exothermic peak at temperatures around 90 °C for LinB, which may be indicative of aggregation[39]. The proportion of the native state after cooling from the peak temperature in these two cases was as little as 20%. When intermediate levels of the native state are observed during reheating (20–99%), model C can be applied. It exhibits a slight decrease of the peak in the reheated run, as shown in the example above (Fig. 8B), where about 80% of the protein was conserved in the native state after reheating from 50 °C.

However, it should be noted that the abovementioned two points needed for reheating fail to discriminate between models B or C and D. Therefore, an additional point is needed. Based on the native fraction analysis (Fig. 8), the most sensitive point appears to be halfway between the peak temperature and end temperature of the respective peak (Fig. 8, point IV). In this case, model D should be applied if the decay between point III and point IV is different from that predicted by model B or C. This is obvious from comparison of the results for the DhaA115 mutant in Fig. 2 with those for DbjA and LinB in Figs 9 and 10, respectively. Considering these two points also improves estimation of the effect of ΔC_p described above.

In summary, we suggest the following experimental procedure to discriminate between the four basic models: (i) obtain the whole thermogram for as high temperature as possible with reheating; (ii) determine the temperature at the foot of the peak after the respective transition for each peak (point V in Fig. 8); (iii) conduct one more experiment with cooling and reheating from this temperature to check reversibility (discriminates between model A and models B/C/D); (iv) if reversibility is only partial, determine the peak temperature (point III in Fig. 8) and

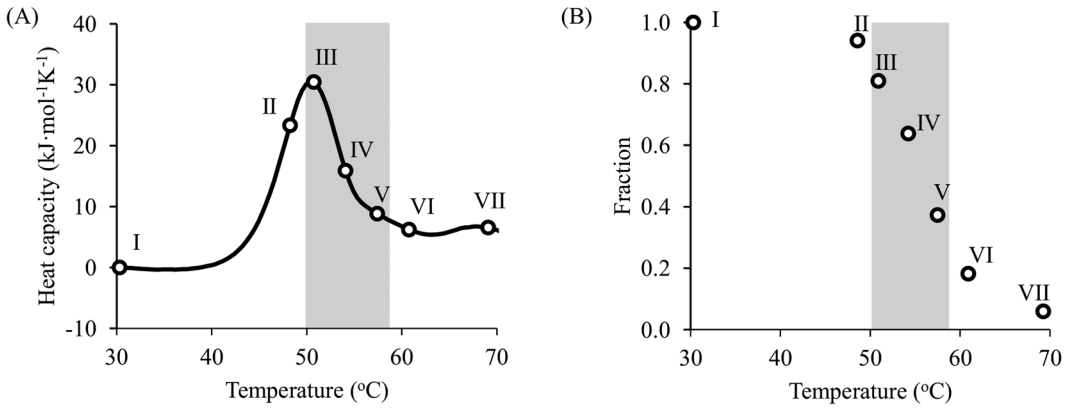


Figure 8. Fraction of the native state during reheating as a function of end temperature of the first run. (**A**) DSC data for the denaturation of DhaA wild type with terminal points for reheated runs (circles); (**B**) heights of reheating peaks as a fraction of those in the first run for different terminal temperatures. The steepest slope was observed after the summit of the peak, indicating high sensitivity of the reheating data to these temperatures (grey area).

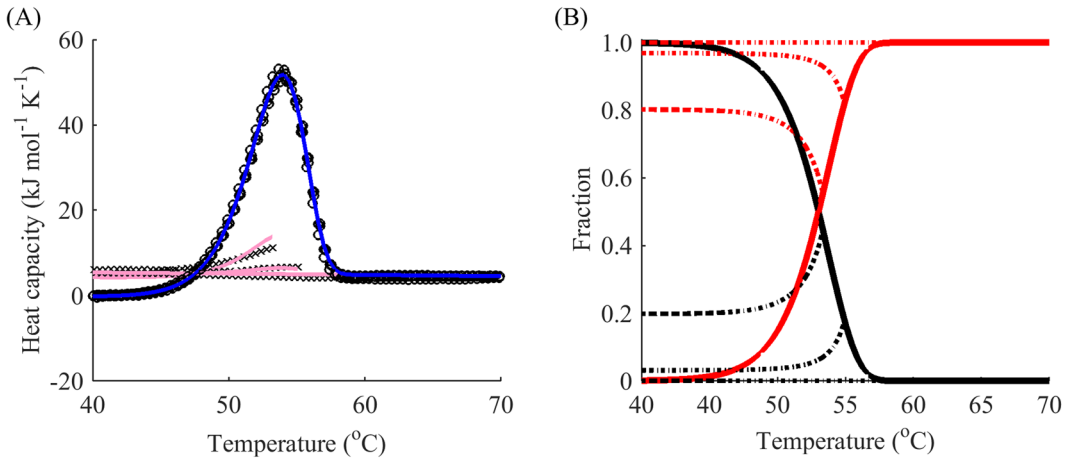


Figure 9. One-step irreversible unfolding of DbjA. (**A**) DSC data (black) for the denaturation of DbjA: first runs (circles), reheated runs for terminal temperatures 53, 55, and 59 °C (crosses), fitted curves for the first run (blue) and reheated runs (pink) from model B. (**B**) Respective modelled fractions of states for a given temperature: native folded (black) and denatured (red) states, cooling for both states is showed by dotted lines. The scan rate was 1 °C min^{-1}. Cooling from the peak temperature resulted in 20% of the protein in the native state.

the temperature half-way between the summit of the peak and the base after the transition (point IV in Fig. 8) – perform cooling and reheating for these temperatures (discriminates between models B, C, and D).

Case study: lysozyme. Models B and D have been presented with regard to the data in Figs 1, 9 and 10. This subsection presents the results of modelling and fitting DSC thermograms from reheated runs for the widely-researched chicken egg lysozyme, for which model C was chosen. Although early studies suggested that its unfolding is fully reversible[40], later research showed that there might be aggregation taking place in the immediate vicinity of the melting temperature[41]. Our calorimetry experiment revealed a single peak with a substantial degree of irreversibility (Fig. 11). Model C provided a reasonably good fit to the data with the addition of one more negative peak at temperatures higher than the first peak. As can be seen from the graph, the actual data indicated a slightly lower fraction of the native state in the first round of reheating and a higher fraction in the second and third rounds of reheating than those predicted by the model. This may serve as additional evidence of the aggregation involving both the native and partially unfolded types described in earlier studies[41], which would result in a greater amount of the native state for high temperatures than the amount predicted by model C.

However, almost the whole apparent peak could be attributed to the reversible transition by the fitting. Hence, the second step was mainly characterized by a rate constant of 0.27–0.32 s^{-1} for the temperature range 50–90 °C without well-defined decomposition into the energy of activation, transition temperature and enthalpy, as highlighted by the significant errors of estimation. Table 2 summarizes all the data obtained from fitting the cases presented in this paper.

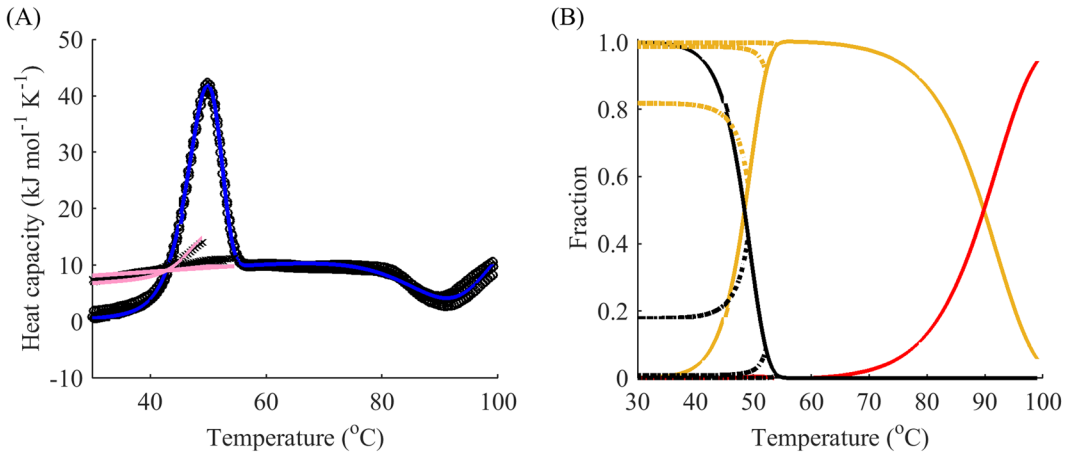


Figure 10. One-step irreversible unfolding of LinB with one additional exothermal peak at high temperatures. (**A**) DSC data (black) for the denaturation of LinB: first runs (circles), reheated runs for terminal temperatures 49, 52 and 54 °C (crosses), fitted curves for the first run (blue) and reheated runs (pink) from model B plus one negative peak at high temperatures. (**B**) Respective modelled fractions of states for a given temperature: native folded (black), intermediate (yellow) and denatured (red) states, cooling for all the states is showed by dotted lines. The scan rate was 1 °C min^{-1}. Cooling from the peak temperature resulted in less than 20% of the protein in the native state.

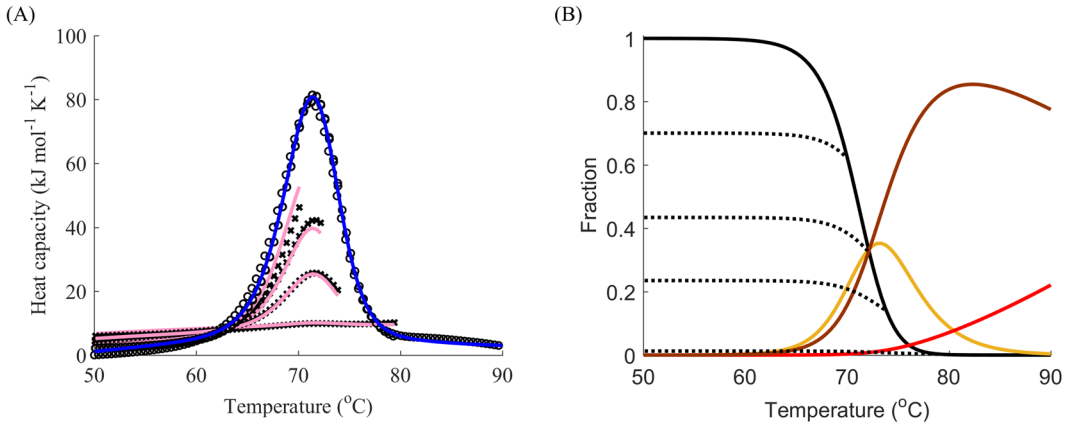


Figure 11. Three-step partially reversible denaturation of lysozyme. (**A**) DSC data (black) for the denaturation of lysozyme: first runs (circles), reheated runs for terminal temperatures 70, 72, 74 and 80 °C (crosses), fitted curves for the first run (blue) and reheated runs (pink). (**B**) Native folded (black), first intermediate (yellow), second intermediate (brown) and denatured (red) states; cooling for the native state is showed by dotted lines. The scan rate was 1 °C min^{-1}.

The values for the reversible component of denaturation of lysozyme ($T_m = 345$ K, $\Delta H = 516$ kJ mol^{-1}, $\Delta C_p = 7.5$ kJ mol^{-1}K^{-1}) are in good agreement with previously published results[41–43] under similar conditions ($T_m = 346$–349 K, $\Delta H = 485$–543 kJ mol^{-1}, $\Delta C_p = 6.2$–10.5 kJ mol^{-1}K^{-1}). The parameters of the irreversible step indicate that the transition rate was 3.5–$4.2 \cdot 10^{-3}$ s^{-1} for the transition temperature range 70–80 °C and its dependence on temperature was not well constrained by the data, as apparent from the low value of E and high T_f. For Dbja and LinB, simple one-step irreversible transitions were observed. DhaA variants exhibited different behaviour: the datasets of the wild type were fitted with two intermediates, whereas data for the stabilized mutant were perfectly fitted to a one intermediate model. This implies that the first two steps of the wild type unfolding were shifted to higher temperatures, "fusing" to produce one apparent peak in the thermograms. This hypothesis was further supported by the fact that only the general three-step model (D) was able to explain the data of the mutant. Hence, the first step for the apparent peak could not be described by a simple equilibrium. Based on this DSC analysis, we concluded that DhaA unfolds according to a rather complex model. Thus, it might be expedient to augment the data analysis with other thermodynamic techniques prior to drawing conclusions about the unfolding model.

Conclusions

In most cases discussed in our manuscript, the main motivation for understanding the unfolding mechanism of proteins is their stability. This stability is determined by the temperature at which the native state of the protein is lost, and which can be different from the apparent peak temperature in the presence of intermediates. Moreover, in order to better engineer stable mutant proteins, the contributions of thermodynamic stability (defined mainly by ΔH, or the Gibbs free energy difference ΔG) and kinetic stability (defined mainly by energy of activation E_a, or the Gibbs free energy barrier $\Delta G^{\ddagger}$) must be defined. The values obtained from the model selection and curve fitting might help quantify those contributions. In addition to that, revealing intermediates on the unfolding pathways might be important for protein function, e.g. to penetrate through cell membranes, or for the determination of protein propensity to aggregate, which is important for industrial production of soluble proteins as well as for neurodegenerative diseases. The method suggested in this paper is to include second runs of DSC into the data modelling.

Reheating provides additional insights for selection of the model of unfolding in complement to experiments with different scan rates. In this paper, we proposed using the whole curves of reheated runs during fitting procedures. Since first runs can easily be superimposed, reheating allows a feasible global fit with no need for additional parameters, in contrast to applying different scan rates. Moreover, the selected models are not affected by a changing k/v ratio, which might be the case when varying scan rates. Hence, fitting of reheated runs should provide more reliable estimations of parameters than obtainable by using different scan rates.

Collection of data for reheated runs should not cause many difficulties as cooling and reheating procedures are frequently included in the software distributed along with DSC devices. We have demonstrated that as few as three final temperature points are needed to produce enough data to discriminate between the four most common models of unfolding. It is worth noting that this approach is by no means limited to the models analyzed in this article as similar modelling of reheated runs can be derived for any mechanism of unfolding provided that complex unfolding is analyzed carefully. Reheating is capable of revealing more information about the ΔC_p effect because stopping the first run at different end points results in a different population of states at the onset of reheating, thus shifting initial points due to the accumulated ΔC_p. Finally, modelling of reheating can be used to quantify the resemblance between the first and second runs in terms of standard deviations of the residuals to rule out the possibility of alternative refolding. To help analyze data and perform a global fit to DSC data with reheating and different scan rates, we provide a computer code and link to the software tool used in the Supplement.

Data Availability. CalFitter software, source code, and binaries were implemented in MATLAB and are freely available for download at http://loschmidt.chemi.muni.cz/peg/software/calfitter.

References

1. Koga, N. *et al*. Principles for designing ideal protein structures. *Nature* **491**(7423), 222–227 (2012).
2. Baker, D. Protein folding, structure prediction and design. *Biochemical Society Transactions* **42**(2), 225–229 (2014).
3. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16**(1), 18–29 (2015).
4. Wang, S. & Kaufman, R. J. The impact of the unfolded protein response on human disease. *The Journal of Cell Biology* **197**(7), 857–867 (2012).
5. Dill, K. A. & MacCallum, J. L. The protein-folding problem. *50 years on. Science* **338**(6110), 1042–1046 (2012).
6. Englander, S. W., Mayne, L., Kan, Z.-Y. & Hu, W. Protein folding—how and why: By hydrogen exchange, fragment separation, and mass spectrometry. *Annual Review of Biophysics* **45**, 135–152 (2016).
7. Zhuravleva, A. & Korzhnev, D. M. Protein folding by NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy* **100**, 52–77 (2017).
8. Johnson, C. M. Differential scanning calorimetry as a tool for protein folding and stability. *Archives of Biochemistry and Biophysics* **531**, 100–109 (2013).
9. Privalov, P. L., Microcalorimetry of Macromolecules: The Physical Basis of Biological Structures (Wiley, Baltimore, 2012).
10. Strucksberg, K. H., Rosenkranz, T. & Fitter, J. Reversible and irreversible unfolding of multi-domain proteins. *Biochimica et Biophysica Acta Proteins and Proteomics* **1774**(12), 1591–1603 (2007).
11. Ibarra-Molero, B., Naganathan, A. N. & Sanchez-Ruiz, J. M. Modern analysis of protein folding by differential scanning calorimetry. *Methods in Enzymology* **567**, 277–314 (2016).
12. Rao, V., Hemanth, G. & Shachi, G. Using the folding landscapes of proteins to understand protein function. *Current Opinion in Structural Biology* **36**, 67–74 (2016).
13. Sanchez-Ruiz, J. M. Protein kinetic stability. *Biophysical Chemistry* **148**(1), 1–15 (2010).
14. Becktel, W. J. & Schellman, J. A. Protein stability curves. *Biopolymers* **26**(11), 1859–1877 (1987).
15. Privalov, P. L. Cold denaturation of protein. *Critical Reviews in Biochemistry and Molecular Biology* **25**(4), 281–306 (1990).
16. Toledo-Núñez, C., Vera-Robles, L. I., Arroyo-Maya, I. J. & Hernández-Arana, A. Deconvolution of complex differential scanning calorimetry profiles for protein transitions under kinetic control. *Analytical Biochemistry* **509**, 104–110 (2016).
17. Freire, E., TheThermodynamic Linkage Between Protein Structure, Stability, and Function. Protein Structure, Stability, and Folding, 37–68 (2001).
18. Makhatadze, G. I. & Privalov, P. L. Energetics of protein structure. *Advances in Protein Chemistry* **5**, 507–510 (1995).
19. Roberts, C. J. Non-native protein aggregation kinetics. *Biotechnology and Bioengineering* **98**(5), 927–938 (2007).
20. Privalov, P. L. & Dragan, A. I. Microcalorimetry of biological macromolecules. *Biophysical Chemistry* **126**, 16–24 (2007).
21. Lyubarev, A. E. & Kurganov, B. I. Modeling of Irreversible Thermal Protein Denaturation at Varying Temperature. *II. The Complete Kinetic Model of Lumry and Eyring. Biochemistry Moscow* **64**, 832–838 (1999).
22. Arroyo-Reyna, A., Tello-Solis, S. R. & Rojo-Dominguez, A. Stability parameters for one-step mechanism of irreversible protein denaturation: a method based on nonlinear regression of calorimetric peaks with nonzero delta Cp. *Analytical Biochemistry* **328**, 123–13 (2003).
23. Lepock, J. R. *et al*. Influence of transition rates and scan rate on kinetic simulations of differential scanning calorimetry profiles of reversible and irreversible protein denaturation. *Biochemistry* **31**, 12706–12712 (1992).
24. Lyubarev, A. E. & Kurganov, B. I. Analysis of DSC data relating to proteins undergoing irreversible thermal denaturation. *Journal of Thermal Analysis and Calorimetry* **62**, 49–60 (2000).
25. Sedlák, E., Schaefer, J. V., Marek, J., Gimeson, P. & Plückthun, A. Advanced analyses of kinetic stabilities of iggs modified by mutations and glycosylation. *Protein Science* **24**(7), 1100–1113 (2015).

26. Vermeer, A. W., Bremer, M. G. & Norde, W. Structural changes of IgG induced by heat treatment and by adsorption onto a hydrophobic Teflon surface studied by circular dichroism spectroscopy. *Biochimica et Biophysica Acta - General Subjects* **1425**(1), 1–12 (1998).
27. Zhadan, G. G. *et al*. Protein involvement in thermally induced structural transitions of pig erythrocyte ghosts. *International Journal of Biochemistry and Molecular Biology* **42**, 11–20 (1997).
28. Singh, N., Liu, Z. & Fisher, H. F. The existence of a hexameric intermediate with molten-globule-like properties in the thermal denaturation of bovine-liver glutamate dehydrogenase. *Biophysical Chemistry* **63**, 27–36 (1996).
29. Markov, D. I., Zubov, E. O., Nikolaeva, O. P., Kurganov, B. I. & Levitsky, D. I. Thermal denaturation and aggregation of myosin subfragment 1 isoforms with different essential light chains. *International Journal of Molecular Sciences* **11**(4194-4226), 4194–4226 (2010).
30. Andrews, J. M. & Roberts, C. J., A Lumry-Eyring nucleated polymerization model of protein aggregation kinetics: 1. Aggregation with pre-equilibrated unfolding. *The Journal of Physical Chemistry B*, 7897–7913 (2007).
31. Privalov, P. L. & Khechinashvili, N. N. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *Journal of Molecular Biology* **86**(3), 665–684 (1974).
32. Milardi, D., La Rosa, C. & Grasso, D. Extended theoretical analysis of irreversible protein thermal unfolding. *Biophysical Chemistry* **52**, 183–189 (1994).
33. Pavlova, M. *et al*. Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nature Chemical Biology* **5**, 727–733 (2009).
34. Stepankova, V., Damborsky, J. & Chaloupkova, R. Organic co-solvents affect activity, stability and enantioselectivity of haloalkane dehalogenases. *Biotechnology Journal* **8**, 719–729 (2013).
35. Sato, S., Religa, T. L. & Fersht, A. R. Φ-Analysis of the folding of the B domain of protein a using multiple optical probes. *Journal of Molecular Biology* **360**, 850–864 (2006).
36. Sanchez-Ruiz, J. M. Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophysical Journal* **61**(4), 921–935 (1992).
37. Privalov, P. L. & Makhatadze, G. I. Contribution of hydration and non-covalent interactions to the heat capacity effect on protein unfolding. *Journal of Molecular Biology* **224**(3), 715–723 (1992).
38. Rodriguez-Larrea, D., Ibarra-Molero, B., de Maria, L., Borchert, T. V. & Sanchez-Ruiz, J. M., Beyond Lumry–Eyring: An unexpected pattern of operational reversibility/irreversibility in protein denaturation. *Proteins: Structure, Function, and Bioinformatics*, 19–24 (2008).
39. Goyal, M., Chaudhuri, T. P. & Kuwajima, K. Irreversible Denaturation of Maltodextrin Glucosidase Studied by Differential Scanning Calorimetry, Circular Dichroism, and Turbidity Measurements. *PLoS ONE* **9**, e115877 (2014).
40. Branchu, S., Forbes, R. T., York, P. & Nyqvist, H. A central composite design to investigate the thermal stabilization of lysozyme. *Pharmaceutical Research* **16**, 702–708 (1999).
41. Sassi, P., Giugliarelli, A., Paolantoni, M., Morresi, A. & Onori, G. Unfolding and aggregation of lysozyme: A thermodynamic and kinetic study by FTIR spectroscopy. *Biophysical Chemistry* **158**, 46–53 (2011).
42. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested and thermostability linked to hydrocarbon packing. *Nature* **345**(6270), 86–89 (1990).
43. Shih, P., Kirsch, J. F. & Holland, D. R. Thermal stability determinants of chicken egg-white lysozyme core mutants: Hydrophobicity, packing volume, and conserved buried water molecules. *Protein Science* **4**(10), 2050–2062 (1995).

Acknowledgements

The work was supported by the Grant Agency of the Czech Republic (GA16–07965S) and the Czech Ministry of Education of the Czech Republic (LO1214, LQ1605, LM2015051 and LM2015055). This work was also supported by the Czech Ministry of Education, Youth and Sports, Programme CETOCOEN UPgrade (CZ.1.05/2.1.00/19.0382). K.B. was supported by the "Employment of Best Young Scientists for International Cooperation Empowerment" (CZ.1.07/2.3.00/30.0037) project co-financed by the European Social Fund and the state budget of the Czech Republic. A.K. is a Brno Ph.D. Talent Scholarship Holder and funded by the Brno City Municipality.

Author Contributions

A.K. and K.B. performed sample preparation and DSC experiments; S.M. performed data analysis and coded the software tool; J.D., Z.P., C.J., S.M., K.B., and A.K. designed research, interpreted data, and contributed to the writing of the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16360-y.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

W344–W349 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 14 May 2018*
doi: 10.1093/nar/gky358

CalFitter: a web server for analysis of protein thermal denaturation data

Stanislav Mazurenko[1,†], Jan Stourac[1,2,†], Antonin Kunka[1,2,†], Sava Nedeljković[1,3], David Bednar[1,2], Zbynek Prokop[1,2,*] and Jiri Damborsky[1,2,*]

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

Received February 07, 2018; Revised April 11, 2018; Editorial Decision April 23, 2018; Accepted April 24, 2018

ABSTRACT

Despite significant advances in the understanding of protein structure-function relationships, revealing protein folding pathways still poses a challenge due to a limited number of relevant experimental tools. Widely-used experimental techniques, such as calorimetry or spectroscopy, critically depend on a proper data analysis. Currently, there are only separate data analysis tools available for each type of experiment with a limited model selection. To address this problem, we have developed the CalFitter web server to be a unified platform for comprehensive data fitting and analysis of protein thermal denaturation data. The server allows simultaneous global data fitting using any combination of input data types and offers 12 protein unfolding pathway models for selection, including irreversible transitions often missing from other tools. The data fitting produces optimal parameter values, their confidence intervals, and statistical information to define unfolding pathways. The server provides an interactive and easy-to-use interface that allows users to directly analyse input datasets and simulate modelled output based on the model parameters. CalFitter web server is available free at https://loschmidt.chemi.muni.cz/calfitter/.

INTRODUCTION

Proteins are the main building blocks of living organisms and are widely used in numerous biomedical and biotechnological applications. Since they are made up of only 20 amino acids, the enormous variety of their functions mainly stems from their unique structures. The interest in protein spatial organization is usually twofold: the exact position of active sites and connected residues can shed light on protein function such as enzymatic activity, intracellular transport or molecular signalling (1,2), and exact knowledge of structural elements provides methods of locating the possible sources of protein (in)stability and designing more stable variants using protein engineering (3,4). Moreover, protein misfolding and aggregation have also been reported as primary causes of several neurodegenerative diseases (5).

Streamlined protein denaturation experimental techniques to study protein (un)folding, misfolding, and aggregation include differential scanning calorimetry (DSC), fluorescence/absorbance spectroscopy, light scattering, and circular dichroism (CD) (6–10). They allow the recording of corresponding signals when a protein undergoes denaturation, e.g. due to an increase in temperature. Since those techniques produce an aggregated output from a highly complex underlying process of unfolding, they necessitate the development of software for data modelling and analysis (11,12). Such analysis usually involves the selection of an appropriate unfolding model that best fits the observed data and allows quantification of unfolding pathways in terms of the number of intermediate states, Gibbs free energy barriers separating those states, and corresponding melting temperatures (7,13,14). The importance of such information can hardly be overestimated: intermediates are often the culprits of aggregation, and energy barriers directly define protein half-lives. Hence, both provide attractive targets for protein engineering (15). Moreover, as far as molecular dynamic simulations are concerned, the number of unfolding intermediates can be used as input to cluster analysis in Markov state models, while experimental half-lives can provide guidance for the necessary length of simulations and their conditions (16,17).

Apart from general purpose but programming-intensive tools for data analysis, such as Matlab, Origin or Igor Pro,

*To whom correspondence should be addressed. Tel: +420 549 4930 41; Fax: +420 549 4962 03; Email: jiri@chemi.muni.cz, Website address: https://loschmidt.chemi.muni.cz/calfitter/ (type of web server: data analysis and visualization).
Correspondence may also be addressed to Zbynek Prokop. Email: zbynek@chemi.muni.cz
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

there are only a few software packages that handle different types of thermal denaturation experiments. These tools were designed to study kinetics or chemical equilibrium — KinTek Explorer, DynaFit — or are linked to a specific type of measurement — MicroCal DSC Origin and CDpal (18–20). Although the former are based on equations that can be adapted for unfolding, data and differential equations with a gradual temperature change are not supported, and the set of parameters offered requires additional manipulations for translation into those usually used to describe unfolding, e.g. Gibbs free energy differences and enthalpy changes. The latter tools offer a limited set of models for fitting, e.g. only reversible denaturation despite increasing cases of proteins unfolding irreversibly. Moreover, such tools are unable to fit global data from different sources, e.g. equilibrium and kinetic data, which may sometimes lead to simplified conclusions.

There are several advantages of global fitting over analyses of separate data sets. Fitting curves to multiple-step models is error-prone because the signal measured can be insensitive to some of the intermediates on the unfolding pathway. Moreover, consecutive steps sometimes overlap significantly and produce apparent single transition which cannot be resolved by fitting into just one data type, and different types of experimental signals must be analyzed simultaneously to overcome this problem (21). Modelling and data fitting of single experimental types individually may also fail to separate parameters enforcing reparametrization of a model for available combinations of constants, e.g. the equilibrium constant K_{eq} instead of the rate ratios k_{fwd}/k_{rev}. Simultaneous fitting into a combination of different data types eventually leads to a single set of the original parameter values without any need for reparametrization (22,23). Finally, many independent variables that have to be introduced in separate data fitting contributes to increased uncertainty and can be avoided in the global data fitting.

There is a need for a single, universal platform that handles thermal denaturation data analysis with multiple test models, a user-friendly interface, and the option to join different types of experimental data in one data fitting session. We have developed the web server CalFitter for an interactive data analysis of the commonly used thermal denaturation techniques, such as DSC, CD, and kinetic temperature-jump (*T*-jump) experiments. This first-of-a-kind web server offers flexible visualization of the data, quick data pre-treatment for removal of irrelevant and poor-quality data, data simulation, and fitting based on a wide range of fully reversible, irreversible, and partially-reversible unfolding models, as well as statistical data analysis of the goodness of fit. Its data fitting functionality was validated using denaturation data for six wild-type proteins from different structural families, and seven mutant proteins.

WORKFLOW

The basic workflow of CalFitter is outlined in Figure 1. There are three main phases in the process. First, the user uploads experimental datasets, plots them, and treats the data using built-in data pre-treatment options. Then the user selects a model for the data fitting, supplies initial parameter estimates, simulates the modelled dynamics, and starts the fitting process. The server performes numerical data fitting by minimizing the normalized sums of squared residuals. Once the data fitting is complete, the server returns optimal parameter values, confidence intervals calculated from asymptotic normal distribution, and statistical information about the goodness of fit. This information can be further used to estimate the outcome of the fitting and help identify the necessary adjustments to the selected model required for refitting. Different fitting sessions can be conducted in independent tabs after data uploading to compare the outcomes merely by switching the tabs. Once the fitting is complete, the results of the analysis can be exported to a single excel file for further use. Moreover, each session is given a unique ID and thus can be repeatedly accessed at the server.

Step 1: Data upload and pre-treatment

The input to the server consists of experimental datasets of three types: (i) temperature-dependent heat capacity data from DSC, (ii) temperature-dependent spectroscopic signals (ellipticity, fluorescence, or absorbance) from spectroscopic scanning measurements, and (iii) time-dependent spectroscopic signals from folding/unfolding T-jumps. In the first step, the user interactively uploads datasets obtained from experiments and specifies the corresponding data columns, units, and experimental setup parameters such as scan rates for DSC and spectroscopy measurements or temperatures for *T*-jumps. In order to eliminate systematic errors during global fitting, all datasets must be collected under the same experimental conditions, e.g. pH, ionic strength, and buffer composition. The concentration dependence of protein unfolding must be verified to avoid aggregation or other association/dissociation effects.

The data files can be uploaded in either CalFitter native format (24), plain comma-separated values format (CSV), or several formats exported directly from the build-in software that comes with instruments, e.g. Chirascan or BioKine. The user can select the units from the most widely used ones for temperature (°C or K), heat capacity difference (J, kJ, cal or kcal/mol K), and time (ns, μs, ms, s, min, h). There is no upper limit on the number of points in the datasets, although larger datasets take longer to calculate and can bias the fitting statistically. The web server provides a detailed Help page with guidelines about data formatting and uploading.

The user can then plot the uploaded datasets and select those that will be used for data analysis. Any combination of the dataset types can be used for global fitting. Moreover, the user can exclude some parts of the datasets such as temperature ranges with a poor experimental signal. Finally, visual data normalisation can be carried out at this step, which is of great importance in global fits, because collected experimental values usually have different units. In particular, DSC data for each scan rate can be superimposed using vertical shifts, spectroscopy data can be normalized by subtracting signal means and dividing by signal standard deviations for each dataset, and *T*-jump traces can be shifted vertically to the same starting point. This has only a visual effect since all the datasets are normalized automat-

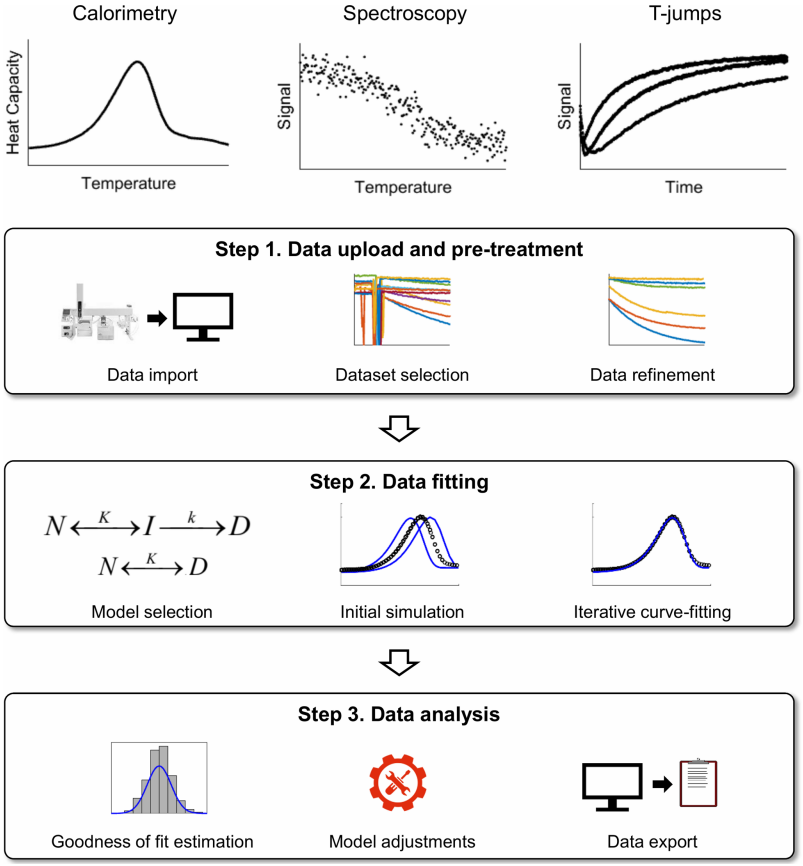


Figure 1. CalFitter workflow. The software provides an integrated analysis of data using three different types of experimental techniques: (I) calorimetry, (II) spectroscopy and (III) T-jumps. The procedure consists of three steps: (1) data upload and pre-treatment, (2) data fitting and (3) data analysis. A detailed description of the individual steps is provided in the text.

ically during the fitting as described in the section Global fitting of the Supplementary Data.

Step 2: Data fitting

Once the user is satisfied with the datasets selected and their quality, data fitting can be carried out. The procedure is similar to existing fitting software such as Origin or KinTek Explorer with a special focus on the determination of thermodynamic and kinetic parameters. First, a potential model is selected based on the desired number of steps on the unfolding pathway and their reversibility (Table 1). CalFitter currently offers models that are based on discreet macrostates on the unfolding pathways, i.e. native, intermediate, denatured, etc. Analysis based on the statistical free energy surface models of microstates (25) is beyond the scope of the web server. Second, initial parameter estimates are specified (Figure 1). The server produces initial values, however, the user needs to check and modify those values as there are currently no algorithms providing reliable initial parameter estimation for thermal denaturation models. The web server provides the option to simulate the output datasets based on the input initial parameters to assist the user at this stage. This allows the display of modelled and input data together on one graph. Finally, the user specifies whether some parameters should remain fixed during the fitting with an adjustable number of iterations. More details on mathematical modelling and data fitting can be found in the Supplementary Data and the relevant literature (11,12,24,26–28).

Step 3: Analysis of the results

Once the fitting is complete, the web server updates the parameter values, their confidence intervals and provides statistical information from the fitting such as Akaike (AIC) and Bayesian (BIC) information criteria and residual plots. At this point, the user can either accept the model or change it and carry out refitting. The most common strategy is to start from the simple model with a few steps and then add additional steps while checking the goodness of fit visually or using the AIC and BIC values (29). The common sign of over-parameterization is drifting parameter values and large confidence intervals. In this case, either the model should be simplified by removing some steps, or some fitting parameters should be held constant. Finally, the user can also check the sensitivity of the output to input parameters or undo the last fit before exporting the results in standard formats.

The output consists of the unfolding pathway, the updated parameter values that best describe the data included, their confidence intervals, and other statistical information from the fitting such as goodness of fit. There are four data

Table 1. The description of the models and the corresponding parameters implemented in CalFitter

Model	Description	Model parameters[a]	Data sets
1 step			
N –> D	A fully irreversible transition	T_f, E_a, $\Delta H^‡$, ΔC_p	All
N = D	A fully reversible transition with equilibrium	T_m, ΔH, ΔC_p	Calorimetry & spectroscopy
N = D (Van't Hoff's)	A fully reversible transition with equilibrium and van't Hoff's enthalpy	T_m, ΔH, ΔH_{vh}, ΔC_p	Calorimetry & spectroscopy
N <=> D	A general transition with forward and reverse components	T_{fwd}, E_{fwd}, T_{rev}, E_{rev}, ΔC_p	All
2 steps			
N –> I –> D	A fully irreversible transition	Step 1: T_f, E_a, $\Delta H^‡$, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p	All
N = I –> D	A transition with a reversible step in equilibrium and an irreversible step	Step 1: T_m, ΔH, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p	Calorimetry & spectroscopy
N <=> I –> D	A general Lumry-Eyring model	Step 1: T_{fwd}, E_{fwd}, T_{rev}, E_{rev}, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p	All
N = I = D	A fully reversible transition	Step 1: T_m, ΔH, ΔC_p Step 2: T_m, ΔH, ΔC_p	Calorimetry & spectroscopy
3 steps			
N –> I_1 –> I_2 –> D	A fully irreversible transition	Step 1: T_f, E_a, $\Delta H^‡$, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p Step 3: T_f, E_a, $\Delta H^‡$, ΔC_p	All
N = I_1 –> I_2 –> D	A transition with the reversible first step in equilibrium and the irreversible second and third steps	Step 1: T_m, ΔH, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p Step 3: T_f, E_a, $\Delta H^‡$, ΔC_p	Calorimetry & spectroscopy
N <=> I_1 –> I_2 –> D	A general Lumry-Eyring model with two intermediates	Step 1: T_{fwd}, E_{fwd}, T_{rev}, E_{rev}, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p Step 3: T_f, E_a, $\Delta H^‡$, ΔC_p	All
N = I_1 = I_2 –> D	A transition with two reversible steps in equilibrium and an irreversible step	Step 1: T_m, ΔH, ΔC_p Step 2: T_m, ΔH, ΔC_p Step 3: T_f, E_a, $\Delta H^‡$, ΔC_p	Calorimetry & spectroscopy
4 steps			
N –> I_1 –> D N –> I_2 –> D	A two-branch irreversible unfolding pathway	Step 1: T_f, E_a, $\Delta H^‡$, ΔC_p Step 2: T_f, E_a, $\Delta H^‡$, ΔC_p Step 3: T_f, E_a, $\Delta H^‡$, ΔC_p Step 4: T_f, E_a, $\Delta H^‡$, ΔC_p	All

[a] T_m – the melting temperature, T_f – the reference temperature of an irreversible step at which the corresponding rate is 1 (fwd. – forward rates; rev. – reverse rates), ΔH – the enthalpy change (at T_m if ΔC_p is nonzero; vh – van't Hoff's); $\Delta H^‡$– the activation enthalpy change (at T_f or T_m for irreversible and general steps, respectively, if ΔC_p is nonzero); E_a – the activation energy; ΔC_p – the heat capacity change. Since *T*-jumps are based on the relaxation kinetics, they cannot be simulated by the models with reversible steps assumed in equilibrium.

types available for visualising the output (Figure 2). The first type is raw experimental and pre-treated data, each dataset being separate or combined with other datasets of the same type. The second type uses modelled signals matching the selected datasets given the current parameter values. A state decomposition may be carried out with this type whereby signal contributions from each step of the unfolding pathway to the overall modelled signal are shown. The third type uses the modelled protein state fractions as functions of temperature or time. Finally, the fourth type uses residuals from the fit, which might shed further light on the quality of the fit and main discrepancies between the modelled and experimental data. The user can then export the graphs and the corresponding datasets as an archive with figures and settings files or as an Excel file.

EXPERIMENTAL VALIDATION

The CalFitter performance was thoroughly validated with all the three data types. DSC data analysis was tested on previously published datasets: the thermograms of wild-type hen egg lysozyme, wild-type haloalkane dehalogenases LinB, DbjA, DhaA, fibroblast growth factor FGF2 and variants of DhaA and FGF2 engineered for higher thermostability (24,30). Calorimetry curves were curve-fitted and compared with the output from MicroCal DSC Origin and the standalone Matlab-based CalFitter 1 (24). Spectroscopy data analysis was tested on four variants of DhaA and compared with the output from CD-pal. Finally, *T*-jump data analysis was validated using four different global datasets consisting of several traces of DhaA wild-type and compared against the values obtained using KinTek Explorer. In all the cases, the output values produced by the CalFitter web differed from the previously published by <0.1% on average for temperature related parameters, and by <6.8% on average for energy and heat-related parameters (Table 2). Maximal discrepancies for the latter were mainly due to smaller parameter values and/or wide confidence intervals. Moreover, in those cases, the CalFitter web simulation with the parameters from the other tools produced visually the same quality of the fit suggesting that the differences stem from the numerical procedures used for fitting rather than from a different model behaviour.

The global fitting provided by the web server was also used to analyse experimental data for stable variants of FGF2 protein designed recently using computer-assisted protein engineering (30). This analysis revealed new biophysical insights, namely the presence of an unfolding inter-

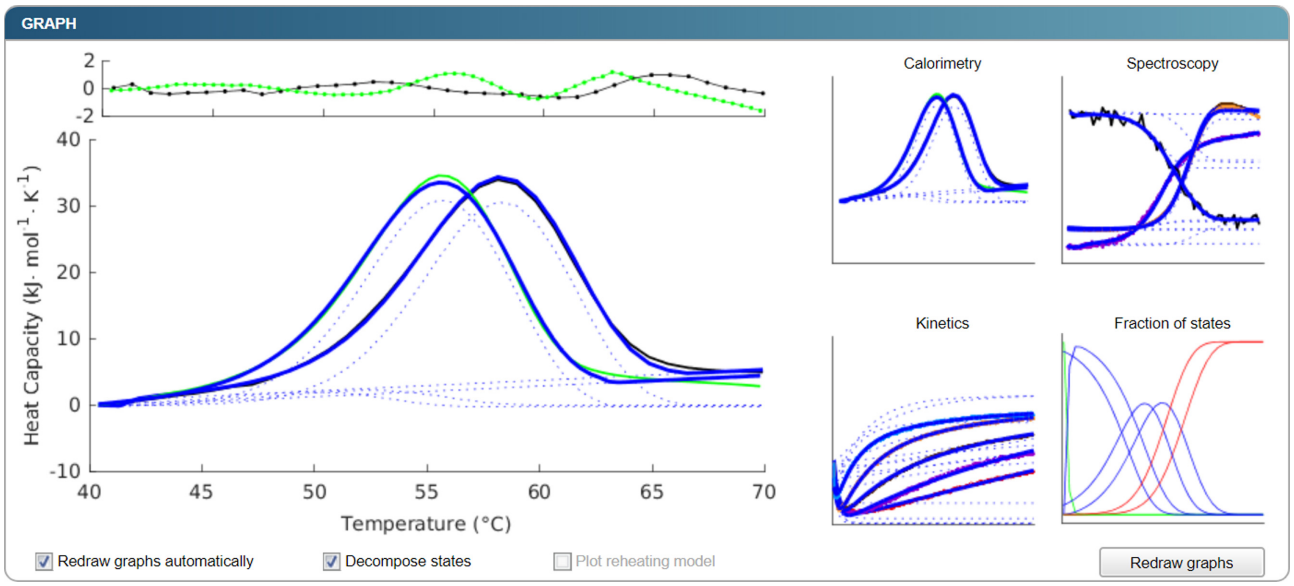


Figure 2. An example of the graphical output. Initial datasets with modelled signal in blue are depicted as icons on the right-hand side and can be zoomed-in and displayed on the left-hand side. The zoomed-in version also depicts the residuals at the top so that a user can estimate the quality of the fit and the presence of any systematic errors or unexplained data variation. In the presented case, waves are apparent in the residual plot that are indicative of an approximately 5% misfit at high temperatures. When the option 'Decompose states' is selected, the contribution of each step to the overall signal is plotted in dotted lines. Apart from the plots of data and corresponding modelled signals, modelled fractions of states are presented in one of the graphs.

Table 2. Experimental validation of CalFitter web server. Discrepancies are given in terms of absolute % difference for parameters obtained for energies and temperatures

			Temperature variables (T_m, T_f)		Energy variables (E_a, ΔH, ΔH_{vh})	
Data type	Software used for comparison	Number of datasets	Average discrepancy	Maximal discrepancy	Average discrepancy	Maximal discrepancy
DSC[a]	MicroCal DSC Origin	67	0.06%	0.42%	3.03%	15.06%
	Matlab-based CalFitter 1		0.00%	0.01%	0.00%	0.07%
CD[b]	CD-pal	44	0.01%	0.10%	0.05%	1.18%
T-jumps[c]	KinTek Explorer	35	0.09%	0.21%	6.74%	10.87%

[a]based on ΔH, ΔH_{vh}, and T_m from a non-two state model with ΔH_{vh}.
[b]based on T_m and ΔH for a one-step fully reversible model.
[c]data from global fitting based on E_a and T_a for a two-step fully irreversible model.

mediate, and demonstrated a good agreement between the *in silico* predicted Gibbs free energy differences and the differences in the transition barriers for the first unfolding step estimated from experiments. Another case study of thermal denaturation of haloalkane dehalogenase DhaA112 engineered for stability is described in the Supplementary Data. This new case reveals an unfolding intermediate and provides quantitative estimates of unfolding rates.

CONCLUSIONS AND OUTLOOK

CalFitter is a web server that offers users a one-stop-shop for data analysis from commonly used temperature denaturation experiments. Not only does it offer a wider range of models for each separate data type when compared to most of the existing analogues, but it also enables the combination of different dataset types, such as equilibrium and T-jump data, in a single global data analysis. This feature has never been implemented for thermal unfolding studies before, to the best of the authors' knowledge. The fitting procedures used were optimised and validated using several dozen datasets from different sources, including recently published data as well as cross-validation using the existing software for each data type analysis.

The server is complemented by an easy-to-use graphical interface that allows users to interactively pre-treat the data by excluding irrelevant parts or artefacts, selecting the desired subset for analysis and fitting, and simulating the behaviour of the models when parameters change. The hidden mathematical calculations and fitting makes the process of data analysis accessible to users without any prior expertise of mathematical modelling. The web server graphical output consists of four different plot types to provide the user with a full image of the modelled pathway and its correspondence to the supplied data.

In the future, we will implement an 'advanced mode' with a model editor that the users can manually input any model of their choice using a simple text entry with an in-

tuitive syntax similar to KinTek Explorer or DynaFit. We are also working on a module for rapid initial parameter estimations. This module will estimate starting values based on the shape of the curves, rendering the web server even more user-friendly for researchers with limited experience in modelling data from thermal unfolding studies. Finally, we plan to add additional modelling capabilities to the existing modules, e.g., double T-jumps and singular value decomposition of CD spectra, as well as to develop modules for analysis of other types of experimental data, e.g., protein chemical denaturation and hydrogen-deuterium exchange mass spectrometry.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Ministry of Education of the Czech Republic [LO1214, LQ1605, LM2015051, LM2015047 and LM2015055]; Czech Grant Agency [GA16-07965S]; European Union [720776 and 722610]; A.K. is a Brno Ph.D. Talent Scholarship Holder and funded by the Brno City Municipality. Funding for open access charge: Ministry of Education of the Czech Republic.
Conflict of interest statement. None declared.

REFERENCES

1. Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.
2. Orengo,C.A., Pearl,F.M.G., Bray,J.E., Todd,A.E., Martin,A., Lo Conte,L. and Thornton,J.M. (1999) The CATH database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.
3. Fersht,A.R., Matouschek,A. and Serrano,L. (1992) The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.*, **224**, 771–782.
4. Yang,H., Liu,L., Li,J., Chen,J. and Du,G. (2015) Rational design to improve protein thermostability: recent advances and prospects. *ChemBioEng Rev.*, **2**, 87–94.
5. Knowles,T.P., Vendruscolo,M. and Dobson,C.M. (2014) The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.*, **15**, 384–396.
6. Sancho,J. (2013) The stability of 2-state, 3-state and more-state proteins from simple spectroscopic techniques… plus the structure of the equilibrium intermediates at the same time. *Arch. Biochem. Biophys.*, **531**, 4–13.
7. Temel,D.B., Landsman,P. and Brader,M.L. (2016) Orthogonal methods for characterizing the unfolding of therapeutic monoclonal antibodies: differential scanning calorimetry, isothermal chemical denaturation, and intrinsic fluorescence with concomitant static light scattering. *Methods Enzymol.*, **567**, 359–389.
8. Gelman,H. and Gruebele,M. (2014) Fast protein folding kinetics. *Q. Rev. Biophys.*, **47**, 95–142.
9. Goyal,M., Chaudhuri,T.K. and Kuwajima,K. (2014) Irreversible denaturation of maltodextrin glucosidase studied by differential scanning calorimetry, circular dichroism, and turbidity measurements. *PloS One*, **9**, e115877.
10. Dimitriadis,G., Drysdale,A., Myers,J.K., Arora,P., Radford,S.E., Oas,T.G. and Smith,D.A. (2004) Microsecond folding dynamics of the F13W G29A mutant of the B domain of staphylococcal protein A by laser-induced temperature jump. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 3809–3814.
11. Lepock,J.R., Ritchie,K.P., Kolios,M.C., Rodahl,A.M., Heinz,K.A. and Kruuv,J. (1992) Influence of transition rates and scan rate on kinetic simulations of differential scanning calorimetry profiles of reversible and irreversible protein denaturation. *Biochemistry*, **31**, 12706–12712.
12. Sanchez-Ruiz,J.M. (1992) Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys. J.*, **61**, 921–935.
13. Privalov,P.L. and Dragan,A.I. (2007) Microcalorimetry of biological macromolecules. *Biophys. Chem.*, **126**, 16–24.
14. Tsytlonok,M. and Itzhaki,L.S. (2013) The how's and why's of protein folding intermediates. *Arch. Biochem. Biophys.*, **531**, 14–23.
15. Neudecker,P., Robustelli,P., Cavalli,A., Walsh,P., Lundstrom,P., Zarrine-Afsar,A., Sharpe,S., Vendruscolo,M. and Kay,L.E. (2012) Structure of an intermediate state in protein folding and aggregation. *Science*, **336**, 362–366.
16. Bowman,G.R., Beauchamp,K.A., Boxer,G. and Pande,V.S. (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, **131**, 124101.
17. Wei,G., Xi,W., Nussinov,R. and Ma,B. (2016) Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chem. Rev.*, **116**, 6516–6551.
18. Johnson,K.A. (2009) Fitting enzyme kinetic data with KinTek global kinetic explorer. *Methods Enzymol.*, **467**, 601–626.
19. Kuzmič,P. (2009) DynaFit—a software package for enzymology. *Methods Enzymol.*, **467**, 247–280.
20. Niklasson,M., Andresen,C., Helander,S., Roth,M.G., Zimdahl Kahlin,A., Lindqvist Appell,M., Mårtensson,L. and Lundström,P. (2015) Robust and convenient analysis of protein thermal and chemical stability. *Prot. Sci.*, **24**, 2055–2062.
21. Harder,M.E., Deinzer,M.L., Leid,M.E. and Schimerlik,M.I. (2004) Global analysis of threestate protein unfolding data. *Prot. Sci.*, **13**, 2207–2222.
22. Li,A., Ziehr,J.L. and Johnson,K.A. (2017) A new general method for simultaneous fitting of temperature and concentration dependence of reaction rates yields kinetic and thermodynamic parameters for HIV reverse transcriptase specificity. *J. Biol. Chem.*, **292**, 6695–6702.
23. Yi,Q., Scalley,M.L., Simons,K.T., Gladwin,S.T. and Baker,D. (1997) Characterization of the free energy spectrum of peptostreptococcal protein L. *Fold. Des.*, **2**, 271–280.
24. Mazurenko,S., Kunka,A., Beerens,K., Johnson,C.M., Damborsky,J. and Prokop,Z. (2017) Exploration of protein unfolding by modelling calorimetry data from reheating. *Sci. Rep.*, **7**, 16321.
25. Ibarra-Molero,B., Naganathan,A.N., Sanchez-Ruiz,J.M. and Muñoz,V. (2016) Modern analysis of protein folding by differential scanning calorimetry. *Methods Enzymol.*, **567**, 281–318.
26. Rodriguez-Larrea,D., Ibarra-Molero,B., de Maria,L., Borchert,T.V. and Sanchez-Ruiz,J.M. (2008) Beyond Lumry–Eyring: an unexpected pattern of operational reversibility/irreversibility in protein denaturation. *Prot. Struct. Funct. Bioinf.*, **70**, 19–24.
27. Lyubarev,A.E. and Kurganov,B.I. (2000) Analysis of DSC data relating to proteins undergoing irreversible thermal denaturation. *J. Therm. Anal. Cal.*, **62**, 51–62.
28. Milardi,D., La Rosa,C. and Grasso,D. (1994) Extended theoretical analysis of irreversible protein thermal unfolding. *Biophys. Chem.*, **52**, 183–189.
29. Kirk,P., Thorne,T. and Stumpf,M.P. (2013) Model selection in systems and synthetic biology. *Curr. Opin. Biotechnol.*, **24**, 767–774.
30. Dvorak,P., Bednar,D., Vanacek,P., Balek,L., Eiselleova,L., Stepankova,V., Sebestova,E., Kunova Bosakova,M., Konecna,Z., Mazurenko,S. *et al.* (2018) Computer-assisted engineering of hyperstable fibroblast growth factor 2. *Biotechnol. Bioeng.*, **115**, 850–862.

Published online 9 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D319–D324***
doi: 10.1093/nar/gkaa981

FireProt^{DB}: database of manually curated protein stability data

Jan Stourac[1,2,†], Juraj Dubrava[1,3,†], Milos Musil[1,2,3], Jana Horackova[1], Jiri Damborsky[1,2], Stanislav Mazurenko[1,*] and David Bednar[1,2,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Masaryk University, Brno, Czech Republic, [2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic and [3]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

Received August 14, 2020; Revised September 18, 2020; Editorial Decision October 09, 2020; Accepted October 12, 2020

ABSTRACT

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.

INTRODUCTION

Proteins play essential roles in many biotechnological and biomedical applications, where they are often subjected to extreme environments, e.g. elevated temperatures or the presence of various salts. However, naturally occurring proteins have mostly evolved to function in the mild environmental conditions, and therefore their applicability is limited in the industrial applications. For this reason, protein engineers generally aim to improve protein stability, and thermostability is one of their primary targets (1) as it is correlated with serum survival time (2), half-life (3), expression yield (4) and activity in the presence of denaturants (5). A reliable assessment of the effect of a mutation on protein stability is often performed experimentally. Extensive experimental screening, however, is slow and costly, prompting the use of *in silico* approaches for the pre-selection of promising mutations. These methods are usually based on one of the three principles: (i) free energy calculations, (ii) phylogenetics or (iii) machine learning. With the recent advances in artificial intelligence, tool developers increasingly resort to the third group of methods. However, the accuracy of the machine learning-based predictors is still severely limited by the lack of high-quality data (6). Experimental characterizations are usually not capable of producing large amounts of data, and the majority of these measurements are scattered in the scientific literature. Thus, there is a strong demand for systematic collection, validation, and organization of such data in a database.

Two attempts have been made to establish a systematic and extensive collection of thermostability data so far. The first and largest database is the Thermodynamic Database for Proteins and Mutants–ProTherm (7). It was first released in 1999 with the aim to collect experimentally determined thermodynamic parameters for wild-type proteins

To whom correspondence should be addressed. Tel: +420 605 143 394; Email: davidbednar1208@gmail.com
Correspondence may also be addressed to S. Mazurenko. Email: mazurenko@mail.muni.cz
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Website address: https://loschmidt.chemi.muni.cz/fireprotdb.

and their mutants from the published literature. Its latest version contains >25 000 entries from 740 proteins, and it serves as the primary source of protein stability data for the development of new predictors. However, ProTherm was last updated in 2013 so the database is already out-of-date. Moreover, several critical issues have been reported, such as inaccurate annotations or wrong signs of values (6,8–10). This makes ProTherm even more difficult to use as time-demanding manual filtering and validation steps are required to confirm the values in the original articles. This manual filtering led to the construction of many different, often overlapping, subsets with corrected values and occasionally new data. Some of these derivative datasets were deposited to the VariBench database (11) without any attempts to reintegrate the changes into ProTherm or create an improved database. This changed in 2018 when ProtaBank (12) was released. This database aims to collect a wide range of protein engineering data such as thermostability, activity, expression, binding and several others. The developers imported all the data from ProTherm, yet they did not seem to perform any manual curation. Therefore, the critical issues listed above were not resolved. And while ProtaBank enriched the ProTherm data with recent experimental studies, the database does not offer any advanced searching and filtering capabilities, at least in its non-commercial version. This makes the data extraction and processing tedious by necessitating many manual steps and hindering the application of such data-driven methods as machine learning.

To overcome these limitations, we established the FireProt^{DB} database that holds manually curated thermostability data for single-point mutants. The database contains the data available in ProTherm, ProtaBank, and our extensive manual literature search. Its user-friendly interface allows easy and interactive browsing through the experimental data and provides links to the corresponding UniProt and PDB entries. Moreover, advanced searching and filtering capabilities, the ability to download the data in a simple table format, and meticulous labelling of data entries used for training and testing of published tools prompt the further application of machine learning.

MATERIALS AND METHODS

Database architecture and data model

The top-level entity of the FireProt^{DB} database is a unique protein sequence entry with the assigned UniProt ID (13). Protein sequences were preferred to structures due to the broader availability of the former. Each sequence is a string of amino acids in specified positions. Multiple mutations can be assigned to a single position, and each mutation can be evaluated by multiple measurements and derived values. The measurements represent the experimental values of the Gibbs free energy changes upon mutation ($\Delta\Delta G$) or changes in melting temperatures (ΔT_m). The derived values stand for averages or medians of multiple measurements for a particular mutation. Each measurement is also accompanied by a curation flag that indicates whether the value was manually validated against the original publication to guarantee its correctness. Furthermore, each measurement and derived value can be assigned to multiple published datasets to promote accurate validation and benchmarking of computational tools.

From the structural point of view, each sequence can have one or more assigned biological units that denote biologically relevant quaternary structures of asymmetric units stored in the PDB database (14). For representative biological units, the HotSpot Wizard 3.0 (15) calculation was executed to compute additional sequential and structural annotations. These annotations can help with the analysis of selected mutations and serve as pre-calculated features applicable in machine learning models.

Stability data acquisition and curation

FireProt^{DB} is composed of the data from four sources: the ProTherm database, the ProtaBank database, manual mining of the scientific literature, and data collected in our laboratory (Figure 1). The primary data source was ProTherm. Due to the multiple problems mentioned in the introduction, we followed several filtering steps. In the first step, we retained only those entries that met the following four criteria: (i) they have a single-point mutation; (ii) the mutation is not an insertion or deletion; (iii) the protein has a SwissProt accession code and/or a PDB identifier; (iv) the entry includes a measured $\Delta\Delta G$ and/or ΔT_m. Secondly, we performed a validity check of SwissProt accession codes and updated obsolete entries. ProTherm references mutations by their structure index, i.e., the residue number in the structure, which in many cases does not match their sequence index, i.e. the position in the sequence. To overcome this issue, we used a similar approach as in PDBSWS (16): use the Needleman-Wunsch algorithm (17) to construct the global sequence alignment of sequences extracted from PDB and UniProt entries and map the mutations onto the UniProt sequences. In the next step, we confirmed that the reported wild-type amino acids are in the correct positions in the structures and unified the reported units. Finally, we matched the data with the manually curated entries in the FireProt dataset (18), updated the values, and marked them as 'curated'.

In addition to ProTherm, we explored the studies reported in the ProtaBank database, extracted the thermostability data, and integrated them into our database. We also performed a manual literature search using stability-based keywords such as 'protein stability', 'thermostability', 'free energy upon mutation', 'protein stabilization'. We mined the recent scientific articles reporting mutants with measured stability data and contacted the authors of the publications when the relevant data were not available in the article. All such entries were marked as 'curated' as we extracted them directly from the original publications. Finally, we reviewed the thermostability data collected in our lab throughout the last few years and added them to the database. We perform experimental protein characterization in our protein engineering projects on a regular basis, and measuring protein stability is an essential part of such characterization. In total, the three sources led to a significant enlargement of the data size by 62% in terms of all the entries. The number of curated entries more than dou-

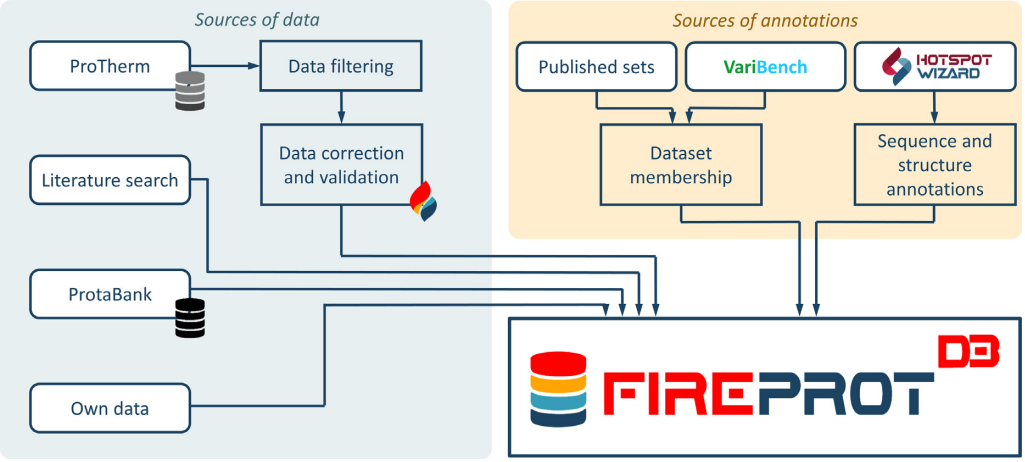


Figure 1. A schematic representation of the data comprising FireProt^{DB}. The primary source of data is filtered ProTherm (7). The FireProt data subset (18) was manually curated, compared to the source publications, and marked with the 'curated' flag. The publications from ProtaBank (12) and manual literature search were also used to deposit the data. Each mutation in the deposited data was annotated according to its membership in the published datasets and those deposited on VariBench (11). The HotSpot Wizard 3.0 (15) annotation tool was applied to each protein entry with a known tertiary structure.

bled compared to the previously collected cleaned FireProt subset of ProTherm.

Dataset assignment

In the second acquisition step, we collected 40 datasets from the VariBench database (11) and literature (18), which were used previously for training or testing of existing predictors. Since all these datasets are at least partially derived from ProTherm, we could label each measurement in FireProt^{DB} by its membership in the datasets. These labels are particularly useful for the comparison of new prediction models to the existing tools. This task is usually done by the performance evaluation of predictors on a dataset that is entirely independent of the training and test sets used for the development of the tools. Since the dataset construction is often laborious and consists of a manual data processing, the possibility to directly exclude the data present in given datasets significantly simplifies and speeds up the construction process.

Calculation of additional annotations

To provide our users with a more advanced description of their proteins of interest, we enriched the database by several important sequence- and structure-related information. These calculations were performed by HotSpot Wizard 3.0 (15), which is currently the only tool capable of deriving all these features in a single calculation (19) and provides machine-readable results. HotSpot Wizard was executed on a representative biological unit of each protein and provided the annotations for a structure, such as the residues located in protein pockets and tunnels, and a sequence, such as catalytic residues, evolutionary conservation scores, back-to-consensus mutations, and correlated pairs. These annotations can be helpful for a better understanding of structure-function relationships as well as for generating features for machine learning.

RESULTS

Web interface

The web interface was designed for both types of expected users—protein chemists and software developers. Protein chemists are often looking for the thermostability evidence for their protein of interest, and they will benefit from its interactivity and details pages with additional information. Machine learning experts and bioinformaticians will be more interested in advanced filtering capabilities facilitating the process of construction of highly customized datasets for the training or assessment of various predictors. The entry point to the database is the search form, which allows browsing in two major ways: (i) a simple full-text search for querying the database using protein name, UniProt accession codes, PDB identifiers, protein names, publications, authors or organisms and (ii) an advanced search allowing the users to construct complex rules based on the relational algebra and all available database fields. The latter is one of the key features of FireProt^{DB} as it facilitates the construction of highly customized datasets needed for the development of new predictors.

Once the user clicks on the 'Search' button, they are redirected to the page with the result table. This table contains a list of available experiments, their basic annotations, and measured values. The table is paginated to eliminate possible performance issues and allows further interactive filtering of displayed values. The user can then easily export the search results in the CSV format using the 'Export' button at the top or the bottom of the page.

Clicking on a mutation name leads to a page with a more detailed view, showing all the data entries and datasets that include the selected mutation. Clicking on a protein name leads to a page providing the basic information such as UniProt accession code, organism and Enzyme Commission number, as well as detailed annotation of secondary structure, catalytic sites, natural variants and amino acid charges derived from UniProt database using interactive

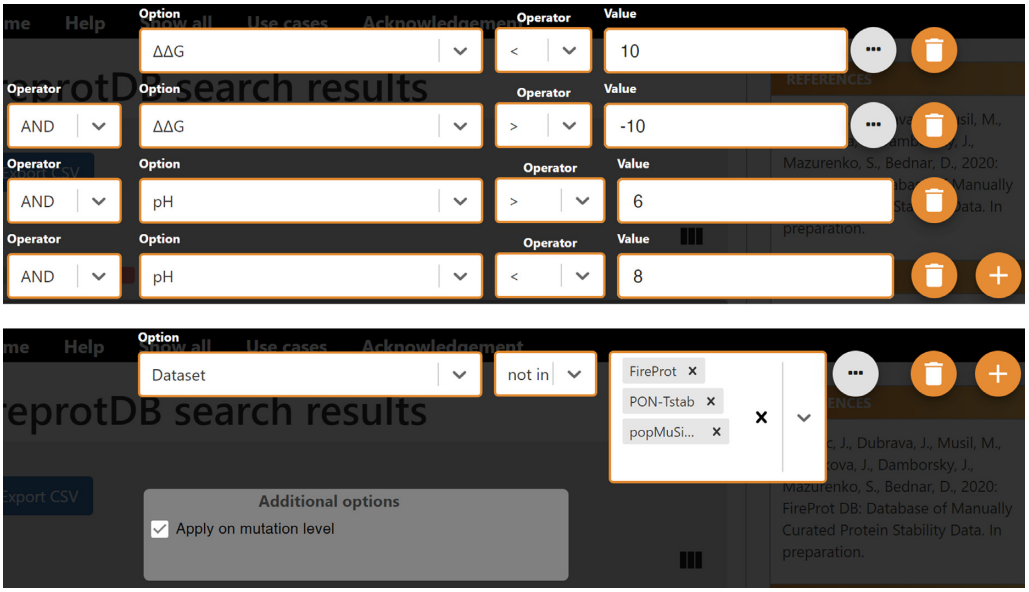


Figure 2. Examples of filtering protocols in FireProt^{DB}. **Top**: The request filters out the data collected at extreme pH or with extreme $\Delta\Delta G$ values, resulting in >3500 data points left. **Bottom**: An example of excluding all the mutations that appear in PopMuSiC, FireProt, or PON-Tstab datasets.

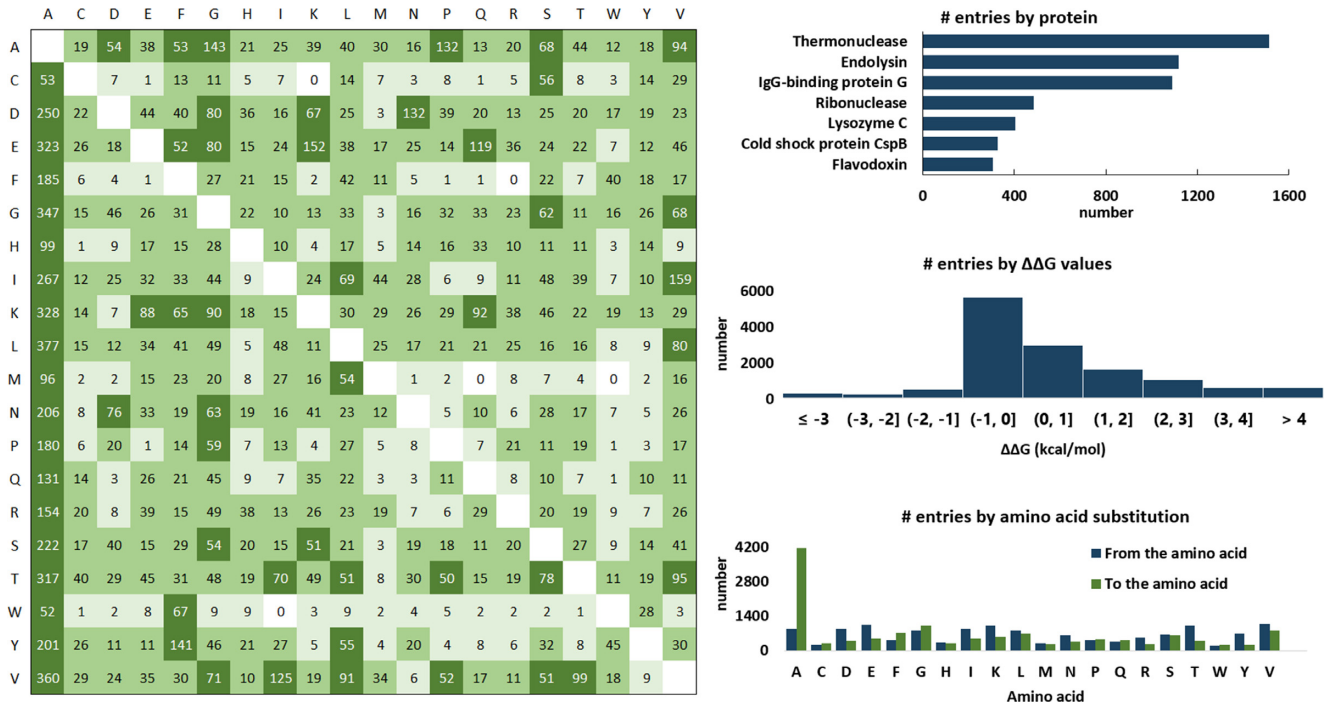


Figure 3. An overview of the data deposited to FireProt^{DB}. **Left**: The table shows the total number of each substitution pair with the wild type amino acids in rows, mutant amino acids in columns, and the coloring according to the thresholds of 1 (light green), 10 (medium green) and 50 (dark green) entries for the corresponding substitution. **Right**: Histograms showing the top seven proteins by their UniProt IDs, the $\Delta\Delta G$ values, and the cumulative number of amino acid substitutions.

ProtVista tracks (20). This page also contains a list of all known biological units and a table with all experimental measurements.

Search queries

Several types of search queries may be of interest to the users. The first one relates to data filtering by values (10). Typically, software developers filter out the data collected at extreme pH (<6 or >8) due to changes in charged states for ionizable residues. The entries with large absolute $\Delta\Delta G$ or ΔT_m are also sometimes excluded due to likely higher measurement errors, and also because dramatic changes to the stability may indicate significant structural alterations to the wild type, which may become a problem for structure-based features. The second type is relevant for benchmark-

ing of a newly designed predictor against the existing tools or creating a meta predictor. In either case, one usually needs to derive a data subset that has not been used by the existing predictors for training. The main reason is the robust performance estimate, which is typically over-optimistic for these sets (6). Two corresponding examples of such filtering protocols are shown in Figure 2.

Database dump

For the users requesting even higher control over the data and filtering capabilities, we offer the possibility to download the complete dump of the database in the SQL format. This data file can be easily imported to any modern MariaDB server, version 10.2, and higher. Since the database structure is complex and any custom query requires joining of multiple tables, the dump also contains a pre-defined view 'mutation_experiments_summary'. The summary combines all the tables and provides the data in a similar structure as the CSV export from the user interface. This view or its definition can serve as a useful starting point for additional filtering or creating custom queries.

Data statistics

Currently, FireProt^{DB} contains 13274 entries for 237 proteins (Figure 3), from which 8189 measurements originated from ProTherm. The remaining 5085 entries were added from our literature search (18%), publications from ProtaBank (28%), VariBench (53%), and our own records (1%). In total, 43% entries are destabilizing mutations ($\Delta T_m \leftarrow 1$ or $\Delta\Delta G > 1$ kcal/mol), 14% stabilizing ($\Delta T_m > 1$ or $\Delta\Delta G \leftarrow 1$ kcal/mol), and 43% considered neutral ($-1 \leq \Delta T_m \leq 1$ or $-1 \leq \Delta\Delta G \leq 1$ kcal/mol). The database also includes annotations for 40 various published datasets derived from ProTherm, deposited to VariBench (11), or available in the corresponding articles and web servers. As far as enzymes are concerned, those collected in the database cover the first six EC classes, three of which by >40% on the second level.

DISCUSSION

The availability of large high-quality datasets is one of the critical requirements for the advancement of machine learning-based *in silico* predictors. While some promising high-throughput experimental methods have been released recently (21,22), their validation is still ongoing, and protein stability experiments are still time-consuming and expensive. Building training and testing datasets is hindered by the data being hidden in the original articles, generating a strong demand for their systematic mining, collection, validation, and homogenization. The existing databases are not fulfilling all the requirements as ProTherm is outdated and contains incorrect data, and ProtaBank does not provide advanced search and export tools and is partly commercial.

FireProt^{DB} is a novel database for experimental thermostability data of protein single-point mutants. It consists of the data manually extracted from ProTherm, articles from ProtaBank, new data obtained by mining the recent literature, and the data collected in our laboratory. The database is accessible via a user-friendly graphical web interface allowing the users to search and browse the data interactively. Moreover, all the entries are annotated to indicate whether they belong to the already published datasets. These annotations, combined with the advanced searching and filtering capabilities, make FireProt^{DB} a valuable data resource for machine learning developers interested in constructing highly customized datasets.

In the future, we will improve our searching queries and employ automatic text-mining machine learning-based approaches (23–25) to accelerate literature mining and data collection, which will be followed by manual curation. We will also prepare an interactive form for data submissions by the users. Finally, we will extend the set of automatically generated features for mutations and add sequence similarity filtering to improve the data usability by the community of engineers applying machine learning to predict changes in protein stability.

FUNDING

Czech Ministry of Education, Youth and Sports [LQ1605, LM2015047, LM2018121, 02.1.01/0.0/0.0/18_046/0015975 to J.D.]; Operational Programme Research, Development and Education project MSCAfellow@MUNI [CZ.02.2.69 /0.0/0.0/17_050/0008496 to S.M.]; Brno University of Technology [FIT-S-20-6293 to M.M.]; CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 of the European Union [857560 to J.D.]; Czech Science Foundation [20-15915Y to D.B.]. Funding for open access charge: Czech ministry of Education, Youth and Sports [LM2015047].
Conflict of interest statement. None declared.

REFERENCES

1. Modarres,H.P., Mofrad,M.R. and Sanati-Nezhad,A. (2016) Protein thermostability engineering. *RSC Adv.*, **6**, 115252–115270.
2. Gao,D., Narasimhan,D.L., Macdonald,J., Brim,R., Ko,M.-C., Landry,D.W., Woods,J.H., Sunahara,R.K. and Zhan,C.-G. (2009) Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, **75**, 318–323.
3. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
4. Ferdjani,S., Ionita,M., Roy,B., Dion,M., Djeghaba,Z., Rabiller,C. and Tellier,C. (2011) Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnol. Lett.*, **33**, 1215–1219.
5. Polizzi,K.M., Bommarius,A.S., Broering,J.M. and Chaparro-Riggers,J.F. (2007) Stability of biocatalysts. *Curr. Opin. Chem. Biol.*, **11**, 220–225.
6. Musil,M., Konegger,H., Hon,J., Bednar,D. and Damborsky,J. (2019) Computational design of stable and soluble biocatalysts. *ACS Catal.*, **9**, 1033–1054.
7. Kumar,M.D.S., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
8. Pucci,F., Bernaerts,K.V., Kwasigroch,J.M. and Rooman,M. (2018) Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, **34**, 3659–3665.
9. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.

10. Mazurenko,S. (2020) Predicting protein stability and solubility changes upon mutations: data perspective. *Chem. Cat. Chem.*, **12**, doi:10.1002/cctc.202000933.
11. Sasidharan Nair,P. and Vihinen,M. (2013) VariBench: a benchmark database for variations. *Hum. Mutat.*, **34**, 42–49.
12. Wang,C.Y., Chang,P.M., Ary,M.L., Allen,B.D., Chica,R.A., Mayo,S.L. and Olafson,B.D. (2018) ProtaBank: a repository for protein design and engineering data. *Protein Sci.*, **27**, 1113–1124.
13. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
14. Jefferson,E.R., Walsh,T.P. and Barton,G.J. (2006) Biological units and their effect upon the properties and prediction of protein-protein interactions. *J. Mol. Biol.*, **364**, 1118–1129.
15. Sumbalova,L., Stourac,J., Martinek,T., Bednar,D. and Damborsky,J. (2018) HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res.*, **46**, W356–W362.
16. Martin,A.C.R. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297–4301.
17. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
18. Musil,M., Stourac,J., Bendl,J., Brezovsky,J., Prokop,Z., Zendulka,J., Martinek,T., Bednar,D. and Damborsky,J. (2017) FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res.*, **45**, W393–W399.
19. Sequeiros-Borja,C.E., Surpeta,B. and Brezovsky,J. Recent advances in user-friendly computational tools to engineer protein function. *Brief. Bioinform.*, doi:10.1093/bib/bbaa150.
20. Watkins,X., Garcia,L.J., Pundir,S., Martin,M.J. and UniProt Consortium (2017) ProtVista: visualization of protein sequence annotations. *Bioinformatics*, **33**, 2040–2041.
21. Bunzel,H.A., Garrabou,X., Pott,M. and Hilvert,D. (2018) Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Curr. Opin. Struct. Biol.*, **48**, 149–156.
22. Matreyek,K.A., Starita,L.M., Stephany,J.J., Martin,B., Chiasson,M.A., Gray,V.E., Kircher,M., Khechaduri,A., Dines,J.N., Hause,R.J. *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.*, **50**, 874–882.
23. Naderi,N. and Witte,R. (2012) Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics*, **13**, S10.
24. Witte,R. and Baker,C.J.O. (2007) Towards a systematic evaluation of protein mutation extraction systems. *J. Bioinform. Comput. Biol.*, **5**, 1339–1359.
25. Wei,C.-H., Harris,B.R., Kao,H.-Y. and Lu,Z. (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433–1439.

Published online 17 May 2022 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W145–W151***
https://doi.org/10.1093/nar/gkac378

CalFitter 2.0: Leveraging the power of singular value decomposition to analyse protein thermostability

Antonin Kunka[1,2,†], David Lacko[3,†], Jan Stourac[1,2], Jiri Damborsky[1,2], Zbynek Prokop[1,2,*] and Stanislav Mazurenko[1,2,*]

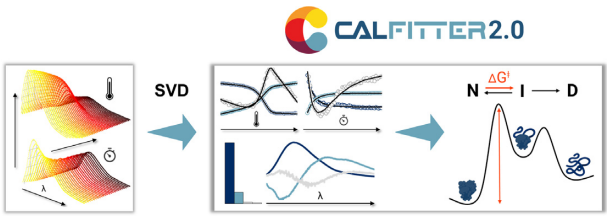
[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic, [2]International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno, Czech Republic and [3]Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

Received March 23, 2022; Revised April 20, 2022; Editorial Decision April 29, 2022; Accepted May 14, 2022

ABSTRACT

The importance of the quantitative description of protein unfolding and aggregation for the rational design of stability or understanding the molecular basis of protein misfolding diseases is well established. Protein thermostability is typically assessed by calorimetric or spectroscopic techniques that monitor different complementary signals during unfolding. The CalFitter webserver has already proved integral to deriving invaluable energy parameters by global data analysis. Here, we introduce CalFitter 2.0, which newly incorporates singular value decomposition (SVD) of multi-wavelength spectral datasets into the global fitting pipeline. Processed time- or temperature-evolved SVD components can now be fitted together with other experimental data types. Moreover, deconvoluted basis spectra provide spectral fingerprints of relevant macrostates populated during unfolding, which greatly enriches the information gains of the CalFitter output. The SVD analysis is fully automated in a highly interactive module, providing access to the results to users without any prior knowledge of the underlying mathematics. Additionally, a novel data uploading wizard has been implemented to facilitate rapid and easy uploading of multiple datasets. Together, the newly introduced changes significantly improve the user experience, making this software a unique, robust, and interactive platform for the analysis of protein thermal denaturation data. The webserver is freely accessible at https://loschmidt.chemi.muni.cz/calfitter.

GRAPHICAL ABSTRACT



INTRODUCTION

The thermal stability of proteins is imperative for their correct biological function, and its disruption often has devastating effects on the host organism. Protein instability leads to misfolding and aggregation that are associated with many severe human diseases, such as Alzheimer's, Parkinson's or Amyotrophic Lateral Sclerosis (1), and that gravely limit the efficient application of proteins in biotechnological, pharmaceutical, and other industries (2). Our general knowledge of the key structural and energetic basis of protein stability originates predominantly from the mutational unfolding studies (3,4). Although the framework for the proper analysis of thermodynamic and kinetic stability of proteins has a long history (5,6), experimental output from many stabilization studies is often limited to only a few empirical parameters, e.g. apparent melting temperatures (7). Considering the significant advancement in high-throughput biophysical techniques and a growing number of data-driven machine learning tools for protein stability prediction (8), the need for a robust, easy-to-use, and freely available platform for analysis of protein thermal denaturation data is therefore pressing.

To address this, we have previously developed the CalFitter webserver (9), which enables a global analysis of temperature-induced protein unfolding data measured with commonly used biophysical techniques, including differ-

*To whom correspondence should be addressed. Tel: +420 549 4930 41; Fax: +420 549 4962 03; Email: mazurenko@mail.muni.cz
Correspondence may also be addressed to Zbynek Prokop. Email: zbynek@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.

ential scanning calorimetry (DSC), fluorescence, circular dichroism (CD), Fourier-transform infrared (FTIR) spectroscopies, and temperature jumps. The software integrates thirteen unique unfolding models, involving a various number of defined macrostates and different combinations of reversible or irreversible transitions between them. CalFitter 1.0 compiles the conventionally used reversible models as well as more complex partially or fully irreversible models collected from more recent literature (6,10). The former analyse the data based on the principles of equilibrium thermodynamics, whereas the latter treat the data from temperature scanning experiments as a dynamic process under kinetic control, sensitive to a particular scan rate, and integrate the equations describing the fractions of states numerically. The detailed mathematical description of these models can be found in the original publications (9,11). Experimental data can be interactively modelled based on the defined parameters, which allows users to easily test the validity of the selected model and make verifiable quantitative predictions about protein unfolding behavior. The output of the analysis is provided in an easily processible format, as physically relevant energy parameters derived based on the Eyring formalism of the transition state theory, e.g. Gibbs free energy differences (ΔG), which are being actively used as training data for recent machine learning stability predictors (12,13). To our best knowledge, it is the only tool that allows simultaneous fitting of data from temperature scanning experiments together with unfolding kinetics. The recent examples of CalFitter use include decoding the mechanism of domain-swapping of computationally stabilized haloalkane dehalogenase (14), explaining the kinetic stability of cold adapted subtilase (15), elucidating the aggregation propensity of polyketide cyclase (16), or study of dihydrofolate reductase evolution (17).

While CalFitter 1.0 has proved integral to the global data analysis of a wide range of experimental signals, recent technological advancements in massive data collection offer new opportunities for analysis yet to be fully exploited in the pipeline. Spectroscopic techniques are conveniently used to monitor protein unfolding due to their low sample requirements, moderate to high-throughput, and rich informational output. Earlier measurements were limited to an intensity change at a single wavelength (e.g. CD ellipticity) or the wavelength of the maximum intensity (fluorescence). However, such simple signals provide an incomplete picture of the unfolding process and are prone to misinterpretation (18). In contrast, recent technologies enable monitoring the entire protein spectra, which directly report on the local and global conformational changes during the unfolding. Yet this tremendous informational potential has not been fully exploited as it was not accompanied by the development of a suitable analytical toolbox for researchers without the advanced data analysis background.

In this work, we present a major update of the original CalFitter that addresses the current needs of the field in complete spectral data analysis using singular value decomposition (SVD). SVD is a powerful mathematical tool for data dimensionality reduction and has been exploited in several mechanistic studies of protein folding and unfolding using time-resolved fluorescence (19), small angle X-ray scattering (20–22), FTIR (23) and other advanced biophysical techniques (24,25). It is widely used for the detection of potential (un)folding intermediates due to its ability to extract spectral fingerprints of the protein states contributing to the overall signal (26–31). CalFitter 2.0 newly features (i) an easy upload of protein spectra recorded as a function of temperature (scanning experiments), time (kinetics), or other parameters (e.g. denaturant concentration, pH), (ii) the automated SVD analysis of these spectra, (iii) the interactive interface for dynamic visualization of the results and data pre-processing, (iv) the readily available export of the results in the excel format and (v) the global fitting of the SVD components from temperature scanning and unfolding kinetics experiments along with other signals, e.g. from DSC. The addition of the SVD analysis to the CalFitter pipeline greatly enhances the informational output of the software by providing spectral fingerprints of the relevant macrostates populated during protein unfolding. Additionally, based on the users' feedback, we have completely reworked the data uploading procedure to accommodate various input file formats. The newly introduced changes significantly expand the applicability of the CalFitter 2.0 and make it a unique platform for global analysis of protein denaturation experiments.

NOVEL FEATURES

The original CalFitter has been described elsewhere (9), and its schematic overview, together with the novel functions introduced to the new version, are shown in Figure 1. The main feature of CalFitter 2.0 is the new SVD analysis module that is used as a data pre-processing step prior to the global fitting or as an independent tool for SVD analysis of virtually any multi-wavelength datasets. Another critical feature is a completely reworked uploading wizard supporting various input data formats and uploads from multiple files. Its interactive interface allows the quick and intuitive selection, labelling, visualization, and pre-processing of the input datasets. We provide a detailed description of the uploading wizard in the help section of the webserver https://loschmidt.chemi.muni.cz/calfitter/?action=help.

SVD ANALYSIS

The input to the singular value decomposition consists of multi-wavelength data organized in a rectangular $m \times n$ matrix in which the m rows represent the wavelengths, and the n columns represent the experimental points, e.g. spectra at different times or temperatures. The SVD is a factorization of the original matrix to three matrices in the form of $U\Sigma V^{\mathrm{T}}$ (Figure 1), where the columns of the U are the left singular vectors (basis spectra), Σ contains singular values (component amplitudes) on its diagonal, and the rows of the V^{T} are the right singular vectors (time or temperature components). The detailed mathematical description of the algorithm procedure, together with the results of its thorough validation, can be found in the Supplementary Data (Table S1). CalFitter 2.0 performs the SVD automatically upon the data upload and displays the results in an interactive interface (Figure 2).

The graphical representation of the SVD results is displayed on the right side of the interface (section 4 in Figure 2). The first ten normalized singular values are shown

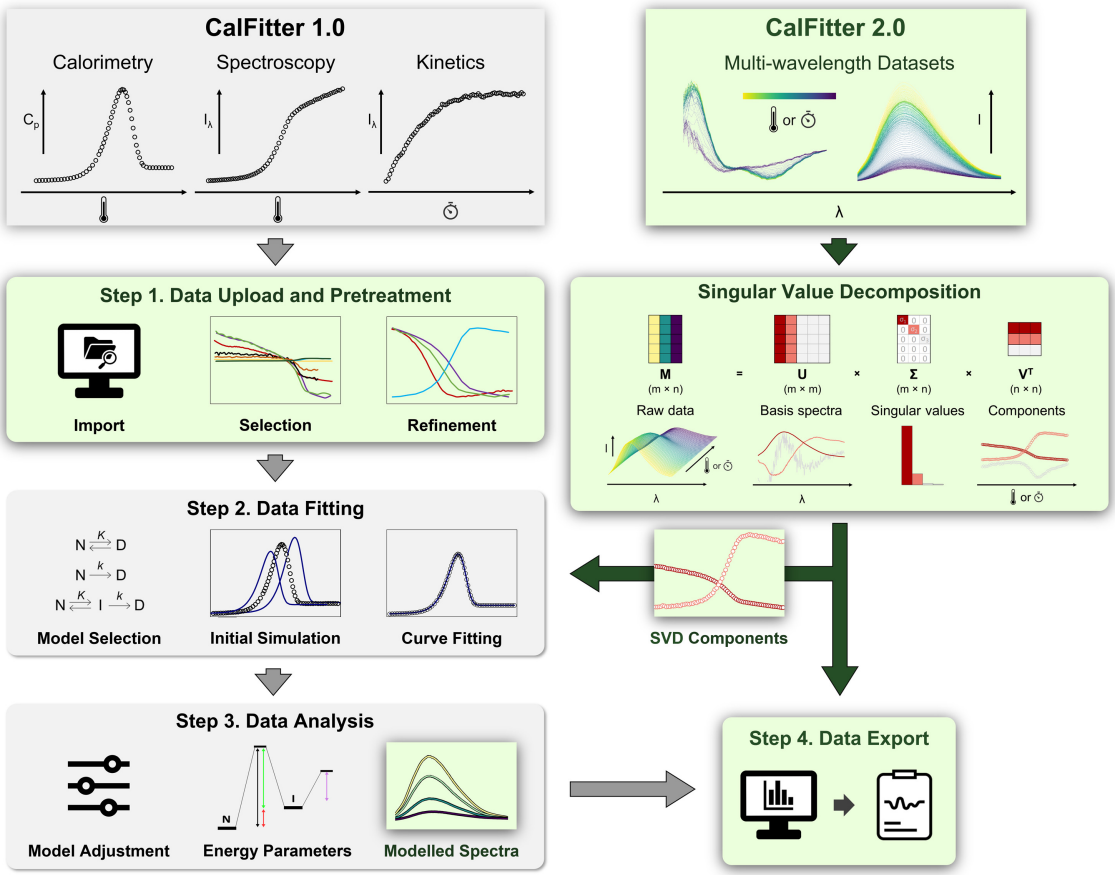


Figure 1. Overview of CalFitter workflow and newly introduced features. The features implemented into the original version 1.0 are shown in grey, while the novel features introduced into the version 2.0 are depicted in green. The details of individual steps and procedures are provided in the text or can be found in the original publication (9).

in the bar graph. While their total number corresponds to the number of wavelengths or experimental points of the original dataset (whichever is lower, i.e. $\min\{m,n\}$), generally fewer than ten components are sufficient to confidently reconstruct the original data. The numbers in the bar graph translate to the variation within the dataset that is explained by the respective component (>98% of data variation is sufficiently explained by the first two components in the example in Figure 2). The determination of the correct number of significant components for further analysis must be done with great care so that they truly represent all important features of the original dataset. Usually, the visual inspection of the shape of the basis spectra and regular patterns of the SVD components is the most robust yet subjective criterion. Alternatively, one can apply a cumulative threshold for the explained variation in the data (e.g. 98%) and keep only the components that are above it. Several statistical measures can also aid in the decision. The autocorrelations of each component basis spectrum and amplitude vector have been shown to provide practical guidance in determining whether a particular component captures the meaningful signal or noise in the data (32,33). To aid the users in the selection, CalFitter 2.0 marks the components whose autocorrelation coefficients are above the 0.8 threshold by an asterisk in the Singular values graph. Their exact values for each component are provided in the export Excel file, and a detailed description of how these values are calculated can be found in the Supplementary Data. In general, the explained variation and the autocorrelation methods can be applied when a more rigorous quantitative assessment of the SVD results needs to be carried out. However, the visual inspection of the components and their singular values usually suffices to make the decision.

Basis spectra (wavelength loadings) are depicted in the middle panel of section 4 in Figure 2. Typically, only the first few of them correspond to the meaningful signal components, while the others represent the experimental noise (Figures S1 and S3). The number of components to display and use in the original data reconstruction and subsequent global data analysis can be changed in the settings section (section 3 in Figure 2). The unique feature of CalFitter 2.0 is the possibility to assign the basis spectrum of the first component to that of a known protein state (typically native state, but others can be used) by uploading its spectrum as a reference. This increases the interpretability and biological relevance of the SVD results by providing spectral fingerprints of other relevant protein species populated during unfolding. The SVD is automatically recalculated when the reference spectrum is uploaded.

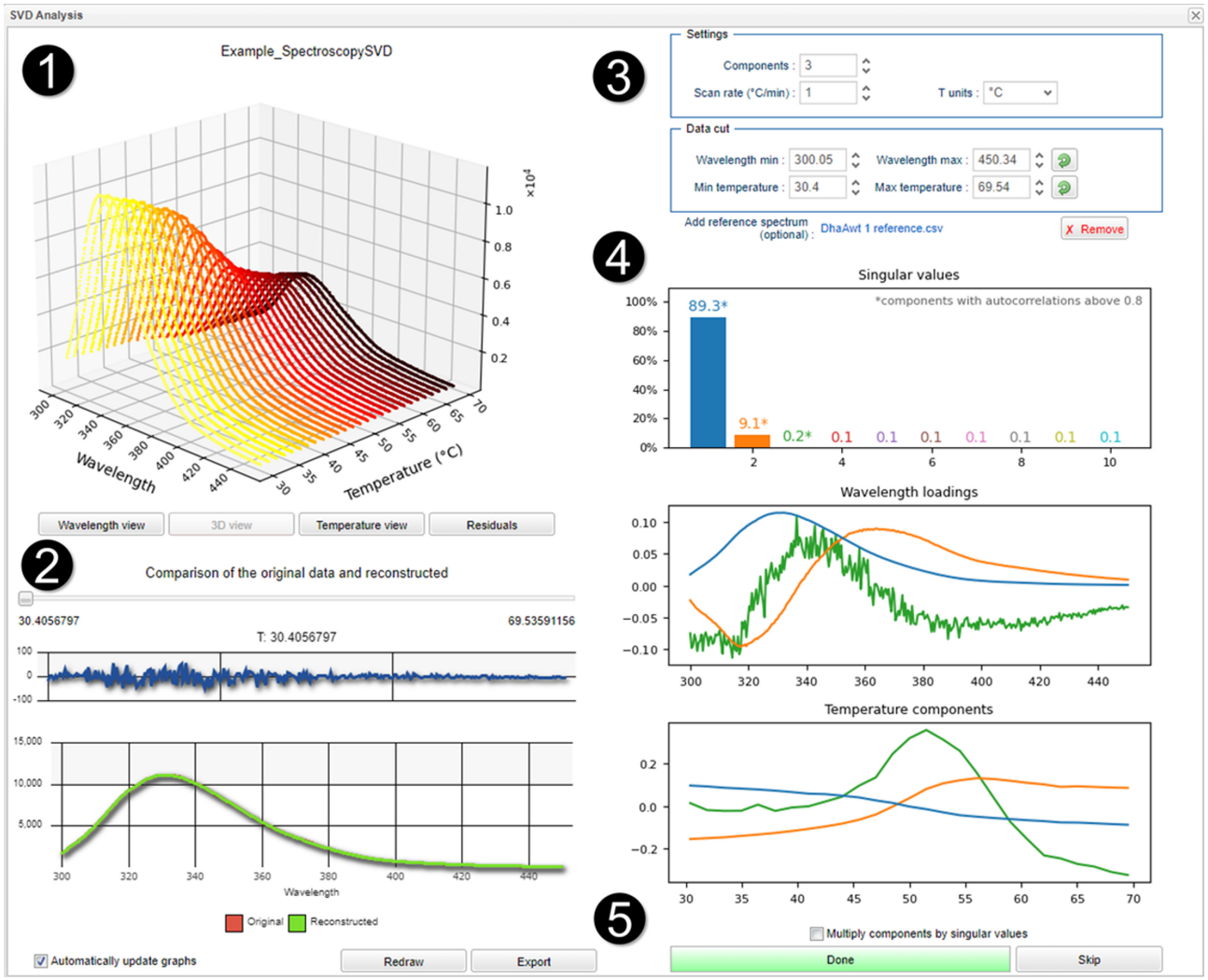


Figure 2. Interactive CalFitter 2.0 SVD analysis interface. The interface sections include (1) raw data visualization, (2) spectral reconstruction, (3) experimental parameter specification and data range settings, (4) SVD analysis results graphs and (5) export and upload options. The example data depict the thermal denaturation of the haloalkane dehalogenase DhaA (UniProt ID: P0A3G2), measured by following the changes in intrinsic protein fluorescence at the heating rate of 1°C/min. The asterisks in the Singular values plot indicate that the first three components have the autocorrelations of both the wavelength loadings and amplitude vectors above 0.8.

Finally, the changes of the component amplitudes with time (Kinetics SVD) or temperature (Spectroscopy SVD) are shown in the bottom right graph. These progress curves report on the evolution of the components throughout the course of the experiment and can be subjected to the subsequent global analysis of denaturation experiments. These curves are fully integrated into the workflow of CalFitter 1.0, i.e. they are modelled and fitted analogously and alongside the other two-dimensional signals such as calorimetry, spectroscopy, and kinetics (see Global Fitting of SVD Datasets).

The SVD analysis is fully automated in CalFitter 2.0, and all graphs dynamically change in response to the changes in parameter settings or dataset range. Spectral reconstruction of the raw data based on the selected number of components can be investigated by moving the slider below the raw data display on the left-hand side of the interface (section 2 in Figure 2). Export of the SVD results to an excel file is readily available. In principle, the SVD module can be used to analyse any type of multi-wavelength data regardless of the dynamic component (e.g. pH, denaturant, salts). However, the use of the SVD components in subsequent global fitting is restricted to the time- or temperature-dependent multi-wavelength spectral datasets collected at fixed temperatures or scan rates, respectively.

GLOBAL FITTING OF SVD DATASETS

The global analysis interface and general procedure of CalFitter have not changed significantly since the first version, and their description is provided in the original publication (9). The *Data pre-treatment* panel of the global analysis interface has been newly expanded by two additional tabs devoted to *Spectroscopy SVD* and *Kinetics SVD* datasets. The data treatment options are identical to the corresponding non-SVD data types, i.e. specification of temperature range

and normalization is possible for spectroscopy data, and collation and endpoint selection for kinetics data. The SVD components available for fitting are restricted to those selected during the SVD analysis. We recommend that only the non-noise SVD components are used for the global analysis to avoid overfitting. These are fitted similarly to other spectroscopic signals using a weighting procedure based on the number of points to ensure the balanced contribution of datasets to the penalty function of the fitting procedure (9).

The SVD output is a more accurate and unbiased representation of the original dataset compared to the conventional two-dimensional signals, e.g. using intensity change at fixed wavelengths or the area under the spectrum. The SVD preserves the informational content of the raw data while reducing its dimensionality. In contrast, the selection of an appropriate 2D signal reflecting the spectral changes during denaturation is made empirically, typically by comparing differences between spectra of the native and denatured states. As a result, these signals are often insensitive to potential intermediates that can be only scarcely populated during unfolding. For example, in Figure 3, we show the analysis of the unfolding of haloalkane dehalogenase DhaA monitored by fluorescence spectroscopy. The denaturation curves constructed from the conventionally used signals reporting on the redshift of the fluorescence maximum (the ratio of intensities at 350 nm and 350 nm, barycentric mean – BCM), or overall intensity (the area under the spectrum) both show a single transition with the overlapping midpoint temperature around 50°C, which can be fitted into a simple two-state unfolding model (Figure S4A). However, the SVD of the raw data results in three significant components, indicating the presence of an intermediate state. A closer inspection reveals that while the first two components reflect the changes captured by the two-dimensional signals, the third component, albeit less significant in explained variance (∼0.2%), has the autocorrelations above 0.8 and shows two distinct transitions. In fact, all the three components fit well to the models involving an intermediate state (Figures S4B, C). Since the singular value of the third component is low, we confirmed the presence of the intermediate by an additional measurement using another experimental technique. In our model case, DSC thermograms consisted of two transitions and were fitted alongside the SVD components to the three-state partially unfolding model (Figure S5).

Another major advantage of fitting SVD components over the conventional two-dimensional signals is the ability to back-calculate the full spectra based on the modelled parameters and compare them to the original data (Figure S6). The reconstruction of the original spectra is carried out by a linear product of the modelled SVD components and the original SVD basis spectra. The visual comparison of data reconstructed from the fitting of a different number of components, therefore, provides additional means for model validation, identification of potential deviations from the data, and evaluation criteria of potential data overfitting. In the example case study, the first two components fit well to the two-state model, but the reconstructed spectra deviate from the raw data (Figure S4A). Only the global fit of all three components to the three-state model provides satisfactory spectral reconstruction (Figures S4B, C). The detailed description of the global analysis of multiple thermal denaturation experiments, including different SVD datasets, is shown and discussed in detail in the Supplementary Data (Section Use case Figure S1–S6).

DATA INPUT AND OUTPUT

We have completely reworked the uploading procedure of the non-SVD datasets based on user feedback to improve its flexibility and user-friendliness. The software newly supports a variety of input formats, including Excel .xlsx files with multiple spreadsheets and fewer requirements on the data organisation. The new uploading wizard enables numerous interactive data pre-treatment options, including dataset visualisation, removal, column designation and parameter specification. Simultaneous upload and quick processing of multiple SVD datasets from single or different Excel files is also supported. At the same time, the input procedure is backward-compatible, i.e. when the legacy data format is recognised, the uploading wizard automatically prefills all the parameters accordingly. Similarly, the results of the SVD and global analyses can be easily exported at different stages of the process. Output datasets are organized logically in multiple spreadsheets within a single Excel .xlsx file. A detailed step-by-step description of the uploading interfaces and exporting options is provided in the help section of the webserver, which can be found at https://loschmidt.chemi.muni.cz/calfitter/?action=help. Altogether, all data manipulation steps have been significantly improved to ensure fast and intuitive application of the CalFitter and promote its wider use in the scientific community.

CONCLUSIONS AND OUTLOOK

In summary, the main new features and improvements introduced to CalFitter 2.0 include: (i) automated SVD of multi-wavelength data in an interactive interface, (ii) global fitting of time- and temperature-dependent SVD components with other types of data from protein thermal denaturation experiments, (iii) spectral reconstruction of data based on the modelled parameters, (iv) the option of uploading a reference spectrum of a known protein state in the SVD analysis, (v) the improved data uploading procedure from multiple data formats and (vi) the flexible and intuitive uploading wizard with variety of data pre-treatment options. The implementation of SVD into CalFitter 2.0 provides an extra resolution to its informational output. We hope that this unique combination of the two complex mathematical analyses, i.e. SVD and global fitting, in the single, highly interactive, and freely available platform greatly diminishes the expertise requirements for their routine application. CalFitter strives to be the gold standard for the analysis of thermal denaturation experiments, providing invaluable quantitative parameters of protein thermostability, which are crucial for the development of efficient and accurate protein engineering tools.

In the future, we plan to introduce new algorithms for automatic initialization of model parameter values based on the input data, which will make the fitting procedure much easier, especially for first-time users. Moreover, we in-

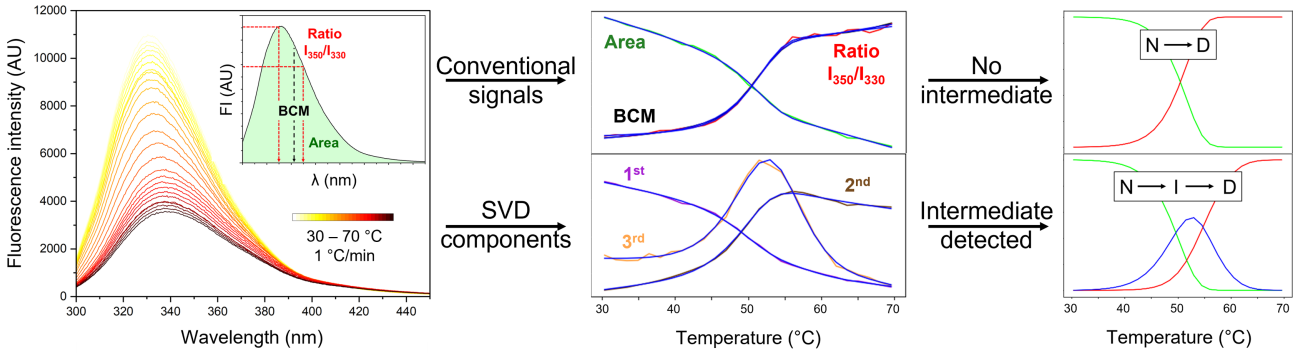


Figure 3. Differences between global fitting of single wavelength datasets and SVD components. Left: Thermal unfolding of DhaA monitored by fluorescence spectroscopy at the 1°C/min scan rate. Inset: The derivation of the conventional signals commonly used for representation of the changes in fluorescence spectra during protein denaturation: the ratio of fluorescence intensities at 350 nm and 330 nm (I_{350}/I_{330}), the barycentric mean of the spectrum (BCM, also referred to as the average emission wavelength), or integrated area of the spectrum. Middle: Comparison of the stability curves derived using the normalized single variables, and the normalized amplitude changes of the first three SVD components calculated from the dataset (corresponding to the SVD analysis shown in Figure 2). Right: The fraction of the states calculated from the global fit (blue lines in the middle panel) of the two-dimensional variables and the SVD components to the two- and three-state unfolding models, respectively. N, native; I, intermediate; D, denatured.

tend to extend the analytical scope of CalFitter by introducing models involving temperature-induced concentration-dependent aggregation and an entirely new module for analysis of chemical denaturation experiments. This will allow users to analyse the effects of various protein perturbants (e.g. solvents, additives, pH) on protein energetics in combination with temperature and extract valuable thermodynamic and kinetic parameters from multi-dimensional energy landscapes, which is particularly relevant for studying complex phenomena, e.g. cold denaturation. Another promising direction is an interactive model editor that will enable users to schematically draw any unfolding scenario, for which the software will automatically derive the underlying mathematical description and respective parameters. These changes will make CalFitter the ultimate one-stop shop for the analysis of protein stability.

DATA AVAILABILITY

CalFitter 2.0 is freely available at https://loschmidt.chemi.muni.cz/calfitter/. The datasets used for the case study and numerical validation are provided in the Supplementary data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Czech Ministry of Education, Youth and Sports [INBIO – CZ.02.1.01/0.0/0.0/16_026/0008451; ELIXIR – LM2018131; eINFRA – LM2018140; RECETOX research infrastructure LM2018121; Cetocoen Plus project – CZ.02.1.01/0.0/0.0/15_003/0000469; CETOCOEN EXCELLENCE project – CZ.02.1.01/0.0/0.0/17_043/0009632]; European Union's Horizon 2020 research and innovation programme [857560]; this publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains. Funding for open access charge: EU Horizon 2020 research and innovation programme [857560].
Conflict of interest statement. None declared.

REFERENCES

1. Chiti,F. and Dobson,C.M. (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.*, **86**, 27–68.
2. Bommarius,A.S. and Paye,M.F. (2013) Stabilizing biocatalysts. *Chem. Soc. Rev.*, **42**, 6534.
3. Nisthal,A., Wang,C.Y., Ary,M.L. and Mayo,S.L(2019) Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis*PNAS*, **116**, 16367–16377
4. Beerens,K., Mazurenko,S., Kunka,A., Marques,S.M., Hansen,N., Musil,M., Chaloupkova,R., Waterman,J., Brezovsky,J., Bednar,D. *et al.* (2018) Evolutionary analysis as a powerful complement to energy calculations for protein stabilization. *ACS Catal.*, **8**, 9420–9428.
5. Sanchez-Ruiz,J.M. (1992) Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys. J.*, **61**, 921–935.
6. Ibarra-Molero,B., Naganathan,A.N., Sanchez-Ruiz,J.M. and Muñoz,V. (2016) Modern analysis of protein folding by differential scanning calorimetry. In: *Methods in Enzymology*. Elsevier, Vol. **567**, pp. 281–318.
7. Stourac,J., Dubrava,J., Musil,M., Horackova,J., Damborsky,J., Mazurenko,S. and Bednar,D. (2021) FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res.*, **49**, D319–D324.
8. Musil,M., Konegger,H., Hon,J., Bednar,D. and Damborsky,J. (2019) Computational design of stable and soluble biocatalysts. *ACS Catal.*, **9**, 1033–1054.
9. Mazurenko,S., Stourac,J., Kunka,A., Nedeljković,S., Bednar,D., Prokop,Z. and Damborsky,J. (2018) CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res.*, **46**, W344–W349.
10. Nemergut,M., Zoldak,G., Schaefer,J.V., Kast,F., Miskovsky,P., Plückthun,A. and Sedlak,E. (2017) Analysis of IgG kinetic stability by differential scanning calorimetry, probe fluorescence and light scattering: kinetic stability analysis of igG. *Protein Sci.*, **26**, 2229–2239.
11. Mazurenko,S., Kunka,A., Beerens,K., Johnson,C.M., Damborsky,J. and Prokop,Z. (2017) Exploration of protein unfolding by modelling calorimetry data from reheating. *Sci. Rep.*, **7**, 16321.

12. Samaga,Y.B.L., Raghunathan,S. and Priyakumar,U.D. (2021) SCONES: self-consistent neural network for protein stability prediction upon mutation. *J. Phys. Chem. B*, **125**, 10657–10671.
13. Marques,S.M., Planas-Iglesias,J. and Damborsky,J. (2021) Web-based tools for computational enzyme design. *Curr. Opin. Struct. Biol.*, **69**, 19–34.
14. Markova,K., Kunka,A., Chmelova,K., Havlasek,M., Babkova,P., Marques,S.M., Vasina,M., Planas-Iglesias,J., Chaloupkova,R., Bednar,D. *et al.* (2021) Computational enzyme stabilization can affect folding energy landscapes and lead to catalytically enhanced domain-swapped dimers. *ACS Catalysis*, **11**, 12864–12885.
15. Óskarsson,K.R., Sævarsson,A.F. and Kristjánsson,M.M. (2020) Thermostabilization of VPR, a kinetically stable cold adapted subtilase, via multiple proline substitutions into surface loops. *Sci. Rep.*, **10**, 1045.
16. Valadares,V.S., Martins,L.C., Roman,E.A., Valente,A.P., Cino,E.A. and Moraes,A.H. (2021) Conformational dynamics of tetracenomycin aromatase/cyclase regulate polyketide binding and enzyme aggregation propensity. *Biochim. Biophys. Acta (BBA) - Gen. Subj.*, **1865**, 129949.
17. Rodrigues,J.V. and Shakhnovich,E.I. (2019) Adaptation to mutational inactivation of an essential gene converges to an accessible suboptimal fitness peak. *Elife*, **8**, e50509.
18. Zoldak,G., Jancura,D. and Sedlak,E. (2017) The fluorescence intensities ratio is not a reliable parameter for evaluation of protein unfolding transitions. *Protein Sci.*, **26**, 1236–1239.
19. Laptenok,S.P., Visser,N.V., Engel,R., Westphal,A.H., van Hoek,A., van Mierlo,C.P.M., van Stokkum,I.H.M., van Amerongen,H. and Visser,A.J.W.G. (2011) A general approach for detecting folding intermediates from steady-state and time-resolved fluorescence of single-tryptophan-containing proteins. *Biochemistry*, **50**, 3441–3450.
20. Wang,I., Chen,S.-Y. and Hsu,S.-T.D. (2016) Folding analysis of the most complex Stevedore's protein knot. *Sci. Rep.*, **6**, 31514.
21. Kim,T.W., Lee,S.J., Jo,J., Kim,J.G., Ki,H., Kim,C.W., Cho,K.H., Choi,J., Lee,J.H., Wulff,M. *et al.* (2020) Protein folding from heterogeneous unfolded state revealed by time-resolved X-ray solution scattering. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 14996–15005.
22. Henry,L., Panman,M.R., Isaksson,L., Claesson,E., Kosheleva,I., Henning,R., Westenhoff,S. and Berntsson,O. (2020) Real-time tracking of protein unfolding with time-resolved X-ray solution scattering. *Struct. Dyn.*, **7**, 054702.
23. Luan,B., Shan,B., Baiz,C., Tokmakoff,A. and Raleigh,D.P. (2013) Cooperative cold denaturation: the case of the C-terminal domain of ribosomal protein L9. *Biochemistry*, **52**, 2402–2409.
24. Dingfelder,F., Benke,S., Nettels,D. and Schuler,B. (2018) Mapping an equilibrium folding intermediate of the cytolytic pore toxin ClyA with single-molecule FRET. *J. Phys. Chem. B*, **122**, 11251–11261.
25. Liu,H., Yin,P., He,S., Sun,Z., Tao,Y., Huang,Y., Zhuang,H., Zhang,G. and Wei,S. (2010) ATP-Induced noncooperative thermal unfolding of hen lysozyme. *Biochem. Biophys. Res. Commun.*, **397**, 598–602.
26. Galantini,L., Leggio,C., Konarev,P.V. and Pavel,N.V. (2010) Human serum albumin binding ibuprofen: a 3D description of the unfolding pathway in urea. *Biophys. Chem.*, **147**, 111–122.
27. Harder,M.E., Deinzer,M.L., Leid,M.E. and Schimerlik,M.I. (2004) Global analysis of three-state protein unfolding data. *Protein Sci.*, **13**, 2207–2222.
28. Fotouhi,L., Yousefinejad,S., Salehi,N., Saboury,A.A., Sheibani,N. and Moosavi-Movahedi,A.A. (2015) Application of merged spectroscopic data combined with chemometric analysis for resolution of hemoglobin intermediates during chemical unfolding. *Spectrochim. Acta Part A*, **136**, 1974–1981.
29. Jaumot,J., Vives,M. and Gargallo,R. (2004) Application of multivariate resolution methods to the study of biochemical and biophysical processes. *Anal. Biochem.*, **327**, 1–13.
30. Enoki,S., Maki,K., Inobe,T., Takahashi,K., Kamagata,K., Oroguchi,T., Nakatani,H., Tomoyori,K. and Kuwajima,K. (2006) The equilibrium unfolding intermediate observed at pH 4 and its relationship with the kinetic folding intermediates in green fluorescent protein. *J. Mol. Biol.*, **361**, 969–982.
31. Curnow,P. and Booth,P.J.Combined kinetic and thermodynamic analysis of Alpha-helical membrane protein unfolding. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 18970–18975.
32. Segel,D.J., Bachmann,A., Hofrichter,J., Hodgson,K.O., Doniach,S. and Kiefhaber,T. (1999) Characterization of transient intermediates in lysozyme folding with time-resolved small-angle X-ray scattering. *J. Mol. Biol.*, **288**, 489–499.
33. Henry,E.R. and Hofrichter,J. (1992) Singular value decomposition: application to analysis of experimental data. In: *Methods in Enzymology*. Elsevier, Vol. **210**, pp. 129–192.

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

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Database Article

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

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

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

* Corresponding authors at: Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlarska 2, Brno 61137, Czech Republic.
E-mail addresses: 222755@mail.muni.cz (D. Bednar), mazurenko@mail.muni.cz (S. Mazurenko).

https://doi.org/10.1016/j.csbj.2022.11.009

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

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

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

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

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

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

2. Materials and methods

2.1. Data from small-scale experiments

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

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

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

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

2.2. Deep mutational scanning data

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

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

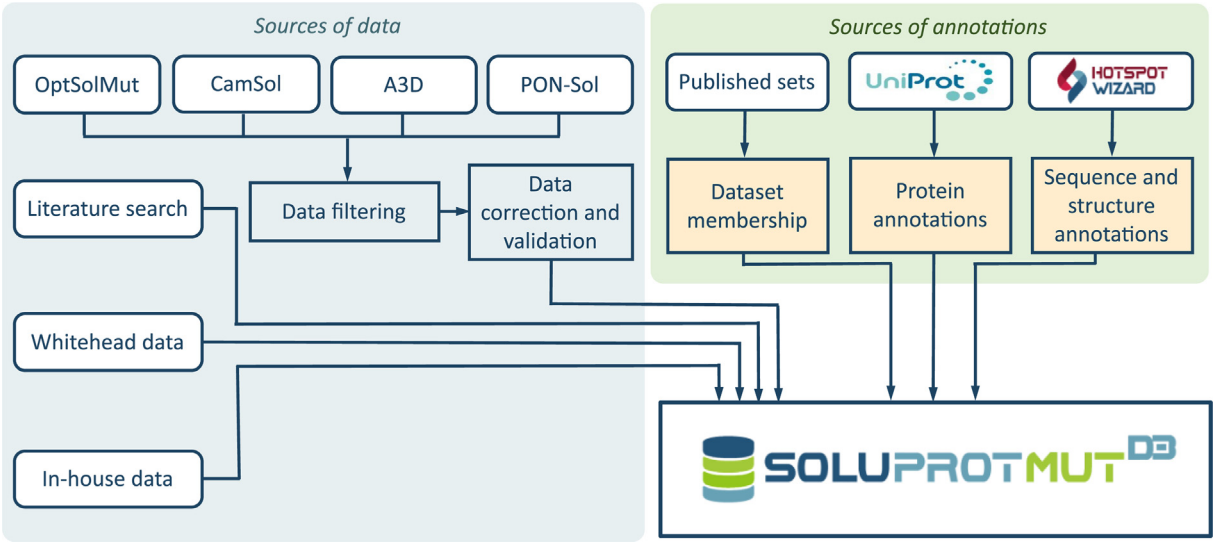


Fig. 1. The data sources of SoluProtMut^{DB} and their processing. The primary sources are the merged data from the earlier published datasets of protein-solubility predictors and the high-throughput data from Whitehead's group [12,13]. The datasets have been manually checked with the original publications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

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

2.3. In-house data

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

2.4. Systematization of values

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

We distinguish the orientation (positive, negative, or neutral) of an effect and, whenever applicable, also its significance (slight or significant). Altogether, up to five discrete values are defined: *significantly/slightly deteriorating*, *neutral*, and *slightly/significantly enhancing*. This system suggests different resolutions in different

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

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

experiments, e.g., a value from the 5-value system should be more precise than from the 3-value system. Hence, if one mutation is enhancing in the 3-value system and another is slightly enhancing in the 5-value system, we can assume the former to be at least as enhancing as the latter, and possibly substantially more.

2.5. Annotations

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

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

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

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

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

2.6. Database structure

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

3. Results

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

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

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

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

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

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

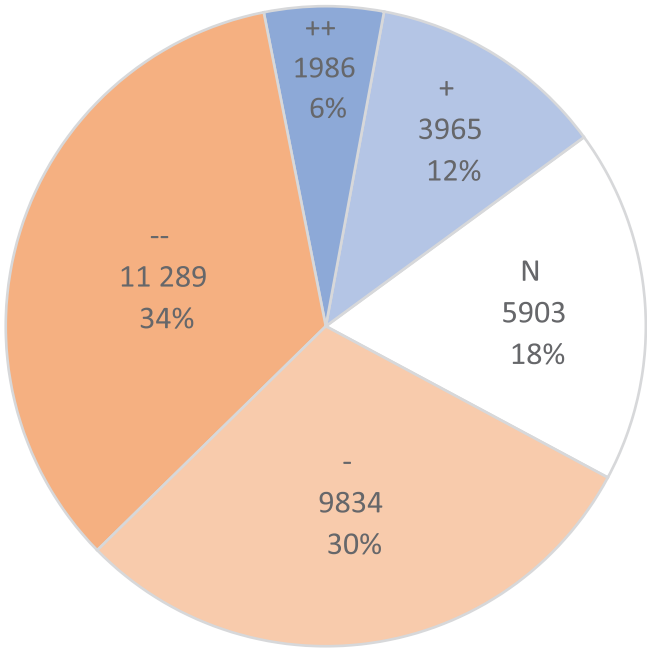


Fig. 2. The distribution of protein variants in the database by their mutational effects on solubility. The distribution is divided into 5 levels: neutral (N), slightly/significantly desolubilizing (−/−−) and solubilizing (**+/++**). Notably, two thirds of the mutants show a deteriorating effect.

3.1. Interface

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

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

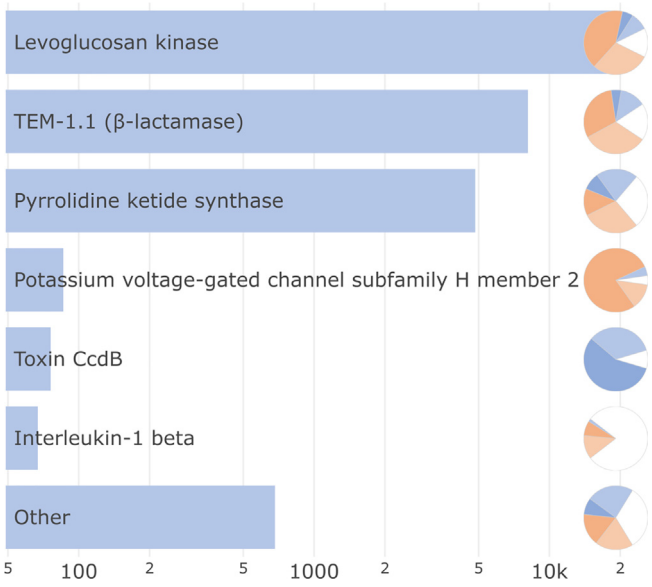


Fig. 3. The six most represented proteins in the database by their entry counts. The data for the first three proteins come from deep mutational scanning experiments. The 'Other' category contains the remaining 97 proteins. The horizontal axis has a logarithmic scale, and the pie charts on the right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

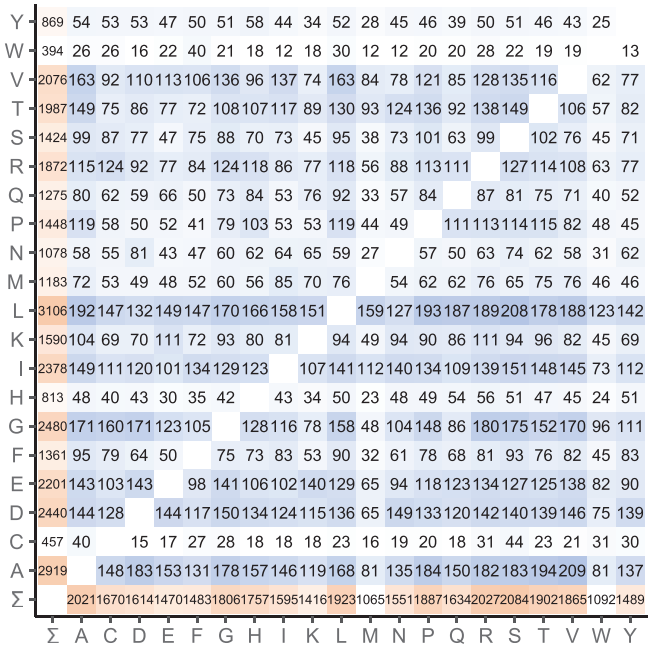


Fig. 4. A matrix showing the numbers of mutation occurrences in the database 'from' (rows) and 'to' (columns) specific amino acids. The Σ column and row represent sums of mutations 'from' and 'to' given amino acids, respectively. A cell color saturation shows the abundance of the corresponding combination.

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

3.2. Data export

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

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

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

4. Discussion

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

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

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

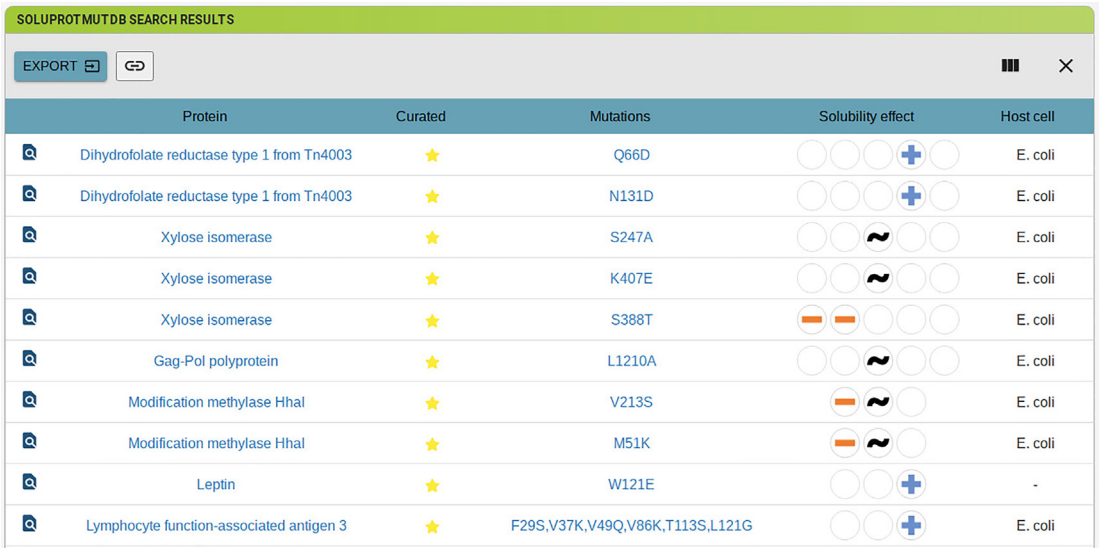


Fig. 5. An example of a result table. For clarity, only the most important columns are displayed by default: protein names, curation flags, mutations, solubility effects, and host cells. The table is paginated to avoid performance issues. A solubility effect graphic depicts both an effect and a value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

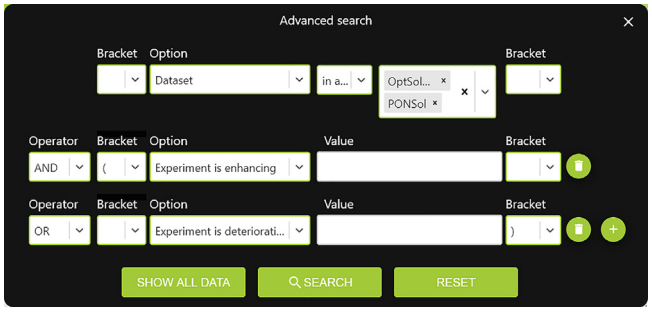


Fig. 6. The advanced search with an example of a filtering protocol. In this example, the database will find measurements from OptSolMut and PON-Sol datasets with enhancing or deteriorating solubility effect.

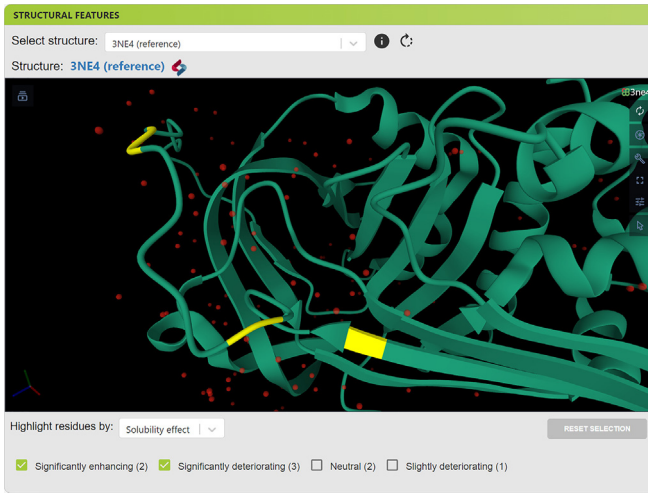


Fig. 7. The visualization of mutations in a protein with a known 3D structure. User-selected mutations can be highlighted in the structure. In this example, the mutated positions resulting in a significant change in solubility are highlighted in yellow.

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

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

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

CRediT authorship contribution statement

Jan Velecký: Software, Visualization, Data curation, Writing – original draft, Writing – review & editing. **Marie Hamsikova:** Soft-

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

Declaration of Competing Interest

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

Acknowledgments

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

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

The authors thank the anonymous reviewers for their valuable suggestions.

Appendix A. Supportive information

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

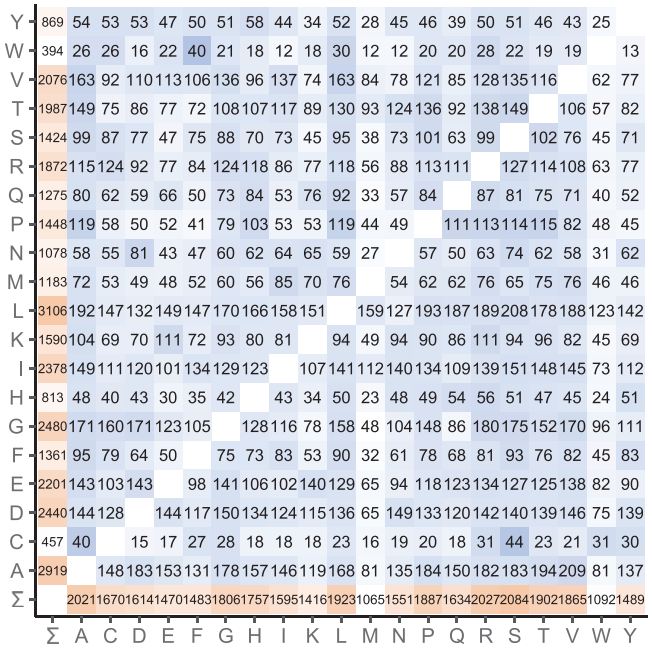


Fig. A.1. A matrix showing the numbers of mutation occurrences in the database 'from' (rows) and 'to' (columns) specific amino acids. The Σ column and row represent the total numbers of mutations 'from' and 'to' given amino acids, respectively. The matrix is *row-weighted* – blue saturation corresponds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

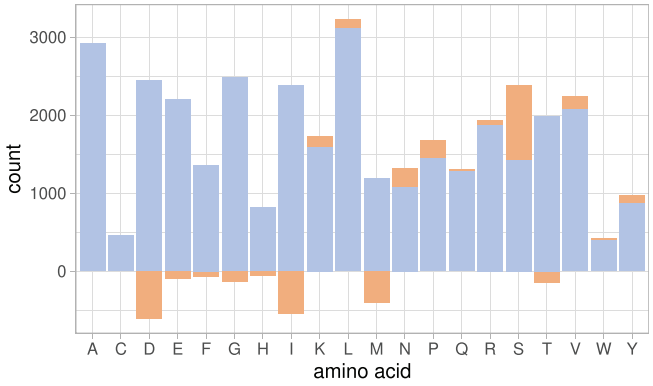


Fig. A.2. The histogram of the substituted amino acids in the database. The red bar shows the deviation from the natural distribution of amino acids, as in sequences of all kingdoms of life [39].

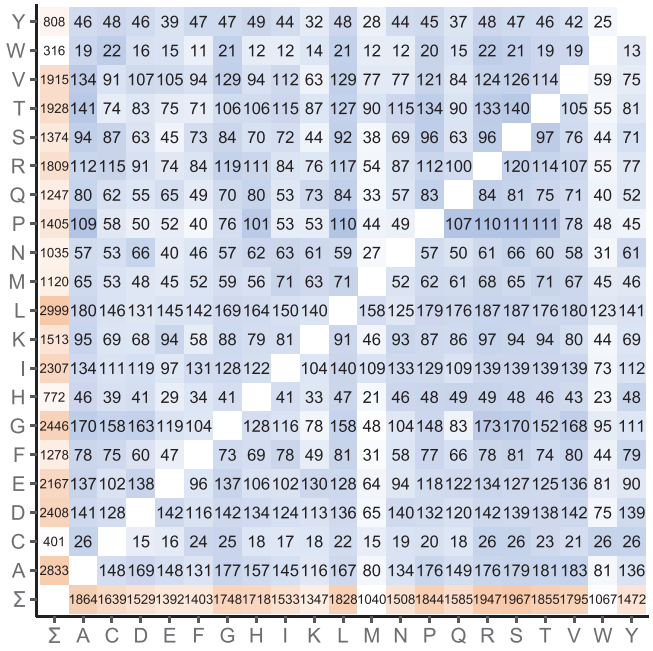


Fig. A.3. A row-weighted substitution matrix for Whitehead's data. It shows some anomalies, such as visible under-representation of substitutions to methionine (M) or tryptophan (W).

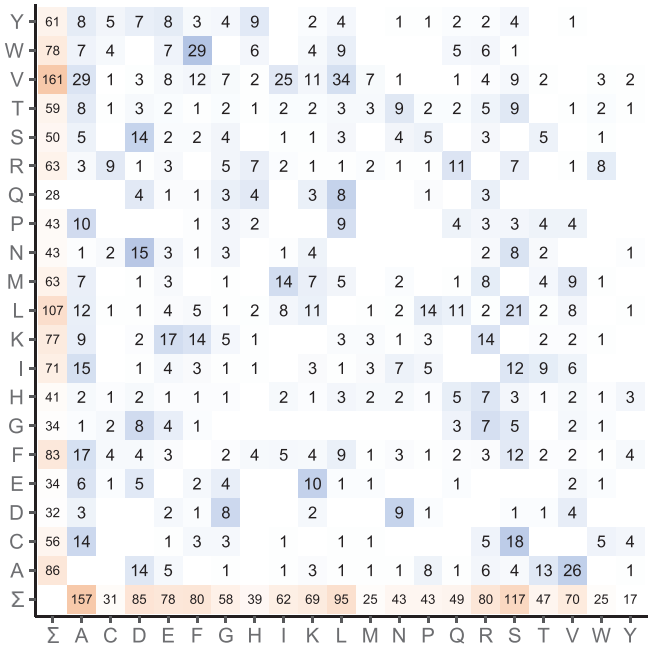


Fig. A.4. A row-weighted substitution matrix for all but Whitehead's data. It shows the selection bias in the small-scale experiments. For example, alanine (A) or serine (S) is chosen as a substituent more frequently than other amino acids. Some of the biases are apparently due to avoidance of introducing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

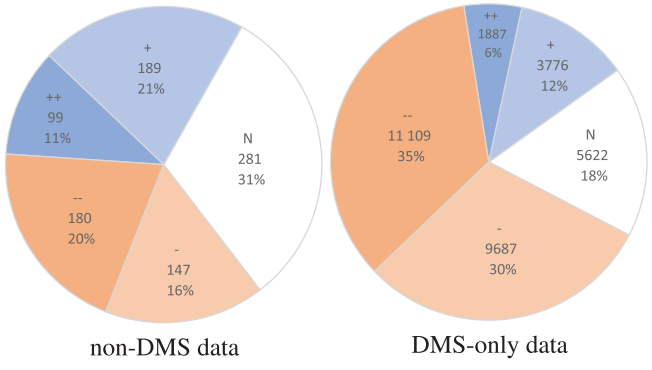


Fig. A.5. A comparison between the distributions of effects in the non-DMS and DMS-only datasets. The latter is skewed towards mutations having desolubilizing effect.

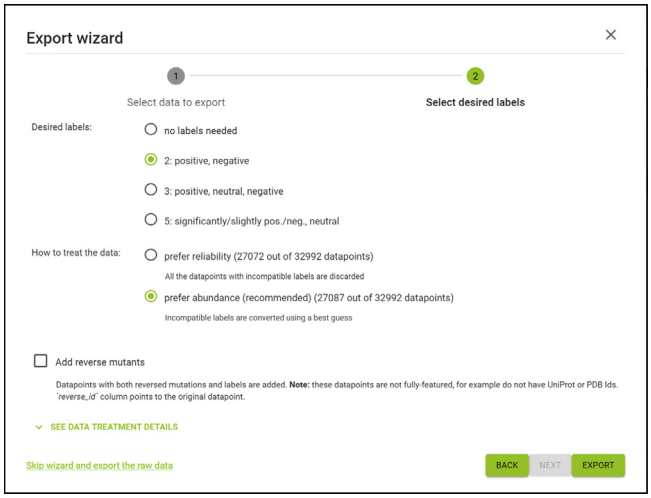


Fig. A.6. An example of the 2nd step of Export Wizard. Here, the solubility effect of all selected entries will be converted into the 2-value system using a best guess, and datapoints will be exported into a CSV file upon clicking on 'Export'. There is also an option to skip the wizard and export the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

References

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

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

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

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

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

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

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

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

[9] Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3d (a3d): server for prediction of aggregation properties of protein structures. Nucleic Acids Res 2015;43(W1):W306–13. https://doi.org/10.1093/nar/gkv359.

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

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

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

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

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

[15] T.U. Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res 49 (D1) (2020) D480–D489. doi:10.1093/nar/gkaa1100. URL https://doi.org/10.1093/nar/gkaa1100.
[16] Sumbalova L., Stourac J., Martinek T., Bednar D., Damborsky J. HotSpot wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information, Nucleic Acids Res 46 (W1) (2018) W356–W362. https://doi.org/10.1093/nar/gky417.
[17] Kaur J, Kumar A, Kaur J. Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. Int J Biol Macromol 2018;106:803–22. https://doi.org/10.1016/j.ijbiomac.2017.08.080.
[18] Slanská K. Study of protein solubility [online] Master's thesis, Faculty of Science, Masaryk University, Brno (2021). URL Availableat<https://is.muni.cz/th/e3jlf/>
[19] Bendl J., Stourac J., Sebestova E., Vavra O., Musil M., Brezovsky J., Damborsky J. HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering, Nucleic Acids Res 44 (Web Server issue) (2016) W479–W487. doi:10.1093/nar/gkw416. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987947/.
[20] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
[21] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics (Oxford, England) 2015;31(6):926–32. https://doi.org/10.1093/bioinformatics/btu739.
[22] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 2010;26(19):2460–1. https://doi.org/10.1093/bioinformatics/btq461.
[23] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539. https://doi.org/10.1038/msb.2011.75.
[24] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics (Oxford, England) 2007;23(15):1875–82. https://doi.org/10.1093/bioinformatics/btm270.
[25] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.
[26] Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79(2):351–71. https://doi.org/10.1016/0022-2836(73)90011-9.
[27] Reetz M.T., Carballeira J.D., Vogel A. Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability, Angewandte Chem Int Ed 45(46) (2006) 7745–7751, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200602795. doi:10.1002/anie.200602795. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200602795.
[28] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinform 2009;10:168. https://doi.org/10.1186/1471-2105-10-168.
[29] Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, Gora A, Sustr V, Klvana M, Medek P, Biedermannova L, Sochor J, Damborsky J. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 2012;8(10):. https://doi.org/10.1371/journal.pcbi.1002708e1002708.
[30] Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin M-J, Kleywegt GJ. SIFTS: Structure integration with function, taxonomy and sequences resource. Nucleic Acids Res 2012;41(D1):D483–9. https://doi.org/10.1093/nar/gks1258.
[31] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Sci Data 3(1) (Mar. 2016). doi:10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.
[32] Watkins X, Garcia LJ, Pundir S, Martin MJ. the UniProt Consortium, Protvista: visualization of protein sequence annotations. Bioinformatics 2017;33(13):2040–1. https://doi.org/10.1093/bioinformatics/btx120.
[33] Sehnal D., Bittrich S., Deshpande M., Svobodova R., Berka K., Bazgier V., Velankar S., Burley S.K., Koca J., Rose A.S. Mol* viewer: modern web app for 3d visualization and analysis of large biomolecular structures, Nucleic Acids Res 49(W1) (2021) W431–W437. https://doi.org/10.1093/nar/gkab314.
[34] Pucci F, Schwersensky M, Rooman M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr Opin Struct Biol 2022;72:161–8. https://doi.org/10.1016/j.sbi.2021.11.001. URL https://www.sciencedirect.com/science/article/pii/S0959440X21001445.
[35] Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Briefings Bioinform 2020;21(4):1285–92. https://doi.org/10.1093/bib/bbz071.
[36] Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotechnol J 2020;18:1968–79. https://doi.org/10.1016/j.csbj.2020.07.011.
[37] Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. Trends Biotechnol 2004;22(7):346–53. https://doi.org/10.1016/j.tibtech.2004.04.006. URL https://www.sciencedirect.com/science/article/pii/S0167779904001118.
[38] Kuroda Y. Biophysical studies of protein solubility and amorphous aggregation by systematic mutational analysis and a helical polymerization model. Biophys Rev 2018;10(2):473–80. https://doi.org/10.1007/s12551-017-0342-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5899702/.
[39] Kozlowski LP. Proteome-pI: proteome isoelectric point database. Nucleic Acids Res 2017;45(D1):D1112–6. https://doi.org/10.1093/nar/gkw978.

Journal of Cheminformatics

RESEARCH

Open Access

A computational workflow for analysis of missense mutations in precision oncology

Rayyan Tariq Khan[1,3], Petra Pokorna[5,7], Jan Stourac[1,2,3], Simeon Borko[2,3,4], Ihor Arefiev[1,2], Joan Planas-Iglesias[1,2,3], Adam Dobias[1,2], Gaspar Pinto[1,2,3], Veronika Szotkowska[1,2], Jaroslav Sterba[6], Ondrej Slaby[5,7], Jiri Damborsky[1,2,3], Stanislav Mazurenko[1,2,3*] and David Bednar[1,2,3*]

Abstract

Every year, more than 19 million cancer cases are diagnosed, and this number continues to increase annually. Since standard treatment options have varying success rates for different types of cancer, understanding the biology of an individual's tumour becomes crucial, especially for cases that are difficult to treat. Personalised high-throughput profiling, using next-generation sequencing, allows for a comprehensive examination of biopsy specimens. Furthermore, the widespread use of this technology has generated a wealth of information on cancer-specific gene alterations. However, there exists a significant gap between identified alterations and their proven impact on protein function. Here, we present a bioinformatics pipeline that enables fast analysis of a missense mutation's effect on stability and function in known oncogenic proteins. This pipeline is coupled with a predictor that summarises the outputs of different tools used throughout the pipeline, providing a single probability score, achieving a balanced accuracy above 86%. The pipeline incorporates a virtual screening method to suggest potential FDA/EMA-approved drugs to be considered for treatment. We showcase three case studies to demonstrate the timely utility of this pipeline. To facilitate access and analysis of cancer-related mutations, we have packaged the pipeline as a web server, which is freely available at https://loschmidt.chemi.muni.cz/predictonco/.

Scientific contribution

This work presents a novel bioinformatics pipeline that integrates multiple computational tools to predict the effects of missense mutations on proteins of oncological interest. The pipeline uniquely combines fast protein modelling, stability prediction, and evolutionary analysis with virtual drug screening, while offering actionable insights for precision oncology. This comprehensive approach surpasses existing tools by automating the interpretation of mutations and suggesting potential treatments, thereby striving to bridge the gap between sequencing data and clinical application.

Keywords Bioinformatics, Cancer, Function, High-performance computing, Machine learning, Molecular modelling, Oncology, Personalised medicine, Single nucleotide polymorphism, Stability, Treatment

*Correspondence:
Stanislav Mazurenko
mazurenko@mail.muni.cz
David Bednar
davidbednar1208@gmail.com
Full list of author information is available at the end of the article

Introduction

More than 19 million cancer cases were diagnosed in 2020 [10] with a projected load of 28.4 million cases in 2040 [44]. The three traditionally used approaches to treat cancer, namely chemotherapy, surgery, and radiotherapy, generally result in higher mortality rates compared to the less adopted precision medicine-based techniques [27]. Next-generation sequencing technologies form the basis of precision oncology and can help generate a large amount of transcriptomic and genomic data. On the other hand, these technologies often do not provide clinically actionable data. This leads to a divide between generation of the said data and their utility, as mutants with unknown effects are often found during clinical testing [9].

There are not many tools that can help bridge the gap between data generation and creation of actionable insights. Swiss-PO, an online tool, allows for mapping experimentally determined mutations on a curated list of 50 genes and their various associated 3D structures. It also allows users to visualise multiple molecular interactions; however, it leaves it to the user to intuitively assess the structural implications of mutations that have not been experimentally determined [25] and it can also not predict patient survival outcomes. PSnpBind, a database, catalogues changes to binding affinities of ligands due to binding site single-nucleotide polymorphisms (SNPs), however this database is limited to 26 human proteins and is limited to interactions between ligands and binding site residues [2]. We sought to overcome some of these limitations by creating a robust pipeline that can predict the effects of missense mutations, even for ones which are not experimentally determined, on cancer-related proteins.

The pipeline relies on advances in fast protein modelling, such as AlphaFold [23], prediction of the effect of missense mutations on a protein structure [4], and protein stability prediction [5, 24]. This allows harvesting much more information from mutations identified by exome sequencing, which can then be used for actionable decision making. Additionally, coupling fast ligand docking in proteins [48] with the availability of multiple drug libraries online, such as ZINC [20], it is possible to screen novel potential inhibitors for the mutated proteins.

As the interpretation of large-scale genomic and transcriptomic data is limited due to the need to utilise multiple computational tools, performing the aforementioned analysis on exome sequences can take time if done manually. After a cancer diagnosis, treatment is generally a race against time, and with the variable success rates of conventional "one size fits all" therapies, fast and accurate interpretation of molecular findings and assessment of their actionability are of vital importance, especially in difficult-to-treat cases. This is where an automated precision oncology approach will be most useful as it can optimise treatment strategies, improve outcomes, and increase the quality of life for many patients [30].

Here we introduce a bioinformatics pipeline for the analysis of the effect of mutations on stability and function in cancer-related proteins. The pipeline applies in silico methods of molecular modelling, structural bioinformatics, and machine learning, and outputs actionable data which can be used for decision making. The coupled predictor produces a decision on the oncogenicity of the protein mutation by utilising the outputs derived at various stages of the pipeline. Moreover, we show the application of the pipeline on three use case studies and highlight the importance of advanced bioinformatics in precision oncology.

Methodology

Manual curation, structure repairs and geometry optimization

A list of 44 cancer-related proteins (including one isoform of a selected protein) were chosen as targets for the manual curation. The selection was based on the importance of the respective proteins for cancer diagnostics and, notably, in cancer treatment. The vast majority of curated proteins are either direct targets of therapeutic agents or, despite not being targets themselves, represent established predictive biomarkers for administering targeted treatments aimed at downstream members of the same pathway. Additionally, we included proteins that are frequently altered across various cancer types and are relevant to both diagnostics and cancer research (e.g., p53). The proteins with their various annotations are listed in the Supplementary material SI 1.

The 44 protein sequences and their annotations were fetched from the UniProt database [47]. In the case of KRAS, two isoforms are provided, including the canonical isoform and an isoform that is commonly utilized accross clinical databases of genetic variants. The essential residues were re-confirmed in the literature as well as in the Mechanism and Catalytic Site Atlas (M-CSA) [38] and the SWISS-PROT [6] databases. For the purposes of this study, in the case of multi-domain proteins, only the catalytic cytoplasmic domains of the proteins were considered. The best available structure from the wwPDB database [51], the ideal biological assembly, as well as the relevant chain (in multimeric structures) were selected based on resolution and missing parts. Canonical co-factors for structures were established using the UniProt database; these were retained in the structure, and all other ligands, ions, and water molecules were removed from the structure (SI 1). The residue indexes were mapped using the SIFTS

database [13]. After a visual inspection of each target protein, the following four key problematic regions/positions were identified: (i) missing regions, i.e., low resolution regions in the crystal structure,(ii) long, missing, and/or intrinsically disordered regions not influencing the catalytic site of the protein; (iii) missing atoms in the side chain; (iv) amino acids requiring identity correction, i.e., the sequence in the 3D structure did not correspond to that recorded in UniProt.

Each protein structure that required any of these structural improvements (for the aforementioned problematic regions/positions i, iii, or iv) was modelled using MODELLER version 9.24, 2020/04/06, r11614 [16]. The modelling was guided by the UniProt-PDB alignment provided by SIFTS. Regions identified as intrinsically disordered (repair ii) were omitted from the modelling. Custom extensions of three MODELLER Python classes (Environment, Model, and AutoModel) were developed to ensure the following: (i) the produced models incorporated any relevant co-factor from the template, (ii) the produced models were not optimised on the regions that did not require repairs, and (iii) structures containing multiple chains could be modelled and minimised at once. If no experimental structure was available, the AlphaFold database [23] was searched. The mutant structure was generated by introducing the desired mutation in the target wild type structure by MODELLER, and it was guided by a trivial alignment between the wild type and the mutant sequences.

For each protein structure, inconsistent torsion angles, total energy, or Van der Waals clashes were reduced using RepairPDB feature of FoldX 4.0 [5]. Then minimization of structures was performed in Rosetta 3.11-static [24] with constraints using the Talaris2014 force field [33]. The wild type and mutant structures were then aligned using DeepAlign 1.135-2-foss-2018b [22] to ensure that their coordinates match for further analysis.

Protein stability prediction

The impact of the missense mutation on the stability of the protein structure was calculated using Rosetta and FoldX. For FoldX the PssmStability command was used, water molecules were only taken from the 'crystal', pH was set to 7, and the number of runs was set to 5. Rosetta calculations were made on the minimised structures using the ddg_monomer command, following protocol 3 [24], for which the extent of sidechain repacking was set to within 8 Å while using the soft-rep energy function and the Talaris2014 force field.

Protein function prediction, phylogenetic analysis, and consensus classification

Additionally, PropKa 3.4.0 [40] was used to predict the impact of the mutation on the pKA values of the proteins, using the propka3 command. Homologous sequences with sufficient identity (more than 50%) and coverage (± 20% of the query sequence), i.e., UniRef50 sequences, were downloaded from the UniRef database [45], and multiple sequence alignment were generated using Clustal-Omega [42] tool from the EMBL-EBI web server [32]. This was used for conservation analysis using Jensen-Shannon Divergence algorithm [11] and transformed to mutability grades by using HotSpot Wizard [43] thresholding. The mutations were also submitted to the HOPE [49] web server to collect information from a multitude of information sources, including calculations on the 3D coordinates of the protein, sequence annotations from the UniProt database, and predictions by DAS (Distributed Annotation System services [37]. Furthermore, PredictSNP [4] was used to predict the effect of the amino acid substitution on the target protein function through consensus classification.

Pocket analysis and virtual screening

Potential binding pockets within the structures of the analysed proteins were calculated using the prank predict command in P2Rank 2.3 [26], the resulting pockets were visually analysed and manually optimised to cover the entire binding sites. Selected pockets were listed in SI 2 according to their colour codes.

Virtual screening was performed on both the wild type and the mutant protein structure. A set of 4380 small molecules that were approved by the Food and Drug Administration and European Medicines Agency was taken from the ZINC database [20]. AutoDock Vina 1.1.2 [48] was run using the standard vina command, within a parameterized grid within each protein. The grid coordinates (SI 1) were created by visually placing the grid on the protein structure in PyMOL using the ADPlugin [41] and ensuring that the binding pockets with essential residues were completely within the grid. The values for the binding energy of each small molecule to a wild type structure as well as its mutant structure were used to calculate the impact of the mutation on the binding energy.

Machine learning predictor development

The predictive part of the pipeline is a machine-learning based tool that was trained on 1073 single-point mutants whose effect was classified as Oncogenic or Benign. The variants for the Benign class were selected from the gnomAD and ClinVar [29] databases. Variants with > 1% population frequency in gnomAD, variants annotated as

"benign" or "likely benign" in the ClinVar database, and variants without ClinVar annotation, for which the classification as "benign" or "likely benign" is at the same time supported by applicable ACMG criteria [39], were utilised. The variants for the Oncogenic class were collected in expert-curated precision oncology knowledge bases, mainly, but not limited to, precision oncology knowledge base OncoKB by Memorial Sloan Kettering Cancer Center [12], as well as The JAX Clinical Knowledgebase by The Jackson Laboratory [35], Personalized Cancer Therapy Knowledge Base by MD Anderson Cancer Center [28], cBioPortal [18], and the DoCM database [1]. Variants with conflicting interpretations across multiple sources were not included in the list. Both subsets were manually filtered for any possible overlaps with the datasets used in the PredictSNP consensus predictor and its constituents.

The entire dataset (SEQ: 509 oncogenic and 564 benign data points) was further annotated by the pipeline of PredictONCO. The following six features were calculated regardless of the structural information available: essentiality of the mutated residue (yes-1/no-0), the conservation of the position (the conservation grade and MSA score), the domain where the mutation is located ("cytoplasmic", "extracellular", "transmembrane", "other"-one-hot encoded), the PredictSNP score, and the number of essential residues in the protein. For approximately half of the data (STR: 377 oncogenic and 76 benign data points), the structural information was available, and six more features were calculated: FoldX and Rosetta ddg_monomer scores, whether the residue is in the ligand-binding pocket obtained from P2Rank (yes-1/no-0), and the pKa changes of essential residues obtained from PROPKA3. The dataset is available at https://zenodo.org/records/10013764.

For the training protocol, 20% of the data in each of the two sets was kept aside for testing, chosen randomly but grouped by positions to ensure that no specific position in a protein from the test set appears in the training set. The following types of predictors were tested: the support vector machine (SVM), decision tree (DT), and XGBoost classifier (XGB), taken as they are implemented in the scikit-learn 1.2.0 and xgboost 1.7.3 libraries for Python 3.8.15. We also used the PredictSNP score alone as a baseline. For each method, we tested a set of hyperparameters based on 5-fold cross-validation implemented on the training data and receiver operating characteristic (ROC) area under the curve (AUC) as the metric (Table S1 in SI 3).

The final evaluation consisted of constructing the ROC and Precision-Recall curves. Furthermore, a round of 100 random-state re-initialisations with different random seeds was performed to evaluate the robustness of the final models. For the area under the ROC curve and the average precision values, we also reported the average and standard deviation obtained by bootstrapping (N=1000). Since any change to the predictor or data split results in a different set of x-axis coordinates in the ROC and Precision-Recall curves, we used a fixed grid of 30 points and applied 1D linear interpolation to obtain the y-axis value for each iteration. These values were then plotted as 10% and 90% quantiles.

All the training scripts, the model files, and the scripts for reproducing the model evaluations are available at https://github.com/loschmidt/predictonco-predictor/. The versions of the software tools and Python packages that were used are provided in SI 4.

Results

Development of a fully automated computational workflow

We created a bioinformatics pipeline for structure and sequence based analysis of the effects of missense mutations on cancer-related proteins (Figure S1). Since the pipeline requires curated protein structures, a method for curation was developed and applied to a list of 44 proteins (SI 1), which were then tested to ensure they can be handled in the pipeline. The pipeline was assembled using multiple bioinformatics tools, databases, and techniques. Figure 1 represents a schematic outline of the pipeline, the output of which ultimately feeds into the machine learning predictor. The predictor gives a binary decision on the effect of mutation with confidence score which is helpful in the summation and comprehension of results. Three cases of oncological interest were then studied using the developed method.

Training of sequence-based and structure-based machine learning predictors

Initially, we trained three different types of predictors, covering different trade-offs between explainability and flexibility, and compared their performance with the baseline model using the PredictSNP score alone. After optimising the hyperparameters (Table S1 in SI 3), we evaluated the performance on the held-out 20% of the dataset split by position in a protein. The support vector machines and XGBoost classifiers showed superior yet similar performance based on the area under the ROC curve and the average precision from the Precision-Recall curve (Fig. 2), also confirmed statistically (Figure S2 in SI 3). We selected the XGBoost predictor for the final model due to the interpretability of its scores: the SVM model evaluation is based on the signed distance to the separating hyperplane, without intuitive interpretation. On the other hand, the XGBoost classifier directly returns the probability that

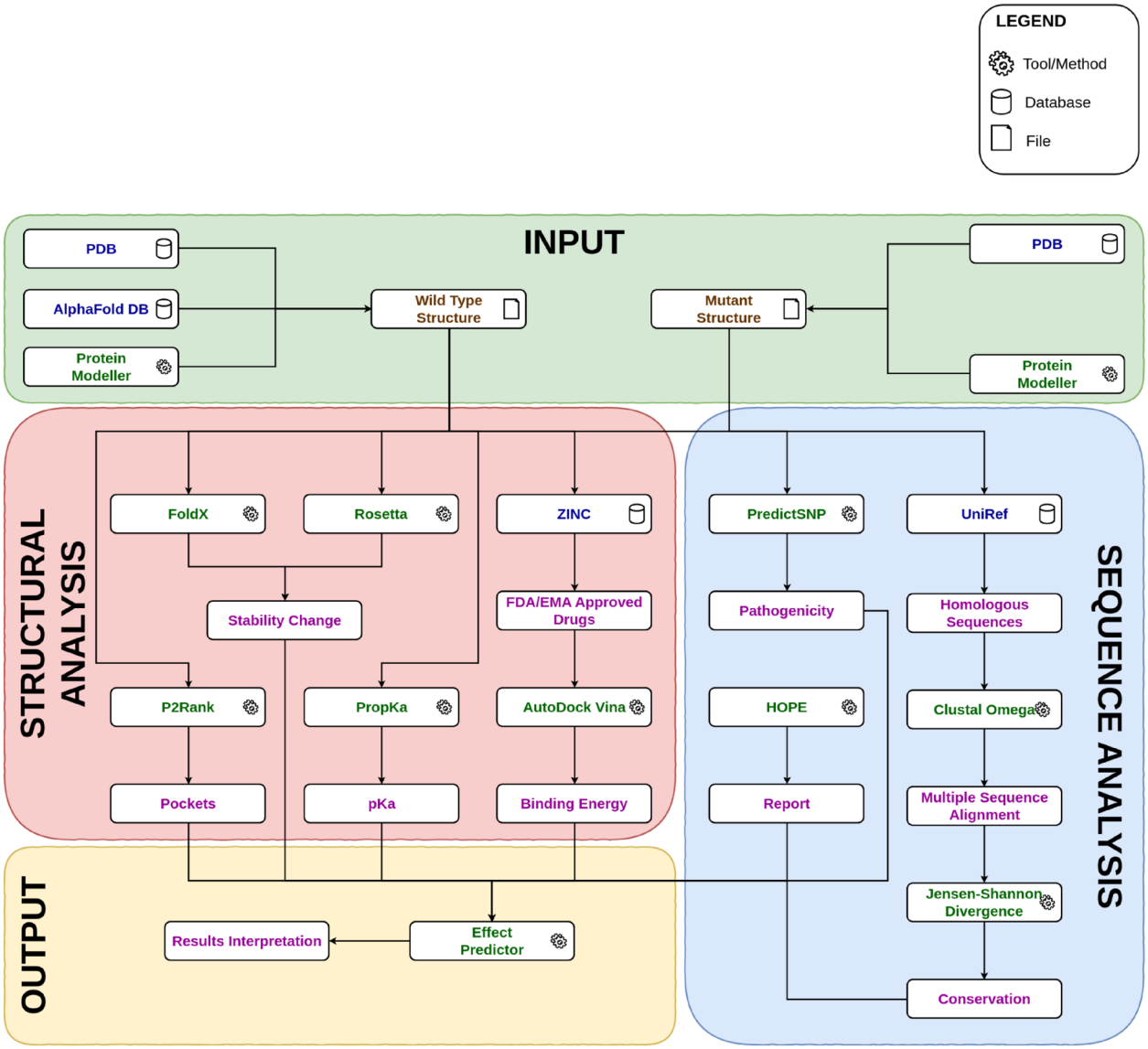


Fig. 1 A schematic representation of the bioinformatic pipeline used to predict the effect of a missense mutation on the oncogenicity of the protein

a particular mutation is oncogenic. The final XGBoost predictor is made up of 15 and 9 decision trees of the depth of 1 for structure and sequence data sets, respectively. The feature importance scores revealed that the PredictSNP score and conservation had the highest information gains (Figure S3 in SI 3). We also tested if using the train/test split by proteins would compromise the performance and saw only a marginal decrease (Figure S4 in SI 3), indicating the significant potential of the pipeline for other protein targets. The balanced accuracy for the sequence-based XGBoost predictor is 87%, and for the structure-based XGBoost predictor is 90%.

We also compared the performance of our predictor on the test set against several other models (Table 1). We evaluated the following individual scores as baselines: conservation, PredictSNP, FoldX, and Rosetta. In addition, we evaluated the performance of the ESM variants model, a recently published workflow based on the 650-million-parameter protein language ESM1b, which was used to score all possible missense variant effects in the human genome [8]. In both settings (SEQ and STR), PredictONCO showed superior performance.

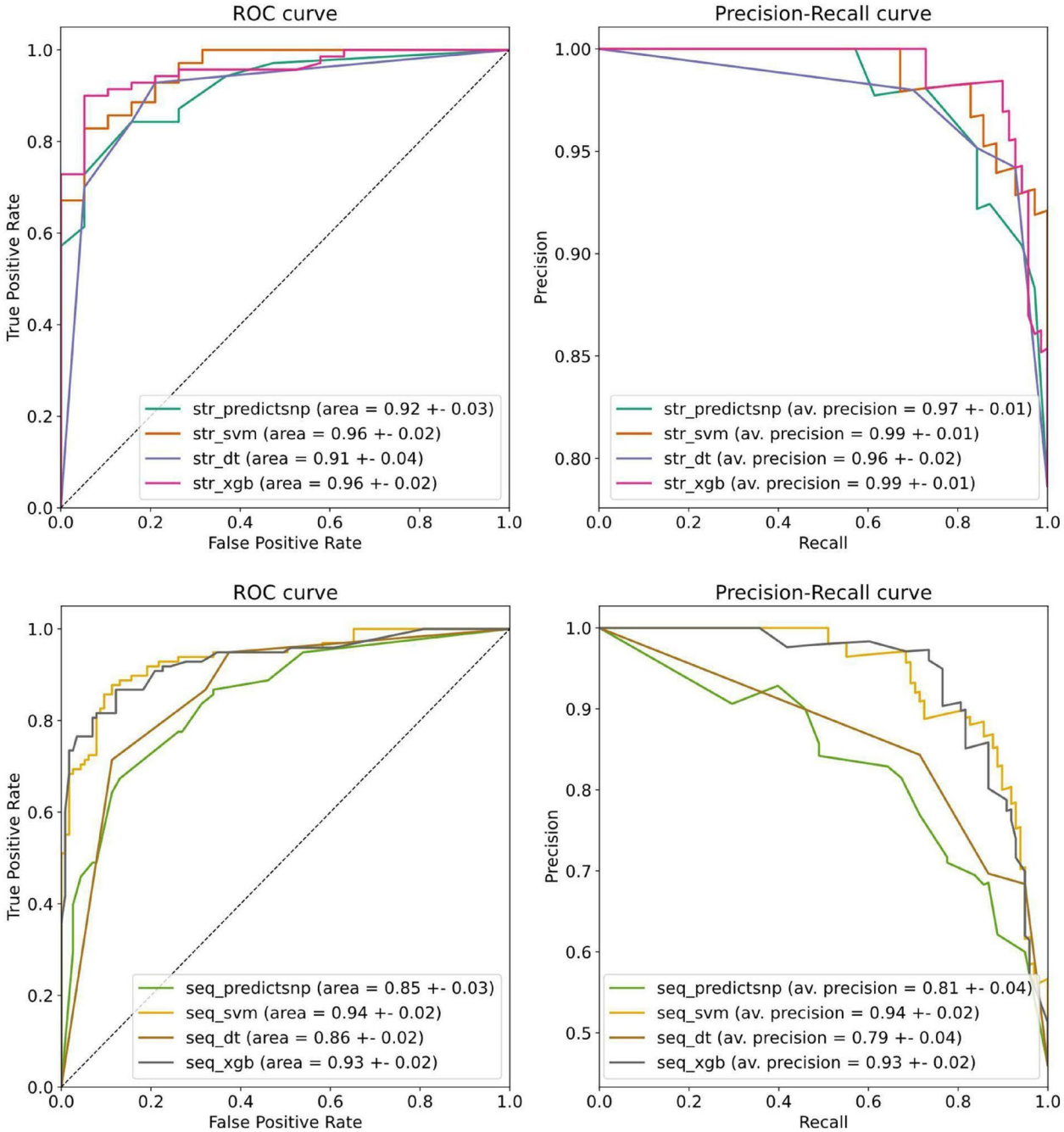


Fig. 2 The Receiver Operating Characteristic and Precision-Recall curves based on held-out test sets. Top: classifiers trained on the dataset with the structural features available (STR). Bottom: classifiers trained on the dataset with the sequence-only features (SEQ). Both the support vector machine (SVM) and XGBoost (XGB) showed comparable performance superior to the baseline model and decision tree (DT). The reported errors are standard deviations obtained by bootstrapping (N = 1000). The PredictSNP score was used as the baseline

Case studies with selected *cancer*-associated proteins

The following case studies demonstrate scenarios in which the tool has helped to facilitate further clinical decision-making. The respective variants featured in the case studies were identified across research projects utilizing high-throughput DNA sequencing techniques, which were conducted by the co-authors of this manuscript.

Table 1 Comparison of PredictONCO with other models on the test set

	Predictor	ROC AUC↑	Avg. Precision↑
SEQ	**PredictONCO**	**0.932 ± 0.018**	**0.934 ± 0.018**
	conservation	0.872 ± 0.026	0.802 ± 0.042
	predictSNP	0.845 ± 0.027	0.808 ± 0.041
	ESM variants	0.923 ± 0.018	0.911 ± 0.023
STR	**PredictONCO**	**0.955 ± 0.020**	**0.988 ± 0.006**
	FoldX	0.575 ± 0.064	0.867 ± 0.037
	Rosetta	0.628 ± 0.064	0.876 ± 0.039
	conservation	0.937 ± 0.037	0.970 ± 0.020
	predictSNP	0.918 ± 0.030	0.973 ± 0.011
	ESM variants	0.929 ± 0.027	0.981 ± 0.009

PredictONCO values are in bold

The models selected for comparison were individual features and the ESM variants predictor. The reported errors are standard deviations obtained by bootstrapping (N = 1000).

Case study 1-platelet derived growth factor receptor *beta* PDGFRB N666T

In a patient with myofibroma, sequencing analysis revealed an N666T variant of the PDGFRB protein (UniProt ID: P09619). Even though some mutations of the N666 residue, including N666K [21], N666H [36], or N666S [34], have already been documented in myofibroma patients, N666T, in particular, lacks published functional evidence and was reported in a total of one patient in combination with another mutation. Therefore, a comprehensive assessment of its effect would provide further confirmatory evidence on the variant's pathogenicity, which is substantial, given the therapeutic implications of receptor tyrosine kinase inhibition. Conservation status showed high evolutionary conservation of mutated position. For amino acid 826, one of the essential catalytic residues, a large increase in dissociation constant was predicted, suggesting a significant functional impact. Both stability predictors suggested a deleterious effect, which is also in agreement with the deleterious effect on protein function predicted by PredictSNP. Given all this data, the oncogenic effect was predicted by the XGBoost classifier with 100% confidence. Furthermore, in virtual screening, Sunitinib showed a slightly better increase in binding affinity compared to Imatinib, which was used as a drug of choice in different myofibroma preclinical studies, making Sunitinib a suitable alternative option for therapeutic planning. The full report can be accessed at https://loschmidt.chemi.muni.cz/predictonco/job/pdgfrb_N666T

Case study 2-angiopoietin-1 receptor TIE2 G1036D

In a patient with a vascular tumour, sequencing analysis revealed a G1036D variant in the TIE2 (UniProt ID: Q02763) gene. The G1036D variant represents a previously undescribed alteration, which has not been documented in the literature, clinical, or population databases of genetic variants. Given the rapidly evolving field of vascular tumour genetics and the possibility of targeted therapeutics administration, identifying novel potentially activating alterations is vastly important. Although the residue is non-essential, moderately evolutionarily conserved, and only moderate changes were predicted for the catalytic residues, the overall impact was evaluated by the XGBoost classifier as oncogenic with a 99% confidence score and was based on a deleterious prediction by both the PredictSNP algorithm and stability predictors FoldX and Rosetta. This could be approached as a basis to facilitate further functional tests to measure mutant receptor phosphorylation and, if proven as activating, introduce a considerable therapeutic opportunity (by potentially using one of the suggested inhibitive compounds such as Ecteinascidin, Ponatinib, etc., or other inhibitors of downstream signalling cascade) as well as an addition to the knowledge on disease pathogenesis. The full report can be accessed at https://loschmidt.chemi.muni.cz/predictonco/job/tie2_G1036D

Case study 3-tumour protein p53 K101Q

In next-generation sequencing screening for cancer predispositions, the K101Q variant of p53 (UniProt ID: P04637) was identified in an individual with a negative family history of cancer. p53 represents the most commonly altered gene in all cancers, and p53 variants predispose to cancer development when of germline origin. Therefore, a careful assessment must be performed for further genetic counselling. The respective variant has not been documented in the literature or functionally characterised. With lacking evidence from literature and databases of genetic variants, typically only prediction algorithms that employ sequence-based information without structural data are available. Therefore, combining both structural and sequence-related perspectives might yield a more accurate prediction. The XGBoost classifier predicted the mutation as neutral with an 81% confidence score, supported by both the PredictSNP prediction and the stability predictors. Information on evolutionary conservation showed that the wild-type residue is not conserved at this position, which may suggest that the variant is not damaging to the protein. Based on these results and no family history of cancer, the variant should not influence subsequent clinical management. Given

the importance of p53 variants in both somatic and germline contexts and their same functional impact, this case study exemplifies the utility of the tool in the assessment of hereditary cancer predisposition. The full report can be accessed at the following link—https://loschmidt.chemi.muni.cz/predictonco/job/p53_K101Q

Discussion

Prediction of the effect of missense mutations on cancer-related protein structures is a complicated task. This paper presents our pipeline for tackling this problem, thus allowing clinical bioinformaticians to easily run multiple cancer-related analyses for their target mutations on a curated list of proteins.

A major part of the pipeline capitalises on structural bioinformatics, and it requires the presence of good quality protein structures for accurate analysis. However, a high number of cancer-associated structures are transmembrane channels and thus only have fragmented domain-level structures. Some of them can be multimeric, and thus modelling proves a challenge. Despite AlphaFold [23] being touted as a major groundbreaker in the field of protein structure modelling, it proves inefficient in modelling large multi-subunit, multimeric proteins as quaternary domain level interactions are difficult to model. Thus the structural bioinformatics part of the pipeline is limited to working with high-quality structures at the domain level. AlphaFold-Multimer [17] can be used to predict the multimeric conformation in 70% of heteromeric cases and 72% of homomeric cases to limit this problem, and it is unclear whether this accuracy of predictions is viable for working with oncogenic or tumour suppressor proteins, especially when the final prediction will likely be used in a medical context.

Currently, the web server provides predictions for 44 target proteins, which were selected based on their relevance to the field of oncology. Appropriate processing of a new structure to be used in the pipeline requires expert-level knowledge of multiple bioinformatic tools. Curation in this field is a recognized bottleneck, especially in the case of the interpretation of results [7].) The addition of new target proteins to the internal database connected to the PredictONCO web server is possible and it is offered to the user community based on direct requests. Once a protein is curated, all mutations in its structure can be easily analysed. Moreover, the pipeline can also work with sequence-only data, and the trained XGBoost classifier can also reliably predict using only the sequence-based features, with only a 4% drop in average precision.

The pipeline has no standard run time as it mostly depends on whether structural analysis needs to be performed along with sequence-based analysis or not. The structural analysis increases the computational load, and the complexity of the structure can further increase the run time. However, the calculations generally do not take more than two days to complete. It is unclear whether this time frame is long or short as run time benchmarking would require the existence of other similar tools, techniques or pipelines for comparative purposes, and specialised methodologies that deal with the same case do not exist. However, this time window meets the initial requirements for the use of the web server in clinical practice as well as for research and educational purposes. Furthermore, it helps assist in making the result interpretation step easier as interpretation itself is a recognized bottleneck [7].)

Comparison to other similar tools is difficult as, as of this writing, we did not come across a pipeline integrating multiple approaches to predict the effect of a missense mutation on a cancer-related protein. However several databases and online data integrating tools do exist. The two most prominent of these databases are the International Cancer Genome Consortium (ICGC) [46] and The Cancer Genome Atlas (TCGA) [50]. Furthermore, survival analysis tools also exist and are primarily based on 4 types of data: (i) mRNA data, such as PRECOG [19], (ii) ncRNA data, such as OncoLnc [3], (iii) DNA methylation and mutation data, such as cBioPortal [18], and (iv) Protein data, such as TCPA [31]. Additionally, the Swiss-PO web tool for mapping gene mutations on the 3D structure can be used, but it only allows for intuitive and qualitative analysis of mutations that have already been experimentally determined [25]. In comparison to the aforementioned database, PSnpBind is also difficult as it only catalogues changes to binding affinities of ligands due to binding site single-nucleotide polymorphisms (SNPs) [2].

Our pipeline currently only supports missense mutations, as it is unable to handle insertions, deletions, or fusions of oncogenic proteins because individual tools in the pipeline are not able to analyse them. However, substitutions do make up a large number of cancer-associated mutations as a large number of genes associated with various cancer types contain single nucleotide variants [15]. For common solid tumours, 95% of cancer driver mutations in humans are single-base substitutions. Approximately, 90.7% of these result in the amino acid being substituted for another, non-synonymous one [14]. Thus, even though insertions, deletions, and fusions cannot be analysed using the pipeline, it still provides predictions for a significant majority of cancer-related alterations. The tool is freely accessible to the community of bioinformaticians and

medical doctors and will provide fast and useful analysis of data from the sequencing of patient samples.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00876-3.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Author contributions

Conceptualization: JSte, OS, JD, SM, DB Methodology: RTK, JS, JPI, GP, JD, SM, DB Data analysis: RTK, PP, JS, SB, IA, JPI, VS, SM, DB Software development: JS, SB, JPI, AD Writing the main manuscript: RTK, PP, JS Supervision: JSte, OS, JD, SM, DB All authors contributed to the final version of the manuscript.

Funding

The authors would like to express their thanks to the Czech Ministry of Education [ESFRI CZECRIN LM2023049; ESFRI eINFRA LM2018140, ESFRI RECETOX LM2023069]; the Technology Agency of the Czech Republic [TREND FW03010208; PERMED TN02000109]; the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 857560 (CETOCOEN Excellence); Brno University of Technology [FIT-S-23-8209]; Ministry of Health [NU20-03-00240]. The research was further supported by the project National Institute for Oncology Research [Programme EXCELES, ID Project No. LX22NPO5102 funded by the European Union—Next Generation EU]. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Availability of data and materials

The pipeline is available as a web server, at https://loschmidt.chemi.muni.cz/predictonco/. The list of proteins, definition of binding pockets, and ML model validation are attached as supplementary files. The training and testing datasets are available at https://zenodo.org/records/10013764.

Declarations

Competing interests

None declared.

Author details

[1]Loschmidt Laboratories, Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic. [2]Loschmidt Laboratories, RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic. [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic. [4]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic. [5]Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [6]Department of Paediatric Oncology, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic. [7]Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic.

Received: 22 November 2023 Accepted: 26 June 2024
Published online: 29 July 2024

References

1. Ainscough BJ et al (2016) DoCM: a database of curated mutations in cancer. Nat Method 13(10):806–807. https://doi.org/10.1038/nmeth.4000
2. Ammar A et al (2022) PSnpBind: a database of mutated binding site protein–ligand complexes constructed using a multithreaded virtual screening workflow. J Chemin. https://doi.org/10.1186/s13321-021-00573-5
3. Anaya J (2016) OncoLnc: linking TCGA survival data to MRNAs, MiRNAs, and LncRNAs. PeerJ Comput Sci 2:e67. https://doi.org/10.7717/peerj-cs.67
4. Bendl J et al (2014) PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1003440
5. Blanco JD et al (2018) FoldX accurate structural protein–DNA binding prediction using PADA1 (protein assisted DNA assembly 1). Nucl Acid Res 46(8):3852–3863. https://doi.org/10.1093/nar/gky228
6. Boeckmann B (2003) The SWISS-PROT protein knowledgebase and its Supplement TrEMBL in 2003. Nucl Acid Res 31(1):365–370. https://doi.org/10.1093/nar/gkg095
7. Bungartz KD et al (2018) Making the right calls in precision oncology. Nat Biotechnol 36(8):692–696. https://doi.org/10.1038/nbt.4214
8. Brandes N et al (2023) Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. https://doi.org/10.1038/s41588-023-01465-0
9. Buzdin A et al (2021) Editorial: next generation sequencing based diagnostic approaches in clinical oncology. Front Oncol. https://doi.org/10.3389/fonc.2020.635555
10. "Cancer Today." *Iarc.fr*, 2020, https://gco.iarc.fr/today/home.
11. Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23(15):1875–1882. https://doi.org/10.1093/bioinformatics/btm270
12. Chakravarty D et al (2017) OncoKB: a precision oncology knowledge base. JCO Precis Oncol. https://doi.org/10.1200/po.17.00011
13. Dana JM et al (2018) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucl Acid Res. https://doi.org/10.1093/nar/gky1114
14. Darbyshire M et al (2019) Estimating the frequency of single point driver mutations across common solid tumours. Sci Rep. https://doi.org/10.1038/s41598-019-48765-2
15. Deng N et al (2017) Single nucleotide polymorphisms and cancer susceptibility. Oncotarget. https://doi.org/10.1632/oncotarget.22372
16. Eswar N et al (2008) Protein structure modeling with MODELLER. Method Mol Biol. https://doi.org/10.1007/978-1-60327-058-8_8
17. Evans R et al (2021) Protein complex prediction with AlphaFold-Multimer. BioRxiv. https://doi.org/10.1101/2021.10.04.463034
18. Gao J et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the CBioPortal. Sci Signal. https://doi.org/10.1126/scisignal.2004088
19. Gentles AJ et al (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. https://doi.org/10.1038/nm.3909
20. Irwin JJ et al (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model. https://doi.org/10.1021/ci3001277
21. Iwamura R et al (2023) PDGFRB and NOTCH3 mutations are detectable in a wider range of pericytic tumors, including myopericytomas, angioleiomyomas, glomus tumors, and their combined tumors. Mod Pathol. https://doi.org/10.1016/j.modpat.2022.100070
22. Jiménez-Moreno A et al (2021) DeepAlign, a 3D alignment method based on regionalized deep learning for Cryo-EM. J Struct Biol 213(2):107712. https://doi.org/10.1016/j.jsb.2021.107712
23. Jumper J et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2
24. Kellogg EH et al (2010) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Protein Struct Funct Bioinform 79(3):830–838. https://doi.org/10.1002/prot.22921
25. Krebs FS et al (2021) Swiss-PO: a new tool to analyze the impact of mutations on protein three-dimensional structures for precision oncology. NPJ Precis Oncol 5(1):19. https://doi.org/10.1038/s41698-021-00156-5
26. Krivák R, Hoksza D (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminformatics. https://doi.org/10.1186/s13321-018-0285-8
27. Krzyszczyk P et al (2018) The growing role of precision and personalized medicine for cancer treatment. Technology. https://doi.org/10.1142/s2339547818300020
28. Kurnit KC et al (2017) 'Personalized cancer therapy': a publicly available precision oncology resource. Cancer Res 77(21):e123–e126. https://doi.org/10.1158/0008-5472.can-17-0341

29. Landrum MJ et al (2017) ClinVar: improving access to variant interpretations and supporting evidence. Nucl Acid Res 46(D1):D1062–D1067. https://doi.org/10.1093/nar/gkx1153
30. Lassen UN et al (2021) Precision oncology: a clinical and patient perspective. Futur Oncol 17(30):3995–4009. https://doi.org/10.2217/fon-2021-0688
31. Li J et al (2013) TCPA: a resource for cancer functional proteomics data. Nat Method 10(11):1046–1047. https://doi.org/10.1038/nmeth.2650
32. Madeira F et al (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. Nucl Acid Res 50(W1):W276–W279. https://doi.org/10.1093/nar/gkac240
33. O'Meara MJ et al (2015) Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. J Chem Theor Computation. 11(2):609–622. https://doi.org/10.1021/ct500864r
34. Ortiz E et al (2020) Invasive myofibromatosis with visceral involvement in a term newborn: a case report. Am J Pediatr 6(2):173–173. https://doi.org/10.11648/j.ajp.20200602.30
35. Patterson SE et al (2016) The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genom. https://doi.org/10.1186/s40246-016-0061-7
36. Pond D et al (2018) A patient with germ-line gain-of-function PDGFRB P.N666H mutation and marked clinical response to imatinib. Genet Med 20(1):142–150. https://doi.org/10.1038/gim.2017.104
37. Prlić A et al (2007) Integrating sequence and structural biology with DAS. BMC Bioinform. https://doi.org/10.1186/1471-2105-8-333
38. Ribeiro AJM et al (2017) Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucl Acid Res 46(D1):D618–D623. https://doi.org/10.1093/nar/gkx1012
39. Richards S et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. Genet Med 17(5):405–424. https://doi.org/10.1038/gim.2015.30
40. Rostkowski M et al (2011) Graphical analysis of PH-dependent properties of proteins predicted using PROPKA. BMC Struct Biol. https://doi.org/10.1186/1472-6807-11-6
41. Seeliger D, de Groot BL (2010) Ligand docking and binding site analysis with PyMOL and autodock/vina. J Comput Aided Mol Des 24(5):417–422. https://doi.org/10.1007/s10822-010-9352-6
42. Sievers F et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 7(1):539. https://doi.org/10.1038/msb.2011.75
43. Sumbalova L et al (2018) HotSpot wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. Nucl Acid Res 46(W1):W356–W362. https://doi.org/10.1093/nar/gky417
44. Sung H et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin 71(3):209–249. https://doi.org/10.3322/caac.21660
45. Suzek BE et al (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739
46. The International Cancer Genome Consortium (2010) International network of cancer genome projects. Nature 464(7291):993–998. https://doi.org/10.1038/nature08987
47. The UniProt Consortium (2022) UniProt: the universal protein knowledgebase in 2023. Nucl Acid Res 51(D1):D523-531. https://doi.org/10.1093/nar/gkac1052
48. Trott O, Olson AJ (2009) AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. https://doi.org/10.1002/jcc.21334
49. Venselaar H et al (2010) Protein structure analysis of mutations causing inheritable diseases. An e-science approach with life scientist friendly interfaces. BMC Bioinform. https://doi.org/10.1186/1471-2105-11-548
50. Weinstein JN et al (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113–1120. https://doi.org/10.1038/ng.2764
51. wwPDB Consortium (2018) Protein data bank: the single global archive for 3D macromolecular structure data. Nucl Acid Res 47(D1):D520–D528. https://doi.org/10.1093/nar/gky949

Publisher's Note

JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau
Article

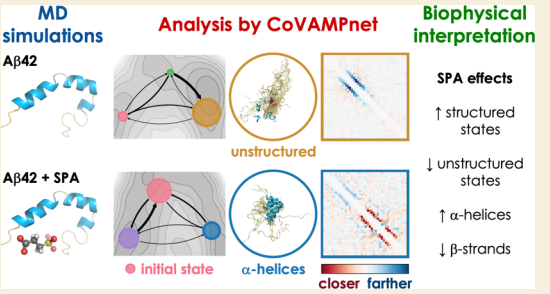
CoVAMPnet: Comparative Markov State Analysis for Studying Effects of Drug Candidates on Disordered Biomolecules

Sérgio M. Marques, Petr Kouba, Anthony Legrand, Jiri Sedlar, Lucas Disson, Joan Planas-Iglesias, Zainab Sanusi, Antonin Kunka, Jiri Damborsky, Tomas Pajdla, Zbynek Prokop, Stanislav Mazurenko,* Josef Sivic,* and David Bednar*

Cite This: JACS Au 2024, 4, 2228–2245
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computational study of the effect of drug candidates on intrinsically disordered biomolecules is challenging due to their vast and complex conformational space. Here, we developed a comparative Markov state analysis (CoVAMPnet) framework to quantify changes in the conformational distribution and dynamics of a disordered biomolecule in the presence and absence of small organic drug candidate molecules. First, molecular dynamics trajectories are generated using enhanced sampling, in the presence and absence of small molecule drug candidates, and ensembles of soft Markov state models (MSMs) are learned for each system using unsupervised machine learning. Second, these ensembles of learned MSMs are aligned across different systems based on a solution to an optimal transport problem. Third, the directional importance of inter-residue distances for the assignment to different conformational states is assessed by a discriminative analysis of aggregated neural network gradients. This final step provides interpretability and biophysical context to the learned MSMs. We applied this novel computational framework to assess the effects of ongoing phase 3 therapeutics tramiprosate (TMP) and its metabolite 3-sulfopropanoic acid (SPA) on the disordered Aβ42 peptide involved in Alzheimer's disease. Based on adaptive sampling molecular dynamics and CoVAMPnet analysis, we observed that both TMP and SPA preserved more structured conformations of Aβ42 by interacting nonspecifically with charged residues. SPA impacted Aβ42 more than TMP, protecting α-helices and suppressing the formation of aggregation-prone β-strands. Experimental biophysical analyses showed only mild effects of TMP/SPA on Aβ42 and activity enhancement by the endogenous metabolization of TMP into SPA. Our data suggest that TMP/SPA may also target biomolecules other than Aβ peptides. The CoVAMPnet method is broadly applicable to study the effects of drug candidates on the conformational behavior of intrinsically disordered biomolecules.



KEYWORDS: *soft Markov state models, intrinsically disordered proteins, adaptive molecular dynamics, Alzheimer's disease, Aβ42 peptide, drug candidates, tramiprosate, 3-sulfopropanoic acid*

INTRODUCTION

Alzheimer's disease (AD) is globally the fifth leading cause of death and fourth cause of disability in people aged 75 years and above and thus represents an enormous societal burden.[1] Amyloid-beta (Aβ) peptides play a major role in the development of AD, although the mechanism behind their toxicity is still debated.[2,3] A model of toxicity known as the oligomer hypothesis states that Aβ oligomerizes into toxic pore-forming oligomers at the neuronal plasma membrane, which ultimately leads to cell death. Among the different Aβ peptides, the 42-residue long peptide (Aβ42; Figure 1A) is the most aggregation-prone isoform.[4,5]

The Aβ peptides are intrinsically disordered, which makes them difficult to study both experimentally and computationally. Intrinsically disordered proteins do not adopt a single well-defined structure, but rather exist as ensembles of conformations with similar energies. These ensembles are best characterized by their population distributions and probabilities of several properties or descriptors (e.g., radius of gyration and secondary structure).[6,7] The disordered nature of Aβ42 significantly complicates the analysis of its molecular dynamics (MD) trajectories, namely, the definition of conformational states, which is an important step toward a deeper understanding of the system and its slowest transitions.[8] A popular approach for identifying notable conformational states in MD simulations involves building so-called Markov state models (MSMs). Under the assumption of the dynamics being Markovian (memoryless), these models cluster the conformational space

Received: February 28, 2024
Revised: April 24, 2024
Accepted: May 13, 2024
Published: May 28, 2024

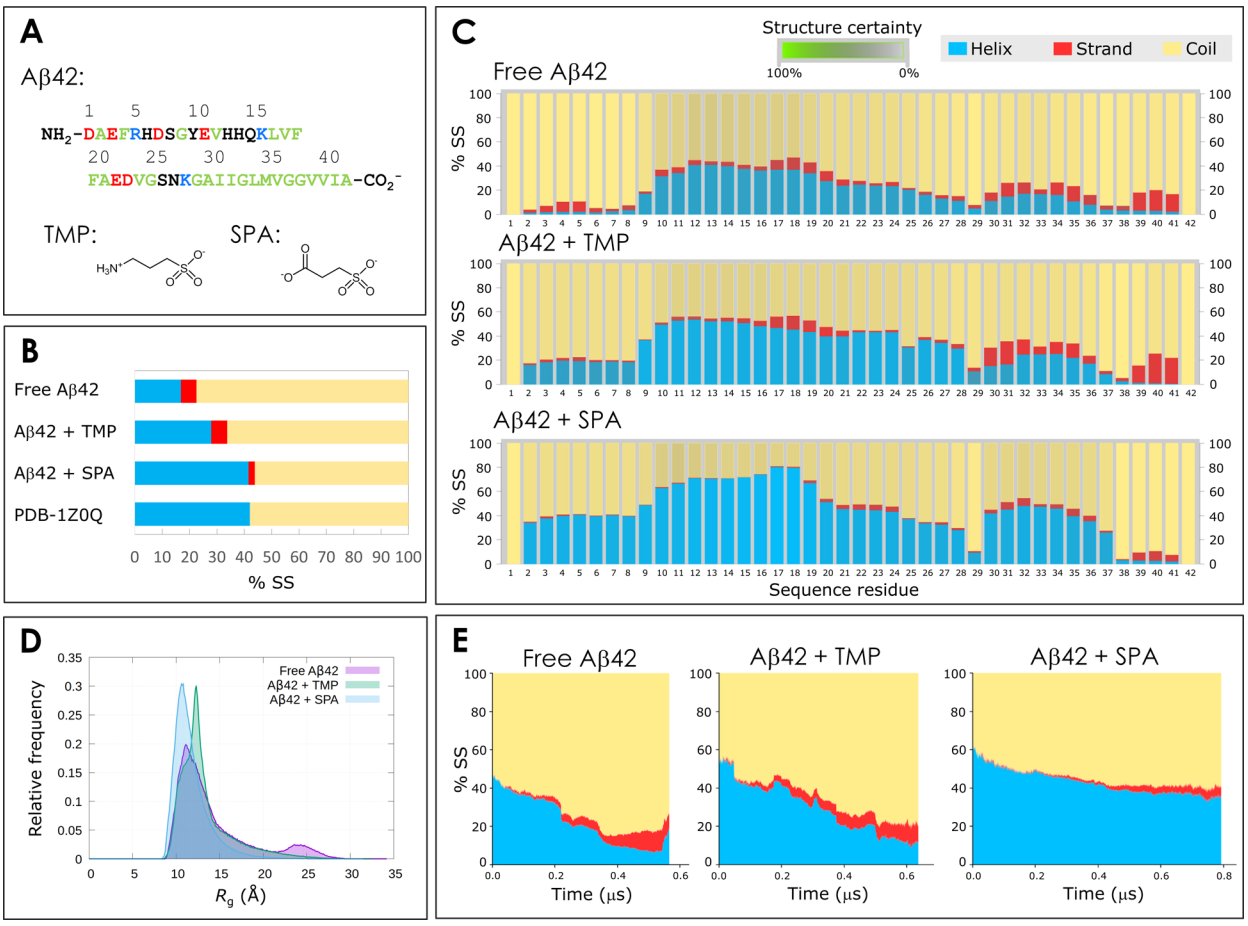


Figure 1. Structures of Aβ42 peptide and the studied small molecules, and properties of the ensembles from the adaptive simulations for the free Aβ42, Aβ42 + TMP, and Aβ42 + SPA. A) Sequence of the Aβ42 peptide and chemical structures of tramiprosate (TMP) and 3-sulfopropanoic acid (SPA) in the dominant protonation states at the physiological pH 7.4. The sequence residues are color-coded as follows: *red* for negatively charged; *blue* for positively charged; *green* for hydrophobic; and *black* for polar neutral residues. B) Total secondary structural propensity (% SS) of Aβ42 during the adaptive MDs, in the original NMR ensemble (PDB 1Z0Q with 30 structures), and from the experimental measurements of free Aβ42 in aqueous solution. C) Secondary structure propensity of Aβ42 by residue, obtained for the global ensembles from the adaptive simulations. The certainty of the secondary structure assignment was obtained by the statistical variance among ten randomized bins of frames and is represented by the saturation of the secondary structure color (the more saturated the color, the more certain the assignment, as indicated by the legend). D) Distribution of the radius of gyration (R_g) of the ensembles from the same adaptive simulations. E) Time evolution of the secondary structure of Aβ42 during the time-based aligned adaptive sampling MD simulations. The secondary elements are aggregated across all 42 residues, averaged at each time over all the trajectories parallel in time according to the time-based alignment. Only the timespan covering at least 20 parallel trajectories is plotted.

into states preserving the Markovianity of the transitions and estimate the equilibrium distribution and transition rates between the states. The conventional methods for building MSMs typically rely on a selection of collective variables, compressing the high-dimensional MD data and simplifying the clustering. Recent progress in variational approaches for conformational dynamics has further allowed scoring different MSMs, e.g., based on their ability to approximate the slowest modes of the dynamics, thus facilitating the development of automatic frameworks for the identification of Markov states.[9] Although some of these procedures are quite advanced and enable, e.g., an accurate estimation of transition rates even from biased simulation data,[10] the manual selection of the collective variables is typically laborious and can often cause the resulting models to fail the tests for Markovianity. While MSMs are extremely valuable tools, they possess certain limitations, such as the assumption of Markovianity, constraints on state representation granularity, reliance on extensive sampling, and relatively rapid relaxation dynamics.[11−14] Several alternative methodologies exist to address these shortcomings. These include hidden Markov models (HMMs) to relax the Markovian assumption,[12] approaches incorporating memory effects such as the generalized master equation (GME) and the generalized Langevin equation (GLE) for more effective dynamic property assessment,[11] and methods rooted in deep learning.[15]

A powerful framework based on deep learning is VAMPnet, a neural network that learns a probabilistic assignment of each simulation frame to individual states in an unsupervised manner by maximizing a variational score.[16] In contrast to the other methods, the VAMPnet approach does not relax the Markovianity assumption but rather combines the search of collective variables with the optimization of a cost function to efficiently identify the slowest modes of the system. The application of VAMPnets to the analysis of Aβ42 trajectories has already shown great potential in producing robust MSMs for quantification of the Aβ42 kinetics and equilibrium properties.[17] Several recent methods build on the VAMPnet approach to address the efficiency of protein representation,[18,19] scalability to multidomain protein systems,[20] stability of the training process,[21] sampling of rare conformations,[22] or the importance

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

of residues based on the attention mechanism.[18,23] However, to the best of our knowledge, a method for aligning and comparing ensembles of learned MSMs across different systems that would simplify the biophysical interpretation of the conformational states by identifying their distinctive features is still missing. In this work, we have developed such a method to help understand and quantify the effects of drug candidates on the conformational space of the analyzed system.

This problem is important in many fields of research, particularly in AD. Due to the prevalence and severity of the disease, there is a growing interest in pharmaceuticals capable of preventing the early stages of the Aβ42 oligomerization and stopping the pathogenic amyloid cascade.[3,4,24] Tramiprosate (TMP), also known as homotaurine or 3-amino-1-propanesulfonic acid, is a naturally occurring aminosulfonate. Even at high concentrations, it is well tolerated in the human brain, where it is metabolized into 3-sulfopropanoic acid (SPA) (Figure 1A). TMP has been reported to prevent the formation of fibrillar forms of Aβ, reduce the Aβ-induced death rate of neuronal cell cultures, and lower the amyloid plaque deposition in the brain.[25−27] Clinical trials have shown its ability to slow down the cognitive decline in patients with homozygous expression of the apolipoprotein E gene *APOE4*, similarly to FDA-approved aducanumab.[24,28] TMP can act not only on Aβ, but also on other pathways that contribute to cognitive impairment in AD and other neurologic disorders.[29,30] ALZ-801 is a valine-conjugated prodrug of TMP that is currently in phase 3 of clinical trials for early stage AD patients bearing the *APOE4/4* genotype (NCT04770220).[31,32] Preliminary *in vitro* and *in silico* studies suggested that both TMP and SPA can lock the Aβ peptides in monomeric conformations that are less prone to oligomerization, thus inhibiting the first step in the pathological pathway of Aβ.[33−35] However, these studies do not provide sufficient insights to fully explain the mechanism of action of these molecules on Aβ. At the moment, it is still unclear whether TMP or its metabolite SPA can exert a stronger biological effect, and this was one of our motivations to carry out this study.

To analyze the effect of TMP and SPA on Aβ and understand how these small molecules may prevent the formation of Aβ oligomers and fibrils, we developed a new computational framework called comparative Markov state analysis (CoVAMPnet). The CoVAMPnet framework reveals the impact of a small molecule (in our case, TMP or SPA) on the conformational space and dynamics of an intrinsically disordered biomolecule (in our case, Aβ) in three steps. First, molecular dynamic trajectories are generated using enhanced sampling, and an ensemble of soft MSMs is computed for each system by training VAMPnet neural networks.[17] In particular, we simulated the monomeric Aβ42 peptide in its free form and in the presence of drug candidates TMP or SPA. Second, using our novel alignment method, these ensembles are aligned to identify similar conformational states across the different systems based on a solution to an optimal transport problem. This proved useful in quantifying the similarities and differences in Aβ42 conformations in response to the presence or absence of the small molecules. Finally, our new approach based on analyzing gradients of the trained neural networks is used to elucidate the patterns underlying the learned MSMs and to understand the biophysical relevance of the molecular features, namely, the directional inter-residue distances, for the classification into each state. To our knowledge, this is the first time that such a biomolecular relevance analysis has been used to compare and interpret MSMs built by unsupervised machine learning methods and quantify the effects of drug candidates on the conformational space of a disordered protein. Experimental comparison of Aβ42 in its free form and in the presence of TMP or SPA by circular dichroism (CD), Fourier-transform infrared spectroscopy (FTIR), nuclear magnetic resonance (NMR), and fluorometry has further shown the effects of the small molecules on longer time scales, complementing our computational findings.

MATERIALS AND METHODS

Here, we present only a concise description of the methods used, focusing mainly on the novel methodology. A complete and detailed description is provided in Supporting Information and Methods.

Molecular Dynamics (MD) Simulations

System Preparation. The structures of tramiprosate (TMP) and 3-sulfopropanoic acid (SPA) were constructed and minimized using Avogadro 2.[36] During the calculation of partial charges, the structures were further optimized by Gaussian 09,[37] and the *antechamber* module of AmberTools 16[38] was then used to prepare the force field-compatible parameters. The three-dimensional structural data of the Aβ42 peptide were obtained from the RCSB Protein Data Bank[39] (PDB entry 1Z0Q). It resulted from NMR experiments and contains 30 structures, which were saved separately. The Aβ42 peptide was protonated using PROPKA[40] at physiological pH 7.4, the small molecules were embedded (when appropriate), the systems solvated, and their topologies built using high-throughput molecular dynamics (HTMD)[41] in combination with the CHARMM36m[42] (C36m) force field. We used a stoichiometry of 100 molecules of TMP or SPA per molecule of Aβ42. This ratio approximates the experimental conditions (1000:1) without compromising the computational costs of the simulations.

MD Simulation Protocols. All the systems were equilibrated using HTMD.[41] The end point of the equilibration cycle was taken as a starting point for subsequent MD simulations, either classic or adaptive sampling ones. The simulations employed the same settings as the last step of the equilibration, and their trajectories were saved every 0.1 ns. HTMD was used to perform adaptive sampling of the Aβ42 conformations. Due to the conformational complexity of Aβ42, three protocols (namely, A, B, and C) were assessed. Each protocol differed from the others in the starting structure set, the adaptive metric, the number of adaptive epochs and replicas, and the total cumulative MD time (Table S1). Protocols A and B were only applied to free Aβ42, while protocol C was applied to free Aβ42, Aβ42 + TMP, and Aβ42 + SPA.

Classical MD simulations were also performed using HTMD, where only the structure of the first model of the PDB entry 1Z0Q was used as the starting point. The free Aβ42, Aβ42 + TMP, and Aβ42 + SPA systems were prepared and equilibrated as described above. These MDs were performed using only the C36m force field. Each MD was run in sequential batches of 200 ns each, for a total of 5 μs, and 10 independent replicates were performed for each system.

Analyses of Properties in Combined MD Ensembles. In order to analyze the produced MD simulations, their topologies were converted from CHARMM to AMBER using ParmEd[43] when required. Water molecules and ions were filtered out from the resulting MDs, which were then compiled into a simulation

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

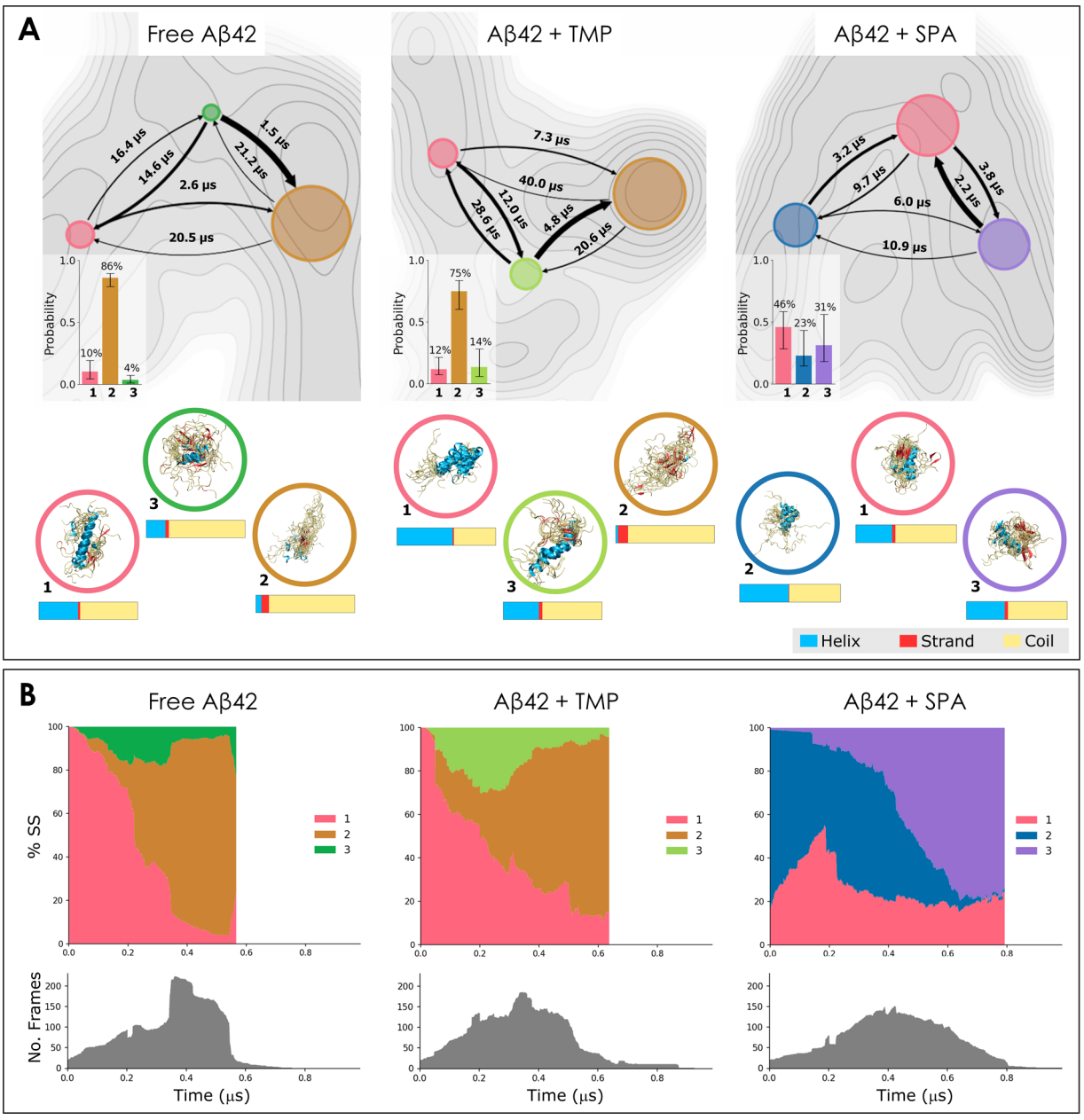


Figure 2. Analysis of conformational states learned using the variational approach to Markov processes on the adaptive simulations and their evolution in time. A) Properties of the states. For each system, we report: (i) the free energy surface (FES) projected on the first two tICA dimensions (gray maps), where darker shades correspond to more negative energy regions; (ii) flux diagrams overlapping the FES and projected on the same tICA space, where each state is represented by a colored circle with the area proportional to the state probability, and the arrows indicate the mean first-passage times T_M between the states, with the thickness proportional to the transition probability; (iii) equilibrium distribution of the states (bottom-left corner of FES; the bars represent the 95th percentile of values centered around the median from the ensemble of 20 learned models; see Supplementary Note 7 for details); (iv) superimposition of 20 representative structures from each state, selected based on the highest assignment probability (below FES, enclosed in colored circles); (v) global mean secondary structure content of each state (below the respective structures). B) Distribution of the learned states in time (top) and the number of frames available at each time point (bottom). The adaptive sampling trajectories were aligned in time and concatenated. The state probability at a given time point was computed as the average soft assignment of all available frames at this time point. From left to right, the state assignments evolve from the beginning to the end of the simulation time. All plots are shown for the free Aβ (left), Aβ + TMP (middle), and Aβ + SPA (right). The states are numbered and color-coded consistently across the entire panel; the same colors across different systems indicate aligned states.

list using HTMD. The *cpptraj*[44] module of AmberTools 16[38] was used to compute several properties in the combined ensembles: root-mean square deviation (RMSD), radius of gyration (R_g), and linear interaction energy (LIE)[45] between Aβ42 and TMP or SPA. DSSP 3.0[46] was used to assign a secondary structure to every residue in every snapshot of the combined trajectories, and the default DSSP seven-letter alphabet was converted to the three main secondary elements (α-helix, β-strand, and coil, see MD analysis section in Supporting Information and Methods). Accounting for all the residues of each secondary structure type in the peptide for all the analyzed snapshots resulted in the total secondary structure

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

content of the ensemble. Mechanics/generalized Born solvent accessible surface area (MM/GBSA)[47,48] calculations were performed with the MMPBSA.py.MPI[47] module of AmberTools 14 to obtain the free energy of the peptide for every frame of the ensemble, from which the peptide intramolecular interactions were derived.

Comparative Markov State Model Analysis (CoVAMPnet)

This section describes our comparative Markov state analysis (CoVAMPnet) of adaptive sampling MD simulations of the free Aβ42, Aβ42 + TMP, and Aβ42 + SPA systems. CoVAMPnet builds on the variational approach to Markov processes by VAMPnet neural networks, followed by two new analyses: (i) alignment of the learned MSM ensembles across different systems based on a solution to an optimal transport problem and (ii) characterization of the learned states by the inter-residue distances based on the neural network gradients.

Learning Markov State Models Using Neural Networks. The variational approach to Markov processes (VAMP)[49] was used to learn Markov state models (MSMs) via unsupervised training of VAMP neural networks (VAMPnets)[16] with physical constraints.[50] VAMPnet learns a nonlinear function that maps the peptide tertiary structure to a vector of state probabilities. The physical constraints ensure that the learned MSM is reversible and that the elements of the matrix representing the governing Koopman operator[16] (a linear operator propagating the state probabilities in time) are nonnegative. In this work, we used the VAMPnet implementation by Löhr et al.,[17] including the self-normalizing setup.[51]

The VAMPnet architecture consists of two parallel weight-sharing lobes: one for a frame at time t and the other for a frame at time $t + \tau$ in the same trajectory, where τ is a fixed lag time. Each frame was represented on the input as a vector (780 elements) of the upper triangular part of the peptide inter-residue heavy atom distance matrix without the diagonal and the first two subdiagonals (i.e., without the distances to the first and second neighboring residues). The output nodes in each lobe measure the probabilities of the constructed MSM states for the input frame. The network was trained on pairs of MD simulation frames separated by a selected lag time τ. To obtain the probabilities of the learned states, the frames were run through one of the lobes. For each system, an ensemble of 20 models was built. The pairs of frames were divided into 20 random splits (90% training and 10% validation) and for each split, three VAMPnet models were trained with different initialization and the one with the highest VAMP-E score[49] was selected for the MSM ensemble. The soft assignment of a frame was defined as the average of its state probabilities across the ensemble, whereas the hard assignment was defined as the state with the highest probability in the soft assignment of the frame. Throughout this work, the soft assignments were used everywhere unless it was necessary to select example frames from a particular state (such as the example structures in Figure 2A or the frames representing the states for the columns of the matrix in Figures S22). Further details on our VAMPnet setup are described in Supporting Information and Methods.

Alignment of Learned States for Comparative Analysis. The order of the states on the output of a trained VAMPnet is not well-defined and may thus vary. To construct an MSM from multiple models or compare MSMs of different systems, a correspondence between states across the models had to be established. In this work, we generalized the approach from Löhr et al.[17] for the alignment of states within a single system to obtain an ensemble of aligned MSMs. Then, we introduced a new method for the alignment of ensembles of MSMs between different systems to compare the systems and further understand the effects of the small molecules on the conformational dynamics of Aβ42.

Aligning States within a Single System. The states from the 20 models within an ensemble were aligned by a constrained k-means clustering algorithm[52] using the average inter-residue distance matrices D_m^n, where n indexes the models in the ensemble and m indexes the states in each model. The cluster centers were initialized by the $D_m^{n_0}$ matrices of a randomly selected model n_0 in the ensemble. The clustering iterated in two steps: 1) for each model n, its states were sequentially assigned to different clusters in the order of the proximity of the D_m^n matrix to the closest unassigned cluster center and respecting the constraint that two matrices from the same model cannot be assigned to the same cluster; 2) each cluster center was recomputed as the mean of the D_m^n matrices of the corresponding states. These two steps were iterated until the cluster assignment did not change. The states in each model were then renumbered according to the final assigned cluster. The method by Löhr et al.[17] is equivalent to performing only one iteration of our method. Our approach is thus less susceptible to incorrect initialization and can lead to a better alignment.

Aligning Ensembles of Markov State Models Between Different Systems. With each system described by an ensemble of N mutually aligned MSMs after the single system state alignment (see above), we proposed a novel method for aligning ensembles of MSMs across different systems. In particular, we (i) characterized each state of the given system by a nonparametric distribution over the ensemble, (ii) defined a distance metric to compare such distributions, and finally, (iii) computed an alignment of the ensembles of MSMs between the two systems by solving an optimal matching problem. Details of these steps are given next. The N instances of the VAMPnet network learned for a given system s output N different feature matrices $\{D_m^{sn}\}_{n=1}^N$ (average inter-residue matrices, see Supporting Information and Methods for a formal definition of the feature matrix) describing each of the M states of the system. Each state m was, therefore, characterized by the distribution $\theta_m^s(\xi)$ of the features over the different VAMPnet instances as

$$\theta_m^s(\xi) = \frac{1}{N}\sum_{n=1}^{N} \delta(\xi - D_m^{sn}) \tag{1}$$

where δ is the Dirac delta function defined over the feature space of inter residue distances in which the simulation frames are represented and D_m^{sn} is the inter-residue distance matrix representing state m of the learned model n for system s. θ_m^s thus represents the state m of system s with a nonparametric distribution given by the set of Dirac functions centered at the feature matrices D_m^{sn} obtained by the instances of the learned ensemble.

To exploit the entire distribution of the features of each state, the distance between two different states was evaluated by comparing their respective distributions. In particular, we employed the Wasserstein distance of two distributions as a distance measure quantifying the cost of aligning two states from different MSMs as

$$c_{ml}^{s_1 s_2} = d_W(\theta_m^{s_1}, \theta_l^{s_2}) \tag{2}$$

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

where $c_{ml}^{s_1s_2}$ is the cost of aligning state m of system s_1 with state l of system s_2 and $d_W(\theta_m^{s_1}, \theta_l^{s_2})$ is the Wasserstein-1 distance of the two respective distributions defined as

$$d_W(\theta_m^{s_1}, \theta_l^{s_2}) = inf_{\gamma \in \Gamma(\theta_m^{s_1}, \theta_l^{s_2})} \int \|\xi, \xi'\| \gamma(\xi, \xi') \quad (3)$$

where $\Gamma(\theta_m^{s_1}, \theta_l^{s_2})$ is the set of joint distributions whose left and right marginals are $\theta_m^{s_1}$ and $\theta_l^{s_2}$, respectively, and $\|\xi, \xi'\|$ is the Euclidean distance of the two feature vectors ξ, ξ' distributed according to the joint distribution $\gamma(\xi, \xi')$. In the case of empirical nonparametric distributions (such as in our case), the problem of Wasserstein-1 distance computation has an equivalent linear program formulation and it was solved using an optimal transport algorithm.[53]

Finally, the alignment of MSM ensembles was formulated as an optimization problem. Without the loss of generality, let us assume that the MSM representing system s_1 does not have more states than the MSM representing system s_2. The problem was defined as

$$\hat{\pi}^{s_1s_2} = \arg \min_{\pi^{s_1s_2} \in \Pi^{s_1s_2}} \sum_{m=1}^{M_{s_1}} c_{m\pi(m)}^{s_1s_2} \quad (4)$$

where M_{s_1} is the number of states of the MSM estimated for system s_1, $\Pi^{s_1s_2}$ is the set of all bijections from the states of system s_1 into any M_{s_1}-sized subset of states of system s_2, and the bijection $\hat{\pi}^{s_1s_2}$ is the optimal mapping of states of system s_1 onto the states of system s_2. This optimization problem, and thus also the alignment of MSM ensembles, was solved using the Hungarian algorithm.[54]

Gradient-Based Characterization of Learned States. The differentiability of the VAMPnet model enables interpretation of the states by investigating the feature importance, which is hard to do using classical Markov state models. This analysis aimed to understand how important the different parts of the protein structure (here represented by the peptide inter-residue distances) are for the definition of different states. While there exist different methods to investigate the importance of features in neural networks,[55,56] they are usually applied to single models for simple tasks, such as the classification of individual images. The challenge of adopting those methods for the current study was in calculating the feature importance for an ensemble of MSMs. We proposed a method to identify which features were important for the classification of the simulation frames into the learned states, building on the gradient-based method proposed for image classification.[56] In our approach, we computed the gradients for each of the models in the MSM ensemble separately and aggregated their results over the ensemble. To this end, the MSMs produced by the models needed to be aligned, which we did by using our state alignment method discussed earlier (see Aligning states within a single system). The gradients for individual Markov states were computed as follows:

$$g_m(\xi) = \frac{1}{N} \sum_{n=1}^{N} \nabla_\xi \chi_{nm} \quad (5)$$

where g_m is a 780-dimensional vector containing the ensemble-averaged gradient of the output probability of state m computed with respect to the input features ξ; N is the number of models in the ensemble; ∇_ξ is the operator of gradient with respect to the coordinates of the network input features ξ; and χ_{nm} represents the output node corresponding to state m of n^{th} VAMPnet model in the ensemble. Here, the 780-dimensional network input vector was obtained by vectorizing the upper triangular inter-residue distance matrix and removing the diagonal and two subdiagonals. The intuition is that the i^{th} entry of vector g_m expresses the change in the probability of the assignment of the given frame of the simulation to state m induced by an increase in the distance of the i^{th} pair of residues at the input of the VAMPnet network. The above definition computes the gradient value for an individual frame of the system. To aggregate the gradient value over a representative set of frames from the investigated system, we evaluated the gradient vector $\widehat{g_m}$ as the average of g_m over 10,000 randomly selected simulation frames ξ. For visualization purposes, we took the 780-dimensional vector of evaluated gradients $\{\widehat{g_m}\}_{m=1}^{M}$ and arranged it back into a 42 × 42 matrix corresponding to the shape of the inter-residue distance matrix. These gradients evaluated and averaged over randomly selected frames should express the importance of particular residues on average for the classification into a specific state without any particular assumptions about the input frame.

Estimation of the Free Energy Landscape. We estimated the free energy landscape of Aβ42 for each of the studied systems, projected on the first 2 time-lagged independent component analysis (tICA) dimensions, by performing Gaussian kernel density estimation on 10% of the simulated frames.[17]

Experimental Validation

Aβ42 in its monomeric form (N-methionine-Aβ42 or N-Met-Aβ42) was produced and purified following an adapted version of the protocol by Cohen et al.[57] Spectroscopic properties of N-Met-Aβ42 alone or in the presence of TMP, SPA, and the membrane-mimicking hexafluoroisopropanol (HFIP) were measured using circular dichroism (CD), Fourier-transformed infrared spectroscopy (FTIR), and nuclear magnetic resonance (NMR). Aggregation kinetics were recorded using thioflavin T (ThT) assays.[58] A 1000-fold molar excess of TMP or SPA with respect to the concentration of N-Met-Aβ42 was used, to replicate the experimental conditions previously reported to exert biological effects from those molecules.[33]

RESULTS

Selection of the Computational Protocol for the Simulation of Aβ42

We aimed to query, by molecular dynamics (MD) simulations, the conformational diversity and dynamics of Aβ42 (the most aggregation-prone and the second-most abundant isoform of Aβ[4,5]) and the effect of small molecules on such dynamics. The molar excess of small molecules with respect to Aβ42 was lower in the simulations (100-fold) than in the experiments (1000-fold), but it ensured sufficient interactions with the peptide (see Supplementary Note 1). Some of the key parameters to consider in any MD simulation are (i) the starting conformation, (ii) the MD technique and its length, and (iii) the force field. For the starting conformation, we chose a structure of the full-length peptide obtained from liquid state NMR (PDB ID 1Z0Q;[59] see Supplementary Note 2 and Figure S1). Because of its enhanced ability to sample events occurring in longer timescales,[60−62] we applied adaptive sampling. This method consists of several MD trajectories simulated in parallel and over multiple consecutive epochs, in an adaptive approach. The MDs from each epoch are iteratively seeded from selected snapshots from previous MDs,

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

according to a predefined criterion. This criterion defines a feature (also called *metric* or *collective variable*), and the objective is to maximize the variability of that feature sampled in the overall simulation (in this case, the secondary structures).[41,63] Based on the literature,[64] we explored the AMBER ff14SB[65] (hereafter termed A14SB) and CHARMM36m[42] (C36m) force fields as the ones likely to provide reasonable ensembles to study Aβ42. Notably, C36m was developed specifically for intrinsically disordered proteins and has already been used with Aβ42.[35] We tested different combinations of parameters in three adaptive sampling protocols and compared the results to the initial structure, experimental data,[59,66] and previous reports.[8] The goal was to obtain conformations of Aβ42 diverging from the initial NMR structure (membrane-like environment) and to reach average secondary structure ratios that approximate the experimental ones (in aqueous environment). The selected protocol used the C36m force field and A14SB was discarded (protocol C; see Supplementary Note 2, Table S1, and Figures S2−S4). The respective MD ensembles seemed to be well converged (Figure S5).

Secondary Structure Content in Simulations of Free Aβ42 and Aβ42 with Ligands

To compare the simulations of Aβ42 alone and in the presence of an excess of TMP and SPA (Figure 1A), we first analyzed the global secondary structure content of the peptide in the three systems (Figure 1B). In the adaptive simulations of free Aβ42, the peptide showed a larger ratio of coils (77.5%), followed by the α-helices (16.8%) and finally the β-strands (5.7%). In the presence of TMP, the α-helix content of Aβ42 peptide increased by 11.1 p.p. to 27.9%, while the ratio of β-strands remained unchanged (5.8% vs 5.7%). In the presence of SPA, the differences in the secondary structure were more striking. In this case, the content of α-helices was nearly the same as in the original NMR structure (41.6% vs 42.1%), the ratio of coils was slightly lower (56.1% vs 57.9%), and the β-strands were half of those in free Aβ42 (2.3% vs 5.7%). This remarkable result suggests a strong effect of SPA in preserving the α-helical structures of Aβ42.

We analyzed the secondary structures in more detail, dissecting the different propensities by the sequence residues (Figure 1C). The results showed that Aβ42 could adopt a coiled structure over its entire sequence, with the highest fractions in the N-terminal residues 1−8. Helical structures were most significant for residues 10−20, with α-helical structures near and above 40% and decreasing in further residues. The β-strands were the least frequent element, present at the C-terminal tail of the peptide (residues 30−41) and, to a lesser degree, also around residues 2−8 and 17−20. This is in agreement with Tomaselli et al., who reported the formation of an antiparallel β-sheet made of two β-strands containing amino acids 18−22 and 37−41.[59] TMP had little effect on the secondary structure distribution, only slightly increasing the frequency of helical structures in the regions that already had a propensity for it (residues 9−28) and reducing the β-strands in the N-terminal residues 2−8. However, the inclusion of SPA resulted in a substantial reduction of the β-strand content in residues 2−20 and 30−41 and in a significant increase of helical propensity in residues 9−28 and 30−37. Thus, we observed that both studied Aβ modulators (TMP and SPA) could increase the regular structures, specifically protecting the α-helix content of the Aβ42 peptide. The effect was notably stronger with SPA, which also prevented or slowed down the transitions from helices into coils and β-strands.

We further analyzed the different MD ensembles and calculated the radius of gyration (R_g) to assess the compactness of the Aβ42 peptide in the three systems. We found that the free Aβ42 alone had a significantly (with p value < 10^{-4} from the t-test) broader and more skewed distribution of R_g (average R_g = 14.2 ± 4.3 Å) than in the presence of TMP or SPA (R_g = 13.3 ± 3.1 and 11.8 ± 2.1 Å, respectively; Figure 1D). This indicates that the free Aβ42 had a population of extended conformations that was not found in the presence of TMP or SPA. SPA showed a particularly strong effect on shifting Aβ42 toward more compact conformations, compared to the other two systems. Interestingly, Löhr and coworkers recently reported an aggregation inhibitor that presented the opposite effect and stabilized the extended, higher-entropy conformations of Aβ42.[67]

Effects of Ligands on the Evolution of Secondary Structure Elements Over Time. To understand the evolution of secondary structure elements in the adaptive sampling simulations, we first performed the time-based alignment and concatenation of the MDs (Supplementary Note 3 and Figure S6). We computed the evolution of the mean secondary structure content along the continuous simulation time of the aligned and concatenated simulations (Figure 1E). We observed that the different secondary structure ratios evolved quickly in the free Aβ42, decreasing for α-helices and increasing for coils and β-strands. In the presence of TMP, those values changed similarly but more slowly, while SPA induced the slowest changes. Classical MDs showed similar trends toward the apparition of coils and strands over time. However, the capacity of the small molecules to preserve helical elements was not as pronounced as in adaptive-sampling MDs (Supplementary Note 4, Figures S7−S9, Table S2). We can speculate that performing longer simulation times might result in a further decrease in the levels of α-helices and an increase of β-strands.

Conformational Analysis of Ligand Effects Using Markov State Models

Initially, we tried to construct conventional Markov state models (MSMs) to analyze the adaptive sampling simulations and characterize the conformational states of Aβ42. Different metrics and settings were tested, namely, the *RMSD* of the Cα atoms, the *secondary-structure*, the *self-distance* of all Cα atoms, and combinations of those metrics (Supporting Information and Methods). However, none of these analyses produced reliable models (see example in Figures S10−S12), so we decided to use the recently published method for MSM construction using artificial neural networks. We further extended that method with new analyses, which proved highly useful for comparing different systems and improving the interpretability of the results.

Construction of Variational Markov State Models. We approached the construction of MSMs with VAMPnet[16] by testing several lag times (25, 50, 75, and 100 ns) and different numbers of Markov states (2, 3, 4, and 5). Since we are interested in identifying the major differences among the three systems (free Aβ42, Aβ42 + TMP, Aβ42 + SPA), we prioritized the characterization of a few major macrostates rather than many microstates. For this reason, we explored only a relatively small number of states, as done previously by Löhr et al.[17] According to the implied time scales plots (Figure S13) and the Chapman−Kolmogorov tests (Figure S14), we selected τ = 25 ns as the final lag time. By evaluating the impact of the additional states on the

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

change in the frame classification (Figure S15), together with considering the transition rates for each state, we decided to use the 3-state MSM for all the studied systems. Using the selected parameters, we re-estimated the MSMs for MD simulations generated by protocol C. We first constructed 16 subsets of data by gradual addition of epochs to the training and validation data. From the models, we calculated the exact transition probabilities, mean first-passage times, and transition rates (Figure S16), as well as the respective structural propensities (Figures 2A and S17). Finally, we verified that additional data did not significantly affect the estimated implied time scales and that the size of our data sets was thus sufficient for VAMPnet training (Figure S18).

Evaluation of the Effect of Using the Soft versus the Hard Assignment. Interestingly, we found the models to be quite certain about the classification of frames into the learned states, thus diminishing the differences between the hard and soft assignment. For free Aβ42, Aβ42 + TMP, and Aβ42 + SPA, we found 99%, 99%, and 98% of the frames, respectively, to be classified into one of the states with probability higher than 95%.

Alignment of Learned States Across Systems with and without Ligands. To automatically detect similar conformational states across different systems and compare the estimated MSMs, we developed and applied a novel alignment method. This method aligns different states, by minimizing the global cost of alignment of MSM ensembles and produces alignment costs for each pair of matched states T_e (see Alignment of learned states). To distinguish truly aligned states from those without a counterpart in the other system, we considered two states as aligned only if their alignment cost was lower than the threshold $T_e = 6$ (see Supplementary Note 5). This threshold was selected empirically by comparing the visualized structures (Figure 2A), the secondary structure content, and contact maps (Figure S17) of the states proposed for mutual alignment. This approach allowed us to find two similar states between free Aβ42 and Aβ42 + TMP (states 1 and 2), and one similar state between free Aβ42 and Aβ42 + SPA (state 1; see Figure S19).

Comparison of Learned States Across Systems with and without Ligands. The evolution and kinetics of the constructed MSMs for the studied systems are shown in Figure 2, as well as a representative ensemble of structures for every state. The free Aβ42 system (Figure 2, left) was characterized by a sparsely populated source state (state 1, pink, 10% equilibrium probability), a dominant sink state (state 2, orange, 86% equilibrium probability), and a metastable transition state between them that was the least populated of all (state 3, green, 4% equilibrium probability). The kinetic roles (source and sink) were derived from the transition kinetic rates and the mean first-passage times, and from the secondary structure contents of each state. Hence, the source state (1, pink), with the structural content most resembling the starting NMR structure (ca. 58% coil, 40% α-helices, and 2% β-strands), converted fast into the sink state (2, orange; $T_M = 2.6$ μs), and could be reasonably formed from the transition state (3, green; $T_M = 14.6$ μs). The sink state was characterized by disorder, with the highest contents of coils and β-strands and the lowest contents of α-helices. The transition state represented a middle point in terms of secondary structure content, and it converted faster into the source or sink states than it was formed. This kinetic ensemble is in good agreement with the results previously described by Löhr et al. for the monomeric Aβ42, namely, in terms of microsecond transition times between the states, the presence of one dominant state that was mainly disordered, and the inexistence of long-lived folded states.[17]

According to our alignment method, the Aβ42 + TMP system (Figure 2, center) had counterparts in the free Aβ42, namely, the disordered sink state (orange) and the helical-rich source state (pink). The equilibrium probability of the sink was slightly reduced (state 2, orange, 75%), and the more helical source was slightly increased (state 1, pink, 12%). A new transition state appeared in this system (lime, 14% equilibrium probability), with intermediate secondary structure propensities and a higher α-helical content compared to the transition state in the free Aβ42. Perhaps for this reason, the cost of their alignment was above the selected threshold (Figure S19), and the state was thus considered a newly formed state. This was supported by the visualized structures (Figure 2A) and the detailed secondary structure and contact maps for the respective states (Figure S17). Overall, the MSM ensemble for the Aβ42 + TMP system showed higher variability of the equilibrium distribution. Interestingly, the kinetics of this system was rather similar to that of the free Aβ42 but significantly slower, generally with higher transition mean-times. As in the case of the free Aβ42, the formation rates of the disordered sink state 1 were higher than its conversion into the other states.

The simulations of Aβ42 + SPA produced a clearly distinct MSM (Figure 2, right), with the equilibrium distribution more uniform than in the other two systems. Furthermore, the confidence intervals of the equilibrium probabilities were even wider, and the free energy landscape appeared more homogeneous, implying that the states in Aβ42 + SPA were less clearly defined compared to the other systems. According to our alignment procedure, only the source state of Aβ42 + SPA (state 1, pink, 46% equilibrium probability) found its counterpart in the free Aβ42 system. The secondary structure content of this state was similar to the corresponding one in the free Aβ42 and the starting NMR structure (61% coil, 36% α-helices, and 3% β-strands). It is noteworthy how the addition of SPA disrupted the kinetic ensemble: the remaining two states differed significantly from those of the free Aβ42, as demonstrated by the high alignment costs (Figure S19) and the secondary structure contents. Strikingly, in contrast with the previous two systems, the unstructured sink state disappeared as the two new unmatched states with high α-helix contents occurred. This was especially the case of state 2 (blue, 23% equilibrium probability), which contained more α-helices (48.7%) and fewer coils (50.6%) than the initial NMR structure (42.1% and 57.9%, respectively). This state 2 evolved over time into state 3 (purple, 31% equilibrium probability; Figure 2B), which had the fastest conversion to the source state, and thus could hardly be considered a "sink" state. All three states interconverted between each other rather quickly, with T_M values in the low microsecond range, suggesting a dynamical metastable equilibrium around the source state. All these observations are supported by the study of the time-evolution of the states in the different simulations (Supplementary Note 6, Figure S20).

We also calculated the radius of gyration (R_g) of the different states (Figure S21). The free Aβ42 system presented the largest dispersion of R_g values, with its states showing peaks at higher values, while for Aβ42 + SPA, all the states displayed low R_g dispersion and peaks at low values (between 10.6 and 11.0 Å). This observation is in agreement with the R_g calculations on the global MD ensembles, discussed above, suggesting that the systems differ intrinsically in their degrees of structural order and compactness.

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

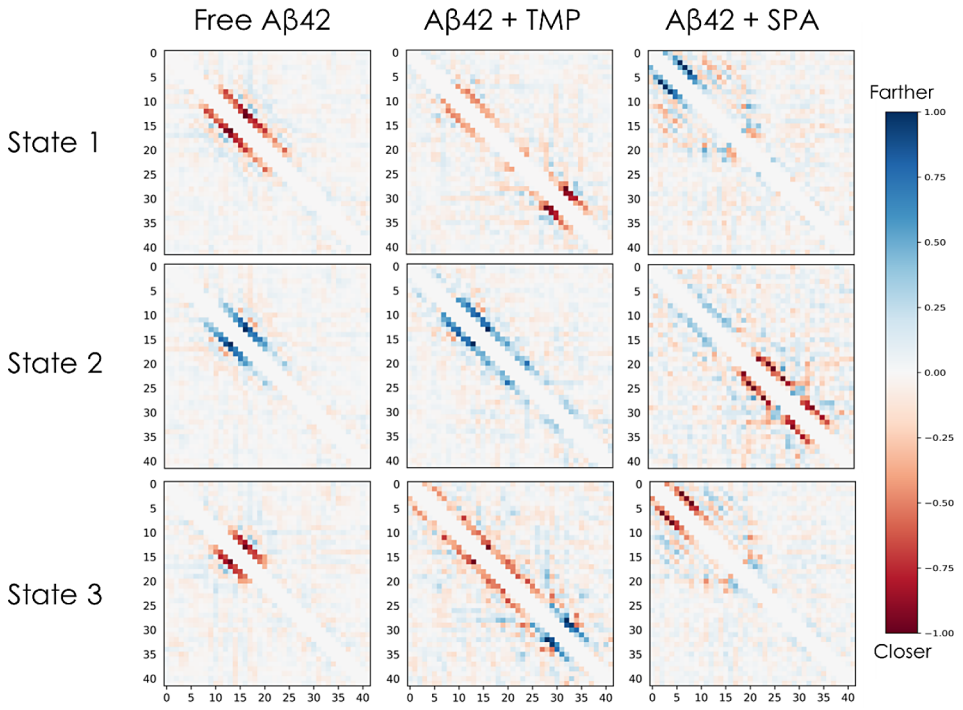


Figure 3. Gradients of the state assignment probabilities of the learned variational Markov state models. Each 42 × 42 heatmap shows the ensemble-averaged gradients of the model probabilities for the corresponding system and state with respect to the input inter-residue Cα distances. The color indicates how the probability of the particular state would change for an input frame if the distance between the particular pair of residues increased: blue indicates that the probability of the state assignment would increase if the distance between the Cα atoms increased whereas red indicates that the probability would increase if that distance decreased. The presented visualizations correspond to ensemble-averaged gradients evaluated and aggregated over 10,000 randomly selected simulation frames. Columns: MSMs for the free Aβ42 (left), Aβ42 + TMP (middle), and Aβ42 + SPA (right) systems. Rows: states 1 (top), 2 (middle), and 3 (bottom) of each model.

Characterization of Learned Conformational States via Network Gradients. To better understand the differences between the states in each MSM, we attempted to interpret the molecular features that were determinant to the assignment of each state. For that, we visualized the ensemble-averaged gradients of the state assignment probabilities obtained from the learned neural network models. Figure 3 shows that the elements near the diagonal were the most important for the classification into the respective states. As our representation does not consider the distances of the residues to their first and second neighbors in the primary sequence, the colored pixels along the empty diagonal in each heatmap correspond to the distances of the residues to their third neighbors in the sequence. Since this roughly corresponds to the length of one turn in an α-helix (ca. 4 residues), the consistently red or blue color of the two subdiagonals closest to the white diagonal to the presence or absence of helices, respectively. This interpretation is also supported by the average secondary structure content per residue and the average contact maps (Figure S17).

For the free Aβ42 system, the peptide residues around positions 10−25 seem to be crucial for the state classification. The results in the free Aβ42 state 1 heatmap imply that if the red colored residues in this region got closer to their third and fourth sequence neighbors in a particular snapshot, the probability of classifying that snapshot into state 1 (source state) would increase. This means that state 1 prefers a helical conformation in this region. On the contrary, the "state 2" heatmap shows that the probability of classification into state 2 would increase if the blue-colored residues in this region got farther from their third and fourth sequence neighbors, i.e., state 2 (sink state) prefers disorder in this region. The classification into state 3 relies on the same region (residues 10−25) but is split into two parts: residues 13−19 (red) and the rest (gray). This implies that state 3 (transition state) prefers a short helix only in residues 13−19.

For the Aβ42 + TMP system, the corresponding heatmaps show that the presence (red) or lack (blue) of a helix at positions 29−36 are important for distinguishing between states 1 and 3, respectively, while state 2 can be discriminated based on the lack of a helix at positions 10−25. For Aβ42 + SPA, the lack (blue) or presence (red) of a helix at positions 3−12 is relevant for discriminating states 1 and 3, respectively. State 2 differs by the presence of two helices at positions 20−27 and 30−35 (red) as well as by long distances between residues in positions 10−17 (blue pattern).

The states can be compared in more detail by evaluating the gradients on sets of state-specific frames (Figure S22). Conversely, the gradient matrices can also be aggregated by residue into simpler but still very informative plots (Figure S23). These can help to readily assess the most influential regions defining the states, compare different systems, and potentially cross-validate the results with other residue-based analyses, e.g., from experimental data (see below).

Molecular Interactions

Ligand−Peptide Interactions. The interactions of TMP and SPA with Aβ42 were assessed by the linear interaction energy (LIE)[45] and computed for all the 100 ligand molecules with each peptide residue during the adaptive sampling simulations. For this purpose, all the snapshots in the simulations were used. The electrostatic component ($\Delta G_{\text{bind}}^{\text{elec}}$) dominated the interactions formed by Aβ42 with both TMP and SPA, overshadowing the van der Waals component (Figure

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

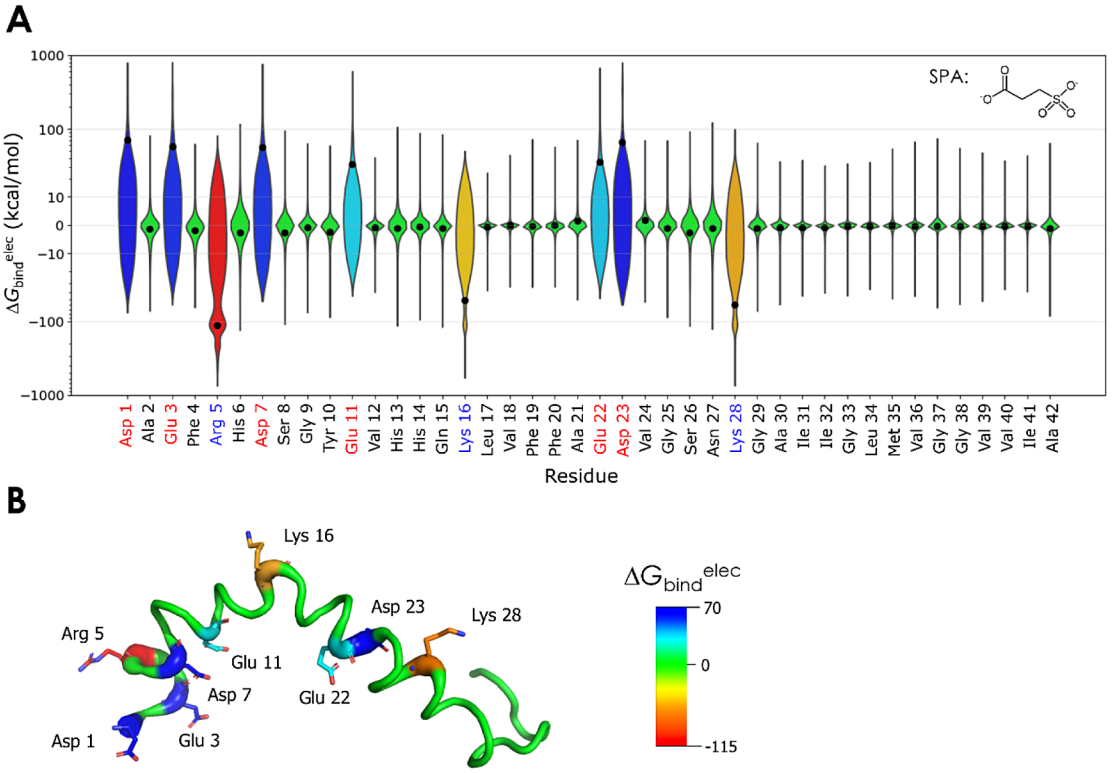


Figure 4. Interactions of SPA with Aβ42 studied by molecular dynamics. A) Violin plot of the binding energy of SPA with each residue of Aβ42. The electrostatic component (ΔG_{bind}^{elec}) was calculated for all the 100 molecules in every snapshot of the adaptive simulation of Aβ42 + SPA. The plot shows the distribution of the energy values; the black dots show the mean values; the *y*-axis uses a quasi-logarithmic scale based on the inverse hyperbolic sine to highlight the higher absolute values. The residue labels are colored by charge: black for neutral, blue for positive, and red for negative. The chemical structure of SPA is shown in the upper-right corner. B) Structure of Aβ42 with the main interacting residues. Aβ42 is shown as the putty cartoon, and the main interacting residues are represented by sticks (structure from PDB ID 1Z0Q). The colors reflect the mean ΔG_{bind}^{elec} (in kcal/mol) and range from the most positive (blue) to the most negative (red) values obtained for SPA.

S24). Those interactions were, on average, much stronger with the charged residues (Figures 4 and S25). This was expected, considering that both TMP and SPA bear two charges at physiological pH, separated by only a short alkyl chain (positive and negative charges in TMP, and two negative charges in SPA). SPA showed both attractive and repulsive interactions (respectively, positive and negative ΔG_{bind}^{elec}; Figure 4); TMP showed mostly favorable interactions (negative ΔG_{bind}^{elec}; Figure S25). The absolute mean interaction energies were also higher with SPA (from −114 to 71 kcal/mol) than with TMP (from −50 to 0 kcal/mol). Moreover, the interactions were highly variable due to the rapid exchange of the TMP and SPA molecules, which formed unspecific short-lived interactions with Aβ42. This explains the large populations of snapshots with a lower range of interaction energies and the smaller populations of snapshots with strong interactions with the charged residues.

Although TMP and SPA have quite similar structures, the global effects of SPA on Aβ42 were more striking than those of TMP. This is probably due to the fact that SPA has a double negative charge, which reverses the charge of positive groups it interacts with. Conversely, TMP is zwitterionic (with positive and negative charges) and thus preserves the charge around the interacting residues. A comprehensive comparison of the properties of TMP and SPA and their effects on the simulations of Aβ42 is presented in Table S3.

Intramolecular Interactions of Aβ42. The interactions within the Aβ42 peptide were calculated using the molecular mechanics/generalized Born solvent accessible surface area (MM/GBSA) method.[47,48] Interestingly, the electrostatic energy prevailed over the van der Waals, but the polar solvation energy outweighed all the other contributions to the internal free energy of Aβ42 (Table S4 and Supplementary Note 8). The peptide was more stable (lower mean total free energy) in the presence of TMP or SPA than alone in solution. This stabilization was mainly due to the solvation energy, which indicates a higher exposure of polar residues to the solvent than the free Aβ42. This effect is concomitant with an increase of the internal hydrophobic contacts in the presence of TMP or SPA, which is consistent with an increase of the compactness of the peptide, according to the R_g values reported above (Figure 1D). Intramolecular salt bridges E22-K28 and D23-K28 have been reported to be important for the conformational transition, oligomerization, and toxicity of Aβ42.[68,69] Analysis of the three ensembles showed that these salt bridges occurred considerably less often in the presence of TMP than in the free Aβ42, and even less with SPA (Figure S26). This suggests a lower propensity of Aβ42 to form oligomers in the presence of those small molecules. Due to their charged moieties, TMP and SPA induce electrostatic dispersion on the residues involved in the salt bridges, thus weakening those interactions (Table S3). Similar observations have previously been reported for apolipoprotein E (ApoE) interacting with SPA.[70]

Experimental Validation

To validate our computational findings described above, we experimentally characterized the conformations of N-methionine-Aβ42 (N-Met-Aβ42) alone and in the presence of TMP

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

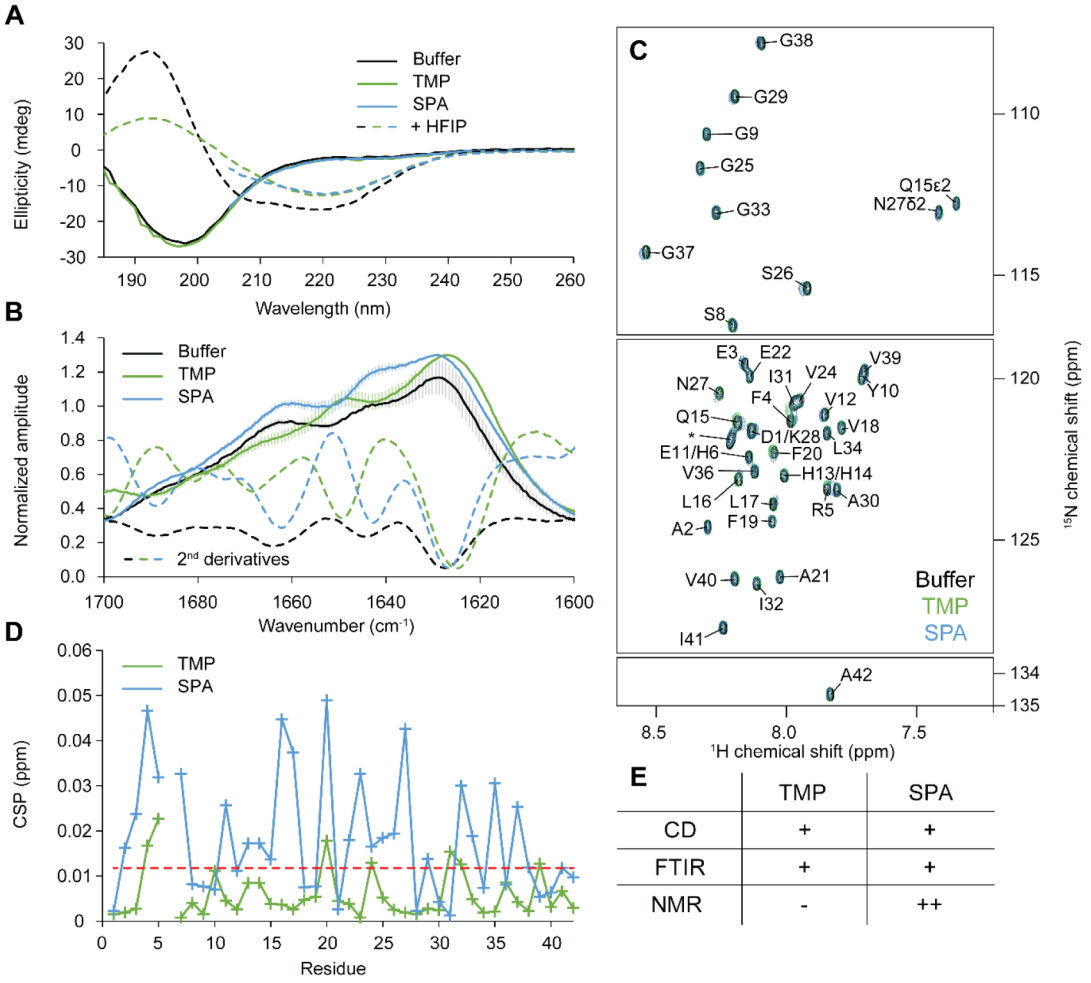


Figure 5. Experimental validation of computational data using biophysical techniques. A) Circular dichroism spectra of Aβ42. N-Met-Aβ42 (37 μM) was studied in the absence (black) or presence of a 1000-fold excess of TMP (green), SPA (blue), or 20% HFIP (dashed curves). The curves for SPA were trimmed below 205 nm to remove the signal from SPA. B) FTIR spectra of Aβ42. N-Met-Aβ42 (60 μM) was studied in the absence (black) or presence of a 1000-fold excess of TMP (green) or SPA (blue). The bars represent the standard deviations from successive acquisitions. The second derivatives are drawn as dashed curves. Offset was shifted to improve readability. C) NMR analysis of Aβ42. ^{1}H−^{15}N HMQC NMR spectra of ^{15}N-labeled N-Met-Aβ42 were determined alone (black, 69 μM) and in the presence of a 1000-fold excess of TMP (green, 58 μM) or SPA (blue, 55 μM). Assignment is given for free N-Met-Aβ42 (black); the assignment of His6 was ambiguous, thus no CSP was calculated for this residue. D) NMR chemical shift perturbation (CSP) of Aβ42. N-Met-Aβ42 in the presence of a 1000-fold excess of TMP (green), or SPA (blue) with respect to the free N-Met-Aβ42. The red dashed line represents the threshold for significance, taken as the standard deviation of all CSPs. E) Summary of the effects of small molecules on Aβ42 conformations studied by three different biophysical techniques: - indicates that no significant effect was detected, + indicates a mild effect, and ++ a stronger effect.

and SPA. The presence of N-terminal methionine was necessary for the Aβ42 recombinant expression and does not influence its aggregation behavior. This is demonstrated by the routine use of N-Met-Aβ42 in aggregation studies.[71,72] Circular dichroism (CD) of N-Met-Aβ42 in aqueous buffer revealed that the peptide was mainly disordered (68% of coils, 29% of β-strands, and 3% of α-helices; Figures 5A and S27A). To replicate the NMR structure obtained in 20% (v/v) of hexafluoroisopropanol (HFIP), used herein as the starting conformation for the computational analysis, we titrated the N-Met-Aβ42 with increasing concentrations of HFIP. At 20% HFIP, the secondary structure content of N-Met-Aβ42 was heavily changed in favor of the α-helices, in agreement with the literature[59] (Figure S28). We repeated the titrations in the presence of a 1000-fold excess of TMP or SPA. In all cases, no major changes in the CD spectra were induced by the small molecules during the titrations (Figures S27A and S28). N-Met-Aβ42 remained mostly disordered at 0% HFIP and had almost similar helical and strand content at 20% HFIP, independently of the presence of TMP or SPA. This is not in agreement with the computational results, which predicted a significant increase of the helical content of Aβ42 with the small molecules, especially with SPA.

To determine whether the molecules induced subtle changes in secondary structure that are below the resolution limit of CD spectroscopy, we analyzed the N-Met-Aβ42 in buffer and in the presence of the small molecules using Fourier-transformed infrared spectroscopy (FTIR). Based on the secondary structure deconvolution of the amide I bands,[73] the FTIR spectra of free N-Met-Aβ42 and N-Met-Aβ42 + SPA showed fingerprints from both helical (peak at around 1660 cm^{-1}) and strand contributions (peak below 1650 cm^{-1}) (Figures 5B and S27B and S29). At 1000-fold excess of TMP, a shift of the peak wavenumbers was observed (Figure 5B). The spectrum for N-Met-Aβ42 + TMP had one peak centered around 1650 cm^{-1} instead of 1660 cm^{-1}, which might suggest more random conformation (coils) of N-Met-Aβ42 in the presence of TMP

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

compared to the free peptide. Nonetheless, the large overlap of the two peaks casts doubts on such interpretations. Further remarks on differences in secondary structure propensities are discussed in Supplementary Note 9.

To gain deeper insights into conformational changes of N-Met-Aβ42 upon the addition of the small molecules, we employed nuclear magnetic resonance (NMR). The $^{1}H-^{15}N$ HMQC spectral fingerprint of N-Met-Aβ42 revealed a narrow distribution in $\delta(^{1}H)$ of the backbone amides (from 7.5 to 8.5 ppm), a characteristic of intrinsically disordered peptides (Figures 5C and S30). Using $^{1}H-^{1}H$ NOESY and $^{1}H-^{1}H-^{15}N$ NOESY-HMQC spectra, we assigned the spectral fingerprint and computed the secondary structure propensities using chemical shift indexing.[74,75] This method is based on the published NMR statistics, where each residue is expected to have a chemical shift within a certain region of the spectrum that is a function of its local secondary structure. The resulting global secondary structure propensity was much higher in α-helices than what was previously obtained by CD (29.6% vs 3%, respectively; Figure S27A,C). The secondary structure probabilities of the different residues showed the highest β-strand propensity for the C-terminal tail, and the highest helical propensity of residues 15−25 (Figure S27D). This is in agreement with the results from our simulations for the free Aβ42 (Figure 1C). We titrated N-Met-Aβ42 with increasing concentrations of TMP or SPA, up to a 1000-fold excess (Figures 5C and S30) and measured the chemical shift perturbation (CSP) in the $^{1}H-^{15}N$ HMQC spectral fingerprint (Figure 5D). The threshold for the CSP significance was taken as the standard deviation of all chemical shifts.[76] Only small CSPs were observed when adding SPA, which were not sufficient to indicate a shift in the global secondary structure (Figure S27C). This is not unprecedented, as others have also reported minimal changes in the NMR spectrum of Aβ42 upon the binding of small molecules.[77] CSP was observed across most of the peptide sequence in the presence of SPA, namely in regions 2−7, 11−17, 20, 22−27, and 32−37. Strikingly, these regions correspond to peptide ranges that emerged in the gradient-based analysis of learned conformational states (namely, regions 3−12, 10−17, 20−27, 30−35; Figures 3 and S23). In the presence of SPA, close distances (structural order) between residues 2−7 are characteristic of the transition between states 1 (pink in Figure 2A) and 3 (purple in Figure 2A). Similarly, close distances in residues 22−27 and 32−37 are characteristic hallmarks of state 2 (blue), which is also determined by long distances (disorder) in the range 11−17. It is noteworthy that states 2 and 3 in this system are distinctively different from the other two systems. Thus, gradient-based analysis of learned states was able to pinpoint similar conformational events as the ones captured by NMR. Moreover, regions 22−27 are neighboring the salt bridges between 22 and 28 and 23−28, which are relevant to the conformational transition, oligomerization, and toxicity of Aβ42,[68,69] as pinpointed in the Intramolecular interactions of Aβ42 section.

Finally, we assessed the fibril formation of N-Met-Aβ42 using the well-known thioflavin T (ThT) fluorometric assay with and without the small molecules. We found that neither TMP nor SPA seemed to significantly reduce the N-Met-Aβ42 fibril formation rates, as observed by other groups.[78] This is in contrast with HFIP, which is a known solubilizing agent of Aβ42 and a crude membrane mimetic[59] (Figure S31). In fact, a change in the CD spectrum was observed in the presence of HFIP and either TMP or SPA (Figures 5A,E and S28).

DISCUSSION

Alzheimer's disease drug candidate TMP and its metabolite SPA are thought to modify the conformational dynamics of the Aβ42 peptide and decrease its propensity to form toxic oligomers.[33,34] The conformational diversity of Aβ42 has been previously explored by exploiting the variational approach to Markov processes in VAMPnets[16] to construct Markov state models (MSMs), to better capture the slowest processes in MD simulations.[17,67] However, the exact mechanism of action of TMP, and particularly SPA, on Aβ42 was still unclear. To fill this gap, we first applied the variational approach to Markov processes on adaptive sampling MD simulations using VAMPnets,[17] and then ran our newly developed comparative Markov state analysis (CoVAMPnet) pipeline to (1) align the learned conformational states across ensembles of different MSMs, and (2) based on the learned VAMPnet gradients, to characterize these states by the inter-residue distances. The CoVAMPnet alignment method proved a powerful approach to (i) quantitatively compare the different conformational states of Aβ42, (ii) identify which states were preserved across different systems, and (iii) identify which states were unique. The CoVAMPnet gradient-based characterization of the learned ensembles of Markov states utilizes the end-to-end differentiability of the neural network-based MSMs, i.e., a property that the conventional methods for MSM estimation lack. The analysis of gradients allowed us to reason, at the molecular level, which residues are responsible for the assignment to a specific state obtained from the variational Markov state analysis. We expect these newly developed methods, i.e., (i) the alignment of ensembles of variational Markov state models across different systems, and (ii) the gradient-based characterization of learned states, to become valuable for studying the impact of small molecules on the conformational dynamics of intrinsically disordered proteins and peptides.[79,80]

The newly developed analyses were applied to MD simulations of Aβ42. It is known that the sampling protocol (namely, the force field, the length of the simulations, the adaptive metrics, and the simulation method) can highly influence the global results.[81,82] This is largely due to the intrinsically disordered nature of the Aβ42 peptide, which has a rather shallow energy landscape with many energy minima separated by small energy barriers.[6,79] For this reason, the conformational sampling of Aβ42 remains a challenge.[81,82] Starting from a helix-rich Aβ42 structure, biased toward the conformation in the membrane environment[59,83] (PDB ID 1Z0Q), we identified the most suitable adaptive protocol to simulate Aβ42, according to the secondary structure contents expected in aqueous phase (dominated by coils and β-strands). In this way, we sampled the conformations and transitions occurring immediately after the release of Aβ42 from the transmembrane region to the extracellular fluid. After approximately 64 μs of adaptive MDs, the free Aβ42 diverged substantially from the initial structure, increasing the total amount of random coils and β-strands (as expected) while decreasing the ratio of α-helices, and became closer to experimental values and previous reports.[59,66,84] We identified two regions of Aβ42 that were more prone to form β-strands (mainly residues 2−8, 17−20, and 30−41). The MSMs learned from the variational Markov state analysis revealed that the most populated state of Aβ42 is highly disordered and contains some β-strands. This state is in equilibrium with two other states with higher contents of α-helices, but still bearing mainly coils. These

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

results are in good agreement with recent reports by Löhr et al., obtained from much longer simulation times (315 μs).[17]

The presence of TMP and SPA shifted Aβ42 toward more structured conformations (less coils and higher content of α-helices) and reduced the propensity of regions 2−8, 17−20, and 30−41 to form β-strands. This behavior is similar to what has previously been reported for some aggregation inhibitors[85−87] and is in contrast with some others.[67] The variational Markov state analysis showed that TMP and SPA induced a change in the equilibrium distribution and interconversion rates of the Aβ42 conformational states. SPA exerted a much stronger effect, stabilizing new conformational states that were richer in α-helices than in the other systems. Since β-strand structures lead to the formation of β-sheets, the precursors that prompt the oligomerization and fibrillation of Aβ,[2,4,5] these results suggest the potential of TMP and SPA to inhibit or delay both processes. This can be particularly relevant if we consider previous studies suggesting that oligomers may start by the formation of β-hairpins made of β-strands of residues 16−24 and 28−35,[88] and that α-helices in regions 10−21[84] or 17−21[89] may prevent the formation of higher oligomers and aggregation. While Aβ42 is preserved in its monomeric form, it should not be harmful until it is cleared from the brain, namely, through the binding to apolipoprotein E (ApoE).[90−92] Our simulations suggest that TMP and SPA may affect the conformational equilibrium of Aβ42 in the brain and prolong its monomeric soluble state, thus allowing to extend the effective time of the clearance mechanisms. Due to their charged terminal moieties, both TMP and SPA formed mainly electrostatic interactions with the charged residues of Aβ42. These interactions were nonspecific and short-lived, but they promoted the exposure of polar residues (similar to a "solvation" effect), induced Aβ42 to be more compact, and weakened intramolecular electrostatic interactions (as previously observed[70]). Importantly, some of the intramolecular salt bridges (E22-K28, D23-K28) considered to promote the formation of β-sheets, aggregation, and neurotoxicity of Aβ42[68,69,88] were disrupted by the presence of those small molecules. The difference between TMP and SPA in terms of charge distribution (zwitterionic and doubly negative, respectively) is likely the main factor responsible for the overall stronger effects of SPA (see Table S3). The reasons for the stronger stabilization of α-helices by SPA are not clear. However, it may be due to competition of the densely charged ligand with the water molecules, which may lead to preventing their destabilizing action on the peptide, as previously described for a series of ions at higher concentrations.[93]

The CoVAMPnet algorithm developed for identification of structural features in the learned variational MSMs based on network gradients proved useful. We were able to identify the peptide regions with preferential order or disorder in the different states and pinpoint major differences across the different systems. Remarkably, this analysis showed good agreement with the CSPs in the NMR spectra, correctly predicting the peptide regions most affected by the presence of SPA. These computational findings were in agreement with previous studies involving Aβ, TMP, and SPA, namely: (i) the unstructured nature of the peptide, (ii) shift of the Aβ42 conformations by those ligands toward more compact structures, (iii) reduction of the β-strand propensity, and (iv) nonspecific interactions with charged residues.[33−35] Reports also have shown that both small molecules can interact with the soluble Aβ40 or Aβ42, change their dominant conformation, inhibit the formation of oligomers and fibrils, decrease the Aβ-induced neuronal cell death,[25,33,34] and have protective effects *in vivo*.[30]

We applied several experimental biophysical techniques to validate the computational results described above. Although the experimental outcomes showed only a mild influence of both TMP and SPA on N-Met-Aβ42, several relevant effects were observed (Figure 5E). FTIR revealed slight changes in secondary structure upon the addition of TMP, suggesting higher coil conformation propensity for the peptide. On the other hand, NMR showed a stronger impact of SPA on the ^{1}H−^{15}N NMR spectral fingerprint of N-Met-Aβ42, indicating either direct ligand−peptide interactions, subtle changes in secondary structure, or both. Strikingly, these perturbations were observed in the same peptide regions highlighted by our network gradient analysis. TMP did not produce significant CSPs. Altogether, these results suggest a stronger effect of SPA on Aβ42 than TMP. Yet, the fibril formation kinetics of N-Met-Aβ42 seemed unaffected by TMP or SPA.

The experimental results corroborated several computational findings: (i) the intrinsically disordered Aβ42 interacts with TMP or SPA molecules through many weak interactions, (ii) these interactions induce conformational changes on the peptide, (iii) SPA has stronger influence on Aβ42 than TMP, and (iv) the regions affected could be predicted by the gradient analysis of the learned state probabilities. On the other hand, not all the predictions from our molecular modeling were confirmed experimentally: (i) Aβ42 showed higher β-strand content compared to the computational results, and (ii) TMP and SPA did not change significantly the global secondary structure propensities of Aβ42 and did not prevent fibril formation. The differences in the time scales sampled by the simulations (microseconds) and the experiments (minutes/hours) and the peptide concentration effects may have contributed to this discrepancy. Moreover, the membrane mimetic HFIP modulated the impact of TMP and SPA on N-Met-Aβ42, which may deserve further investigation. An extended discussion of these phenomena is provided in Supplementary Note 10. In further works, the development of *specific binders* able to stabilize α-helices in the regions of Aβ42 mentioned above could be a better approach for designing drugs targeting the neurotoxic oligomerization of Aβ. The interaction of TMP and SPA with other proteins participating in the amyloid cascade,[94] which has been demonstrated in the case of ApoE (especially ApoE4),[70,95] should also be considered and evaluated in future studies. Particularly, we have recently shown the strong impact of TMP and SPA on ApoE4, shifting its structure and properties toward those of ApoE3 and significantly reducing its aggregation.[70] The observation of significant effects of TMP and SPA on ApoE4, but weaker ones on Aβ, is also important in the context of a recently published paper reporting the existence of five subtypes of AD.[96] All subtypes showed a higher prevalence of the APOE e4 genotype, while only selected ones are characterized by modified levels of Aβ. In the future, it will be interesting to relate this information to the data collected within phase 3 of clinical trials, which will reveal the efficacy of TMP on the different subtypes.

In summary, in this work, we introduced CoVAMPnet to compare and interpret learned MSMs across different systems. CoVAMPnet is composed of two methods: (i) the alignment of Markov state models and (ii) characterization of learned conformational states based on network gradients. The CoVAMPnet approach can be applied to study and compare any related molecular systems and extract valuable information.

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

It can be especially useful to study the impact of small molecules on intrinsically disordered proteins and peptides, whose quantitative analysis can be extremely difficult. Furthermore, we applied CoVAMPnet to study molecular effects of potential anti-Alzheimer's drugs on hallmark peptide Aβ42. Our computational results suggested that TMP, and particularly SPA, in short dynamic time windows can stabilize structured helical conformations of Aβ42, potentially preventing its oligomerization. *In vitro* validation confirmed the stronger impact of SPA on Aβ42 and the peptide regions affected by this molecule. However, in long time ranges, the global secondary structure was not significantly modified, neither was the Aβ42 aggregation propensity under the experimental conditions. This suggests the potential existence of additional mechanisms, such as the suppression of ApoE4 aggregation,[81] contributing to the mode of action and the clinical effects of TMP/SPA in AD besides the conformational shift of Aβ42.

■ ASSOCIATED CONTENT

Data Availability Statement

The sampled stripped trajectories and intermediate data, including the trained neural network weights, are available at https://data.ciirc.cvut.cz/public/projects/2023CoVAMPnet/. The code and example data are available at https://github.com/KoubaPetr/CoVAMPnet.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.4c00182.

Detailed materials and methods, complementary to the concise descriptions in this main text, Supplementary discussions (Notes 1−10), Supplementary figures: structure of Aβ42 (Figure S1), comparison of the Amber ff14SB and CHARMM36m force fields (Figures S2 and S3), comparison of different adaptive sampling protocols (Figures S4 and S5), temporal alignment and concatenation of the adaptive sampling and classical MDs (Figures S6−S9), conventional Markov state model analysis (Figures S10−S12), variational Markov state analysis using VAMPnets (Figures S13−S18), comparative Markov state model analysis (Figure S19), time-based evolution of the states (Figure S20), radius of gyration by state (Figure S21), characterization of learned conformational states via network gradients (Figures S22 and S23), interactions of Aβ42 with the small molecules (Figures S24−S26), experimental validation (Figures S27−S31), Supplementary tables: comparison of computational protocols for the simulation of Aβ42 (Table S1), analysis of classical MDs (Table S2), summary of the effects of small molecules (Table S3), and intramolecular interactions of Aβ42 (Table S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Stanislav Mazurenko − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*; Email: mazurenko@mail.muni.cz

Josef Sivic − *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Praha 6 160 00, Czech Republic*; Email: josef.sivic@cvut.cz

David Bednar − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*; orcid.org/0000-0002-6803-0340; Email: 222755@mail.muni.cz

Authors

Sérgio M. Marques − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*

Petr Kouba − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Praha 6 160 00, Czech Republic; Faculty of Electrical Engineering, Czech Technical University in Prague, Praha 6 166 27, Czech Republic*; orcid.org/0000-0002-9979-4159

Anthony Legrand − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*

Jiri Sedlar − *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Praha 6 160 00, Czech Republic*; orcid.org/0000-0002-4704-3388

Lucas Disson − *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Praha 6 160 00, Czech Republic*

Joan Planas-Iglesias − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*; orcid.org/0000-0002-6279-2483

Zainab Sanusi − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*

Antonin Kunka − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*

Jiri Damborsky − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic; International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic*; orcid.org/0000-0002-7848-8216

Tomas Pajdla − *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Praha 6 160 00, Czech Republic*

Zbynek Prokop − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic;*

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic; orcid.org/0000-0001-9358-4081

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c00182

Author Contributions

S.M.M., P.K., and A.L. contributed equally to this study. CRediT: **Sérgio M. Marques** conceptualization, data curation, investigation, methodology, validation, visualization, writing-original draft, writing-review & editing; **Petr Kouba** data curation, methodology, software, validation, writing-original draft; **Anthony Legrand** data curation, investigation, validation, writing-original draft; **Jiri Sedlar** conceptualization, methodology, software, visualization, supervision, writing-review & editing; **Lucas Disson** methodology, software, visualization; **Joan Planas-Iglesias** data curation, investigation, visualization, writing-original draft; **Zainab Sanusi** investigation, validation; **Antonin Kunka** investigation, methodology, validation, writing-original draft; **Jiří Damborský** conceptualization, funding acquisition, project administration, supervision, writing-review & editing; **Tomas Pajdla** conceptualization, supervision; **Zbyněk Prokop** conceptualization, methodology, supervision, writing-review & editing; **Stanislav Mazurenko** conceptualization, data curation, methodology, supervision, writing-review & editing; **Josef Sivic** conceptualization, funding acquisition, project administration, supervision, writing-review & editing; **David Bednar** conceptualization, funding acquisition, project administration, supervision, writing-review & editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (grants ESFRI RECETOX RI LM2023069, e-INFRA CZ LM2018140, ESFRI ELIXIR CZ LM2023055, TEAMING CZ CZ.02.1.01/0.0/0.0/17_043/0009632), the European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468), the National Institute for Neurology Research (EXCELES Neuro nr. LX22NPO5107 MEYS), and the European Union (Next Generation EU, SinFonia 814418, TEAMING 857560 and ERC FRONTIER 101097822). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Petr Kouba is a holder of the Brno Ph.D. Talent scholarship funded by the Brno City Municipality and the JCMM.

■ REFERENCES

(1) Gustavsson, A.; Norton, N.; Fast, T.; Frölich, L.; Georges, J.; Holzapfel, D.; Kirabali, T.; Krolak-Salmon, P.; Rossini, P. M.; Ferretti, M. T. Global Estimates on the Number of Persons across the Alzheimer's Disease Continuum. *Alzheimer's Dementia* **2022**, *19*, 658−670.

(2) Benilova, I.; Karran, E.; De Strooper, B. The Toxic Aβ Oligomer and Alzheimer's Disease: An Emperor in Need of Clothes. *Nat. Neurosci.* **2012**, *15* (3), 349−357.

(3) Karran, E.; De Strooper, B. The Amyloid Hypothesis in Alzheimer Disease: New Insights from New Therapeutics. *Nat. Rev. Drug Discov.* **2022**, *21* (4), 306−318.

(4) Castellani, R. J.; Plascencia-Villa, G.; Perry, G. The Amyloid Cascade and Alzheimer's Disease Therapeutics: Theory versus Observation. *Lab. Invest.* **2019**, *99* (7), 958−970.

(5) Matiiv, A. B.; Trubitsina, N. P.; Matveenko, A. G.; Barbitoff, Y. A.; Zhouravleva, G. A.; Bondarev, S. A. Amyloid and Amyloid-Like Aggregates: Diversity and the Term Crisis. *Biochemistry (Moscow)* **2020**, *85* (9), 1011−1034.

(6) Bhattacharya, S.; Lin, X. Recent Advances in Computational Protocols Addressing Intrinsically Disordered Proteins. *Biomolecules* **2019**, *9* (4), 146.

(7) Saurabh, S.; Nadendla, K.; Purohit, S. S.; Sivakumar, P. M.; Cetinel, S. Fuzzy Drug Targets: Disordered Proteins in the Drug-Discovery Realm. *ACS Omega* **2023**, *8* (11), 9729−9747.

(8) Paul, A.; Samantray, S.; Anteghini, M.; Khaled, M.; Strodel, B. Thermodynamics and Kinetics of the Amyloid-β Peptide Revealed by Markov State Models Based on MD Data in Agreement with Experiment. *Chem. Sci.* **2021**, *12* (19), 6652−6669.

(9) McGibbon, R. T.; Pande, V. S. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. *J. Chem. Phys.* **2015**, *142* (12), 124105.

(10) Spiriti, J.; Noé, F.; Wong, C. F. Simulation of Ligand Dissociation Kinetics from the Protein Kinase PYK2. *J. Comput. Chem.* **2022**, *43* (28), 1911−1922.

(11) Dominic, A. J. I.; Cao, S.; Montoya-Castillo, A.; Huang, X. Memory Unlocks the Future of Biomolecular Dynamics: Transformative Tools to Uncover Physical Insights Accurately and Efficiently. *J. Am. Chem. Soc.* **2023**, *145* (18), 9916−9927.

(12) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139* (18), 184114.

(13) Suárez, E.; Wiewiora, R. P.; Wehmeyer, C.; Noé, F.; Chodera, J. D.; Zuckerman, D. M. What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein-Folding Models. *J. Chem. Theory Comput.* **2021**, *17* (5), 3119−3133.

(14) Dominic, A. J.; Sayer, T.; Cao, S.; Markland, T. E.; Huang, X.; Montoya-Castillo, A. Building Insightful, Memory-Enriched Models to Capture Long-Time Biochemical Processes from Short-Time Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120* (12), No. e2221048120.

(15) Wehmeyer, C.; Noé, F. Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics. *J. Chem. Phys.* **2018**, *148* (24), 241703.

(16) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for Deep Learning of Molecular Kinetics. *Nat. Commun.* **2018**, *9* (1), 5.

(17) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Kinetic Ensemble of the Alzheimer's Aβ Peptide. *Nat. Comput. Sci.* **2021**, *1* (1), 71−78.

(18) Ghorbani, M.; Prasad, S.; Klauda, J. B.; Brooks, B. R. GraphVAMPNet, using graph neural networks and variational approach to Markov processes for dynamical modeling of biomolecules. *J. Chem. Phys.* **2022**, *156* (18), 184103.

(19) Liu, B.; Xue, M.; Qiu, Y.; Konovalov, K. A.; O'Connor, M. S.; Huang, X. GraphVAMPnets for Uncovering Slow Collective Variables of Self-Assembly Dynamics. *J. Chem. Phys.* **2023**, *159* (9), 094901.

(20) Mardt, A.; Hempel, T.; Clementi, C.; Noé, F. Deep Learning to Decompose Macromolecules into Independent Markovian Domains. *Nat. Commun.* **2022**, *13* (1), 7101.

(21) Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear Discovery of Slow Molecular Modes Using State-Free Reversible VAMPnets. *J. Chem. Phys.* **2019**, *150* (21), 214114.

(22) Kleiman, D. E.; Shukla, D. Active Learning of the Conformational Ensemble of Proteins Using Maximum Entropy VAMPNets. *J. Chem. Theory Comput.* **2023**, *19* (14), 4377−4388.

(23) Mardt, A.; Noé, F. Progress in Deep Markov State Modeling: Coarse Graining and Experimental Data Restraints. *J. Chem. Phys.* **2021**, *155* (21), 214106.

(24) Tolar, M.; Abushakra, S.; Hey, J. A.; Porsteinsson, A.; Sabbagh, M. Aducanumab, Gantenerumab, BAN2401, and ALZ-801—the First Wave of Amyloid-Targeting Drugs for Alzheimer's Disease with

Potential for near Term Approval. *Alzheimer's Res. Ther.* **2020**, *12* (1), 95.
(25) Gervais, F.; Paquette, J.; Morissette, C.; Krzywkowski, P.; Yu, M.; Azzi, M.; Lacombe, D.; Kong, X.; Aman, A.; Laurin, J.; et al. Targeting Soluble Aβ Peptide with Tramiprosate for the Treatment of Brain Amyloidosis. *Neurobiol. Aging* **2007**, *28* (4), 537−547.
(26) Caltagirone, C.; Ferrannini, L.; Marchionni, N.; Nappi, G.; Scapagnini, G.; Trabucchi, M. The Potential Protective Effect of Tramiprosate (Homotaurine) against Alzheimer's Disease: A Review. *Aging: Clin. Exp. Res.* **2012**, *24* (6), 580−587.
(27) Zou, X.; Himbert, S.; Dujardin, A.; Juhasz, J.; Ros, S.; Stöver, H. D. H.; Rheinstädter, M. C. Curcumin and Homotaurine Suppress Amyloid-B25−35 Aggregation in Synthetic Brain Membranes. *ACS Chem. Neurosci.* **2021**, *12* (8), 1395−1405.
(28) Abushakra, S.; Porsteinsson, A.; Vellas, B.; Cummings, J.; Gauthier, S.; Hey, J. A.; Power, A.; Hendrix, S.; Wang, P.; Shen, L.; Sampalis, J.; Tolar, M. Clinical Benefits of Tramiprosate in Alzheimer's Disease Are Associated with Higher Number of APOE4 Alleles: The "APOE4 Gene-Dose Effect. *J. Prev. Alzheimers Dis.* **2016**, *3* (4), 219−228.
(29) Tian, J.; Dang, H.; Wallner, M.; Olsen, R.; Kaufman, D. L. Homotaurine, a Safe Blood-Brain Barrier Permeable GABAA-R-Specific Agonist, Ameliorates Disease in Mouse Models of Multiple Sclerosis. *Sci. Rep.* **2018**, *8* (1), 16555.
(30) Manzano, S.; Agüera, L.; Aguilar, M.; Olazarán, J. A Review on Tramiprosate (Homotaurine) in Alzheimer's Disease and Other Neurocognitive Disorders. *Front. Neurol.* **2020**, *11*, 614.
(31) Hey, J. A.; Yu, J. Y.; Versavel, M.; Abushakra, S.; Kocis, P.; Power, A.; Kaplan, P. L.; Amedio, J.; Tolar, M. Clinical Pharmacokinetics and Safety of ALZ-801, a Novel Prodrug of Tramiprosate in Development for the Treatment of Alzheimer's Disease. *Clin. Pharmacokinet.* **2018**, *57* (3), 315−333.
(32) A Phase 3, Multicenter, Randomized, Double-Blind, Placebo-Controlled Study of the Efficacy, Safety and Biomarker Effects of ALZ-801 in Subjects With Early Alzheimer's Disease and APOE4/4 Genotype, ClinicalTrials.gov ID; Clinical trial registration NCT04770220; https://clinicaltrials.gov/ct2/show/NCT04770220 (accessed 2022−07−21).
(33) Kocis, P.; Tolar, M.; Yu, J.; Sinko, W.; Ray, S.; Blennow, K.; Fillit, H.; Hey, J. A. Elucidating the Aβ42 Anti-Aggregation Mechanism of Action of Tramiprosate in Alzheimer's Disease: Integrating Molecular Analytical Methods, Pharmacokinetic and Clinical Data. *CNS Drugs* **2017**, *31* (6), 495−509.
(34) Hey, J. A.; Kocis, P.; Hort, J.; Abushakra, S.; Power, A.; Vyhnálek, M.; Yu, J. Y.; Tolar, M. Discovery and Identification of an Endogenous Metabolite of Tramiprosate and Its Prodrug ALZ-801 That Inhibits Beta Amyloid Oligomer Formation in the Human Brain. *CNS Drugs* **2018**, *32* (9), 849−861.
(35) Liang, C.; Savinov, S. N.; Fejzo, J.; Eyles, S. J.; Chen, J. Modulation of Amyloid-B42 Conformation by Small Molecules Through Nonspecific Binding. *J. Chem. Theory Comput.* **2019**, *15* (10), 5169−5174.
(36) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Cheminf.* **2012**, *4* (1), 17.
(37) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A., et al. Gaussian 09. In *Revision E.01*, Gaussian, Inc., 2009.
(38) Case, D. A.; Betz, R. M.; Cerutti, D. S.; Cheatham, III, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N., et al.; *AMBER 16*, University of California, San Francisco, 2016.
(39) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlić, A.; Quesada, M.; et al. The RCSB Protein Data Bank: New Resources for Research and Education. *Nucleic Acids Res.* **2012**, *41* (D1), D475−D482.
(40) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pKa values for protein−ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2008**, *73* (3), 765−783.
(41) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12* (4), 1845−1852.
(42) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14* (1), 71−73.
(43) Swails, J.; *ParmEd*, GitHub, Inc, 2010. https://github.com/ParmEd/ParmEd. (accessed 2018−03−08).
(44) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084−3095.
(45) Aqvist, J.; Medina, C.; Samuelsson, J. E. A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Eng.* **1994**, *7* (3), 385−391.
(46) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577−2637.
(47) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314−3321.
(48) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discovery* **2015**, *10* (5), 449−461.
(49) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *J. Nonlinear Sci.* **2020**, *30* (1), 23−66.
(50) Mardt, A.; Pasquali, L.; Noé, F.; Wu, H. Deep Learning Markov and Koopman Models with Physical Constraints. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*; PMLR, 2020; pp. 451475.
(51) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S.; Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc, 2017, Vol. *30*.
(52) MacQueen, J.Some Methods for Classification and Analysis of Multivariate Observations*Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics And Probability, Volume 1: Statistics* Le Cam, L. M.; Neyman, J. Project Euclid 1967, *5*; 281298.
(53) Bonneel, N.; van de Panne, M.; Paris, S.; Heidrich, W. Displacement Interpolation Using Lagrangian Mass Transport. *ACM Trans. Graph.* **2011**, *30* (6), 1−12.
(54) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Naval Res. Logistics Quarterly* **1955**, *2* (1−2), 83−97.
(55) Fong, R.; Vedaldi, A. Explanations for Attributing Deep Neural Network Predictions. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. In *Lecture Notes in Computer Science*, Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; Müller, K.-R.; Springer International Publishing: Cham, 2019; pp. 149167. DOI: .
(56) Simonyan, K.; Vedaldi, A.; Zisserman, A.Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, *arXiv* **2014**.
(57) Cohen, S. I. A.; Linse, S.; Luheshi, L. M.; Hellstrand, E.; White, D. A.; Rajah, L.; Otzen, D. E.; Vendruscolo, M.; Dobson, C. M.; Knowles, T. P. J. Proliferation of Amyloid-B42 Aggregates Occurs through a Secondary Nucleation Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (24), 9758−9763.
(58) Arosio, P.; Knowles, T. P. J.; Linse, S. On the Lag Phase in Amyloid Fibril Formation. *Phys. Chem. Chem. Phys.* **2015**, *17* (12), 7606−7618.
(59) Tomaselli, S.; Esposito, V.; Vangone, P.; van Nuland, N. A. J.; Bonvin, A. M. J. J.; Guerrini, R.; Tancredi, T.; Temussi, P. A.; Picone, D. The α-to-β Conformational Transition of Alzheimer's Aβ-(1−42) Peptide in Aqueous Media Is Reversible: A Step by Step Conformational Analysis Suggests the Location of β Conformation Seeding. *ChemBiochem* **2006**, *7* (2), 257−267.

(60) Shamsi, Z.; Moffett, A. S.; Shukla, D. Enhanced Unbiased Sampling of Protein Dynamics Using Evolutionary Coupling Information. *Sci. Rep.* **2017**, *7* (1), 12700.
(61) Zimmerman, M. I.; Porter, J. R.; Sun, X.; Silva, R. R.; Bowman, G. R. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *J. Chem. Theory Comput.* **2018**, *14* (11), 5459−5475.
(62) Betz, R. M.; Dror, R. O. How Effectively Can Adaptive Sampling Methods Capture Spontaneous Ligand Binding? *J. Chem. Theory Comput.* **2019**, *15* (3), 2053−2063.
(63) Kleiman, D. E.; Nadeem, H.; Shukla, D. Adaptive Sampling Methods for Molecular Dynamics in the Era of Machine Learning. *J. Phys. Chem. B* **2023**, *127* (50), 10669−10681.
(64) Man, V. H.; He, X.; Derreumaux, P.; Ji, B.; Xie, X.-Q.; Nguyen, P. H.; Wang, J. Effects of All-Atom Molecular Mechanics Force Fields on Amyloid Peptide Assembly: The Case of Aβ16−22 Dimer. *J. Chem. Theory Comput.* **2019**, *15* (2), 1440−1452.
(65) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696−3713.
(66) Harada, T.; Kuroda, R. CD Measurements of β-Amyloid (1−40) and (1−42) in the Condensed Phase. *Biopolymers* **2011**, *95* (2), 127−134.
(67) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Small Molecule Stabilizes the Disordered Native State of the Alzheimer's Aβ Peptide. *ACS Chem. Neurosci.* **2022**, *13* (12), 1738−1745.
(68) Reddy, G.; Straub, J. E.; Thirumalai, D. Influence of Preformed Asp23-Lys28 Salt Bridge on the Conformational Fluctuations of Monomers and Dimers of Aβ Peptides with Implications for Rates of Fibril Formation. *J. Phys. Chem. B* **2009**, *113* (4), 1162−1172.
(69) Chandra, B.; Bhowmik, D.; Maity, B. K.; Mote, K. R.; Dhara, D.; Venkatramani, R.; Maiti, S.; Madhu, P. K. Major Reaction Coordinates Linking Transient Amyloid-β Oligomers to Fibrils Measured at Atomic Level. *Biophys. J.* **2017**, *113* (4), 805−816.
(70) Nemergut, M.; Marques, S. M.; Uhrik, L.; Vanova, T.; Nezvedova, M.; Gadara, D. C.; Jha, D.; Tulis, J.; Novakova, V.; Planas-Iglesias, J.; et al. Domino-like Effect of C112R Mutation on ApoE4 Aggregation and Its Reduction by Alzheimer's Disease Drug Candidate. *Mol. Neurodegener.* **2023**, *18* (1), 38.
(71) Walsh, D. M.; Thulin, E.; Minogue, A. M.; Gustavsson, N.; Pang, E.; Teplow, D. B.; Linse, S. A Facile Method for Expression and Purification of the Alzheimer's Disease-Associated Amyloid Beta-Peptide. *FEBS J.* **2009**, *276* (5), 1266−1281.
(72) Thacker, D.; Sanagavarapu, K.; Frohm, B.; Meisl, G.; Knowles, T. P. J.; Linse, S. The Role of Fibril Structure and Surface Hydrophobicity in Secondary Nucleation of Amyloid Fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (41), 25272−25283.
(73) Yang, H.; Yang, S.; Kong, J.; Dong, A.; Yu, S. Obtaining Information about Protein Secondary Structures in Aqueous Solution Using Fourier Transform IR Spectroscopy. *Nat. Protoc.* **2015**, *10* (3), 382−396.
(74) Hafsa, N. E.; Arndt, D.; Wishart, D. S. CSI 3.0: A Web Server for Identifying Secondary and Super-Secondary Structure in Proteins Using NMR Chemical Shifts. *Nucleic Acids Res.* **2015**, *43* (W1), W370−W377.
(75) Borcherds, W. M.; Daughdrill, G. W. Using NMR Chemical Shifts to Determine Residue-Specific Secondary Structure Populations for Intrinsically Disordered Proteins. *Methods Enzymol.* **2018**, *611*, 101−136.
(76) Schumann, F. H.; Riepl, H.; Maurer, T.; Gronwald, W.; Neidig, K.-P.; Kalbitzer, H. R. Combined Chemical Shift Changes and Amino Acid Specific Chemical Shift Mapping of Protein−Protein Interactions. *J. Biomol. NMR* **2007**, *39* (4), 275−289.
(77) Heller, G. T.; Aprile, F. A.; Michaels, T. C. T.; Limbocker, R.; Perni, M.; Ruggeri, F. S.; Mannini, B.; Löhr, T.; Bonomi, M.; Camilloni, C.; et al. Small-Molecule Sequestration of Amyloid-β as a Drug Discovery Strategy for Alzheimer's Disease. *Sci. Adv.* **2020**, *6* (45), No. eabb5924.
(78) Habchi, J.; Arosio, P.; Perni, M.; Costa, A. R.; Yagi-Utsumi, M.; Joshi, P.; Chia, S.; Cohen, S. I. A.; Müller, M. B. D.; Linse, S.; et al. An Anticancer Drug Suppresses the Primary Nucleation Reaction That Initiates the Production of the Toxic Aβ42 Aggregates Linked with Alzheimer's Disease. *Sci. Adv.* **2016**, *2* (2), No. e1501244.
(79) Granata, D.; Baftizadeh, F.; Habchi, J.; Galvagnion, C.; De Simone, A.; Camilloni, C.; Laio, A.; Vendruscolo, M. The Inverted Free Energy Landscape of an Intrinsically Disordered Peptide by Simulations and Experiments. *Sci. Rep.* **2015**, *5* (1), 15449.
(80) Chong, S.-H.; Ham, S. Folding Free Energy Landscape of Ordered and Intrinsically Disordered Proteins. *Sci. Rep.* **2019**, *9* (1), 14927.
(81) Saravanan, K. M.; Zhang, H.; Zhang, H.; Xi, W.; Wei, Y. On the Conformational Dynamics of β-Amyloid Forming Peptides: A Computational Perspective. *Front. Bioeng. Biotechnol.* **2020**, *8*, 532.
(82) Grasso, G.; Danani, A. Molecular Simulations of Amyloid Beta Assemblies. *Adv. Phys.: x* **2020**, *5* (1), 1770627.
(83) Haass, C.; Kaether, C.; Thinakaran, G.; Sisodia, S. Trafficking and Proteolytic Processing of APP. *Cold Spring Harbor Perspect. Med.* **2012**, *2* (5), a006270.
(84) Zhou, M.; Wen, H.; Lei, H.; Zhang, T. Molecular Dynamics Study of Conformation Transition from Helix to Sheet of Aβ42 Peptide. *J. Mol. Graphics Modell.* **2021**, *109*, 108027.
(85) Shuaib, S.; Goyal, B. Scrutiny of the Mechanism of Small Molecule Inhibitor Preventing Conformational Transition of Amyloid-B42 Monomer: Insights from Molecular Dynamics Simulations. *J. Biomol. Struct. Dyn.* **2018**, *36* (3), 663−678.
(86) Liu, F.; Ma, Z.; Sang, J.; Lu, F. Edaravone Inhibits the Conformational Transition of Amyloid-B42: Insights from Molecular Dynamics Simulations. *J. Biomol. Struct. Dyn.* **2020**, *38* (8), 2377−2388.
(87) Narang, S. S.; Goyal, D.; Goyal, B. Inhibition of Alzheimer's Amyloid-B42 Peptide Aggregation by a Bi-Functional Bis-Tryptoline Triazole: Key Insights from Molecular Dynamics Simulations. *J. Biomol. Struct. Dyn.* **2020**, *38* (6), 1598−1611.
(88) Cao, Y.; Jiang, X.; Han, W. Self-Assembly Pathways of β-Sheet-Rich Amyloid-β(1−40) Dimers: Markov State Model Analysis on Millisecond Hybrid-Resolution Simulations. *J. Chem. Theory Comput.* **2017**, *13* (11), 5731−5744.
(89) Rojas, A. V.; Liwo, A.; Scheraga, H. A. A Study of the α-Helical Intermediate Preceding the Aggregation of the Amino-Terminal Fragment of the β Amyloid Peptide (Aβ1−28). *J. Phys. Chem. B* **2011**, *115* (44), 12978−12983.
(90) Tarasoff-Conway, J. M.; Carare, R. O.; Osorio, R. S.; Glodzik, L.; Butler, T.; Fieremans, E.; Axel, L.; Rusinek, H.; Nicholson, C.; Zlokovic, B. V.; et al. Clearance Systems in the Brain-Implications for Alzheimer Disease. *Nat. Rev. Neurol.* **2015**, *11* (8), 457−470.
(91) Patterson, B. W.; Elbert, D. L.; Mawuenyega, K. G.; Kasten, T.; Ovod, V.; Ma, S.; Xiong, C.; Chott, R.; Yarasheski, K.; Sigurdson, W.; et al. Age and Amyloid Effects on Human Central Nervous System Amyloid-Beta Kinetics. *Ann. Neurol.* **2015**, *78* (3), 439−453.
(92) Yamazaki, Y.; Zhao, N.; Caulfield, T. R.; Liu, C.-C.; Bu, G. Apolipoprotein E and Alzheimer Disease: Pathobiology and Targeting Strategies. *Nat. Rev. Neurol.* **2019**, *15* (9), 501−518.
(93) Bye, J. W.; Falconer, R. J. Thermal Stability of Lysozyme as a Function of Ion Concentration: A Reappraisal of the Relationship between the Hofmeister Series and Protein Stability. *Protein Sci.* **2013**, *22* (11), 1563−1570.
(94) Martens, Y. A.; Zhao, N.; Liu, C.-C.; Kanekiyo, T.; Yang, A. J.; Goate, A. M.; Holtzman, D. M.; Bu, G. ApoE Cascade Hypothesis in the Pathogenesis of Alzheimer's Disease and Related Dementias. *Neuron* **2022**, *110* (8), 1304−1317.
(95) Chai, A. B.; Lam, H. H. J.; Kockx, M.; Gelissen, I. C. Apolipoprotein E Isoform-Dependent Effects on the Processing of Alzheimer's Amyloid-β. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2021**, *1866* (9), 158980.
(96) Tijms, B. M.; Vromen, E. M.; Mjaavatten, O.; Holstege, H.; Reus, L. M.; van der Lee, S.; Wesenhagen, K. E. J.; Lorenzini, L.; Vermunt, L.;

Venkatraghavan, V.; et al. Cerebrospinal Fluid Proteomics in Patients with Alzheimer's Disease Reveals Five Molecular Subtypes with Distinct Genetic Risk Profiles. *Nat. Aging* **2024**, *4* (1), 33−47.

https://doi.org/10.1021/jacsau.4c00182
JACS Au 2024, 4, 2228−2245

Bioinformatics, 2024, **40(9)**, btae553
https://doi.org/10.1093/bioinformatics/btae553
Advance Access Publication Date: 11 September 2024
Applications Note

OXFORD

Data and text mining

BenchStab: a tool for automated querying of web-based stability predictors

Jan Velecký [1,†], Matej Berezný [2,†], Milos Musil [1,2,3], Jiri Damborsky [1,3], David Bednar [1,3,*], Stanislav Mazurenko [1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic
[2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, 612 00 Brno, Czech Republic
[3]International Clinical Research Centre, St. Anne's University Hospital, 656 91 Brno, Czech Republic

*Corresponding authors. Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5, bld. C13, 625 00 Brno, Czech Republic. E-mails: 222755@mail.muni.cz (D.B.) and mazurenko@mail.muni.cz (S.M.)

[†]Equal contribution.

Associate Editor: Macha Nikolski

Abstract

Summary: Protein design requires information about how mutations affect protein stability. Many web-based predictors are available for this purpose, yet comparing them or using them en masse is difficult. Here, we present BenchStab, a console tool/Python package for easy and quick execution of 19 predictors and result collection on a list of mutants. Moreover, the tool is easily extensible with additional predictors. We created an independent dataset derived from the FireProtDB and evaluated 24 different prediction methods.

Availability and implementation: BenchStab is an open-source Python package available at https://github.com/loschmidt/BenchStab with a detailed README and example usage at https://loschmidt.chemi.muni.cz/benchstab. The BenchStab dataset is available on Zenodo: https://zenodo.org/records/10637728

1 Introduction

Protein stability is one of the key determinants of protein applicability. Stable proteins can withstand harsh industrial conditions such as high temperatures, unfavorable pH, or the presence of denaturing agents. However, most proteins have evolved to function in relatively mild environments (Modarres *et al.* 2016). Therefore, there is a need to engineer proteins to meet the requirements of commercial applications. The laborious and costly process of experimental methods can be partially mitigated using predictive tools that provide fast and inexpensive solutions for mutation prioritization. In recent years, the rise of machine-learning techniques and the availability of experimental data have led to a plethora of predictors of the effect of mutations on protein stability with varying accuracies, strengths, and weaknesses (Planas-Iglesias *et al.* 2021).

These predictors typically predict a change of Gibbs free energy (ΔΔG) or only classify mutations as stabilizing or destabilizing. Prediction may be based on structural information or sequence alone. We distinguish four basic modes of operations: (i) analysis of molecular interactions with force-field calculations (Yin *et al.* 2007), (ii) machine learning on structure-based features (Cheng *et al.* 2006), (iii) machine learning on features derived from a sequence (Folkman *et al.* 2016) or using a language model (Umerenkov *et al.* 2023), and (iv) meta predictions combining multiple other models (Chen *et al.* 2013). Particularly the number of predictors of the third type has risen recently thanks to breakthroughs in structure prediction and large language models for bioinformatic data (Umerenkov *et al.* 2023). We can expect a further increase in the number of predictors with the emergence of very large mutational datasets collected in a high-throughput manner (Tsuboyama *et al.* 2023).

For a selection of the best tools for protein engineering and establishing new predictive methods, proper and independent benchmarking is crucial. However, the large number of existing tools makes their comprehensive evaluation challenging. On the one hand, such evaluation can prove difficult due to the potential overlaps between training and test datasets, various formats of the input data, and provided outputs. On the other hand, a majority of machine learning predictors are only available as web services with limited input size, variable waiting times, and occasional downtimes, thus making a large-scale analysis a troublesome task.

Here, we present BenchStab, a freely available Python package for the swift execution of calculations on web-based predictors and collection of results. Our package currently implements 19 web-based computational tools that we evaluated on the independent dataset (Velecký *et al.* 2024) derived from FireProt^{DB} (Stourac *et al.* 2021).

BenchStab is fully modular, facilitating the integration of new web tools. We offer a straightforward solution for a fast

Received: 9 May 2024; **Revised:** 2 August 2024; **Editorial Decision:** 5 September 2024; **Accepted:** 10 September 2024

and effective benchmarking of well-established and future tools for predicting the effect of mutations on protein stability. BenchStab represents a significant step toward a comprehensive evaluation of computational tools, identification of their limitations, and further advancement of the field of stability prediction using machine learning. We believe that our tool will be particularly useful to the machine learning community, as BenchStab may eliminate some barriers to entry into the competition of stability change prediction.

2 Implementation

We developed BenchStab as a Python library with a command-line interface, fully automating the process of submitting requests to protein stability predictors and retrieving the results. The standalone application consists of multiple clients for distinct web-based predictors and allows adding new predictor clients through its framework, which comprises two main modules: (i) input data preprocessing and (ii) predictor client implementations (Fig. 1). The predictors upon a point mutation may be both classifiers and regression tools. The robustness of our application is proven by an automated test suite of 61 different unit tests. These tests also facilitate future application extensions with new predictors or other improvements.

Every BenchStab run involves preprocessing the input data using the pandas library (McKinney 2010). The input contains the list of mutations defined within a single file that adheres to a fixed column structure. The application accepts common column separators (commas, semicolons, tabs, spaces). Each row may define the target protein by a Protein Data Bank (PDB) or UniProt accession code, PDB file, FASTA file, or raw sequence. Users can also define specific temperature and pH values as per-row optional parameters so the values are forwarded to predictors that support them. Then, the tool performs cascade data acquisition to query each predictor with its required input, e.g. by retrieving a sequence for an entry specified by a PDB code for sequence-only tools. Where needed, SIFTS JSON API (Dana *et al.* 2019) is employed to map a PDB chain to UniProt and altogether with RCSB API (Rose *et al.* 2021), a correct mutation position in the sequence is calculated addressing PDB artifacts, such as insertion codes or expression tags. In the case of PDB files, the sequence is extracted directly from the file using Biopython. The integrity of submitted proteins, chains, and mutations is checked during preprocessing to ensure the predictors are not queried with faulty requests.

A client for a new predictor can be added using the adaptable framework implemented in our tool by following the steps described in the README file. The framework supports various protein data types, payload formats, authentication, and job-waiting loops. Moreover, it leverages both aiohttp and asyncio libraries, enabling a non-blocking communication between a client and the corresponding predictor and parallel processing of the input data, both predictor-wise and entry-wise. Additionally, our tool provides users with a collection of global and per-predictor options through a configuration file described in the documentation.

3 Results

We implemented the clients for 19 web-based tools out of 28 considered Supplementary Table S1. The remaining tools were not implemented due to (i) email-only results: STRUM (Quan *et al.* 2016), (ii) excessive job waiting times: ELASCPIC (Witvliet *et al.* 2016), (iii) malfunctioning prediction submission forms: EASE-MM (Folkman *et al.* 2016), (iv) server discontinuation: ENCoM, (Frappier *et al.* 2015), or (v) frequent outages and failures.

The sequence-based tools implemented in BenchStab are, with one exception, structure enabled. They offer two modes for prediction: from a sequence or a structure. In BenchStab, they are implemented as separate predictors, bringing the total number of available predictors to 25 (Supplementary Fig. S1). BenchStab can be set to query only the sequence-based or structure-enabled predictors.

We tested the proper function of the predictors and their integration within the tool as a potential use case on a crafted dataset. Prediction gathering consisted of several rounds of predictor queries during which we adjusted client parameters per predictor: the status-check delays, number of concurrent queries, and error handling (to avoid causing a denial of service).

4 Use case

BenchStab can be utilized to benchmark the available predictors on a specific mutational dataset. To demonstrate this functionality, we created a new dataset based on FireProt^{DB}, disjoint from the commonly used datasets. We present the results collected using BenchStab on this dataset.

We used only the records with both ΔΔG measurements and PDB accession codes. To prevent data leakage from training datasets, we eliminated records similar to the

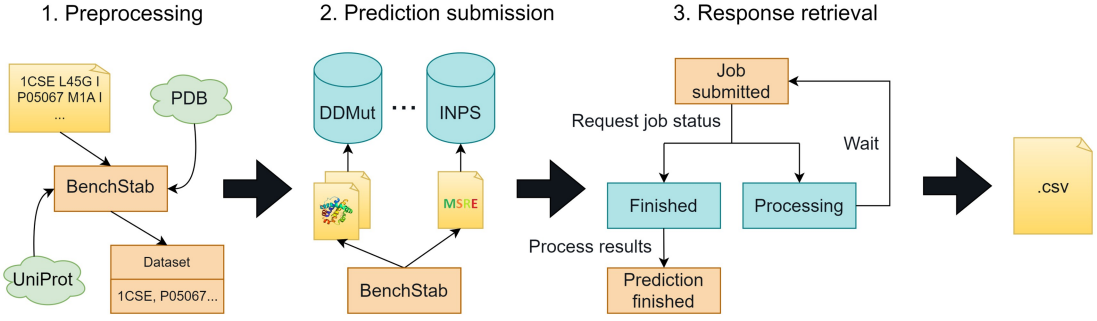


Figure 1. Three stages of the prediction acquisition process. The initial stage is the dataset preprocessing, validation, and enrichment. Every datapoint is then submitted to all selected predictors in the specific format unique to each tool. This is done asynchronously to minimize idling of the program as well as the user's waiting since the responses can be handled immediately as they come (predictors without job queues) or awaited in a non-blocking loop (job-based predictors). Finally, the results are progressively merged as they are processed and periodically exported as a CSV file.

proteins used in the training of the predictors as follows. First, we pooled all training datasets from the implemented predictors (Supplementary Table S2) to create a joint training set. Next, we assigned a UniRef50 cluster (Suzek *et al.* 2015) to each datapoint in both filtered FireProt^{DB} and training set. Finally, with assigned clusters, we eliminated all datapoints assigned any UniRef50 cluster ID appearing in the training set too. The resulting dataset comprises 289 records for 36 proteins (Velecký *et al.* 2024).

To check the structural heterogeneity of this dataset, we employed SCOP (Andreeva *et al.* 2014) for fold-based structure clustering to discover that our dataset contains 25 unique SCOP folds among the 36 proteins. Half of the folds were seen before by at least one of the predictors (Supplementary Table S3). Moreover, a distribution analysis shows that the dataset is not biased to a particular protein, an enzyme class, a particular structural element, or a conservation of mutated residues (Supplementary Fig. S2). However, the alanine-involving mutations make up half of the dataset, and many substitutions are not represented (Supplementary Fig. S3), which is a known problem for protein stability datasets (Caldararu *et al.* 2020). We explored a possible remedy by deriving new datapoints using thermodynamic permutation (Diaz *et al.* 2024), but only two structures for mutants in our dataset were available in the PDB at the time of writing. Further statistics on the produced dataset are presented in Supplementary Tables S3 and S4 and Supplementary Fig. S4.

With the dataset, we benchmarked 24 predictors: 22 of the predictors implemented in BenchStab (Supplementary Table S1) and two standalone tools — FoldX versions 4 and 5 (Schymkowitz *et al.* 2005) — providing a comparison with a popular standalone and force-field-based predictor. We did not include three of the implemented tools in the final results: sRide (Magyar *et al.* 2005), SDM (Worth *et al.* 2011), and PROSTATA (Umerenkov *et al.* 2023). The first does not provide predictions for individual mutants, the second became unavailable during benchmarking, and the last used heterogeneous training data including individual protein domains (Tsuboyama *et al.* 2023); creating a dataset robust to structural leakage via domains to guarantee a fair evaluation was beyond the scope of this study. Supplementary Figure S1 clarifies which tools were implemented and which were benchmarked.

The concise statistics of the results are shown in Fig. 2 for both regression and binary classification (informedness; Powers 2011). Our evaluation revealed that most of the tools can be more or less successfully used for mutation prioritization with balanced accuracy between 51% and 64%. On the other hand, the overall low predictive performance (Supplementary Figs S5 and S6) implied considerable room for improvement. Almost all the tools showed a particularly poor performance in the regression task, i.e. predicting the exact change in the protein stability (the worst and best R^2 equal to 0.01 and 0.15, respectively) with frequent both false positive and false negative errors (Supplementary Fig. S5). Furthermore, the vast majority of tested tools displayed a bias toward destabilizing predictions (Supplementary Fig. S8), also shown by mean signed deviation ranging from −0.79 to −0.11, as has been reported previously (Usmanova *et al.* 2018, Broom *et al.* 2020, Sanavia *et al.* 2020, Pucci *et al.* 2022). The abovementioned metrics, as well as root mean squared error, mean absolute error, accuracy, and Matthews or Pearson correlation coefficients, are reported for individual predictors in Supplementary Table S5. The structure-enabled tools did not perform much better than the sequence-only tools. In the case of precision-recall curves for binary classification, structure-enabled sequence-based predictors performed worse when the structures were provided (Supplementary Fig. S7), as was observed in another recent study (Pancotti *et al.* 2022).

5 Conclusions

We presented BenchStab – a tool that facilitates the use of online stability-change predictors and streamlines the process of benchmarking a new predictor against established competitors. Protein engineers can use it directly on their proteins of interest with a tailored dataset to find the best-working predictor in their use case. Our tool is validated by automated tests. On top of that, we investigated the robustness of our

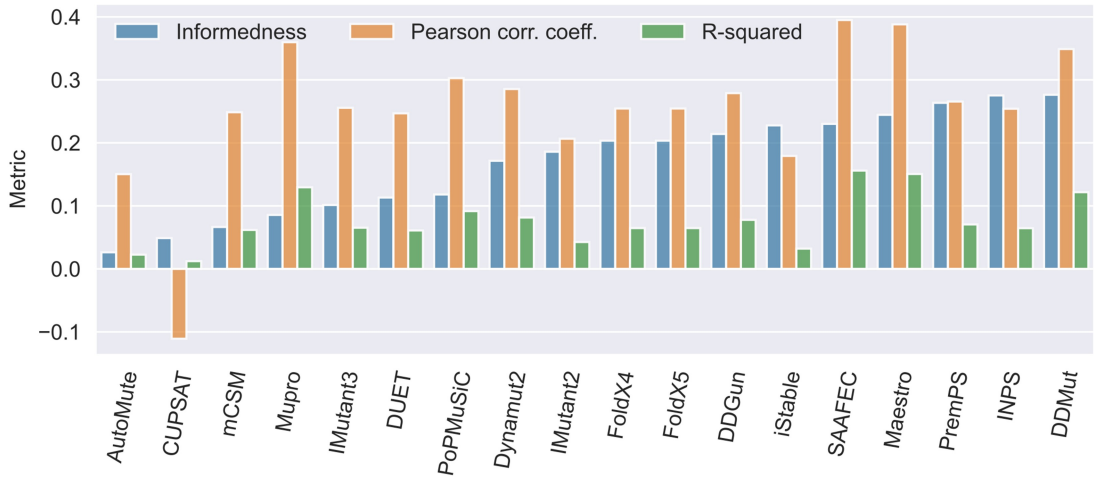


Figure 2. Performance of the predictors as measured on the BenchStab dataset. The tools are compared among themselves by these metrics: informedness, Pearson correlation coefficient, and R^2. Informedness* (Powers 2011), a probability of an informed decision, is used to order the results. For the predictors with two input variants (structure and sequence), we selected the higher-scoring variant. *informedness [−1, 1] = 2 × balanced accuracy − 1 = recall + inverse recall − 1 = TP/P + TN/N − 1 where TP, TN stands for true positives, true negatives, and T, F for all true, false cases, respectively.

tool and of the underlying predictors on a newly created independent dataset.

As we can expect the discontinuation of some of the predictors in the future or breaking changes in their web interfaces, we released BenchStab as an open source to encourage quick updates from the scientific community. In the same way, our application could be extended to incorporate new predictors, including those for other protein properties, e.g. melting temperature or solubility.

We demonstrated the use case of the tool on a benchmarking task. The results revealed that hard cases for the current predictors exist, and therefore there is still a need for more precise tools. Structure-based tools did not beat their sequence-only counterparts. This finding seems consistent with a recent study (Pancotti *et al.* 2022) and may suggest that the structural information may not have been grasped optimally. We also reconfirmed the bias toward destabilizing predictions (Usmanova *et al.* 2018, Broom *et al.* 2020, Sanavia *et al.* 2020, Pucci *et al.* 2022). The dataset consists of proteins unseen by the benchmarked predictors before.

It is important to stress that the purpose of our dataset was to serve as test data and a use case for the BenchStab tool. Our dataset has several limitations, e.g. data from alanine-scanning experiments are overrepresented, which are often employed to identify residues crucially contributing to the protein stability (Caldararu *et al.* 2020), and several mutation types are not represented. Applying thermodynamic permutation (Diaz *et al.* 2024) to recover some mutation types would have a limited effect due to the unavailability of structures needed to query most of the predictors. Therefore, a more robust dataset is required for a comprehensive comparison of the predictors, which is beyond the scope of this study.

In conclusion, we believe BenchStab will motivate computer scientists to enter the domain of stability-change prediction by facilitating the comparison of their predictors to the state of the art.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

The authors thank the RECETOX Research Infrastructure [grant number LM2023069] financed by the Czech Ministry of Education, Youth and Sports for its supportive background. This project was also supported by Brno University of Technology [grant number FIT-S-23-8209]; the European Union's Horizon 2020 Research and Innovation Programme [grant agreement number 857560 (CETOCOEN Excellence)] and the National Institute for Neurology Research [grant number LX22NPO5107 (MEYS—funded by the European Union—Next Generation EU)]. Computational resources were provided by the e-INFRA CZ and ELIXIR-CZ projects [grant numbers LM2018140 and LM2023055]. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Data availability

The data underlying this article are available in its online supplementary material. The BenchStab dataset is available on Zenodo: https://zenodo.org/records/10637728.

References

Andreeva A, Howorth D, Chothia C *et al.* SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* 2014;**42**:D310–4.

Broom A, Trainor K, Jacobi Z *et al.* Computational modeling of protein stability: quantitative analysis reveals solutions to pervasive problems. *Structure* 2020;**28**:717–26.e3.

Caldararu O, Mehra R, Blundell TL *et al.* Systematic investigation of the data set dependency of protein stability predictors. *J Chem Inf Model* 2020;**60**:4772–84.

Chen C-W, Lin J, Chu Y-W *et al.* iStable: off-the-shelf predictor integration for predicting protein stability changes. *BMC Bioinformatics* 2013;**14**:S5.

Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 2006; **62**:1125–32.

Dana JM, Gutmanas A, Tyagi N *et al.* SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* 2019;**47**:D482–9.

Diaz DJ, Gong C, Ouyang-Zhang J *et al.* Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nat Commun* 2024;**15**:6170.

Folkman L, Stantic B, Sattar A *et al.* EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J Mol Biol* 2016;**428**:1394–405.

Frappier V, Chartier M, Najmanovich RJ *et al.* ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Res* 2015;**43**:W395–400.

Magyar C, Gromiha MM, Pujadas G *et al.* SRide: a server for identifying stabilizing residues in proteins. *Nucleic Acids Res* 2005;**33**:W303–5.

McKinney W. Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*. Austin, Texas, SciPy, 2010, 56–61. Doi: 10.25080/Majora-92bf1922-012

Modarres HP, Mofrad MR, Sanati-Nezhad A *et al.* Protein thermostability engineering. *RSC Adv* 2016;**6**:115252–70.

Pancotti C, Benevenuta S, Birolo G *et al.* Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief Bioinform* 2022;**23**:bbab555.

Planas-Iglesias J, Marques SM, Pinto GP *et al.* Computational design of enzymes for biotechnological applications. *Biotechnol Adv* 2021; **47**:107696.

Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Tech* 2011;**2**:37–63.

Pucci F, Schwersensky M, Rooman M *et al.* Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Curr Opin Struct Biol* 2022;**72**:161–8.

Quan L, Lv Q, Zhang Y *et al.* STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* 2016;**32**:2936–46.

Rose Y, Duarte JM, Lowe R *et al.* RCSB Protein Data Bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. *J Mol Biol* 2021;**433**:166704.

Sanavia T, Birolo G, Montanucci L *et al.* Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput Struct Biotechnol J* 2020;**18**:1968–79.

Schymkowitz J, Borg J, Stricher F *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* 2005;**33**:W382–8.

Stourac J, Dubrava J, Musil M *et al.* FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res* 2021;**49**:D319–24.

Suzek BE, Wang Y, Huang H *et al.*; UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;**31**:926–32.

Tsuboyama K, Dauparas J, Chen J *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* 2023;**620**:434–44.

Umerenkov D, Nikolaev F, Shashkova TI *et al.* PROSTATA: a framework for protein stability assessment using transformers. *Bioinformatics* 2023;**39**:btad671.

Usmanova DR, Bogatyreva NS, Ariño Bernad J *et al.* Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* 2018;**34**:3653–8.

Velecký J, Berezný M, Musil M *et al.* The BenchStab dataset: a dataset for comparing mutational predictors of stability. 2024. Doi: 10.5281/zenodo.10637727.

Witvliet DK, Strokach A, Giraldo-Forero AF *et al.* ELASPIC webserver: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics* 2016;**32**:1589–91.

Worth CL, Preissner R, Blundell TL *et al.* SDM—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res* 2011;**39**:W215–22.

Yin S, Ding F, Dokholyan NV *et al.* Eris: an automated estimator of protein stability. *Nat Methods* 2007;**4**:466–7.

Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae553
Applications Note

Journal of Cheminformatics

RESEARCH

Open Access

Large-scale annotation of biochemically relevant pockets and tunnels in cognate enzyme–ligand complexes

O. Vavra[1,2], J. Tyzack[3], F. Haddadi[1,2], J. Stourac[1,2], J. Damborsky[1,2], S. Mazurenko[1,2*], J. M. Thornton[3*] and D. Bednar[1,2*]

Abstract

Tunnels in enzymes with buried active sites are key structural features allowing the entry of substrates and the release of products, thus contributing to the catalytic efficiency. Targeting the bottlenecks of protein tunnels is also a powerful protein engineering strategy. However, the identification of functional tunnels in multiple protein structures is a non-trivial task that can only be addressed computationally. We present a pipeline integrating automated structural analysis with an *in-house* machine-learning predictor for the annotation of protein pockets, followed by the calculation of the energetics of ligand transport via biochemically relevant tunnels. A thorough validation using eight distinct molecular systems revealed that CaverDock analysis of ligand un/binding is on par with time-consuming molecular dynamics simulations, but much faster. The optimized and validated pipeline was applied to annotate more than 17,000 cognate enzyme–ligand complexes. Analysis of ligand un/binding energetics indicates that the top priority tunnel has the most favourable energies in 75% of cases. Moreover, energy profiles of cognate ligands revealed that a simple geometry analysis can correctly identify tunnel bottlenecks only in 50% of cases. Our study provides essential information for the interpretation of results from tunnel calculation and energy profiling in mechanistic enzymology and protein engineering. We formulated several simple rules allowing identification of biochemically relevant tunnels based on the binding pockets, tunnel geometry, and ligand transport energy profiles.

Scientific contributions

The pipeline introduced in this work allows for the detailed analysis of a large set of protein–ligand complexes, focusing on transport pathways. We are introducing a novel predictor for determining the relevance of binding pockets for tunnel calculation. For the first time in the field, we present a high-throughput energetic analysis of ligand binding and unbinding, showing that approximate methods for these simulations can identify additional mutagenesis hotspots in enzymes compared to purely geometrical methods. The predictor is included in the supplementary material and can also be accessed at https://github.com/Faranehhad/Large-Scale-Pocket-Tunnel-Annotation.git. The tunnel data calculated in this study has been made publicly available as part of the ChannelsDB 2.0 database, accessible at https://channelsdb2.biodata.ceitec.cz/.

Keywords Bottleneck, Cognate ligand, Cavity, Enzyme, Tunnel, Machine learning, Pocket, Transport

*Correspondence:
S. Mazurenko
mazurenko@mail.muni.cz
J. M. Thornton
thornton@ebi.ac.uk
D. Bednar
222755@mail.muni.cz
Full list of author information is available at the end of the article

Introduction

Enzymes are biological catalysts that can accelerate chemical reactions, which makes them essential for every living cell. These chemical reactions occur in the active site, which consists of residues with specific physicochemical properties. Active sites can be found either in clefts on the surface of an enzyme or buried inside a cavity shielded from the outer environment. In the latter case, the active site cavity is connected with the surface by access tunnels to enable the passage of ligands, small molecules that interact with the enzyme [1]. This encompasses the exchange of reactant and product molecules or the binding of cofactors. The tunnels also impact the activity and specificity of the enzyme by restricting access to the active site for unfavourable molecules [2]. The introduction of mutations in protein tunnels and channels can affect activity, specificity, promiscuity, enantioselectivity, and stability [3, 4].

Several computational tools were developed for the detection of important cavities and pockets, e.g., Fpocket [5], CASTp [6], and P2Rank [7]. These tools rank all the pockets found in a protein structure by their scoring functions and select the best potential binding pocket for the user. To improve the reliability of the selection, one can use annotations found in structure databases [8, 9]. Unfortunately, these annotations are available only for a limited number of enzymes. The selection of the functionally relevant pocket is also crucial for the calculation of access tunnels. However, currently there is no tool available that would predict the suitability of a pocket for this purpose.

To identify tunnels in enzymes, one may use tools such as CAVER [10], MOLE [11] or MOLAXIS [12]. Similarly, with pocket calculation, these tools can detect multiple tunnels and also provide ways to rank them based on their geometrical properties. In many proteins with buried active sites, multiple tunnels can be identified, which makes it difficult to decide which tunnel is biochemically relevant. This crucial decision could be greatly supported by a large-scale analysis of protein structures. Previous efforts in this matter focused purely on finding tunnels in enzymes [13, 14]. While these studies proved that tunnels appear in all enzyme classes, they did not define how to recognise biochemically relevant tunnels.

The classical computational approach to studying the biological relevance of tunnels is to simulate the interactions between a protein and a ligand with methods based on molecular dynamics [15]. Unfortunately, this time-demanding type of simulation is not feasible for large datasets. More recent tools, such as CaverDock [16], GPathFinder [17], or ART-RRT [18], employ various approximations to simulate ligand transport in short computational times and provide valuable information about the energy profile of the process. These tools are gaining popularity [19] and have successfully been used for screening and identifying novel drugs [20, 21] and engineering proteins [22–26].

In this study, we present a novel strategy for annotating pocket relevance for tunnel calculation and assign biochemical relevance of tunnels based on ligand transport and binding energies. With the growing number of available protein structures [27] and models [28], automatic annotation of binding pockets and tunnels without the dependency on residue annotations would be of great use. Based on the premise that substrate and product molecules are present in relevant pockets in enzyme structures, we created a dataset independent of annotations. We selected experimentally derived enzyme structures with bound molecules that were similar to cognate ligands, i.e., ligands that potentially bind or react with a given enzyme. For this purpose, we used a previously published dataset of enzyme cognate ligand pairs [29–31], which we updated and utilized for structural analyses of pockets and tunnels in this study. We then developed a pipeline combining machine learning, the geometrical analysis of tunnels, and the energy profiling of transported ligands. The pipeline was then validated against molecular dynamics simulations and applied to the large-scale dataset with more than 17,000 protein structures.

Methods

The study used data collected from the publication by Tyzack et al. [31] and updated it for the purposes of our study. After filtering the original dataset, we analysed 17,092 unique protein–ligand pairs (Table 1) The data consists of enzyme–ligand complexes ranked by the similarity of the bound ligand with the cognate ligand from the KEGG [32] database calculated by the PARITY algorithm [31]. To process the data, we designed an automatic pipeline which consists of three parts (Fig. 1): (i) automatic annotation of enzyme–cognate ligand complexes by computational tools (ii) classification of the main binding pocket of the enzyme to buried or surface pocket by machine learning (ML) predictor, and (iii) the energetical analysis of ligand un/binding by CaverDock (Fig. 1). Each part of the pipeline was separately tested and validated. The data provided from all three parts were combined and analysed in the later part of the study. Here we provide a summary of the methodology behind the pipeline. The detailed description of each step with used parameters is part of the supplementary material.

Table 1 The summary of the pipeline proposed in this study and the dataset sizes at various stages of the pipeline execution

Pipeline	Items	Number of cases
1. Automatic annotation	Protein–ligand pairs	35,882
	Unique PDBs	17,092
	Ligand missing in the biological unit	193
	Ligand not present in PDB	133
	Ligand not present in any pocket	1058
	Pocket calculation errors	337
	Successfully calculated pockets	15,697
	Proteins with annotation in CSA and Uniprot	11,046
	Proteins without annotation	4651
	Selected pockets with matching annotated residues	8350
	No matching residues in the selected pocket	2696
	No tunnels found by Caver	526
	Tunnel calculation errors	739
	Successfully calculated tunnels	14,432
2. Machine Learning predictor for pocket annotation	Buried pockets without tunnels	508
	Borderline pockets without tunnels	160
	Surface pockets without tunnels	597
	Buried pockets with calculated tunnels	3552
	Borderline pockets with calculated tunnels	3178
	Surface pockets with calculated tunnels	7702
3. Energy profiles	Protein–ligand pairs for CaverDock calculations	14,432
	Unfinished CaverDock calculations	1274
	Successfully completed CaverDock jobs	13,158
	Successfully calculated energy profiles	29,693

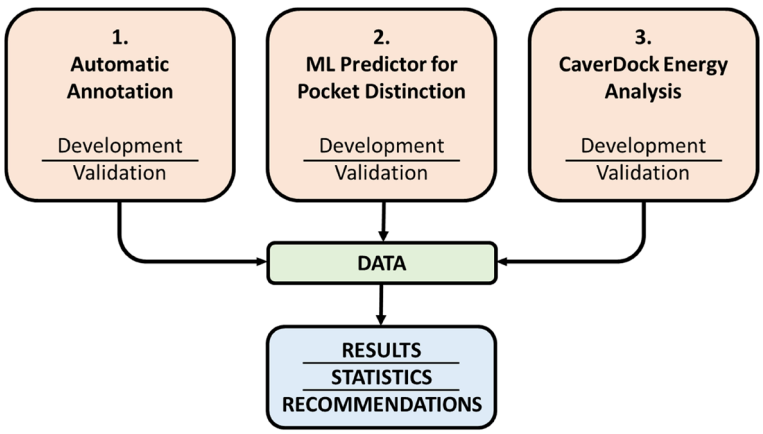


Fig. 1 The overview of the pipeline developed in this study. The pipeline consists of three steps: (i) automatic annotation of enzyme–cognate ligand complexes with computational tools, (ii) classification of ligand binding pocket by machine learning (ML) predictor, and (iii) analysis of ligand transport through enzyme tunnels with CaverDock

Automatic annotation

At the beginning of the pipeline, the biological unit is collected for each enzyme in the dataset [27]. The structures were processed to remove all ligand molecules, while known cofactors [5, 33] were kept in the structure. Next, we calculated pockets in the structure with Fpocket 2 [5] and selected the main pocket based on the location of the bound ligand structurally related to the cognate ligand of the enzyme. The selected pocket was used to define the starting point for tunnel detection using

CAVER 3.02 [10]. The automatic annotation part of the pipeline was validated in two ways. First, we collected annotations from Swiss-Prot, UniProtKB [8], and CSA [9], together with Fpocket and druggability scores calculated for all pockets by Fpocket 2. This data was used to observe whether the selected main pocket contained annotated residues important for the function of the enzyme and to analyse if the main pocket had the best-predicted scores. Second, we studied the impact of the ligand presence in protein structures to determine the changes in tunnel properties. By using the REST API in PDBe [27] and RCSB [34], we collected 2904 pairs of protein–ligand complexes and ligand-free structures. In the next step, we aligned the structures with DeepAlign [35] and calculated tunnels in each pair of structures. Finally, we analysed the changes and differences to determine the properties of potentially relevant tunnels.

Machine-learning predictor for pocket distinction

The main goal of this part of the pipeline was to create a predictor which would be able to assess and differentiate between buried and surface-exposed protein pockets. For the training of the predictor, we manually labelled 200 pockets. We analyzed the distribution of the Enzyme Commission (EC) classes in the dataset and randomly collected samples in quantities that matched the EC class distribution. Features were extracted from Fpocket 2 output, and an additional "Exposed ratio" feature was included, representing the number of solvent-accessible residues. In total, 20 features were used (Table S1). Pockets were categorized into three classes: buried, borderline, and surface, based on manual inspection. The following software was used for the training of the predictor: Python 3.9.7, NumPy 1.26.2, Pandas 1.4.3, Scikit-learn 1.1.1. We tested the Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Shallow Neural Network (ANN), Gaussian Naive Bayes, and Random Forest as classifiers. In each case, we applied a grid search with five-fold cross-validation for tuning hyperparameters of the algorithms (Table S2) and conducted data preprocessing, including Kolmogorov–Smirnov feature filtering [36]. The performance was evaluated using accuracy, precision, recall, FPR, and F1 measures because the dataset was balanced. For validation, we employed an independent test set of additional 100 manually labelled samples, mirroring the class distribution of the training set (Table S3). The best predictor was then used to classify all calculated pockets.

CaverDock energy analysis

CaverDock 1.1 [37] was used to analyse the ligand pathways in all cases in the dataset with successfully calculated tunnels. CaverDock is a tool designed for rapid analysis of ligand transport. It enables fast simulation of the binding and unbinding of ligand molecules through protein tunnels. CaverDock achieves short calculation times which makes it well-suited for virtual screening applications. The current version of CaverDock uses CAVER 3.02 for the pathway identification and AutoDock Vina 1.1.2 as the docking engine, applying its docking algorithm and empirical scoring function without any modifications. Each CaverDock calculation requires the receptor, ligand, and tunnel input files and the configuration. The tunnel is discretized into a set of discs which are used to guide the ligand through the protein during the simulation. To produce the trajectories for the study, we used the lower-bound CaverDock calculations. In each step of the CaverDock lower-bound trajectories, the ligand is constrained to a disc, and the docking algorithm docks the molecule to the disc and optimises the conformation. Apart from the selected drag atom which is constrained to the disc (Table S4), the rest of the molecule can move freely. Then the ligand is moved to the next disc and the process is repeated until the molecule reaches the end of the tunnel. The outputs are the ligand trajectory and the energetic profile of the un/binding process.

The information from the relevant cognate KEGG [32] reaction was used to collect the cognate ligand and to set the drag atom used to guide the molecule during the simulation by processing the information with Reaction Decoder Tool [38] and RDKit (https://github.com/rdkit/rdkit). The processed ligand and enzyme structure files were then converted to PDBQT using the scripts from MGLtools 1.5.6 [39]. The tunnel 3D representations in PDB format were discretized into a set of discs using the Discretizer tool from the CaverDock package. Finally, the grid box around the tunnel and the configuration file were prepared by the prepare-config script from the CaverDock package. The direction of the simulation was defined based on the type of the ligand, binding for substrates and unbinding for products. Only the lower-bound trajectory was calculated and analysed. Important energy values were extracted from the energy profiles manually for the validation dataset and automatically in the annotation pipeline: E_{Bound}, E_{Max}, and $E_{Surface}$. The energy barriers were calculated as $E_a = E_{Max} - E_{Bound}$ for the products and $E_a = E_{Max} - E_{Surface}$ for the reactants.

The CaverDock tool has been tested extensively and used on various datasets in previous publications [20, 21]. However, validation of the quality of predicted trajectories from CaverDock has not been done by any method approaches based on Molecular Dynamics (MD). We validated CaverDock by running classical MD simulations and Adaptive Steered Molecular Dynamics (ASMD) [40]. In contrast with unbiased MD, the ASMD method

applies constant external force on two atoms in the simulated systems. This can be used to simulate unbinding or binding ligands through tunnels. The direction of the movement is set by selecting the steering atoms to move the ligand in the direction of a selected tunnel by lengthening or shortening the distance for unbinding or binding respectively. While changing the distance between those two atoms, the ligand moves in the given direction, but it can follow the curves of the tunnel which allows it to move through the protein. The steering atoms or the direction are not changed during the simulation. In ASMD, the simulation is divided into multiple stages. During each stage, the steered simulation is performed in several parallel replicas, and the Jarzynski average [41] is calculated at the end of that stage. The simulation then proceeds by selecting the single trajectory with a work value closest to the Jarzynski average. The next stage continues from the selected trajectory. The Potential of Mean Force (PMF) is calculated at each stage, and at the end of the ASMD simulation, the segments of the PMF are combined to form the complete PMF. For the validation, we selected eight cases from the dataset with protein structures which had 2–4 well-defined tunnels and the cognate product bound inside (Table S4). To prepare the complexes for the validation unbinding simulations, we selected the lowest-energy binding pose from the CaverDock analysis of the first tunnel, extracted the pose, and saved it in the protein structure. The complexes were then processed by several tools, minimised, and equilibrated before running the MD simulations with AMBER 16 [42–51].

Before we started with the biased unbinding simulations, we ran classical MD simulations of *System #3* and *System #4* (Table 2) to showcase the need for biased MD simulations [15, 52] and approximative methods for the study of ligand unbinding [19]. We used the prepared complexes and ran 3 replicas of 1 μs simulations to study the behaviour of the complexes and the potential unbinding of the ligand molecules. Next, the unbinding trajectories were calculated with ASMD. The following parameters were used: 25 parallel simulations, 2 Å stages, a velocity of 10 Å/ns, and a force of 7.2 N. The protein atom for the steering was different for each tunnel. The ligand atom for steering was selected as the one closest to the centroid of the molecule. Lastly, we ran MD simulations with ligand-free structures to generate ensembles of protein snapshots to study how much CaverDock results change when using dynamic structures. We used the same settings for the preparation of the systems, minimisation, and equilibration. We ran 50 ns of production MD, saved 25,000 snapshots, and from these we collected 100 snapshots covering the entire MD simulation. We calculated the tunnels in selected snapshots using CAVER and the transport of ligands through the snapshots with CaverDock. Then, we collected and averaged the energy values for each snapshot and tunnel in every system. Finally, the Potential of Mean Force profiles from ASMD and CaverDock energy profiles from a single static structure and averaged values were compared. We qualitatively analyzed the results by comparing the order of the calculated profiles based on their maximum energy for each tunnel and the number of matching profiles between the two methods (e.g. if the profile for a tunnel is the first one by ASMD and in CaverDock it is considered as a match). We are aware that both MDs and CaverDock use different methods for both parametrisation and evaluation of the transport energy. Our main aim was the qualitative comparison to see if the molecules could unbind through the selected tunnels.

Table 2 The comparison of Potential of Mean Force profiles obtained from ASMD simulations and energy profiles from single structure or averaged CaverDock calculations over snapshots from MD simulations

Case	Enzyme	Ligand	Number of tunnels	Match with static CaverDock	Match with averaged CaverDock
System #1 (PDB ID 1OTW)	Pyrroloquinoline–quinone synthase	Pyrrolo-quinoline quinone	3	1 out of 3	1 out of 3
System #2 (PDB ID 2BFN)	Haloalkane dehalogenase LinB	*trans*-3-Chloro-2-propene-1-ol	3	3 out of 3	3 out of 3
System #3 (PDB ID 2RFY)	Cellobiohydrolase	Cellobiose	3	0 out of 3	1 out of 3
System #4 (PDB ID 2UWH)	Cytochrome P450 BM3	11,14,15-Trihydroxyicosatrienoic acid	3	2 out of 3	3 out of 3
System #5 (PDB ID 4E2Z)	C-3′-methyltransferase	Se-adenosyl-L-selenohomocysteine	3	3 out of 3	3 out of 3
System #6 (PDB ID 5EDT)	Cytochrome P450 CYP121	(4*S*)-4-(5,5-Dimethylcyclohex-1-en-1-yl) cyclohex-1-ene-1-carboxylate	4	0 out of 4	0 out of 4
System #7 (PDB ID 3ORW)	Phosphotriesterase	*N*-(6-Aminohexanoyl)-6-aminohexanoate	2	2 out of 2	2 out of 2
System #8 (PDB ID 5U6M)	UDP-glucosyltransferase	Uridine 5′-diphosphate	3	3 out of 3	3 out of 3

Results

Automatic annotation

Annotation of the filtered PROCOGNATE dataset

The summary of the filtering of the PROCOGNATE dataset is given in Table 1. Out of the 17,092 unique PDBs, the ligand was not present in the biological unit in 193 cases, so we had to use the asymmetric unit instead. In 133 cases, the ligand was not present in the PDB at all—when three-letter codes for ligands did not match the bound ligand code in the dataset. In 1058 cases, the ligand was not inside any of the calculated pockets but rather at or near the protein surface, therefore it was impossible to define any pocket which contained the ligand. In 337 cases, there were errors in the pocket calculation and the tool failed to predict any pockets. We looked at the representation of the enzyme classes defined by their catalysed reaction and classification by the EC numbers in the 15,697 cases with successfully calculated pockets, and all EC classes were represented in the dataset: EC 1 (25.1%), EC 2 (38.5%), EC 3 (20.4%), EC 4 (7.7%), EC 5 (4.3%), EC 6 (3.6%) and EC 7 (0.4%). Concerning the tunnel detection, no tunnels were found in 526 cases, and 739 cases finished with errors. This could stem from the following: (i) the pocket was at the surface of the protein and the CAVER algorithm was unable to calculate any tunnels, (ii) the automatically set starting point was in an incorrect position, or (iii) the space was too narrow for the 0.9 Å probe during the calculation, and the tunnel calculation failed.

Validation of annotations

The twofold validation was used to evaluate the usability of the proposed pipeline. We analysed the collected annotations for residues essential for function, i.e., catalytic or binding residues, in selected binding pockets. By searching UniProt and CSA, we managed to find annotations for 11,046 protein structures, and for 4651 structures, we found no information on essential residues (Table 1). Out of 11,046 annotated cases, 76% matched the essential residues with the pocket-lining residues.

We further investigated the impact of the selection of the studied pocket on the performance of the pipeline. Using the ligand coverage, i.e., the fraction of the molecule overlapping with a pocket, we discriminated between three scenarios: (i) the ligand belonged only in one pocket (single pocket), (ii) a part of the ligand was found in another pocket, but the ligand was occupying the main pocket by 10% more than other pockets, (iii) the ligand occupied multiple pockets, and the difference was less than 10%. In the third scenario, e.g., when half of a ligand was inside one pocket, and the second part lay in another (Figure S1), we selected the pocket with the highest druggability score. To this end, we looked at how often the selected pocket has the best Fpocket and druggability scores in the matching/mismatching/no annotations subsets (Table S5). In these subsets, the selected ligand-binding pocket was top-ranked by Fpocket scores only in 43%, 27%, and 41% of the cases. In the case of druggability scores, it was 23%, 12%, and 17%, respectively. These values were surprisingly low, implying that selecting the pocket based on calculated scores would lead to a high number of errors. On the other hand, based on the 75% overlap of the selected pockets with annotated essential residues in structural databases, we can say that the approach of selecting the pocket based on the ligand location is significantly better than a blind selection of the best pockets ranked by Fpocket score or druggability. Furthermore, using the same settings for the Fpocket calculation for all proteins in the dataset seems insufficient as it led to cases where the ligand overlapped with multiple pockets. In addition, selecting the pocket by predicted scores is not generally applicable to any ligand-free structure without available essential residue annotations. A solution could be to extrapolate the location of the ligand and selected pocket from structurally similar proteins.

In the second part of the validation, we analysed how the presence of a ligand impacted the geometry of tunnels in proteins in pairs of ligand-bound and ligand-free structures to analyse the potential effect of induced fit in the structures. We used the priority score in CAVER 3.02 to calculate how many out of the top 5 tunnels identified in ligand-bound structures could also be found in the top 5 tunnels of ligand-free structures (Fig. 2A). In the 2904 studied pairs, we found no common tunnels in 24% of the cases. This could be caused by the absence of the ligand in the structure, which led to a narrower binding site and impacted the geometry of calculated tunnels. In this category, no tunnels were found in the ligand-free structure in 146 cases, and only one tunnel, which did not match with any of the tunnels from ligand-bound structures, was found in 139 cases. In the rest of the structures, 35% had one common tunnel. Based on the results, we observed that it was generally rare for a protein to have more than three potentially biologically relevant tunnels. We collected the priority scores for each of the five ranks of common tunnels and calculated the probability distribution to further study the clusters and define a metric for potentially relevant tunnels (Fig. 2B). We concluded that the tunnels with the priority above 0.55, the average priority score of the third tunnel, could be potentially relevant, with geometrical properties suitable for ligand un/binding. We suggest that for screening purposes, users should focus only on the first three tunnels calculated by CAVER 3.02 or use more tunnels with a priority

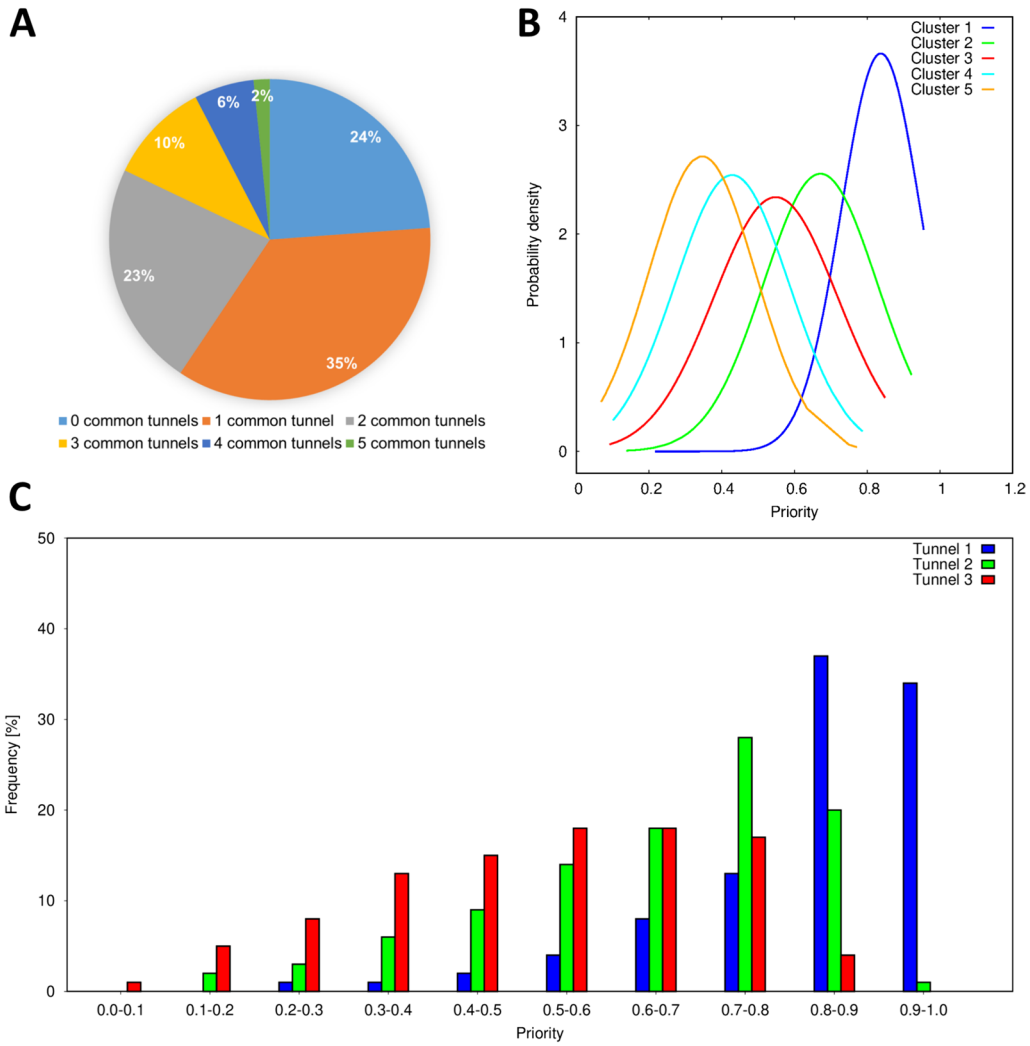


Fig. 2 Analysis of tunnels in pairs of ligand-bound and ligand-free structures and the entire annotated dataset. **A** The number of common tunnels found in both ligand-bound and ligand-free structures. **B** Probability of distribution of CAVER priority score for the best five clusters in pairs of structures. **C** Distribution of the priority score for the first three tunnels in all proteins from the dataset with calculated tunnels. The analyses show that the first three tunnels are commonly present in enzyme–ligand complexes and ligand-free structures. These tunnels have the best geometrical parameters and are suitable for ligand un/binding. Tunnels with the priority above 0.55 could be potentially biologically relevant

score above 0.55. This recommendation is aimed only at the cases in which there is no previous information about the relevancy of tunnels in the studied protein. Based on these findings, we focused only on the first three tunnels in our subsequent data analyses.

Machine-learning predictor for pocket discrimination

Since tunnel calculations are of little use for the surface binding pockets in the annotation pipeline, we trained a machine-learning predictor for identification of such pockets. We used KNN, Random Forest, SVM, ANN, and Naïve Bayes to discriminate between buried and surface binding pockets. We tested two annotation strategies: a three-class problem (buried, borderline, surface) and a two-class problem in which the buried and borderline classes were merged into one (Table S6, Figure S2).

The Naïve Bayesian predictor was used as a simple baseline, and while it showed the highest value of 1-FPR of 90% and 93% on the training dataset for three- and two-class problems, respectively, it failed to identify any buried samples in the test dataset. For the three-class problem, the ANN achieved the highest accuracy (54%) and F1 score (50%), and the second-highest 1-FPR score (67%) on the test set. ANN was also among the top-performing models for the two-class prediction, with all three metrics of 70% on the test set. Despite featuring

lower absolute values, the three-class prediction results were similar to those for a two-class predictor if the baseline accuracy of a completely random prediction was taken into account (33% vs. 50%). Therefore, we selected the ANN-based three-class predictor to annotate the successfully calculated pockets (Table 1).

To get a better understanding of our results, we conducted several additional analyses. Since KNN achieved the highest 1-FPR score on the three-class dataset and performed similarly to ANN on the two-class dataset, we further examined whether the misclassified cases differed between the two models. There was almost no overlap in misclassifications in the three-class dataset, except for a few cases (i.e., 8 buried pockets classified as surface) in the test set. Moreover, while both predictors showed low performance in borderline cases and similar performance in buried cases, the ANN predicted most surface pockets correctly. Furthermore, in addition to evaluating our predictors on the test set, we also constructed learning curves to determine whether expanding the training set (beyond 160 training data points in each fold) could enhance the performance of the predictors (Figure S3). However, the curves did not provide any evidence that the accuracy would increase if more data points were added for training. Finally, the feature pre-selection based on the two-sample Kolmogorov–Smirnov test did not improve the results, so we used the entire set of features in our final predictor. The python code for the pocket discrimination predictor is available at https://github.com/Faranehhad/Large-Scale-Pocket-Tunnel-Annotation.git and as a part of the supplementary material.

CaverDock energy analysis

CaverDock annotation results

We analysed 14,432 proteins with calculated tunnels with CaverDock. We were not able to produce ligand trajectories for 1244 protein–ligand systems (Table 1) due to several factors: (i) we had problems with the automatic parsing of ligand data from KEGG, (ii) protein structures contained parts of DNA or RNA which caused the receptor preparation to fail, (iii) we failed to discretize the tunnels for CaverDock because they were extremely short, represented by only one dummy sphere or one sphere encompassed by another, or (iv) we discarded the cases in which the lower-bound CaverDock calculation did not finish within 48 h on 4 CPUs. Based on the tunnel priority distribution, we analysed the energies of ligand un/binding in up to three tunnels found in each protein. In 13,188 successfully calculated protein–ligand systems, we produced 29,752 energy profiles: 12,804 trajectories for the tunnel 1, 9465 for the tunnel 2, and 7483 for the tunnel 3.

MD simulations for validation of CaverDock trajectories

Both unbiased and biased MD simulations were used to validate the quality of CaverDock results. We simulated three replicas of 1 μs unbiased MD simulations for cellobiohydrolase with cellobiose and cytochrome P450 BM3 with 11,14,15-trihydroxyicosatrienoic acid (*System #3* and *#4* in Table 2, respectively). The ligand remained in the binding site, and we did not observe unbinding in any replicas. This result showed the importance of applying bias in MD to study events such as ligand unbinding. Furthermore, it demonstrated the applicability of approximative methods for the simulation of un/binding to save computational time and effort since unbinding was not observed even in these long simulations. We qualitatively compared the match between the Potential of Mean Force profiles (PMF) from ASMD and CaverDock calculations. We used the CaverDock trajectories from the single static structure and the averaged CaverDock results from 50 ns MD snapshots (Table 2). We show the highest energy value in the profile E_{Max} for the static and the averaged CaverDock calculations in Table S7.

In the case of *System #1* (Figure S4), the energies for tunnels 1 and 2 were similar, but the order was swapped compared to the ASMD simulations. Both tunnels were not frequently open in the 100 snapshots (Table S7). Moreover, the priority of tunnel 1 was lower in MD snapshots, so both tunnels 1 and 2 seem to be feasible for ligand binding. The ligand was not able to unbind through tunnel 3 in ASMD simulations, which agrees with the large barrier found in CaverDock energy profiles. *System #2* had a match for all three tunnels (Fig. 3). In ASMD the ligand was able to unbind with difficulties in tunnel 3, but the force started to unfold the part of the protein that was used for steering the simulation. This result is in accord with the large CaverDock barriers. In *System #3*, there was no matches between ASMD and CaverDock results for the static structure (Figure S5). The use of averaged results from MD snapshots improved the results, as the energy profile for the tunnel 2 was the highest. We concluded that the loops around tunnel 2 made it too wide open in the static structure and biased the results. The ligand in *System #4* unbound successfully in both tunnels 1 and 2 (Figure S6). On the other hand, it did not unbind through tunnel 3 and remained stuck in the binding site. Therefore, we deduced that both tunnel 1 and 2 could be preferred by the ligand. In *System #5*, there was no unbinding observed in tunnels 2 and 3 (Figure S7). The results from all the simulations agreed. The inability to pass through the tunnels in ASMD was reflected in the barriers in both types of simulations. *System #6* had no matches between CaverDock and ASMD, and the use of averaged energies from snapshots did not improve the results (Figure S8). The

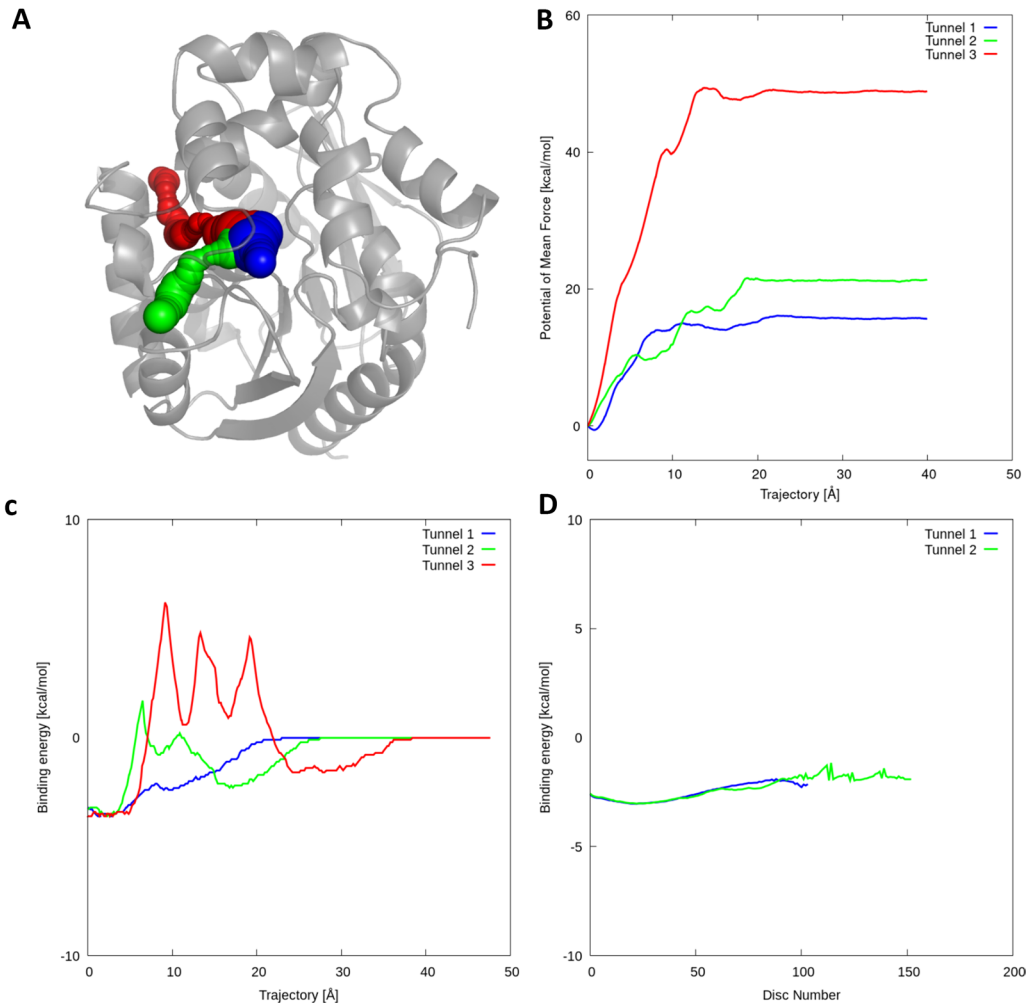


Fig. 3 Results from CaverDock validation for haloalkane dehalogenase LinB with trans-3-chloro-2-propene-1-ol. **A** Visualisation of the protein structure (PDB ID 2BFN) with analysed tunnels showed as spheres: tunnel 1 (blue), tunnel 2 (green), tunnel 3 (red). **B** Potential of mean force profiles from ASMD simulations. **C** Energy profiles from static CaverDock calculations. **D** Averaged CaverDock energy profiles from 50 ns simulation snapshots. The third tunnel was not present in the MD snapshots. The System #2 showed qualitative agreement between the ASMD and CaverDock results

crystal structure seemed too compact and presumably did not have enough time to open during the short MD simulation of the complex. In *System #7*, the ligand was able to unbind successfully through tunnel 1 but was not able to pass through tunnel 2 (Figure S9). CaverDock results agreed with ASMD, so we had a good match across all simulations. In the case of *System #8*, the ligand preferred tunnel 1 over tunnel 2 and was not able to pass through tunnel 3 (Figure S10). Static CaverDock showed similar energies for both tunnels 1 and 2, and the results were improved in MD snapshots, where we saw a slightly higher barrier in tunnel 2. It indicated that both tunnels 1 and 2 could be used by the ligand. Regarding this validation dataset, we observed that some profiles both from PMF and CaverDock were too high in comparison with the other profiles, e.g., *System #5* and *System #6*, suggesting the low probability of these tunnels being used for ligand transport. The RMSD values for 50 ns MD simulations without ligand and ASMD simulations with ligands are listed in Table S11.

Data analysis

The ANN predictor was used to discriminate the pockets based on their type for all cases within the dataset. In the case of the pockets for which we did not manage to calculate tunnels, 508 pockets were predicted as buried, 160 as borderline, and 597 as surface. This was a surprising finding since we expected all these pockets to be predicted as

surface pockets. In the second part of the dataset, i.e., the cases with pockets and successfully calculated tunnels, 3552 cases were predicted as buried pockets, 3178 as borderline, and 7702 as surface (Table 1). In the subset of proteins for which we were able to identify pockets and tunnels, we coupled the predictions with the information about tunnels (Table S8). We binned tunnels similarly as in the study of Pravda et al. [13]: short tunnels under 5 Å, medium-length tunnels between 5 and 15 Å, and long tunnels over 15 Å. In tunnel 1, we see a significant overlap between the categories of pockets with corresponding tunnel lengths. The vast majority of tunnels (75%) in buried cases were either medium or long. Borderline cases were defined as a separate category because during manual annotation, it was difficult to assess if pockets were completely open on the surface or partially buried. For this category, we had 41% short, 49% medium, and 10% long tunnels. For the surface cases, 74% were short tunnels. Thus, our predictor proved successful in its predictions for tunnel 1 and could be a useful tool for assessing whether the calculation of tunnels in a protein makes sense or there is just a surface cavity. We carried out a similar analysis for tunnels 2 and 3, but these tunnels were of lower priority and always longer than tunnel 1. Therefore, almost all the tunnels were either medium or long. Based on this result, we defined tunnel 1 as the only reliable descriptor of the relationship between the predicted pocket type and tunnel length. Moreover, the proteins with tunnels shorter than 5 Å could potentially be discarded since they were calculated for pockets predicted as surface pockets and were, therefore, irrelevant to the tunnel analysis. The main benefit of the predictor is the possibility of pocket annotation in enzyme structures with very narrow tunnels, which would not be found unless the user used a smaller probe during the calculation, or when the tunnel calculation fails. One could also use the predictions to decide whether calculating and analysing tunnels is worthwhile for a particular protein structure. Since the predictor does not require the presence of a ligand in the structure, it is also generally applicable for ligand-free structures.

We studied the geometry of the first three tunnels in more detail. The distribution of tunnel priority scores for all cases with calculated tunnels is presented in Fig. 2C. Importantly, we observed the same trend in the priority scores as in the analysis of pairs of complex and ligand-free structures. The throughput of tunnels 2 and 3 was lower because they were narrower, longer, and more curved than the tunnel 1 (Figure S11). This is not surprising since the priority score is related to the geometrical tunnel properties. Therefore, the priority score should be a sufficient metric for screening purposes. We continued this analysis by separating the dataset based on EC numbers (Figure S12). Tunnels were present in proteins from all EC classes, which was in agreement with previous studies [13]. The tunnel priority followed the same trend in all the classes apart from EC 7 due to the low number of cases in the dataset. We did not observe any major differences in the geometrical properties, which would otherwise indicate that certain EC classes preferred tunnels with specific geometries. We also studied the number of tunnels in each EC class with a priority higher than 0.55 (defined in the analysis of pairs of structures). Apart from EC 7, the results were similar for all EC classes (Figure S13). For future tunnel analyses, it might be worthwhile to compare subclasses to see more significant differences in tunnel geometries.

Next, we studied whether the geometrical bottleneck, i.e., the narrowest part of a tunnel, was the best hot spot for mutagenesis to improve ligand binding and selectivity. For this purpose, we collected the maximum energy E_{Max} from each CaverDock trajectory. In the next step, we compared the location of the energy maximum and the geometrical bottleneck in the tunnel (Fig. 4). We tracked how often the maximum energy was in the disc with the lowest radius or in its vicinity (1.5 Å, 3 Å, and 5 Å). The match between the energy and geometry bottleneck was around 50% for the exact disc and 75% for the 5 Å vicinity (Table S9). The mismatch showed that studying the geometry of the tunnel is a good starting point for quantifying the likelihood of a tunnel being used for ligand transport. Furthermore, the analysis of the energy profiles by approximative methods can be the source of valuable information and help with the identification of other important hot spots for the study and the modification of the ligand transport. The analysis was run with cognate ligands; therefore, these molecules should be recognizable by the enzymes. The results might change for a set of ligands of a larger size or with physicochemical properties different from the cognate ligands.

CaverDock energy profiles were used to analyse the ligand preference of tunnels based on the energy barriers. We compared the maximum energies in up to the three tunnels and selected the best one. The first third of the profiles was removed in order not to include peaks of energy at the beginning of the profiles caused by clashes at the bottom of the tunnel. In the 13,158 proteins with successful CaverDock calculations, tunnel 1 had the most favourable energy in 75% of the cases and tunnel 2 in another 19% of the cases (Table S10). Therefore, for screening purposes, the analysis of tunnel 1 (or at most tunnel 2) would be enough for more than 93% of proteins. Based on these results, tunnel 1 had the best properties for ligand un/binding and would be the most biochemically relevant.

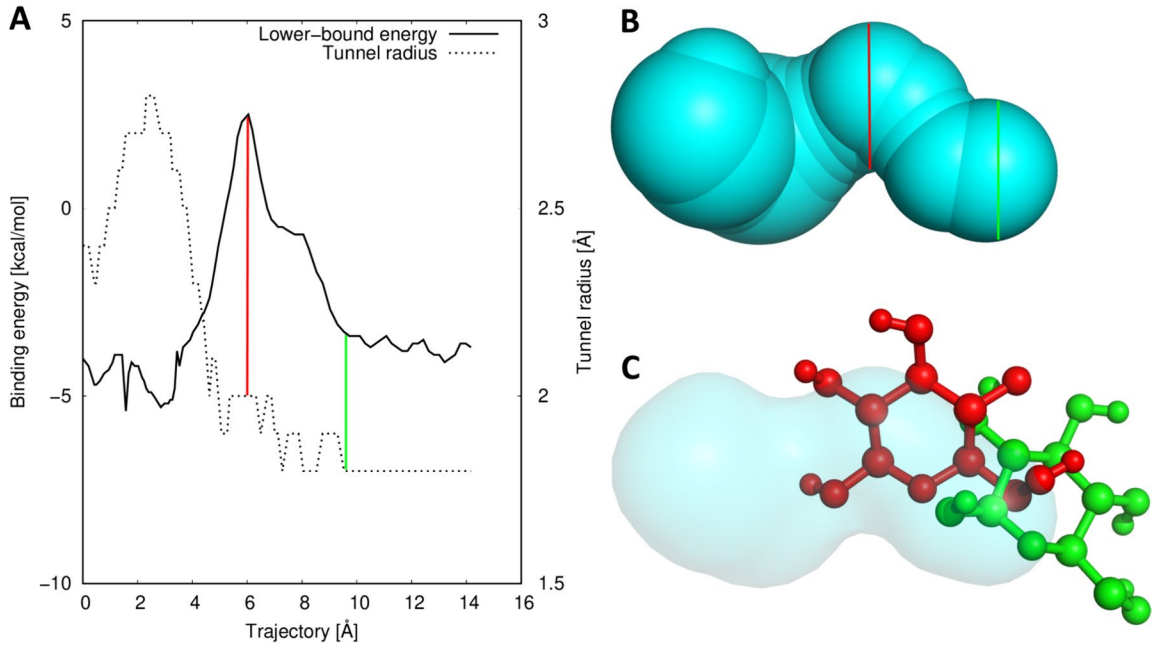


Fig. 4 An example of the case with a large difference between the energetical maximum identified by CaverDock and the geometrical bottleneck identified by CAVER. **A** Energy profile from CaverDock (solid) and the geometric profile from CAVER (dotted). The tunnel region with the energy maximum is highlighted with the red line, and the region with a geometric bottleneck is highlighted with the green line. **B** Visualisation of the tunnel with highlights corresponding to the energy profile. **C** Visualisation of the cognate ligand β-D-glucose conformations extracted from the trajectory from tunnel 1 of the structure of glucose dehydrogenase (PDB ID 2VWG). The binding pose based on the energy maximum (red) and geometric tunnel bottleneck (green)

Finally, we studied how well the cognate ligands were recognised by their receptors. We analysed the distribution of energy maxima (Figure S14) and the energy barrier (Figure S15). In the case of tunnel 1, which was the most preferred tunnel for ligand un/binding, almost 80% of the E_{Max} values were in the range between − 10 kcal/mol to 5 kcal/mol, and the energy barriers E_a were in the range between 0 kcal/mol and 10 kcal/mol. Both values were highly correlated for cognate ligands, and the Pearson's correlation coefficient was 0.98 for energies from all three tunnels. Using E_{Max} seems to be equivalent to E_a for cognate ligands and probably for other natural substrates, which should be transported reasonably fast and bound in the active site. For inhibitors, both values could have different meanings as the molecule does not need to pass the entire way to the active site, but the binding affinity must be much stronger. In such cases, we recommend using E_{Max} values as they are easier to collect and interpret. We analysed the data split by pocket classes, EC numbers, and cognate ligand similarity. Still, all the datasets showed similar trends without major differences (data not shown), implying that ligand trajectories are case-specific rather than showing some general trends in different groups of enzymes.

Discussion and conclusions

We describe the development of an automatic pipeline for the analysis of pockets and tunnels in enzymes and its application to study enzyme–cognate ligand complexes. The results provided a way to select potentially biologically relevant tunnels. The proposed approach can be used for extending large protein datasets for structural analyses and screenings. We analysed more than 17,000 cognate enzyme–ligand complexes. We were able to successfully annotate and analyse structural features and the energetics of ligand passage through tunnels in 13,158 enzyme structures. The tunnel data collected in this study has been made publicly available as part of the ChannelsDB 2.0 database [14]. Each part of the pipeline was thoroughly validated, and the data showed that binding pockets selected based on the location of a bound ligand had a good overlap with catalytic and binding residue annotations from the structural databases. Therefore, bound ligands can be used to extend the datasets for pocket and tunnel analyses. Our experiments showed that selecting the pocket purely by score or druggability from Fpocket would be significantly less precise. On the other hand, our pipeline is limited to enzyme structures with bound ligands, which limits its use. However, this limitation is merely a consequence of being able to

classify enzymes and their cognate ligands based on their reactions, which are available in public databases. Extrapolation of ligand positions among homologous protein structures could remove this limitation for many structurally or functionally related proteins. Furthermore, the use of the pipeline to detect non-cognate ligands would probably provide less precise results as it would be harder to select the correct pocket for the following analyses and calculations. Due to the development of AlphFold [53] and AlphaFill [54] the protein engineering community has access to a staggering amount of new protein models and modelled complexes. As an example of the adaptability of our pipeline, we contributed to the update of ChannelsDB 2.0 database [14]. We calculated tunnels for a dataset based on protein structures from AlphaFill with known cofactors. The position of cofactors was used to define the binding pocket and for the later calculation of tunnels in the model structures.

The presented machine learning predictor for the annotation of pockets has proven to be efficient in deciding on the type of pocket. Based on the test set, the machine learning predictor demonstrated the accuracy of 54% and 1-FPR metric of 75% of buried pockets in the three-class prediction. While there still is room for improvement, the current version shows reasonable performance for selecting whether a particular enzyme and pocket are viable for tunnel calculations. Most importantly, it uses the readily available features from the Fpocket, making it easy to obtain these necessary features. At the same time, we release the training data together with the scripts to encourage follow-up studies to improve the predictor, e.g., by considering other, more discriminative features. The structural analyses revealed that it is possible to select potentially biologically relevant tunnels both in ligand-bound and ligand-free structures. Tunnels are present in the enzymes of all seven EC classes. Strikingly, the ligand transport calculations revealed that the energetic maximum was not in the geometrical bottleneck in 50% of analysed tunnels. Therefore, energy profiling provides a highly relevant information about hot spots for enzyme engineering. The comparison of CaverDock energetic maxima for calculated tunnels in each enzyme structure indicated that tunnel 1 had the lowest energy barrier in 75% of cases. This shows, that the energy analysis by CaverDock is valuable addition to the study of tunnel geometry when multiple tunnels can be relevant for a specific ligand. To improve the predictive power of such analysis, the study of geometrical and energetical bottlenecks should be done on a large set of dynamic snapshots. The results from a single structure may be biased by the enzyme conformation in the crystal structure.

The knowledge and data acquired in this study will be important for future screening studies and the development of computational tools. We showed that the presented pipeline could be used to generate features for machine learning predictors and to provide valuable information for key repositories of biological data, such as PDBe Knowledgebase [55]. The validation of CaverDock against MD simulations proved that approximative methods are precise enough for fast energetical analyses of ligand passages. Approximative methods and enhanced sampling simulations are necessary to simulate ligand transport within reasonable times. Thus, we recommend energy calculations with approximative methods for protein engineering studies. Our comprehensive analysis of protein tunnels and the passages of cognate ligands let us formulate the following recommendations for the protein engineering community:

1. For analysis of tunnels in enzymes, start with the literature search and exploration of databases to determine essential residues, identify the location of the binding pocket, and discover transport pathways, whenever possible.
2. The pocket(s) that contains the essential functional residues should be preferred. In the systems with unknown essential residues, the pocket which contains a bound cognate ligand of the enzyme should be used. If there are no ligand-bound structures for the enzyme of interest, analyse available structures of homologous enzymes which contain the ligand. We recommend caution when selecting the binding pocket based solely on the predicted scores by the tools for pocket calculation.
3. The most important step of the tunnel analysis is to set the starting point correctly. When annotations of essential residues are not available the conserved residues are another possibility. Otherwise, we recommend using the residue inside of the selected pocket, closest to the centre of the biological unit or the analysed protein chain in the asymmetrical unit to start the tunnel calculation from the deep part of the pocket. An incorrectly set starting point may hinder the tunnel calculation and impact the geometry of found tunnels.
4. Selection of the biochemically relevant tunnel(s) should be preferably made based on the experimental literature data. When no such information is available, either focus on the first tunnel in a screening scenario, or the first three tunnels according to the highest priority score. CAVER users are advised to inspect the tunnels with priority scores above 0.55. If none of the found tunnels has a priority score above this value, select a different starting point and redo the calculations.

5. If the starting point for tunnel calculation is selected correctly and the first tunnel is shorter than 5 Å, the binding pocket could be located on the surface and tunnel analysis might not be relevant.
6. Analysis of tunnels should be complemented by the study of substrate or product passage whenever possible.
7. Use the ranges of energy barriers defined in this study to filter out molecules with poor (un)binding (E_{Max}: −10 kcal/mol to 5 kcal/mol, E_a: 0 kcal/mol to 10 kcal/mol) for energetic analyses of ligand passage by the approximative method CaverDock [16]. Other methods available for this purpose are SLITHER [56], MoMA-LigPath [57], GPathFinder [17], and ART-RRT [18].
8. Binding and unbinding studies by the approximative methods can be significantly enhanced by the analysis of an ensemble of structures obtained even from a short molecular dynamics simulations.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00907-z.

Supplementary Material 1. Detailed description of the methods with settings and parameters; list of features and hyperparameters for the predictor; predictor learning curves; setup of ASMD simulations; detailed results from validations; details from structural and energetical analyses; tunnel parameters and presence in EC classes (PDF); filtered input dataset (CSV); information for 8 validation systems (CSV); list of PDB ID pairs of the complexes and ligand free structures (CSV); training dataset (CSV); testing dataset (CSV); predictor Python code (PY); training and testing dataset with labels from predictors (CSV).

Acknowledgements
The authors thank Dr. Sérgio Marques for valuable advice during the design of the pipeline and set up of validation simulations.

Author contributions
OV: conceptualization, methodology, workflow, visualization, data curation, writing—original draft, writing—review; JT: conceptualization, methodology, supervision; FH: methodology, software development, writing—original draft. JS: methodology, workflow, writing—review; JD: conceptualization, supervision, funding acquisition, writing—review; SM: conceptualization, supervision, writing—review; JT: conceptualization, supervision; DB: conceptualization, supervision, project administration, funding acquisition, writing-review. All authors reviewed and approved the manuscript.

Funding
This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 857560 (CETOCOEN Excellence). The authors thank the Czech Ministry of Education (INBIO—CZ.02.1.01/0.0/0.0/16_026/0008451, RECETOX RI—LM2023069, ELIXIR CZ—LM2023055, and e-INFRA—LM2018140), National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102) - Funded by the European Union - Next Generation EU, the Technology Agency of the Czech Republic (Permed—TN02000109), and the Grant Agency of the Czech Republic (20-15915Y) for financial support. OV is the recipient of a Ph.D. Talent award provided by Brno City Municipality. This publication reflects only the authors' view, and the European Commission is not responsible for any use that may be made of the information it contains.

Data availability
Supplementary materials contain: detailed description of the methods with settings and parameters, list of features and hyperparameters for the predictor; predictor learning curves; setup of ASMD simulations; detailed results from validations; details from structural and energetical analyses; tunnel parameters and presence in EC classes (PDF); filtered input dataset (CSV); information for 8 validation systems (CSV); list of PDB ID pairs of the complexes and ligand free structures (CSV); training dataset (CSV); testing dataset (CSV); predictor Python code (PY); training and testing dataset with labels from predictors (CSV). The python code for the pocket discrimination predictor is also available at https://github.com/Faranehhad/Large-Scale-Pocket-Tunnel-Annotation.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5/A13, 625 00 Brno, Czech Republic. [2]International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, 656 91 Brno, Czech Republic. [3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust GenomeCampus, Cambridge CB10 1SD, UK.

Received: 5 June 2024 Accepted: 16 September 2024
Published online: 15 October 2024

References

1. Gora A, Brezovsky J, Damborsky J (2013) Gates of enzymes. Chem Rev 113:5871–5923
2. Brezovsky J, Babkova P, Degtjarik O, Fortova A, Gora A, Iermak I et al (2016) Engineering a de novo transport tunnel. ACS Catal 6:7597–7610
3. Kokkonen P, Bednar D, Pinto G, Prokop Z, Damborsky J (2019) Engineering enzyme access tunnels. Biotechnol Adv 37:107386
4. Marques SM, Daniel L, Buryska T, Prokop Z, Brezovsky J, Damborsky J (2016) Enzyme tunnels and gates as relevant targets in drug design. Med Res Rev. https://doi.org/10.1002/med.21430
5. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinform 10:168
6. Tian W, Chen C, Lei X, Zhao J, Liang J (2018) CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res 46:W363–W367
7. Krivák R, Hoksza D (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform 10:39
8. Consortium U (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169
9. Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 42:D485–D489
10. Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B et al (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 8:e1002708
11. Berka K, Sehnal D, Bazgier V, Pravda L, Svobodova-Varekova R, Otyepka M et al (2017) Mole 25—tool for detection and analysis of macromolecular pores and channels. Biophys J 112:292a–293a
12. Yaffe E, Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R (2008) MolAxis: a server for identification of channels in macromolecules. Nucleic Acids Res 36(Web Server issue):W210–W215
13. Pravda L, Berka K, Svobodová Vařeková R, Sehnal D, Banáš P, Laskowski RA et al (2014) Anatomy of enzyme channels. BMC Bioinform 15:379
14. Špačková A, Vávra O, Raček T, Bazgier V, Sehnal D, Damborský J et al (2024) ChannelsDB 2.0: a comprehensive database of protein tunnels and pores in AlphaFold era. Nucleic Acids Res 52:D413–D418
15. Gelpi J, Hospital A, Goñi R, Orozco M (2015) Molecular dynamics simulations: advances and applications. Adv Appl Bioinform Chem 8:37

16. Filipovic J, Vavra O, Plhak J, Bednar D, Marques SM, Brezovsky J et al (2019) CaverDock: a novel method for the fast analysis of ligand transport. IEEE/ACM Trans Comput Biol Bioinform 17:1–11
17. Sánchez-Aparicio JE, Sciortino G, Herrmannsdoerfer DV, Chueca PO, Pedregal JRG, Maréchal JD (2019) Gpathfinder: identification of ligand-binding pathways by a multi-objective genetic algorithm. Int J Mol Sci 20:3155
18. Nguyen MK, Jaillet L, Redon S (2018) ART-RRT: as-rigid-as-possible exploration of ligand unbinding pathways. J Comput Chem 39:665–678
19. Vavra O, Damborsky J, Bednar D (2022) Fast approximative methods for study of ligand transport and rational design of improved enzymes for biotechnologies. Biotechnol Adv 60:108009
20. Pinto GP, Vavra O, Filipovic J, Stourac J, Bednar D, Damborsky J (2019) Fast screening of inhibitor binding/unbinding using novel software tool CaverDock. Front Chem 7:709
21. Pinto GP, Vavra O, Marques SM, Filipovic J, Bednar D, Damborsky J (2021) Screening of world approved drugs against highly dynamical spike glycoprotein of SARS-CoV-2 using CaverDock and machine learning. Comput Struct Biotechnol J 19:3187–3197
22. Rapp LR, Marques SM, Zukic E, Rowlinson B, Sharma M, Grogan G et al (2021) Substrate anchoring and flexibility reduction in CYP153A M.aq leads to highly improved efficiency toward octanoic acid. ACS Catal 11:3182–3189
23. Papadopoulou A, Meierhofer J, Meyer F, Hayashi T, Schneider S, Sager E et al (2021) Re-programming and optimization of a L-proline cis -4-hydroxylase for the cis-3-halogenation of its native substrate. ChemCatChem 13:3914–3919
24. Knez D, Colettis N, Iacovino LG, Sova M, Pišlar A, Konc J et al (2020) Stereoselective activity of 1-propargyl-4-styrylpiperidine-like analogues that can discriminate between monoamine oxidase isoforms A and B. J Med Chem 63:1361–1387
25. Wang L, Marciello M, Estévez-Gay M, Soto Rodriguez PED, Luengo Morato Y, Iglesias-Fernández J et al (2020) Enzyme conformation influences the performance of lipase-powered nanomotors. Angew Chemie Int Ed 59:21080–21087
26. Singh PP, Jaiswal AK, Kumar A, Gupta V, Prakash B (2021) Untangling the multi-regime molecular mechanism of verbenol-chemotype *Zingiber officinale* essential oil against *Aspergillus flavus* and aflatoxin B1. Sci Rep 11:6832
27. Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ et al (2014) PDBe: protein data bank in Europe. Nucleic Acids Res 42(Database issue):D285–D291
28. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G et al (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444
29. Bashton M, Nobeli I, Thornton JM (2006) Cognate ligand domain mapping for enzymes. J Mol Biol 364:836–852
30. Bashton M, Nobeli I, Thornton JM (2008) PROCOGNATE: a cognate ligand domain mapping for enzymes. Nucleic Acids Res 36(Database issue):D618–D622
31. Tyzack JD, Fernando L, Ribeiro AJM, Borkakoti N, Thornton JM (2018) Ranking enzyme structures in the PDB by bound ligand similarity to biological substrates. Structure 26:565-571.e3
32. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
33. Fischer JD, Holliday GL, Thornton JM (2010) The CoFactor database: organic cofactors in enzyme catalysis. Bioinformatics 26:2496–2497
34. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS et al (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39(Database issue):D392–D401
35. Ma J, Wang S (2014) Algorithms, applications, and challenges of protein structure alignment. Adv Protein Chem Struct Biol 94:121–175
36. Pratt JW, Gibbons JD (1981) Kolmogorov–Smirnov two-sample tests. In: Concepts of Nonparametric Theory. Springer Series in Statistics. Springer, New York, NY, p 318–344. ISBN: 978-1-4612-5931-2. https://doi.org/10.1007/978-1-4612-5931-2_7
37. Vavra O, Filipovic J, Plhak J, Bednar D, Marques SM, Brezovsky J et al (2019) CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels. Bioinformatics 35:4986–4993
38. Rahman SA, Torrance G, Baldacci L, Martínez Cuesta S, Fenninger F, Gopal N et al (2016) Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics 32:2065–2066
39. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
40. Ozer G, Quirk S, Hernandez R (2012) Adaptive steered molecular dynamics: validation of the selection criterion and benchmarking energetics in vacuum. J Chem Phys 136:215104
41. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2690–2693
42. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM et al (2005) The amber biomolecular simulation programs. J Comput Chem 26:1668–1688
43. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform 3:33
44. Vanquelef E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC et al (2011) R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucleic Acids Res 39(suppl_2):W511–W517
45. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A (2005) H++: a server for estimating pKas and adding missing hydrogens to macromolecules. Nucleic Acids Res 33(Web Server issue):W368–W371
46. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput 11:3696–3713
47. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
48. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comput 9:3878–3888
49. Le Grand S, Götz AW, Walker RC (2013) SPFP: speed without compromise—a mixed precision model for GPU accelerated molecular dynamics simulations. Comput Phys Commun 184:374–380
50. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an *N*·log(*N*) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
51. Ryckaert J-P, Ciccotti G, Berendsen HJ (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341
52. Miao Y, Bhattarai A, Wang J (2020) Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. J Chem Theory Comput 16:5526–5547
53. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493–500
54. Hekkelman ML, de Vries I, Joosten RP, Perrakis A (2023) AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Methods 20:205–213
55. Varadi M, Anyango S, Armstrong D, Berrisford J, Choudhary P, Deshpande M et al (2022) PDBe-KB: collaboratively defining the biological context of structural data. Nucleic Acids Res 50:D534–D542
56. Lee PH, Kuo KL, Chu PY, Liu EM, Lin JH (2009) SLITHER: a web server for generating contiguous conformations of substrate molecules entering into deep active sites of proteins or migrating through channels in membrane transporters. Nucleic Acids Res 37(Web Server issue):W559–W564
57. Devaurs D, Bouard L, Vaisset M, Zanon C, Al-Bluwi I, Iehl R et al (2013) MoMA-LigPath: a web server to simulate protein–ligand unbinding. Nucleic Acids Res 41(Web Server issue):W297–W302

Publisher's Note

JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau
Article

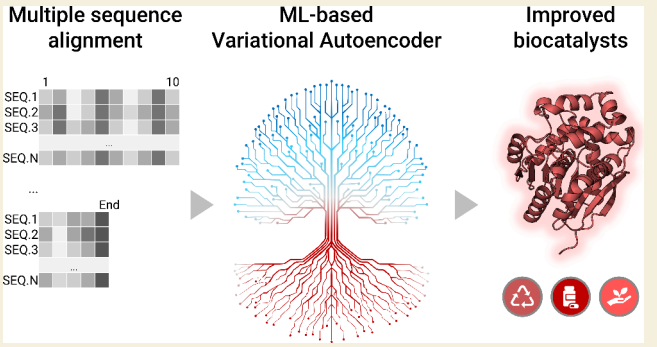
Engineering Dehalogenase Enzymes Using Variational Autoencoder-Generated Latent Spaces and Microfluidics

Pavel Kohout,[#] Michal Vasina,[#] Marika Majerova, Veronika Novakova, Jiri Damborsky, David Bednar, Martin Marek, Zbynek Prokop,[*] and Stanislav Mazurenko[*]

Cite This: *JACS Au* 2025, 5, 838−850
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Enzymes play a crucial role in sustainable industrial applications, with their optimization posing a formidable challenge due to the intricate interplay among residues. Computational methodologies predominantly rely on evolutionary insights of homologous sequences. However, deciphering the evolutionary variability and complex dependencies among residues presents substantial hurdles. Here, we present a new machine-learning method based on variational autoencoders and evolutionary sampling strategy to address those limitations. We customized our method to generate novel sequences of model enzymes, haloalkane dehalogenases. Three design−build−test cycles improved the solubility of variants from 11% to 75%. Thorough experimental validation including the microfluidic device MicroPEX resulted in 20 multiple-point variants. Nine of them, sharing as little as 67% sequence similarity with the template, showed a melting temperature increase of up to 9 °C and an average improvement of 3 °C. The most stable variant demonstrated a 3.5-fold increase in activity compared to the template. High-quality experimental data collected with 20 variants represent a valuable data set for the critical validation of novel protein design approaches. Python scripts, jupyter notebooks, and data sets are available on GitHub (https://github.com/loschmidt/vae-dehalogenases), and interactive calculations will be possible via https://loschmidt.chemi.muni.cz/fireprotasr/.



KEYWORDS: *dehalogenase, protein engineering, machine learning, microfluidics, protein stability, variational autoencoder*

INTRODUCTION

Biocatalysis is a promising field that offers sustainable and environmentally friendly solutions for industries increasingly driven by enzymes. Thanks to millions of years of evolution, enzymes are fine-tuned to carry out specific chemical reactions with high efficiency. This makes them attractive alternatives to traditional catalysis, which often relies on harsh conditions and toxic chemicals.[1] Thus, these biocatalysts find application across various industries, including pharmaceuticals, food production, and sustainability efforts aimed at reducing waste and energy consumption.[2] Since natural enzymes often exhibit suboptimal performance in non-native environments, enzyme engineering is usually required to unlock their full potential.[3,4] In addition to commonly used experimental approaches such as directed evolution, scientists can also expedite the process and reduce associated development costs by incorporating computational methods.[5,6] One group of computational methods rely on physical-based modeling techniques such as Empirical Valence Bond (EVB)[7−9] and hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods, which simulate atomic-level interactions and energy landscapes of enzymes.[10,11] Another group of methods exploit protein sequences. These methods help navigate the vast sequence space, as it is estimated that only a fraction of all possible sequences fold into functional protein structures.[12] Most natural proteins have marginal stability,[13] thus posing a significant risk for any manipulations with their sequences.

Many computational methods aiming to refine the search space of such sequence manipulations rely on homologous sequences.[14,15] These sequences of different but related proteins stemming from a common ancestor contain rich evolutionary information.[16] Homologous protein sequences can be employed to identify conserved and functionally important regions, suggest beneficial mutations, and create phylogenetic trees.[17] Notable examples of approaches in this context include the Maximum Entropy (MaxEnt) model and ancestral sequence reconstruction. The MaxEnt model

Received: November 18, 2024
Revised: January 23, 2025
Accepted: January 30, 2025
Published: February 13, 2025

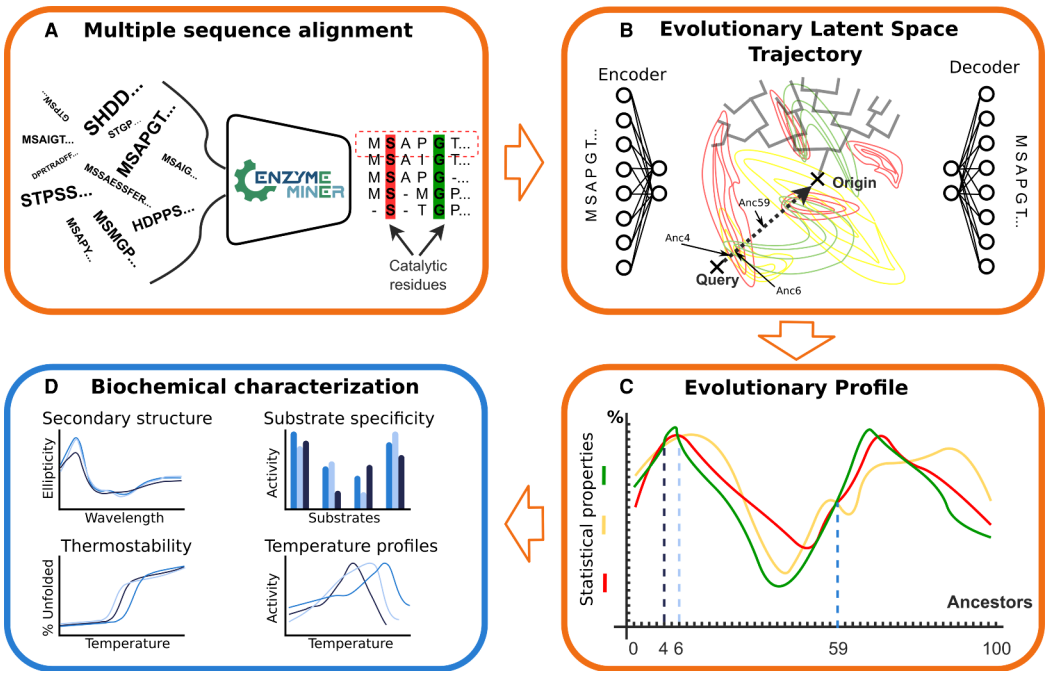


Figure 1. The scheme of the variational autoencoder-based pipeline for the design of novel sequences. (**A**) Advanced sequence search of homologous proteins using EnzymeMiner.[47] (**B**) Optimization of the variational autoencoder architecture to capture the sequence distribution of the MSA and phylogenetic dependencies within the latent space. (**C**) Exploration of the evolutionary dependencies between the sequences extracted from the variational autoencoder and its low-dimensional latent space. This representation is then used to guide the protein design strategy and generate sequences along the trajectory from the query to the latent space origin. The generated sequences are characterized based on their statistical and sequential properties to produce the evolutionary profile. This profile serves as a guide for selecting designs. (**D**) The experimental characterization of the proposed designs is conducted. The orange frames represent the computational steps and the blue frame is the experimental step.

employs statistical energy derived from homologous sequences, applying the maximum entropy principle to establish correlations with enzyme catalysis and stability in both the active site and more distant regions.[18,19] Ancestral sequence reconstruction utilizes phylogenetic trees and sequence alignment techniques to trace evolutionary changes and infer the sequences of ancestral proteins. This method has proven to be a promising strategy for enhancing protein stability.[20−24]

Despite the recent progress in extracting evolutionary information from multiple sequence alignments (MSA) of homologous proteins, analyzing this variability is challenging. Historically, this data was used primarily by looking at only one or two positions at a time.[25,26] More recent approaches extract patterns by deep neural networks, in particular algorithms that map the sequence space onto their internal low-dimensional representation, also referred to as latent spaces. Generative models trained on large data sets of tens of thousands of sequences have shown excellent results in producing highly interpretable embeddings and generating novel protein variants.[27−30] The most recent examples of this class include diffusion models, which are trained to denoise synthetically noised inputs. They were initially used to generate protein backbone structures[31,32] and predict the binding of a flexible ligand to a protein[33] but later have been adapted to generating sequences as well.[34,35] Another example is Generative Adversarial Networks, which learn to generate new data through competitive training involving two artificial neural networks. For instance, ProteinGAN was used to generate functional protein sequences of malate dehydrogenases.[36] Variational autoencoders (VAEs) are a third type of MSA-based models, which shows a particular promise in this domain due to the explicit modeling of the latent space.[37] VAEs have already proven useful in several applications, including predicting protein structures,[38] discovering novel drugs,[39] and predicting protein functions.[40] By learning the latent space representation of a specific family, VAEs provide valuable insights into the evolution of protein families, as demonstrated in recent studies exploring the phylogenetic relationships within the latent space.[41−43] In particular, Ding et al. showed that the latent space of the variational autoencoders can capture the biophysical properties of protein variants and the phylogenetic relationships within protein families.[41] However, the study did not offer a strategy that would allow exploiting these relationships to generate new proteins from the latent space. For a comprehensive overview of generative models, we refer the reader to an excellent recent review.[44]

Here we propose a simple strategy to leverage the evolutionary-shaped geometry of the VAE-learned latent space to design novel ancestral-like variants of haloalkane dehalogenases (HLDs; EC 3.8.1.5). These enzymes cleave the carbon−halogen bonds[45] and are widely used in biocatalysis, biosensing, cell imaging, and protein analysis.[46] The proposed workflow is based on a small number of proteins with known functions and aims to produce new variants that preserve catalytic function and improve stability (Figure 1). First, we mined sequences with preserved catalytic residues using EnzymeMiner[47] to obtain an MSA of functionally related proteins. Second, we trained a VAE and specified several metrics to measure its capacity to generate protein sequences and capture the phylogeny in the constructed latent representations. Third, based on the geometry of the latent space, we developed a sampling strategy and produced a

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

statistical profile of candidate sequences to select promising variants from the evolutionary trajectory. Fourth, we overexpressed and characterized variants experimentally using advanced microfluidics. Three consecutive rounds of experimental characterization and workflow optimization resulted in 20 variants, sharing as little as 67% sequence similarity to known HLDs. Obtained enzymes showed up to a 9 °C increase in melting temperatures and an average improvement of 3 °C across all soluble variants. We also observed a boost in activity, up to 3.5-fold for the most stable variant, whereas most of the other expressed variants showed activity levels comparable to benchmark enzymes.

MATERIALS AND METHODS

MSA and Data Preprocessing

Two data sets, HLDI-IV, and HLDI-II, were created using the EnzymeMiner tool[47] based on haloalkane dehalogenase sequences. Both data sets underwent preprocessing, including sequence filtering, gap reduction, and clustering, to ensure diversity. The second data set also underwent additional adjustments to reduce gaps and improve solubility rates. Detailed descriptions of the preprocessing steps and data set creation are provided in SI Section 1.1. These data sets were then used to train models in multiple experimental rounds, specifically HLDI-IV for rounds I and II, and HLDI-II for round III.

Two sets of experimentally measured stability values were mapped to the latent space of VAE Model 1. The first set consisted of six ancestral sequences from the previous ancestral campaign of the thoroughly characterized dehalogenases DbjA, DbeA, DhaA, DmxA, and DmmA.[48] These sequences were realigned with the original input MSA and preprocessed accordingly. The second set consisted of 24 previously engineered DhaA variants based on the FireProt method,[22,49,50] similarly aligned and preprocessed with the query sequence P59336_S14 of the input MSA.

Variational Autoencoders and Training

Variational autoencoders (VAEs)[37] are a type of deep generative learning model whose goal is to learn the data distribution. VAEs consist of two main components: an encoder and a decoder (Figure 1B). The encoder takes the input sequence and maps it to a lower dimensional representation called a latent space. Within this latent space, the encoded input is modeled as a normal distribution by two parameters, the mean and the variance. Subsequently, the decoder draws samples from this latent space distribution and maps these samples back to sequences. The training of VAEs is based on minimizing the loss function made of the reconstruction term (penalizes incorrect reconstruction of the input data) and the regularization term (serves to constrain the latent space distribution of encoded values). The latter forces the latent space to be close to normal distribution by measuring the Kullback−Leibler divergence. As a result, the individual distributions are forced to overlap within the latent space, ensuring proper alignment of the sequences corresponding to the nearby points in the latent space. We tested several architectures and eventually used one hidden layer in the encoder and decoder, both composed of N neurons, where N is the width of preprocessed MSA (number of positions), the latent space dimensionality of 2, and either zero (rounds I and II) or decreasing (round III) weight decay. We employed the tanh activation function without dropout and assigned equal weighting to the reconstruction and regularization terms in the training objective (SI Section 1.4). The final model had 3 million parameters. We used the Adam optimizer with a learning rate of 0.001 and stopped training after not improving the loss function for more than 3 consecutive rounds.

In the conditional variational autoencoders (CVAEs),[51] a tag (LOW, MEDIUM, HIGH) was added to the encoder and decoder to represent different solubility levels. The tags were generated based on the solubility values predicted by SoluProt[52] and binned to achieve uniform distribution across bins and ensure balanced sample extraction.[53] A detailed description can be found in SI Section 4.

Model Generative Capacity

We performed a first- and second-order statistical analysis to compare the frequency of amino acids at each position in the multiple sequence alignment (MSA) between input and generated data sets. First-order statistics assess the occurrence of each amino acid at a given position, while second-order statistics capture pairwise relationships between two positions. We computed the pairwise covariance scores to evaluate how well the generative model reproduces interactions between amino acids, an essential indicator for the likely stability and function of the generated proteins.[54] For this study, we used 3,000 randomly selected samples from both input and generated data sets for the statistical comparisons. A detailed description is provided in SI Section 1.2.

Average Reconstruction Accuracy and Controls

The average reconstruction accuracy of each sequence was approximated as an average reconstructed sequence identity for 5,000 samples around the original sequence coordinates of its latent space embedding based on the mean and variance returned by the encoder for a given sequence. The negative control subset was generated by sampling sequences from the profile of the input MSA only based on the amino acid frequencies in each position. The positive control subset comprised 5% of preprocessed sequences randomly selected from the MSA and excluded from training. Finally, the ancestral subset was composed of 100 reconstructed sequences by the straight evolutionary strategy (see Construction of the evolutionary trajectory). Except for the ancestral subset, all the subsets contained the number of sequences corresponding to 5% of the preprocessed data set.

Phylogeny Mapping and Evaluation

We generated 13 phylogenetic trees using our input MSA to analyze the relationship between phylogenetic branches and the latent space. Each tree had around 100 randomly sampled nodes, and ancestral sequences were reconstructed using FireProtASR.[22] We explored the correlation between the depth of nodes and their latent space embeddings and analyzed the directionality of tree branches within the latent space similar to.[41] See SI Section 1.3 for more details.

AlphaFold Structure Prediction and Manual Analysis of the Suggested Mutations

For structural predictions of ancestral sequences, the AlphaFold2 Google Colaboratory implementation, ColabFold, using MMseqs2, was used.[55] We predicted structures without providing templates, and we performed amber relaxation with 200 steps on the top-ranked structure. We used the default MSA options with pair sequences from the same species and unpaired sequences from separate MSA for each chain (paired+unpaired option). The optimal structure was selected automatically by ColabFold. The refinement process was repeated over three cycles to improve the structure's accuracy. The relaxed first-ranked structure was used as the result of the prediction.

In round 3, the proposed mutations by VAEs were also curated manually (see SI Section 2 for more detail). The visual inspection of the modeled AlphaFold variants was performed by Pymol,[56] and the MutCompute web server[57] was used to calculate the score per residue (log-likelihood ratio). Thus, a positive score indicates that MutCompute assesses the substituted residue as more likely to occur in the given structural microenvironment than the wild-type residue.

Protein Production, Purification and Whole-Cell Activity Screening

First,*E. coli*BL21(DE3) cells (NEB, USA) were transformed with the pET21b expression plasmid containing the corresponding gene, plated on LB-agar with 100 μg/mL ampicillin, and incubated at 37 °C overnight (12−16 h). Cells transformed with pET21b::DhaAwt, pET21b::RLuc, and empty pET21b served as controls. For small-scale protein overexpression and affinity purification, cultivation in 96-deep well plates, harvesting, SDS-PAGE analysis, and high-throughput affinity purification using TALON SuperFlow Metal Affinity Resin (Takara) were performed (see details in SI Section 7.1). Cell

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

cultivations for enzymatic screenings and halide oxidation (HOX) assay[58] were carried out, including cell cultivation, harvesting, and whole-cell activity screening (see details in SI Section 7.2). For large-scale protein overexpression and purification, selected mutant enzymes were expressed in *E. coli* BL21(DE3), and purification was done using metal affinity resin and gel filtration (see details in SI sections 7.3–7.5).

Secondary Structure Experimental Validation

The secondary structure of the analyzed variants was experimentally verified using circular dichroism (CD) spectroscopy, measured at 15 °C using a spectropolarimeter Chirascan (Applied Photophysics). The samples were dissolved in 1 mM HEPES buffer or in the 50 mM Phosphate buffer, and their concentration was adjusted to ∼0.18 mg/mL. Data were collected from 185 nm to 260 nm with 0.25 s integration time and 1 nm bandwidth using a 0.1 cm quartz cuvette. Each spectrum was obtained as an average of five individual repeats. Prediction of CD spectra was performed by PDBMD2CD[59] (https://pdbmd2cd.cryst.bbk.ac.uk), using either experimental structures from PDB database (1CQW for DhaA) or AlphaFold models. The estimation of secondary structure elements from experimental data and PDB database structures was performed additionally by BeStSel[60] (https://bestsel.elte.hu/).

Thermal Denaturation by CD and NanoDSF

Thermal unfolding of selected enzyme variants was carried out using a Chirascan spectropolarimeter (Applied Photophysics, UK). Each protein sample was diluted in 50 mM Phosphate buffer to the concentration of 0.18 mg·mL^{-1} and measured in a 0.1 cm quartz cuvette. Changes of ellipticity were monitored at three wavelengths (195 nm, 210 nm, and 227 nm) from 15 to 80 °C with a 0.1 °C resolution and 1 °C·min^{-1} heating rate. Recorded data were fitted using the model "Sigmoid curve + slope" in the Pro Data Viewer software (Applied Photophysics, UK). The apparent melting temperature (T_m^{app}) was evaluated as a midpoint of the normalized thermal transition.

Thermal unfolding was further studied using NanoDSF Prometheus NT.48 (NanoTemper, Germany) by monitoring tryptophan fluorescence over the temperature range of 20 to 95 °C, at a heating rate of 1 °Cwith 20% excitation power. The thermostability parameters (T_{on} and T_m^{app}) were evaluated directly by ThermControl v2.0.2.

Dehalogenase Activity Measurements on MicroPEX

Activity measurements for the determination of temperature profiles and substrate specificity were conducted on the capillary-based droplet microfluidic platform MicroPEX,[61] enabling the characterization of specific enzyme activity within droplets for multiple enzyme variants in one run. A detailed description of the microfluidic method can be found elsewhere[62,63] and briefly in SI Section 7.6.

RESULTS

We developed the pipeline to leverage the power of the variational autoencoder and its latent spaces for the design of promising biocatalysts (Figure 1). This pipeline was inspired by the previous studies reporting the connections between latent space geometry and phylogeny for a given MSA,[41,43] however, this connection has not been exploited for generating new protein sequences thus far. We hypothesized that the coordinates within the latent space could serve as a navigational tool for identifying ancestral-like sequences, offering a way to improve the stability of query proteins while maintaining their function. We iteratively executed our pipeline across three rounds, each iteration followed by experimental validation to improve our workflow (SI Section 2). In both the first and second rounds, we utilized the same trained VAEs (Model 1), with the only difference being a revised selection of VAEs ancestors. In the third round, we introduced changes to the MSA preprocessing and additionally manual curation of the generated ancestors by AlphaFold and MutCompute.[16,57] In addition, during the third round of experimental validation, we explored the possibility of conditioning the VAEs on solubility scores returned by the ML-based tool SoluProt.[52] In total, three trained VAEs models were explored in the third round (Models 2–4) to better understand the strengths and weaknesses and refine our approach.

Multiple Sequence Alignment Processing

The Data Collection Is Optimized to Preserve Catalytic Activities. The first step of our pipeline is to construct an MSA. Instead of using Pfam alignments as in,[41] we narrowed the search of relevant sequences to those likely to preserve the dehalogenation activity. Pfam MSAs are sometimes too broad, introducing large-gapped regions and making it difficult to design proteins with desired functions.[64] To overcome this challenge, we used the EnzymeMiner web tool,[47] which generates alignments specifically selected for function and catalytic site similarity (Figure 1A) as recently demonstrated on diverse enzyme families, such as fluorinases, marine bacterial flavozymes, and NADPH-dependent reductive aminases.[61,65−67]

The query of haloalkane dehalogenase (DhaA) from *Rhodococcus* strain TDTM0003 with UniProt ID P59336 yielded 22,567 sequences in EnzymeMiner. This extensive search resulted in the creation of a data set named HLDI-IV. To further refine the results, we preprocessed the resulting MSA against the DhaA query by removing protein sequences and positions with too many gaps. This step narrowed down the size of the alignment to 12,053 sequences and 299 positions, which were used for training. The HLDI-IV data set was utilized in both the first and second rounds of wet lab experiments (SI Section 2, Figure S1).

Based on the low experimental solubility observed in the first two rounds, we implemented a stricter protocol for creating the initial MSA. Inspired by Vasina et al.,[61] we focused on more soluble HLD subfamilies I and II, generating a smaller MSA. This data set included the well-characterized DhaA enzyme from *Rhodococcus sp.* (UniProt ID P0A3G3), which served as the updated query for MSA preprocessing. We applied additional filters to reduce gap frequencies, lowering the threshold for gap column removal and filtering columns with frequent gaps, even if the query had an amino acid in that position, resulting in an MSA width of 293 positions with 4,053 sequences (HLDI-II dataset).

Network Architecture Optimization

Variational Autoencoders Capture Sequence Spaces and Sequence Distribution. Replicating the methodology described by Ding et al.[41] (SI Section 1.3), we demonstrated the capacity of variational autoencoders (VAEs) to delineate phylogenetic relationships among proteins in our HLDI-IV data set (Model 1). By encoding sequences into a latent space where evolutionary-related sequences map to nearby points, we observed a star-like configuration with multiple spikes radiating from a central point, reflecting the evolutionary divergence within the data set (Figure S3). This structure contrasts with the dispersed and unstructured representation of random sequences, highlighting that the latent space for our sequences captures their phylogenetic relationships, consistent with observations described in the original publication.[41]

Before testing our hypothesis of generating ancestral-like sequences, we embarked on selecting the best model

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

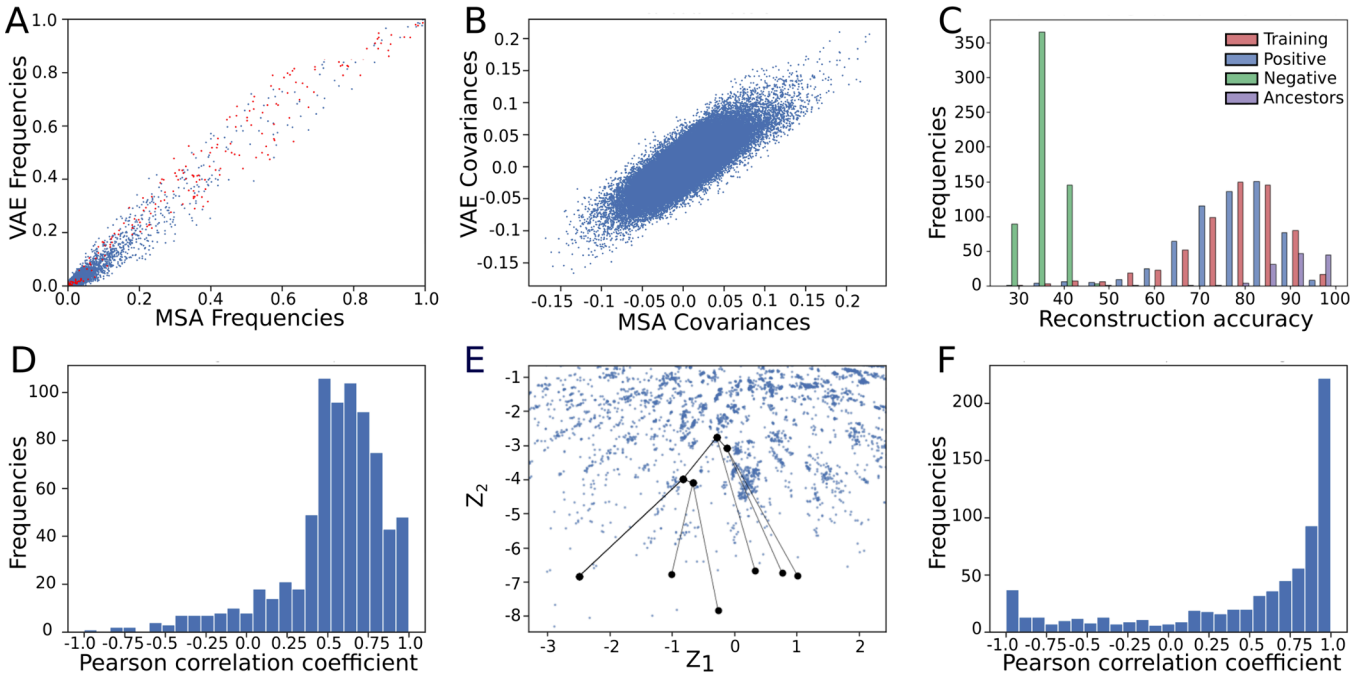


Figure 2. Showcases of the statistics used to measure the generative capacity of the final VAE model (Model 1, see SI Section 1) and the geometric properties of its latent space. (**A**) The first-order statistics for 3000 sequences randomly selected from the input MSA or VAE-generated. The red dots represent the gap symbol frequencies in sequence positions, while the blue points denote amino acids. (**B**) The second-order statistics demonstrate that our model can reconstruct pairwise amino acid occurrences fairly well (ρ = 0.68). (**C**) The average reconstruction accuracy for the negative (green), training (red), positive (blue), and ancestral (violet) control data sets. The shifts in the histograms between the sets imply that the model can distinguish random sequences (negative) from those in the input MSA (training and positive) and those corresponding to the straight-line strategy of generating ancestors (ancestors). (**D**) The Pearson's correlation between depth in phylogenetic trees and latent space origin distance. Most sequences in tree branches have a positive correlation indicating that the latent space captures phylogeny. (**E**) Mapping a small phylogenetic tree onto the latent space. (**F**) Histogram illustrating the directional trends of phylogenetic tree branches projected onto the latent space. In this representation, 1 indicates a straight trajectory toward the latent space origin, while −1 represents the opposite trend. The histogram highlights that the majority of branches tend to align toward the latent space origin.

architecture based on the implementation provided by Ding et al.[41] This involved optimizing the encoder, decoder, and training procedure to minimize the difference between the generated and input sequence distributions (generative capacity)[54] while also preserving the relationship between phylogeny and the latent space (geometric properties). In order to evaluate the model's generative capacity, we implemented several tests. The first test examined how well our model reproduces the statistics of the input data set on the output. To this end, we compared the first and second-order statistics of 3,000 randomly sampled MSA input sequences with those generated by our VAE model (Figure 2A-B) following the approach outlined in previous studies.[26,54,68] The comparison revealed a close match between the two sets. We integrated query reconstruction accuracy[69] as an additional metric in our analysis to ensure that the model was capable of reconstructing the query sequence with minimal mutations. Notably, the final Model 1 demonstrated the ability to reconstruct 98% of the query.

Our second test evaluated the model's statistical profile by measuring the average reconstruction accuracy for sequences from various control sets. The test showed that the model could distinguish random sequences from MSA sequences using a reconstruction accuracy cutoff of around 50%. In other words, all the sequences from the negative set had an average reconstruction accuracy below this threshold together with only 23 of sequences from training and positive control sets, while the remaining 1201 sequences from these two control sets had an average reconstruction accuracy above this threshold (Figure 2C).

VAEs Capture Evolutionary Trends. To preserve evolutionary information in the latent space, we monitored the relationship between phylogeny and latent space geometry (Figure 2D-F). Phylogenetic trees with inferred ancestral sequences were mapped into the latent space, and we quantified the distance between latent space points and their corresponding positions in the phylogenetic tree (Figure 2D). Additionally, we analyzed the angle between vectors from leaf nodes to the origin and the first principal component of the branch's latent coordinates (Figure 2F). Our results show that small dense encoder-decoder architectures capture evolutionary dependencies, while deeper architectures disrupt them (Figure S4). We set the dense layer width to match the protein sequence length and used a latent space dimensionality of 2 for simplicity and effective representation. Testing with higher-dimensional latent spaces did not yield significant improvements in reconstruction performance, further supporting our choice of a two-dimensional latent space (Table S1). Repeating this for Models 2−4 confirmed our findings with a correlation of 0.8, supporting our choices for layer width and latent dimensionality (SI Section 2).

Construction of the Evolutionary Trajectory

The Latent Space Captures Protein Stability. The ancestral sequences are often associated with enhanced stability compared to their extant counterparts.[23,70] We hypothesized that the structure of the latent space might

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

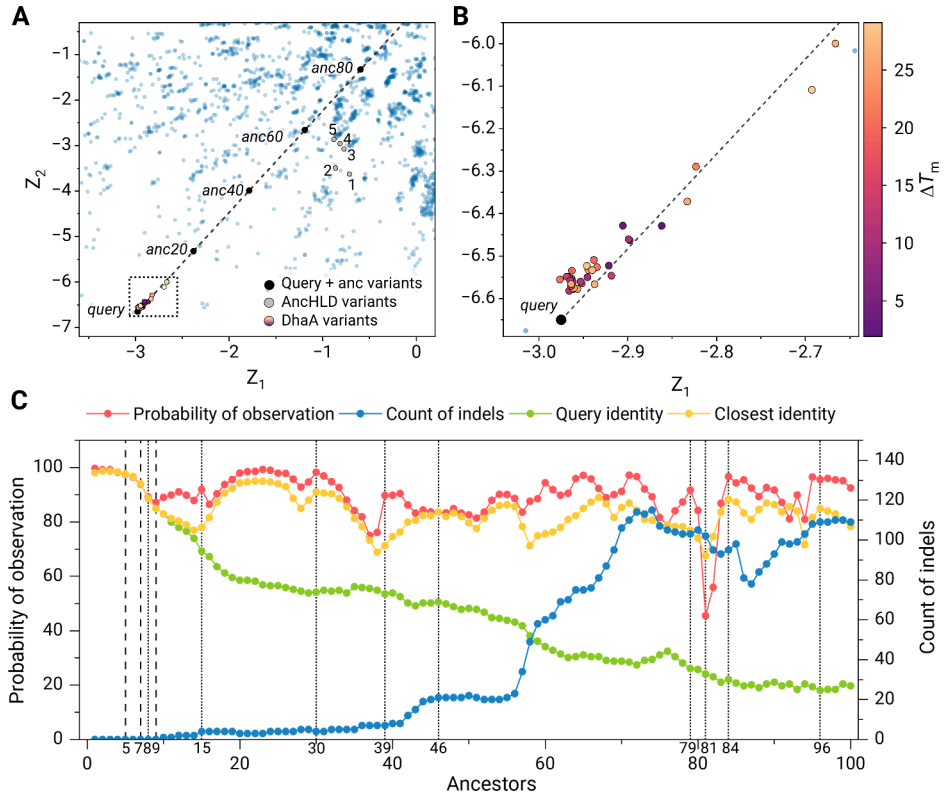


Figure 3. The straight-line evolutionary strategy for Model 1. (**A**) Straight-line evolutionary strategy reconstructed 100 sequences along the trajectory from query embedding to the latent space origin (black dashed line). The embeddings of previously characterized ancestors (gray points 1−5 denoting AncHLD variants of the respective number[48]) and engineered DhaA variants[49] (magma spectrum points) are mapped closer to the latent space origin, supporting the idea behind our ancestral generation strategy. (**B**) A detailed view of the previously engineered DhaA variants. While there is no strong correlation between the positions in the latent space and the stability gain (ΔT_m) of variants up to 28 °C, some of the most stable points are situated closer to the origin. (**C**) The statistical profile of 100 sequences from the straight-line evolutionary strategy. The vertical lines represent sequences selected for experimental characterization for the first and second rounds (Table S2) where dashed line variants were successfully expressed, while for dotted lines, no soluble expression was observed. The ancestors are numbered 1 to 100 based on their order in the VAE-generated latent space, with lower numbers being closer to the starting sequence and higher numbers representing more divergent designs closer to the latent space origin. Number 0/Query represents the reconstruction of the original embedding of the query sequence.

encode stability and place more stable variants of our target protein DhaA closer to the origin compared to the wild type. To test this hypothesis, we mapped two sets of experimentally measured stability values to the latent space of Model 1. The first set consisted of six ancestral sequences from the previous protein engineering campaign (DbjA, DbeA, DhaA, DmxA, and DmmA).[48] The second set consisted of 24 previously engineered DhaA variants based on the FireProt method.[22,49,50] Both data sets had latent space coordinates closer to the origin of the latent space (Figure 3A-B), supporting the notion that the latent space captures the information about stability. The observations were also recapitulated for HLDI−II data set and Model 2.

Latent Space Strategy Guides the Search and Selection of Sequences. The regular distribution of stable sequences in the latent space (Figure 3B) led us to develop the *straight-line evolutionary strategy*. This strategy encodes the query sequence (DhaA in our case) into its latent representation and then follows the straight line connecting that point to the origin of the latent space (Figure 3A), mimicking the mapping of ancestral dependencies into the latent space. In our experiments, the line is divided into 100 equal intervals, whose boundaries are then selected for reconstruction by the decoder. The motivation for the *straight-line evolutionary strategy* is based on the observation by Ding et al.[41] that ancestors tend to be placed closer to the origin of the latent space. Therefore, sequences reconstructed in this direction from the latent space might be ancestor-like, e.g., show increased stability. For the sake of brevity, in what follows we will refer to these ancestor-like designs generated by the decoder from the straight-line evolutionary strategy in the latent space as "ancestors" and use the prefix "Anc".

To represent the designed sequences, we analyzed several statistical parameters for the individual designed sequences: the average reconstruction probability, similarity to the query sequence, similarity to the closest sequence from the training set, and the number of insertions/deletions compared to the query sequence. The values obtained were plotted and visually inspected to identify variants with interesting statistical values. The generated profiles were then used to select suitable variants for subsequent experimental characterization (Figure 3C).

Using the statistical profile, we identified 9 promising designs in the first round for further laboratory experiments to gain deeper insight into the statistical indicators. These designs exhibited a wide range of sequence variability, ranging from 45 substitutions and no insertions or deletions (indels) to 138 substitutions and 109 indels (Table S2) (AncDhaA1−9). The substitutions and indels, with deletions in most cases, covered the entire protein structure. In the second round, we focused

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

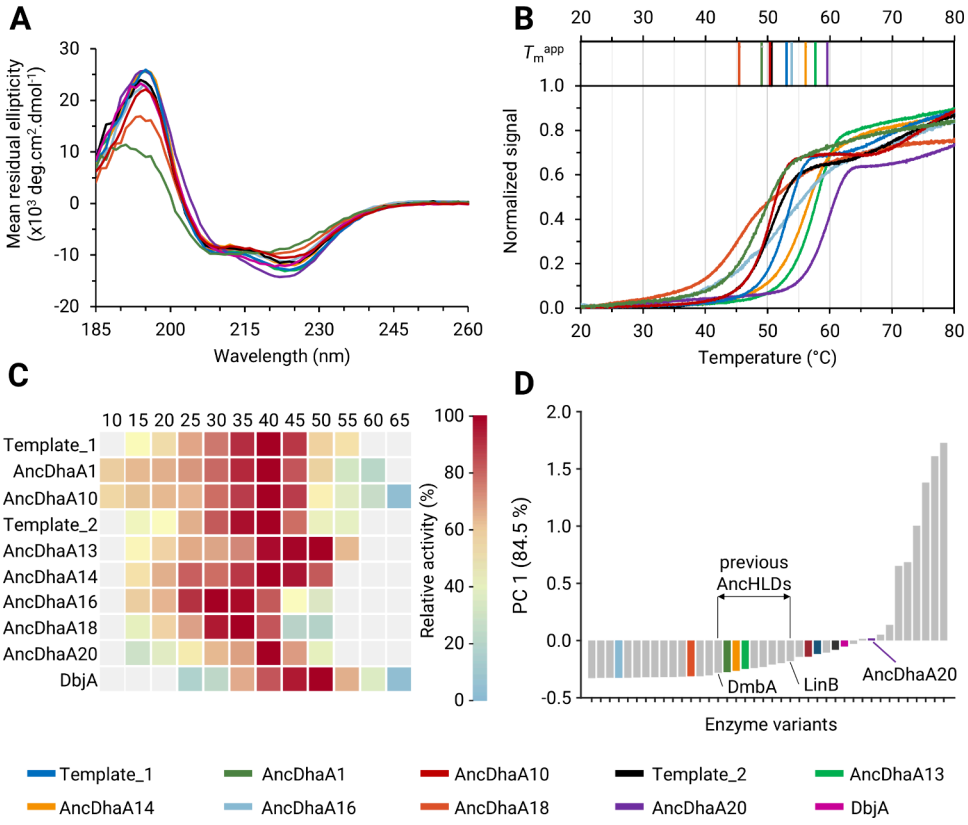


Figure 4. Experimental characterization of selected variants. (**A**) Far-UV circular dichroism spectra probing the correct folding and secondary structure of the variants. (**B**) Normalized thermal denaturation curves from nanoDSF spectroscopy with apparent melting temperatures (T_m^{app}) are shown above the curves. (**C**) The dependence of specific activity on temperature. The heatmap represents the relative activity of individual variants. (**D**) The score plot shows the first principal component PC 1 explaining 84.9% of the data variance, which compares VAE-based designs (in color) with previously characterized wild-type haloalkane dehalogenases (gray)[61] in terms of their activity with 27 substrates being determined by the MicroPEX method.[62] The highlighted range between DmbA and LinB corresponds to the ranges of values observed for previously characterized AncHLD variants.[48] The values of PC1 above this range imply that the overall activity of the corresponding designs was higher than those for previous AncHLD variants. The color code for individual variants is shown at the bottom.

on the more conserved variants (ancestors 5, 7, and 8 in Figure 3C; AncDhaA10−12 for reference in experiments) with 7 to 34 substitutions without indels. Altogether, 12 designs were selected for laboratory expression and biophysical characterization from Model 1.

Variant Selection Conditioned for Soluble Proteins

Solubilization of Designs by Introducing New Knowledge to the Data Set. As we observed low solubility in the first round (see Section Experimental Characterization of Expressed Variants and Figure S10A-B), we incorporated previous findings on the low solubility of HLD subfamilies III and IV[61] in our workflow. To this end, we embarked on a third round of experiments restricted to the HLD subfamilies I and II. This round focused on training additional models (Models 2−3) and designed eight DhaA variants (AncDhaA13−20, Table S2). The first candidate VAE (Model 2) achieved up to 97% similarity in query sequence reconstruction with satisfactory second-order statistics. Considering the often-disruptive impact of indels, we selected and curated five designs from the straight-line evolutionary strategy accumulating at most one indel (AncDhaA13−17). Another VAE was trained from a different initialization (Model 3). During analysis of this model, we found that at the end of reconstructed sequences, it incorporated a His-tag sequence, common peptide tag for protein purification. The inclusion of the His-tag likely happened during the training phase, which influenced the model's generation of sequences. Overall, 42 out of 100 reconstructed sequences from the *straight-line evolutionary strategy* exhibited the closest sequence similarity toward 5 different protein sequences of PDB structures found in HLDI−II data set. Additionally, Model 3 had a curious pattern for the origin of the latent space: the model demonstrated a significant shift in sequence similarity toward DbjA. Therefore, we selected ancestor 99 (AncDhaA20) from this model for further experimental characterization to explore its unique shared sequence similarity to both DhaA (52%) and DbjA (93%).

Solubilization of Designs by Conditional Variational Autoencoder. Finally, in the third round, we also decided to explore one more solution to low solubility, conditional variational autoencoders (CVAEs). To this end we added solubility scores from SoluProt[52] to the training, discretized into three bins for low, moderate, and high solubility values (Model 4) (Figure S6A). We conditioned the sampling process from Model 4 using a straight-line evolutionary strategy on the highest bin label forcing CVAEs to introduce patterns from sequences with predicted high solubility in decoded designs (Figure S5). We took two variants from Model 4: ancestor 0 (AncDhaA18) with 30 mutations as the control of the pattern extracted from highly soluble sequences and ancestor 18 (AncDhaA19), which had 49 mutations and one deletion in

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

the coil region at position 32 (Figure S6B), as it exhibited high confidence in the model (reconstruction probability of 93.28%) and an increased number of high probable residues (85 positions scored above 90%).

Refining the Mutations by Manual Curation of the Structures and Stability Scores. From Model 2, we selected three ancestors with unique features: ancestor 3 closely mimicked the wild type with 98% similarity; ancestor 16 notably introduced a proline at position 75; and ancestor 23 was the last variant starting with a regular sequence pattern (MSEIGT), suggesting high solubility and expression potential based on its 88% similarity to the PDB sequence4WCV. To increase our chances of producing soluble variants, we manually curated proposed mutations based on the structural predictions from AlphaFold[16] and stability assessments using the MutCompute tool[57] (Table S3). Mutation manual curation led to the classification of VAE-proposed mutations into safe and risky categories, respectively (SI Section 3). Starting from ancestor 3, we kept four nonrisky mutations. We removed two structurally and statistically risky mutations P34V and L238F from the original six-point variant proposed by the VAE, producing the design AncDhaA14. As experimental validation of identified risky mutations, we selected ancestor 0 with eight substitutions (AncDhaA13). For the second manually curated variant (AncDhaA16), we selected ancestor 15 as the template, which included as the last VAEs ancestor with no insertions, and we removed the risky mutations P34V and L238F, leaving nine mutations compared to the DhaA wild type. To experimentally determine the effect of suggested proline insertion with risky mutations, we selected ancestor 16 as predicted by VAEs (AncDhaA15) (Figure S2). As a template for the third manual curation target, we selected ancestor 23. We incorporated all mutations suggested by the VAEs except the risky ones (P34V, L238F) and the proline insertions, resulting in a final design carrying 32 mutations (AncDhaA17).

Investigation of Trajectory Mutational Patterns in VAEs Designs. To gain a comprehensive understanding of how VAEs propose mutations across designs, we further analyzed mutational patterns along the evolutionary trajectory generated by VAEs Model 2. The analysis indicated that mutations are not distributed randomly but tend to accumulate at specific positions, suggesting a targeted evolutionary trajectory rather than a stochastic process. Additionally, certain mutations introduced in earlier designs were retained in subsequent designs, while others were reverted, indicating an iterative refinement process. Further details on the mutational profiles and alignment patterns of the generated ancestors are provided in SI Section 6 and Figure S8, S9.

Experimental Characterization of Expressed Variants

Variant Production. Protein overexpression, purification and assessment of solubility and whole-cell activity testing was carried out in three rounds (Table S2, SI Section 7), yielding 9 soluble variants: AncDhaA1 (round 1), AncDhaA10−11 (round 2) and AncDhaA13−16, 18, and 20 (round 3). Thus, the success rate in obtaining soluble variants gradually increased from 11% in round 1 to 67% in round 2 and reaching 75% in round 3. The whole-cell activity screening by HOX assay revealed that 6/9 soluble variants were active with a benchmark substrate 1,2-dibromoethane (Figure S11).

Secondary Structure Analysis. To confirm the proper folding of the studied variants, circular dichroism (CD) spectra were collected for all soluble variants (Figure 4A). Overall, CD spectra of most variants highly resemble those of the templates (typical α/β-hydrolase fold), confirming proper folding. On the contrary, the spectra of AncDhaA1, AncDhaA11, AncDhaA15, and AncDhaA18 (Figure S12) deviated from the templates. To further understand the secondary structure of the variants, BeStSel server[60] was used for fitting experimental data and analysis of PDB structures, and PDBMD2CD[59] was used for predicting CD spectra from experimental structures of templates and AlphaFold models of selected variants. Figure S12B shows that the prediction of CD spectra based on AlphaFold structures did not match the experiments in all the variants. This highlights a limitation in AlphaFold's ability to accurately predict changes in folding and emphasizes the need for further improvements in computational methods. Experimental validation remains essential to address this challenge.

Thermostability. Thermodynamic stability of all variants was assessed by nanodifferential scanning fluorimetry (Figure 2B, Table S4). The apparent melting temperatures for the variants were in the range of 45 °C−60 °C. AncDhaA13, AncDhaA14, AncDhaA16 and AncDhaA20 surpassed the respective query in terms of apparent melting temperature. The highest ΔT_m of 9 °C was measured for AncDhaA20. Protein aggregation was observed for the variants AncDhaA1 and AncDhaA20, showing the onset at 45.5 and 48.5 °C, respectively (Figure S13).

Temperature Profiles. We then proceeded to measure temperature profiles (Figure 4C). Most variants showed the T_{max} (temperature at which maximum activity was detected) of 40 °C, which is in agreement with previously determined temperature profiles for DhaA.[62] Notably, AncDhaA13 showed T_{max} of 50 °C, which aligns with its increased thermostability. The temperature profiles for AncDhaA11 and AncDhaA16 were not obtained, as the activities were below the detection limit. Due to compromised activity and folding, both variants were excluded from the subsequent substrate specificity profiling.

Substrate Specificity. The temperature of 35 °C was selected for the subsequent specificity characterization for being below the onset of denaturation (Table S4) for most of the remaining variants and close to their T_{max} values. To explore the obtained substrate specificities in the context of the haloalkane dehalogenase family, the principal component analysis (PCA) was conducted by augmenting the previously used data set comprising substrate specificities for 32 wild-type dehalogenases[61] with the newly obtained data. The PCA of raw data (Figure 4D) as a standard representation of overall dehalogenase activity[46,61,71] showed that AncDhaA20 surpassed both Templates, DbjA and the previously characterized AncHLD variants.[48] The higher values of PC1 indicate the higher overall activity, as the first principal component corresponds to the weighted average of all the activities, shifted to be centered around zero. The acquired substrate specificity data for previously characterized AncHLD variants are not directly comparable with the MicroPEX data, however, due to different assays used. Therefore, only an approximate area of graph where AncHLD variants stand in terms of overall activity could be determined (Figure 4D). The overall activity of AncDhaA10 was also higher than the previously characterized AncHLD variants, on the level of the wild type. Three more variants showed activity in the ranges of previously characterized AncHLD, and two more below (Figure S14).

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

The analysis of log-transformed activity data (Figure S15) further showed that AncDhaA1, AncDhaA10, AncDhaA13, and AncDhaA14 differed only very slightly from templates in their substrate preferences. The profiles of AncDhaA18 and AncDhaA20 resembled more closely the specificity profile of DbjA, which is not unexpected in the case of AncDhaA20 due to its high sequence similarity. AncDhaA16 differed significantly from all other variants due to the low number of converted substrates (10 out of 27). Probable activity with other substrates could have ended below the limit of detection (different for each substrate, in the range of 10−100 μM).

DISCUSSION

In this study, we utilized variational autoencoders (VAEs) to map functionally related haloalkane dehalogenase sequences from EnzymeMiner onto VAEs latent spaces, following an approach inspired by Ding et al.[41] This process revealed that the latent space could capture the phylogenetic relationships of the sequences, which motivated us to employ the VAE framework to create ancestral-like variants of the haloalkane dehalogenase DhaA. We tuned the network hyperparameters to accurately reflect the statistical frequencies of the input while maintaining the relationship between evolutionary trajectories and latent spaces. We discovered that a simple feed-forward neural network with a single dense layer matched to the input MSA columns and the two-dimensional latent space[42] was enough for this task.

We then introduced a simple strategy to generate novel sequences based on the geometry of the latent space. We achieved this by reconstructing the embeddings along the trajectory toward the origin of the latent space. Employing this strategy, we systematically executed the pipeline in three rounds of laboratory experiments and optimization, leading to four VAEs models producing 20 designs in total. Similarly to the study of computational filter evaluation for synthetical protein designs from generative models,[72] each iteration revealed new insights, allowing for iterative refinement and improvement of our approach. Notably, we increased the success rate of soluble designs from 11% in the first round and 66% in the second round to 75% in the third round, illustrating the effectiveness of applying accumulated knowledge for improvement (SI Section 2).

In the first two rounds, we faced solubility issues, with three designs (AncDhaA1, AncDhaA10, and AncDhaA11) showing sufficient expression for detailed characterization (secondary structure, thermostability, temperature profiles, and substrate specificity) including the in-house microfluidic device MicroPEX[73] . Surprisingly, the predicted CD spectra from AlphaFold structures differed significantly from experimental data (Figure S12). While AlphaFold predicted native-like structures, experiments revealed misfolding, suggesting a bias toward native folds in AlphaFold predictions for synthetic sequences generated by protein language models.[74] Thermostability analysis showed no significant changes, and AncDhaA10 exhibited above-average activity among the haloalkane dehalogenases.

In the third round, we addressed solubility by refining the input MSA to HLD subfamilies I−II, applying stricter preprocessing to suppress indels, and manually curating sequences by incorporating AlphaFold[16] and MutCompute[57] stability assessments. These steps improved solubility in most designs. Interestingly, the noncurated AncDhaA13 showed good solubility and activity despite risky mutations (P34V and L238F), highlighting the limits of current tools in predicting epistasis effects. On the other hand, manual curation rescued poorly soluble AncDhaA17, which informed the design of AncDhaA16 and predicted the disruptive impact of a proline insertion on AncDhaA15's activity. We also observed that even despite a large number of mutations (51.4−98.5% sequence identity to template, Table S4), all soluble designs showed stability at least comparable to that of the WT (the smallest T_m^{app} was 45.4 °C), which further emphasizes the utility of VAEs in suggesting new protein sequences. Furthermore, in terms of overall activity, 5 out of 7 variants showed comparable or higher activity than previously characterized AncHLD variants[48] (Figure 4D).

A different initialization of a model training and stricter threshold of column removal with query amino acid positions in the third round led to a second VAE model, which generated sequences with an implicit His-tag and high similarity to proteins with known experimental structures. Investigating the sequence reconstructed from the origin of the latent space (AncDhaA20) unveiled a notable shift in similarity toward another wild-type dehalogenase, DbjA,[75] altered substrate specificity, increased thermostability (60 °C) and improved activity (3.5-fold), being a top-performing haloalkane dehalogenase (Figure 4D, Table S4). To better understand the sensitivity of the *straight-line evolutionary strategy* to different initializations, we examined the embeddings over an ensemble of four randomly initialized VAEs (SI Section 5, Figure S7). Although we observed a general trend that the straight-line evolutionary trajectories converged toward the origin of the latent spaces of different VAEs, it was also evident that the trajectories exhibited quite wide scatter. This suggests that ensemble learning[76] might be an interesting direction for follow-up research to improve the robustness of our strategy.

Another promising direction for further improvement include developing better scoring methods for the sequences generated by protein language models, particularly those that will allow filtering out misfolded or poorly soluble designs *in silico*, and adopting the recent developments in transformer-based architectures, which have demonstrated a better capacity for learning from amino acid sequences.[77,78] Integrating transformer-based architectures with manifold learning can further enhance their ability to generate sequences of stable and soluble proteins.[42,43] To bolster the robustness of future studies, adopting a generation protocol for ancestral sequences that incorporates an ensemble of models might also be advantageous.[79] This approach addresses the observed instability of ancestral trajectories within the latent space and could establish a more reliable foundation for future investigations.

CONCLUSIONS

Our study demonstrated that the structure of the latent space and the generative potential of VAEs are capable of guiding the sequence search and designing novel soluble and functional proteins with enhanced stability. The workflow underwent systematic improvements through three consecutive design-build-test phases, with each iteration informed by the findings from the previous one. The success rate of soluble designs increased from 11% in the first round to 66% in the second round and 75% in the third round. Through this process, complemented by manual curation of specific variants, we achieve a notable increase in stability—up to 9 °C for the top-

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

performing AncDhaA20 variant, with an average improvement of 3 °C and a significant boost in activity up to 3.5-fold. The frequency and location of indels were the most critical parameters. In general, we recommend selecting designs with a low number of indels or with high protein similarity to natural sequences, preferably to those in PDB. A current limitation of our study is that it was conducted using a single enzyme family. Validation of designs from other protein families will help understand the generalizability of the developed approach. Overall, our study demonstrates that VAEs represent a promising strategy for generating novel soluble, stable, and functional enzymes.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.4c01101.

Additional *in silico* and wet lab experimental details, materials, and methods, including the workflow overview, list of protein designs, analysis of introduced mutations, and detailed experimental characterization (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Stanislav Mazurenko − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic;* orcid.org/0000-0003-3659-4819; Email: mazurenko@mail.muni.cz

Zbynek Prokop − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic;* orcid.org/0000-0001-9358-4081; Email: zbynek@chemi.muni.cz

Authors

Pavel Kohout − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic*

Michal Vasina − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic*

Marika Majerova − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic*

Veronika Novakova − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic*

Jiri Damborsky − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic;* orcid.org/0000-0002-7848-8216

David Bednar − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic*

Martin Marek − *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno 611 37, Czech Republic; International Clinical Research Centre, St. Anne's Hospital, Brno 656 91, Czech Republic;* orcid.org/0000-0001-7220-5644

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c01101

Author Contributions

#PK and MV contributed equally. CRediT: **Pavel Kohout** conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing - original draft, writing - review & editing; **Michal Vasina** conceptualization, data curation, formal analysis, investigation, methodology, resources, validation, visualization, writing - original draft, writing - review & editing; **Marika Majerova** investigation, methodology, validation, visualization, writing - original draft; **Veronika Novakova** investigation, methodology, validation, visualization, writing - original draft; **Jiří Damborský** conceptualization, funding acquisition, methodology, resources, supervision, writing - original draft, writing - review & editing; **David Bednar** conceptualization, investigation, methodology, writing - original draft, writing - review & editing; **Martin Marek** conceptualization, methodology, project administration, resources, supervision, writing - original draft, writing - review & editing; **Zbyněk Prokop** conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing - original draft, writing - review & editing; **Stanislav Mazurenko** conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing - original draft, writing - review & editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Brno Ph.D. Talent Scholarship holders Pavel Kohout and Michal Vasina acknowledge funding from the Brno City Municipality. Computational resources were provided by the e-INFRA CZ [LM2018140] and ELIXIR-CZ [LM2023055] projects, supported by the Ministry of Education, Youth and Sports of the Czech Republic. Core Facility Biomolecular Interactions and Crystallography of CEITEC Masaryk University, Brno is gratefully acknowledged for the obtaining of the differential scanning fluorimetry data presented in this paper. This work was also supported by Czech Ministry of Education, Youth and Sports [ESFRI RECETOX RI LM2023069]; Technology Agency of the Czech Republic [TN02000109]; the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 857560 (CETOCOEN Excellence) and 101136607 (CLARA); by the project National Institute for Cancer Research (No. LX22NPO5102) − Funded by the European

Union – Next Generation EU, and by COST (European Cooperation in Science and Technology) [COST Action COZYME CA21162]. We acknowledge CF Biomolecular Interactions and Crystallography of CIISB, Instruct-CZ Centre, supported by MEYS CR (LM2023042) and European Regional Development Fund-Project "Innovation of Czech Infrastructure for Integrative Structural Biology" (No. CZ.02.01.01/00/23_015/0008175). Martin Marek thanks the Czech Science Foundation (grant no. GA22-09853S). This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

(1) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem., Int. Ed.* **2021**, *60* (1), 88–119.

(2) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; Ryan, K. S.; Turner, N. J.; Flitsch, S. L. Biocatalysis. *Nat. Rev. Methods Primer* **2021**, *1* (1), 1–21.

(3) Silvestre, B. S.; Ţîrcă, D. M. Innovations for Sustainable Development: Moving toward a Sustainable Future. *J. Clean. Prod.* **2019**, *208*, 325–332.

(4) Tiso, T.; Winter, B.; Wei, R.; Hee, J.; de Witt, J.; Wierckx, N.; Quicker, P.; Bornscheuer, U. T.; Bardow, A.; Nogales, J.; Blank, L. M. The Metabolic Potential of Plastics as Biotechnological Carbon Sources - Review and Targets for the Future. *Metab. Eng.* **2022**, *71*, 77–98.

(5) Planas-Iglesias, J.; Marques, S. M.; Pinto, G. P.; Musil, M.; Stourac, J.; Damborsky, J.; Bednar, D. Computational Design of Enzymes for Biotechnological Applications. *Biotechnol. Adv.* **2021**, *47*, 107696.

(6) Marques, S. M.; Planas-Iglesias, J.; Damborsky, J. Web-Based Tools for Computational Enzyme Design. *Curr. Opin. Struct. Biol.* **2021**, *69*, 19–34.

(7) Kamerlin, S. C. L.; Warshel, A. The Empirical Valence Bond Model: Theory and Applications. *WIREs Comput. Mol. Sci.* **2011**, *1* (1), 30–45.

(8) Vardi-Kilshtain, A.; Roca, M.; Warshel, A. The Empirical Valence Bond as an Effective Strategy for Computer-Aided Enzyme Design. *Biotechnol. J.* **2009**, *4* (4), 495–500.

(9) Oanca, G.; van der Ent, F.; Åqvist, J. Efficient Empirical Valence Bond Simulations with GROMACS. *J. Chem. Theory Comput.* **2023**, *19* (17), 6037–6045.

(10) Kubař, T.; Elstner, M.; Cui, Q. Hybrid Quantum Mechanical/Molecular Mechanical Methods For Studying Energy Transduction in Biomolecular Machines. *Annu. Rev. Biophys.* **2023**, *52* (1), 525–551.

(11) van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52* (16), 2708–2728.

(12) Baker, D. What Has de Novo Protein Design Taught Us about Protein Folding and Biophysics? *Protein Sci.* **2019**, *28* (4), 678–683.

(13) Taverna, D. M.; Goldstein, R. A. Why Are Proteins Marginally Stable? *Proteins Struct. Funct. Bioinforma.* **2002**, *46* (1), 105–109.

(14) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16* (8), 687–694.

(15) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (18), 8852–8858.

(16) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.

(17) Jones, B. J.; Kan, C. N. E.; Luo, C.; Kazlauskas, R. J. Chapter Six - Consensus Finder Web Tool to Predict Stabilizing Substitutions in Proteins. In *Methods in Enzymology*; Tawfik, D. S., Ed.; Enzyme Engineering and Evolution: General Methods; Academic Press, 2020; Vol. *643*, pp 129–148. DOI: 10.1016/bs.mie.2020.07.010.

(18) Xie, W. J.; Asadi, M.; Warshel, A. Enhancing Computational Enzyme Design by a Maximum Entropy Strategy. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (7), No. e2122355119.

(19) Gelfand, N.; Orel, V.; Cui, W.; Damborský, J.; Li, C.; Prokop, Z.; Xie, W. J.; Warshel, A. Biochemical and Computational Characterization of Haloalkane Dehalogenase Variants Designed by Generative AI: Accelerating the SN2 Step. *J. Am. Chem. Soc.* **2025**, *147*, 2747.

(20) 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* (W1), W356–W362.

(21) Furukawa, R.; Toma, W.; Yamazaki, K.; Akanuma, S. Ancestral Sequence Reconstruction Produces Thermally Stable Enzymes with Mesophilic Enzyme-like Catalytic Properties. *Sci. Rep.* **2020**, *10* (1), 15493.

(22) Musil, M.; Khan, R. T.; Beier, A.; Stourac, J.; Konegger, H.; Damborsky, J.; Bednar, D. FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction. *Brief. Bioinform.* **2021**, *22* (4), bbaa337.

(23) Livada, J.; Vargas, A. M.; Martinez, C. A.; Lewis, R. D. Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts. *ACS Catal.* **2023**, *13* (4), 2576–2585.

(24) Prakinee, K.; Phaisan, S.; Kongjaroon, S.; Chaiyen, P. Ancestral Sequence Reconstruction for Designing Biocatalysts and Investigating Their Functional Mechanisms. *JACS Au* **2024**, *4* (12), 4571–4591.

(25) Lehmann, M.; Pasamontes, L.; Lassen, S. F.; Wyss, M. The Consensus Concept for Thermostability Engineering of Proteins. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **2000**, *1543* (2), 408–415.

(26) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (49), No. E1293-E1301.

(27) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16* (12), 1315–1322.

(28) Hawkins-Hooker, A.; Depardieu, F.; Baur, S.; Couairon, G.; Chen, A.; Bikard, D. Generating Functional Protein Variants with Variational Autoencoders. *PLOS Comput. Biol.* **2021**, *17* (2), No. e1008736.

(29) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123–1130.

(30) Elnaggar, A.; Essam, H.; Salah-Eldin, W.; Moustafa, W.; Elkerdawy, M.; Rochereau, C.; Rost, B. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv*, January 16, 2023. DOI: 10.48550/arXiv.2301.06568.

(31) Wu, K. E.; Yang, K. K.; van den Berg, R.;et al. Protein structure generation via folding diffusion. *Nat. Commun.*, *15*, 2024. DOI: 10.1038/s41467-024-45051-2.

(32) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.;

DiMaio, F.; Baek, M.; Baker, D. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089−1100.
(33) Corso, G.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. In *Conference on Learning Representations (ICLR 2023)*, 2023.
(34) Alamdari, S.; Thakkar, N.; van den Berg, R.; Lu, A.; Fusi, N.; Amini, A.; Yang, K. Protein Generation with Evolutionary Diffusion: Sequence Is All You Need; *bioRxiv*, 2023. DOI: 10.1101/2023.09.11.556673.
(35) Lisanza, S. L.; Gershon, J. M.; Tipps, S. W. K.; Sims, J. N.; Arnoldt, L.; Hendel, S. J.; Simma, M. K.; Liu, G.; Yase, M.; Wu, H.; Tharp, C. D.; Li, X.; Kang, A.; Brackenbrough, E.; Bera, A. K.; Gerben, S.; Wittmann, B. J.; McShan, A. C.; Baker, D. Multistate and Functional Protein Design Using RoseTTAFold Sequence Space Diffusion. *Nat. Biotechnol.* **2024**, 1−11.
(36) Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Rokaitis, I.; Zrimec, J.; Poviloniene, S.; Laurynenas, A.; Viknander, S.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding Functional Protein Sequence Spaces Using Generative Adversarial Networks. *Nat. Mach. Intell.* **2021**, *3* (4), 324−333.
(37) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv*, December 10, 2022. DOI: 10.48550/arXiv.1312.6114.
(38) Eguchi, R. R.; Choe, C. A.; Huang, P.-S. Ig-VAE: Generative Modeling of Protein Structure by Direct 3D Coordinate Generation. *PLOS Comput. Biol.* **2022**, *18* (6), No. e1010271.
(39) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(40) Lian, X.; Praljak, N.; Subramanian, S. K.; Wasinger, S.; Ranganathan, R.; Ferguson, A. L. Deep Learning-Enabled Design of Synthetic Orthologs of a Signaling Protein. *Cell Syst.* **2024**, *15*, 725.
(41) Ding, X.; Zou, Z.; Brooks, C. L., III Deciphering Protein Evolution and Fitness Landscapes with Latent Space Models. *Nat. Commun.* **2019**, *10* (1), 5644.
(42) Ziegler, C.; Martin, J.; Sinner, C.; Morcos, F. Latent Generative Landscapes as Maps of Functional Diversity in Protein Sequence Space. *Nat. Commun.* **2023**, *14* (1), 2222.
(43) Detlefsen, N. S.; Hauberg, S.; Boomsma, W. Learning Meaningful Representations of Protein Sequences. *Nat. Commun.* **2022**, *13* (1), 1914.
(44) Barghout, R. A.; Xu, Z.; Betala, S.; Mahadevan, R. Advances in Generative Modeling Methods and Datasets to Design Novel Enzymes for Renewable Chemicals and Fuels. *Curr. Opin. Biotechnol.* **2023**, *84*, 103007.
(45) Janssen, D. B. Evolving Haloalkane Dehalogenases. *Curr. Opin. Chem. Biol.* **2004**, *8* (2), 150−159.
(46) Koudelakova, T.; Bidmanova, S.; Dvorak, P.; Pavelka, A.; Chaloupkova, R.; Prokop, Z.; Damborsky, J. Haloalkane Dehalogenases: Biotechnological Applications. *Biotechnol. J.* **2013**, *8* (1), 32−45.
(47) Hon, J.; Borko, S.; Stourac, J.; Prokop, Z.; Zendulka, J.; Bednar, D.; Martinek, T.; Damborsky, J. EnzymeMiner: Automated Mining of Soluble Enzymes with Diverse Structures, Catalytic Properties and Stabilities. *Nucleic Acids Res.* **2020**, *48* (W1), W104−W109.
(48) Babkova, P.; Sebestova, E.; Brezovsky, J.; Chaloupkova, R.; Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem.* **2017**, *18* (14), 1448−1456.
(49) Kunka, A.; Marques, S. M.; Havlasek, M.; Vasina, M.; Velatova, N.; Cengelova, L.; Kovar, D.; Damborsky, J.; Marek, M.; Bednar, D.; Prokop, Z. Advancing Enzyme's Stability and Catalytic Efficiency through Synergy of Force-Field Calculations, Evolutionary Analysis, and Machine Learning. *ACS Catal.* **2023**, *13* (19), 12506−12518.
(50) 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 As a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catal.* **2018**, *8* (10), 9420−9428.
(51) Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation Using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2015; Vol. *28*.
(52) Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: Prediction of Soluble Protein Expression in Escherichia Coli. *Bioinformatics* **2021**, *37* (1), 23−28.
(53) Yao, Y.; Wang, X.; Ma, Y.; Fang, H.; Wei, J.; Chen, L.; Anaissi, A.; Braytee, A. Conditional Variational Autoencoder with Balanced Pre-Training for Generative Adversarial Networks. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*; 2022; pp 1−10. DOI: 10.1109/DSAA54385.2022.10032367.
(54) McGee, F.; Hauri, S.; Novinger, Q.; Vucetic, S.; Levy, R. M.; Carnevale, V.; Haldane, A. The Generative Capacity of Probabilistic Protein Sequence Models. *Nat. Commun.* **2021**, *12* (1), 6302.
(55) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat. Methods* **2022**, *19* (6), 679−682.
(56) *The PyMOL Molecular Graphics System*, Version 1.2r3pre; Schrödinger, LLC.
(57) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollihar, J.; Ellington, A. D.; Thyer, R. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* **2020**, *9* (11), 2927−2935.
(58) Aslan-Üzel, A. S.; Beier, A.; Kovář, D.; Cziegler, C.; Padhi, S. K.; Schuiten, E. D.; Dörr, M.; Böttcher, D.; Hollmann, F.; Rudroff, F.; Mihovilovic, M. D.; Buryška, T.; Damborský, J.; Prokop, Z.; Badenhorst, C. P. S.; Bornscheuer, U. T. An Ultrasensitive Fluorescence Assay for the Detection of Halides and Enzymatic Dehalogenation. *ChemCatChem.* **2020**, *12* (7), 2032−2039.
(59) Drew, E. D.; Janes, R. W. PDBMD2CD: Providing Predicted Protein Circular Dichroism Spectra from Multiple Molecular Dynamics-Generated Protein Structures. *Nucleic Acids Res.* **2020**, *48* (W1), W17−W24.
(60) Micsonai, A.; Moussong, É.; Wien, F.; Boros, E.; Vadászi, H.; Murvai, N.; Lee, Y.-H.; Molnár, T.; Réfrégiers, M.; Goto, Y.; Tantos, Á.; Kardos, J. BeStSel: Webserver for Secondary Structure and Fold Prediction for Protein CD Spectroscopy. *Nucleic Acids Res.* **2022**, *50* (W1), W90−W98.
(61) Vasina, M.; Vanacek, P.; Hon, J.; Kovar, D.; Faldynova, H.; Kunka, A.; Buryska, T.; Badenhorst, C. P. S.; Mazurenko, S.; Bednar, D.; Stavrakis, S.; Bornscheuer, U. T.; deMello, A.; Damborsky, J.; Prokop, Z. Advanced Database Mining of Efficient Haloalkane Dehalogenases by Sequence and Structure Bioinformatics and Microfluidics. *Chem. Catal.* **2022**, *2* (10), 2704−2725.
(62) Buryska, T.; Vasina, M.; Gielen, F.; Vanacek, P.; van Vliet, L.; Jezek, J.; Pilat, Z.; Zemanek, P.; Damborsky, J.; Hollfelder, F.; Prokop, Z. Controlled Oil/Water Partitioning of Hydrophobic Substrates Extending the Bioanalytical Applications of Droplet-Based Microfluidics. *Anal. Chem.* **2019**, *91* (15), 10008−10015.
(63) Vasina, M.; Vanacek, P.; Damborsky, J.; Prokop, Z. Chapter Three - Exploration of Enzyme Diversity: High-Throughput Techniques for Protein Production and Microscale Biochemical Characterization. In *Methods in Enzymology*; Tawfik, D. S., Ed.; Enzyme Engineering and Evolution: General Methods; Academic Press, 2020; Vol. *643*, pp 51−85. DOI: 10.1016/bs.mie.2020.05.004.
(64) Wong, K. M.; Suchard, M. A.; Huelsenbeck, J. P. Alignment Uncertainty and Genomic Analysis. *Science* **2008**, *319* (5862), 473−476.
(65) Jongkind, E. P. J.; Domenech, J.; Govers, A.; van den Broek, M.; Daran, J.-M.; Grogan, G.; Paul, C. E. Discovery and Synthetic Applications of a NAD(P)H-Dependent Reductive Aminase from Rhodococcus Erythropolis. *ACS Catal.* **2025**, *15*, 211−219.
(66) Love, A. C.; Purdy, T. N.; Hubert, F. M.; Kirwan, E. J.; Holland, D. C.; Moore, B. S. Discovery of Latent Cannabichromene

Cyclase Activity in Marine Bacterial Flavoenzymes. *ACS Synth. Biol.* **2024**, *13* (4), 1343−1354.
(67) Pardo, I.; Bednar, D.; Calero, P.; Volke, D. C.; Damborský, J.; Nikel, P. I. A Nonconventional Archaeal Fluorinase Identified by In Silico Mining for Enhanced Fluorine Biocatalysis. *ACS Catal.* **2022**, *12* (11), 6570−6577.
(68) Johnson, S. R.; Monaco, S.; Massie, K.; Syed, Z. Generating Novel Protein Sequences Using Gibbs Sampling of Masked Language Models. *bioRxiv*, January 27, 2021. DOI: 10.1101/2021.01.26.428322.
(69) Costello, Z.; Martin, H. G. How to Hallucinate Functional Proteins. *arXiv*, March 1, 2019. DOI: 10.48550/arXiv.1903.00458.
(70) Spence, M. A.; Kaczmarski, J. A.; Saunders, J. W.; Jackson, C. J. Ancestral Sequence Reconstruction for Protein Engineers. *Curr. Opin. Struct. Biol.* **2021**, *69*, 131−141.
(71) Koudelakova, T.; Chovancova, E.; Brezovsky, J.; Monincova, M.; Fortova, A.; Jarkovsky, J.; Damborsky, J. Substrate Specificity of Haloalkane Dehalogenases. *Biochem. J.* **2011**, *435* (2), 345−354.
(72) Johnson, S. R.; Fu, X.; Viknander, S.; Goldin, C.; Monaco, S.; Zelezniak, A.; Yang, K. K. Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nat. Biotechnol.*, 2024.
(73) Vasina, M.; Kovar, D.; Damborsky, J.; Ding, Y.; Yang, T.; deMello, A.; Mazurenko, S.; Stavrakis, S.; Prokop, Z. In-Depth Analysis of Biocatalysts by Microfluidics: An Emerging Source of Data for Machine Learning. *Biotechnol. Adv.* **2023**, *66*, 108171.
(74) Amani, K.; Fish, M.; Smith, M. D.; Castroverde, C. D. M. NeuroFold: A Multimodal Approach to Generating Novel Protein Variants in Silico. *bioRxiv*, March 14, 2024. DOI: 10.1101/2024.03.12.584504.
(75) Sato, Y.; Natsume, R.; Tsuda, M.; Damborsky, J.; Nagata, Y.; Senda, T. Crystallization and Preliminary Crystallographic Analysis of a Haloalkane Dehalogenase, DbjA, from Bradyrhizobium Japonicum USDA110. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **2007**, *63* (4), 294−296.
(76) Cao, Y.; Geddes, T. A.; Yang, J. Y. H.; Yang, P. Ensemble Deep Learning in Bioinformatics. *Nat. Mach. Intell.* **2020**, *2* (9), 500−508.
(77) Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR, 2021; pp 8844−8856.
(78) Castro, E.; Godavarthi, A.; Rubinfien, J.; Givechian, K.; Bhaskar, D.; Krishnaswamy, S. Transformer-Based Protein Generation with Regularized Latent Space Optimization. *Nat. Mach. Intell.* **2022**, *4* (10), 840−851.
(79) Ganaie, M. A.; Hu, M.; Malik, A. K.; Tanveer, M.; Suganthan, P. N. Ensemble Deep Learning: A Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151.

https://doi.org/10.1021/jacsau.4c01101
JACS Au 2025, 5, 838−850

Journal of Cheminformatics

METHODOLOGY

Open Access

Anticipating protein evolution with successor sequence predictor

Rayyan Tariq Khan[1,2], Pavel Kohout[1,2], Milos Musil[1,2,3], Monika Rosinska[1,3], Jiri Damborsky[1,2], Stanislav Mazurenko[1,2*] and David Bednar[1,2*]

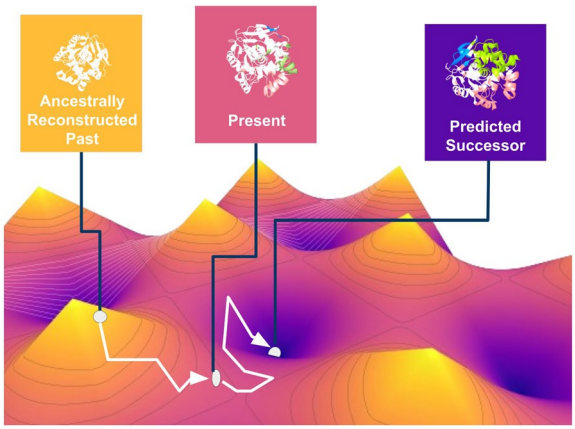
Abstract The quest to predict and understand protein evolution has been hindered by limitations on both the theoretical and the experimental fronts. Most existing theoretical models of evolution are descriptive, rather than predictive, leaving the final modifications in the hands of researchers. Existing experimental techniques to help probe the evolutionary sequence space of proteins, such as directed evolution, are resource-intensive and require specialised skills. We present the successor sequence predictor (SSP) as an innovative solution. Successor sequence predictor is an in silico protein design method that mimics laboratory-based protein evolution by reconstructing a protein's evolutionary history and suggesting future amino acid substitutions based on trends observed in that history through carefully selected physicochemical descriptors. This approach enhances specialised proteins by predicting mutations that improve desired properties, such as thermostability, activity, and solubility. Successor Sequence Predictor can thus be used as a general protein engineering tool to develop practically useful proteins. The code of the Successor Sequence Predictor is provided at https://github.com/loschmidt/successor-sequence-predictor, and the design of mutations will be also possible via an easy-to-use web server https://loschmidt.chemi.muni.cz/fireprotasr/.

Scientific Contribution The Successor Sequence Predictor advances protein evolution prediction at the amino acid level by integrating ancestral sequence reconstruction with a novel in silico approach that models evolutionary trends through selected physicochemical descriptors. Unlike prior work, SSP can forecast future amino acid substitutions that enhance protein properties such as thermostability, activity, and solubility. This method reduces reliance on resource-intensive directed evolution techniques while providing a generalizable, predictive tool for protein engineering.

Keywords Protein design, Activity, Adaptation, Evolution, Thermostability, Solubility, Evolutionary trajectory

*Correspondence:
Stanislav Mazurenko
mazurenko@mail.muni.cz
David Bednar
davidbednar1208@gmail.com
Full list of author information is available at the end of the article

Graphical abstract



Introduction

Evolution is a general term that describes the changes in inherited traits of biological entities through successive generations, generally in response to environmental changes [1]. While it can be modelled or described at many levels of biological organisation and varying levels of accuracy, for this study, we will focus on protein evolution.

Protein evolution can be reduced to two key steps: amino acid mutation and the fixation of the mutated protein in a population [2, 3]. An individual mutation may result from errors in DNA replication during cell division, exposure to mutagens, or a viral infection. The probability of fixation of this new mutation in the population depends on the fitness effect of the mutation itself. The new variant can be neutral, deleterious, or beneficial. While this two-step model is useful, it is only descriptive and not predictive [4]. For this reason, it cannot be used to predict upcoming mutations in the future and their fixation probability [5]. Thus, generally the field of evolutionary predictions has been limited to forecasting adaptive processes, as opposed to amino acid level mutations. Efforts to improve these kinds of predictions are typically focused on the aspect of selection. This neglects the fact that adaptive processes are reliant on new mutations, which in turn do have predictable biases [6]. Yet most evolutionary predictions are focused on evolution of infectious diseases, cancer and or other somatic evolutions at the phenotypic level [7]. An in silico methodology that can predict evolution at the amino acid level can ease our reliance on cost prohibitive methodologies such as those in the realm of directed evolution [8].

Directed evolution refers to experimental techniques used to engineer a protein and possibly understand the effect of various mutations on a protein and their fixation probabilities. These techniques allow a user to probe a protein's evolutionary space. They are used to improve protein characteristics and, sometimes, even to confer new characteristics onto a protein [9] by selecting or screening many variants. The markers for improvement in protein characteristics due to induced mutations can be taken as a proxy for fixation probabilities of the induced mutation in a natural environment if it occurs without human intervention. While this model has not been framed in such a way previously, it closely models the concepts of classic Darwinian/positive selection [10].

However, directed evolution experimental techniques require specialised skills and are both time and resource-intensive. Thus, any in silico technique for predicting and mimicking laboratory-based protein evolution would be of great use for the design of proteins with novel properties. As of this writing, we have only come across one technique, Proseeker, which uses physicochemical characteristics and structure to pick sequences that have higher probabilities of evolving a desired function [11]. However, the technique was designed specifically for binding proteins. It uses smaller peptide sequences (13 amino acids), and it does not filter AAindices, i.e., physiochemical descriptors [12], rather it uses all available AAindices. This leaves room for refinement by selection of more useful indices.

On the other hand, ancestral sequence reconstruction (ASR) complements these approaches by leveraging phylogenetic trees and sequence alignments to trace evolutionary changes and infer ancestral protein sequences [13–15]. By reconstructing evolutionary histories, ASR reveals positions in protein sequences where selective pressures have driven adaptations.

Building on this foundation, Combinatorial Libraries of Ancestors for Directed Evolution (CLADE) was developed to target specific positions identified through ASR [16, 17]. CLADE leverages the uncertainty inherent in ancestral reconstructions by creating combinatorial libraries, focusing on positions with the highest uncertainty for mutagenesis. This strategy enables the exploration of sequence space at evolutionarily significant sites, yielding superior results compared to consensus mutagenesis, which targets conserved residues from sequence alignments [18]. However ASR only lets us explore the evolutionary past of a sequence. Combining evolutionary insights with physicochemical properties through AAindices holds great potential for predicting evolutionary successors that align with physical evolutionary pressures.

To this end, we propose a novel method called Successor Sequence Predictor (SSP), which can mimic laboratory-based protein evolution. It reconstructs the evolutionary history of a protein sequence and then suggests amino acid substitutions based on trends observed in the evolutionary history of the protein when projected through the lens of various, carefully selected, physicochemical descriptors. Introducing the predicted mutations would enhance specific protein properties. For example, if SSP is used on a protein that in the history of its evolution was experiencing a selection pressure towards becoming more thermostable, the predicted substitutions will most likely make the mutant protein even more thermostable, and likewise for other physicochemical properties of the protein. We describe the method in detail and then conduct its critical validation against five different experimental data sets targeting properties such as thermostability, activity, and solubility. A dataset of amino acid sites that were determined to be positively selected by various evolutionary sequence analysis methodologies was also incorporated in the validation [19].

Materials and methods

Selection of AAindices

Nine AAindices were manually selected after consideration, to reflect a variety of possibly relevant physiochemical descriptors (Table 1). While the AAindex stores many more indices, they were considered inappropriate due to factors such as redundancy or context-specific physiochemical descriptions. Correlation analysis ensured that the nine selected indices had significant differences (Fig. 1), and while molecular weight and residue volume indices were similar, they were retained due to the slight nuances of how they evaluated different amino acids. Thus no indices were discarded.

Table 1 The AAindices used to analyse amino acid evolution

Index	Property	Reference
FASG760101	Molecular weight	[20]
FASG760102	Melting point	[20]
GOLD730102	Residue volume	[21]
WOLR790101	Hydrophobicity index	[22]
BHAR880101	Average flexibility indices	[23]
BULH740101	Transfer free energy to the surface	[24]
FAUJ880108	Localised electrical effect	[25]
ZIMJ680103	Polarity	[26]
ZIMJ680104	Isoelectric point	[26]

The correlations among the individual indices are presented in Fig. 1

Successor sequence predictor workflow

Successor Sequence Predictor follows the workflow outlined in Fig. 2. Firstly, the FASTA sequence of a target protein is used to identify a dataset of homologous sequences using BLAST [27]. Only sequences with 30–90% sequence identity to the target are retained. A length filter is then applied to keep sequences within 80–120% of the target protein's length. The remaining sequences are clustered using USEARCH at 90% sequence identity, and one sequence from each cluster is randomly selected (Fig. 2A).

The dataset obtained from these steps is divided to construct multiple phylogenetic trees, each containing 150 sequences (Fig. 2B) with the final number of phylogenetic trees dependent on the dataset size. Before processing, the dataset is amended to ensure sequence headers do not contain problematic special characters (e.g., parentheses, colons, semicolons, or numbers at the start of headers) that could disrupt the function of the utilised tools. Next, sequences are clustered based on similarity using SigClust [28] with the parameter $c = 150$, producing up to 150 clusters. Sequence files are then generated with the following rules:

1. Each target sequence must appear in at least one sequence file.
2. One sequence is randomly selected from each cluster for each file.
3. To maximize diversity, the algorithm avoids reusing sequences from clusters unless all options are exhausted.
4. Every sequence file must contain the target sequence.

Once the sequence files are prepared, ClustalOmega [29] creates a multiple sequence alignment (MSA) for each file.

The MSAs are then processed using the standard FireProt^{ASR} workflow [14]. RAxML [30] is employed to

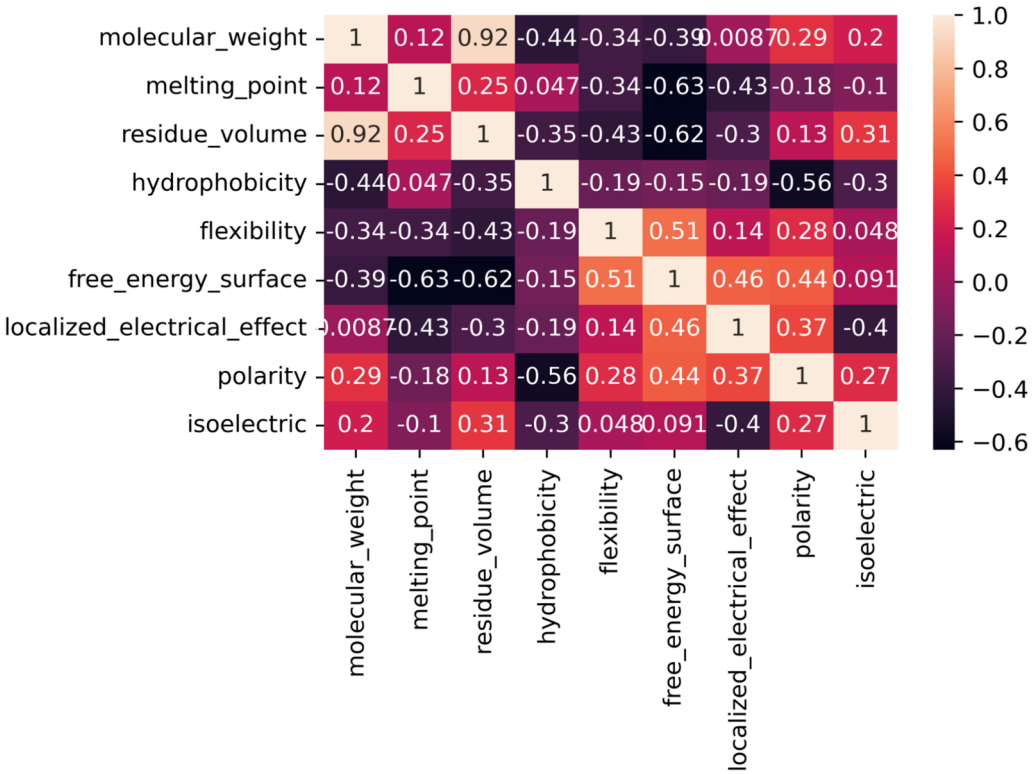


Fig. 1 Pearson correlation matrix of selected AAindices. The correlation coefficients are colour-coded from dark purple at − 0.7 to off-white at 1.0. The indices are summarised in Table 1

construct a phylogenetic tree for each MSA using the maximum-likelihood algorithm. RAxML runs in its SSE3 version, using 50 bootstraps and the best-fit evolution matrix suggested by IQ-TREE. Once the calculation is completed, the minimum ancestral deviation algorithm is used to root the generated trees (Fig. 2C) [31]. This approach generally leads to highly similar trees across multiple runs, yet a level of stochasticity can still be expected as most of the employed tools rely on heuristic algorithms.

Rooted phylogenetic trees and corresponding MSAs are loaded into LAZARUS [32] to calculate posterior probabilities for each sequence file. LAZARUS uses the "codeml" module, the appropriate evolutionary matrix, and a fixed branch lengths, with gap reconstruction disabled (this step is handled by the gap correction algorithm implemented in FireProt^{ASR}). Based on the posterior probabilities and predicted ancestral gaps, ancestral sequences are reconstructed for each node in the phylogenetic tree. The main path from the root to the target sequence is identified for each phylogenetic tree (Fig. 2D).

The sequences from the target and all ancestral nodes to the root are extracted into a separate file and aligned using ClustalOmega. Finally, a Python script employing the "numpy" and "sklearn.linear_model" libraries [33] predicts the successor sequence as the next step along a regression curve, following these steps (Fig. 2E):

1. For each column in the MSA (referred to as a "Trajectory"), a matrix of amino acid physicochemical features is generated, with each column representing one of nine selected AAindices.
2. For each column, a vector of changes in physicochemical features is calculated, weighted by the evolutionary distance from the root node.
3. This vector is used to train a linear regression model to predict the next amino acid in the trajectory, mimicking laboratory-based protein evolution.
4. The distance between consecutive amino acids in the trajectory (based on AAindex values) is calculated as the average distance between nodes along the main path in the phylogenetic tree.
5. Separate regressions for each physicochemical feature are aggregated to assign categories and bundle predictions (Table 2).
6. This process is repeated for every column in the MSA and each sequence file.

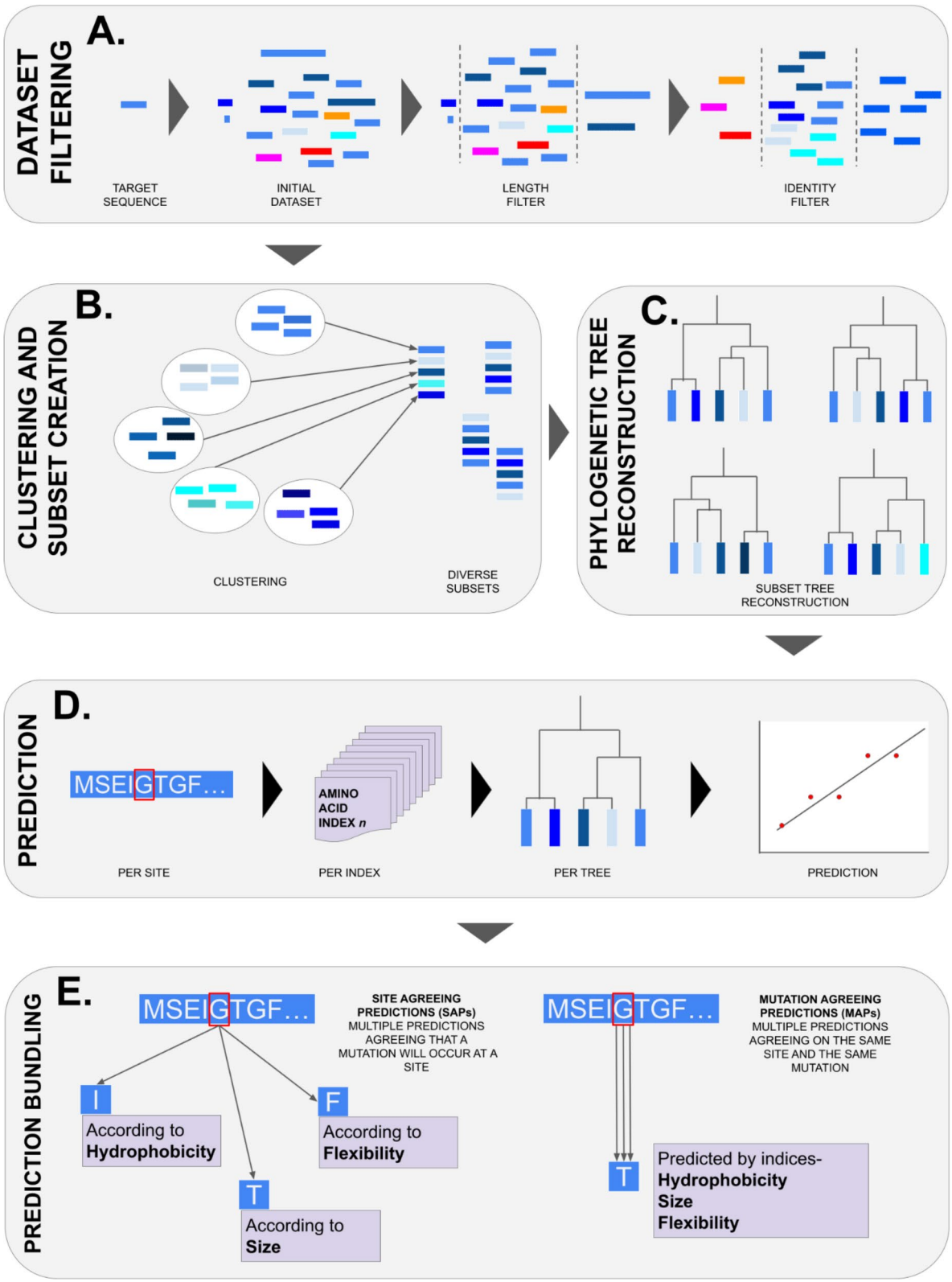


Fig. 2 A generalised overview of the Successor Sequence Predictor (SSP). **A** Initial curation and filtering of the target protein's dataset. **B** Further division of data using a clustering methodology. **C** Phylogenetic tree reconstruction and ancestral sequence reconstruction for the nodes on the trees. **D** Trend construction and amino acid prediction. **E** Prediction bundling

Table 2 An example of the generalised prediction bundling scheme for three different levels of prediction: prediction, site agreeing prediction (SAP), and mutation agreeing prediction (MAP)

Type	Amino acid	Position	Prediction	Index
Prediction	A	12	L	Size
Site agreeing prediction (SAP)	A	12	L	Size
			R	Hydrophobicity
Mutation agreeing prediction (MAP)	A	12	L	Size
				Polarity
				Flexibility

Several precautions are taken to minimize over-interpretation of the linear regression approach. The regression plot is normalized by the number of transitions (amino acid substitutions) in the trajectory. If an amino acid remains unchanged across successive ancestors at a given position, it is treated as part of a group and not penalized in scoring. Transitions between groups are counted only when they occur.

Key metrics include the penultimate transition, which flags any changes inconsistent with the overall trend as a "*break trend*." Trajectory *sequentiality* is scored out of 100, with a perfect score achieved when each transition in a positive trend increases the feature's value compared to the previous one. *Fluctuations* in a trajectory are measured by dividing the number of different amino acids by the number of amino acid groups, reflecting the trajectory's stability or variability. Sites with fewer than three transitions are excluded from predictions.

These scores are used to rank successor amino acid predictions for each site, index, and phylogenetic tree. The highest-ranking predictions are those with high *sequentiality*, high *fluctuation*, and no *break trend* at the penultimate amino acid position. Each amino acid prediction is averaged across sites and trees.

When multiple predictions agree at a specific site but differ in the mutation type, they are referred to as Site Agreeing Predictions (SAPs). Conversely, when predictions from different AAindices align on the same mutation, they are consolidated into a single prediction, known as a Mutation Agreeing Prediction (MAP) (Table 2).

Validation datasets

Mutational datasets investigating a physicochemical property of any specific protein were searched through the literature. The ones with large enough datasets, which also had enough overlap with predicted mutations (thus allowing us to validate them,) were selected. This includes homolog sets for levoglucosan kinase—UniProt ID B3VI55 [34], cold shock protein CspB—UniProt ID P32081 [35], ADP-ribosylarginine hydrolase—Uniprot ID P54922 [19], and aminoglycoside 3'-phosphotransferase—UniProt IS P00552 [36].

Individual datasets were compiled in different ways. The levoglucosan kinase set was found via the *in-house* SoluProtMut^{DB} database [37] by searching for a protein with a large number of experimentally validated single-point mutations and their effects on the solubility of the protein. Similarly, the cold shock protein CspB dataset was found in the *in-house* FireProt^{DB} database [38], by searching for a protein with a large number of experimentally validated single-point mutations and their effects on the thermostability of the protein. In cases where multiple values were available for a single mutation, the mean was taken. ADP-ribosylarginine hydrolase dataset was picked as it was one of the example cases for Slodkowicz and Goldman's online tool [19] for Structure Integrated with Positive Selection. ADP-ribosylarginine hydrolase was picked after a literature review, due to the sheer number of single-point mutations tested (fully site saturated) on the target protein by Melnikov et al. [36]. This naturally presented a perfect test case for SSP. Individual and detailed dataset handling steps are noted in SI 2.

Results

Dataset statistics

We tested the performance of SSP on the homolog sets for levoglucosan kinase (solubility), cold shock protein CspB (thermostability), ADP-ribosylarginine hydrolase (selectivity), and aminoglycoside 3′-phosphotransferase (activity). It is important to note that with the exception of Aminoglycoside 3′-phosphotransferase dataset, none of the other datasets used in the study have the values for the relevant effect for every possible point mutation that SSP predicts. Thus it is not possible to validate all predictions made by SSP. The results section only shows validation based on all mutational data points that SSP predicted and for which experimental labels were available. Figure 3 summarises the total single-point

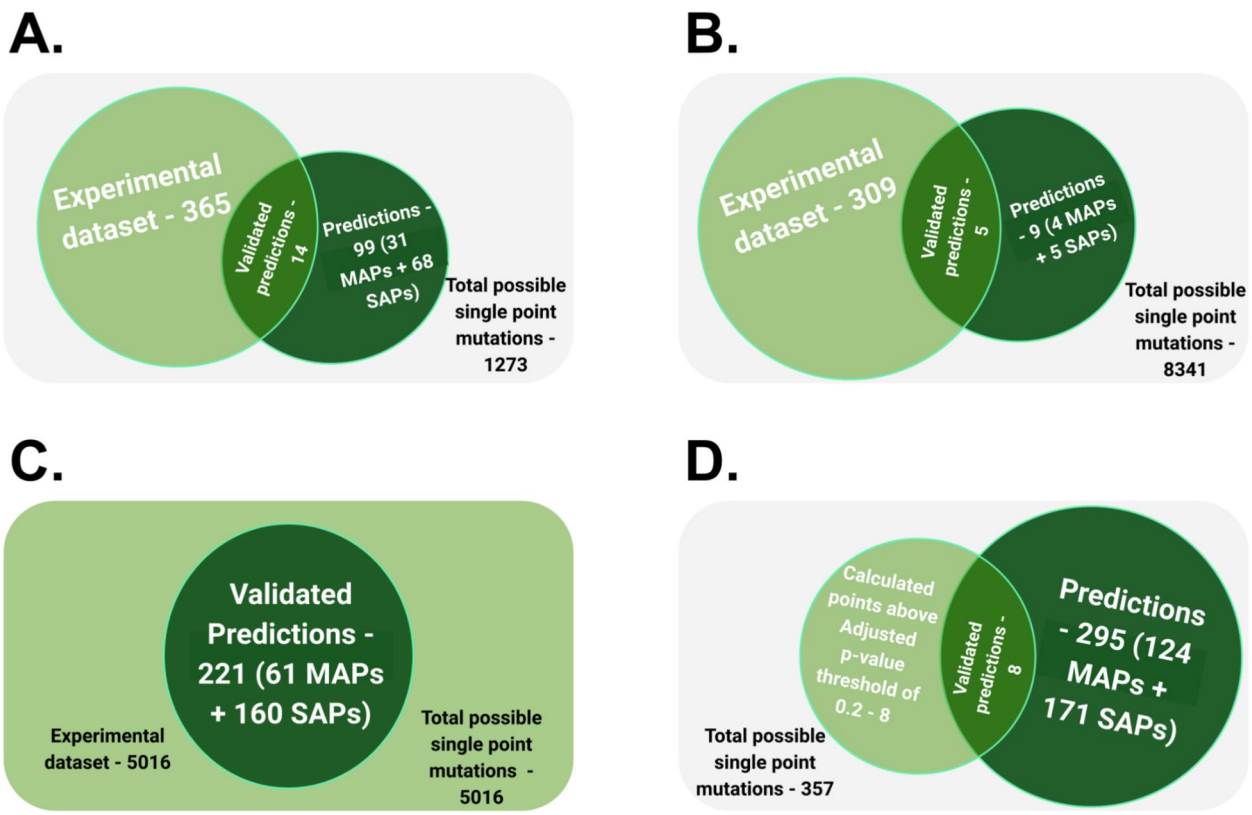


Fig. 3 The visualisation of overlaps between the available experimental data and the predicted data. **A** Overlap metrics for Cold shock protein CspB set (FireProt^{DB} dataset—[38], **B** Overlap metrics for levoglucosan kinase set [34] **C** Overlap metrics for Aminoglycoside 3′-phosphotransferase set [36], and **D** Overlap metrics for ADP-ribosylarginine hydrolase set [19]. The experimental data are represented by a light green circle, while a dark green circle represents predicted data

mutational space, the available experimental values, the number of predictions, and the overlaps between the two.

Engineering thermostability

SSP predictions for Cold shock protein CspB were compared to experimental data points with known effects of the mutation on protein thermostability from a collated dataset stored in the database FireProt^{DB} [38]. In cases where values from multiple datasets were available, the mean values were noted. E3Q was the only MAP that was supported by more than three indices. E3K was supported by 2 indices, and all others were SAPs. The results are provided in Table 3.

There were 365 total mutations in the FireProt^{DB} dataset [38], of which 18% were enhancing mutations in terms of thermostability (ΔΔG lower than − 1 kcal/mol), 55% were neutral (ΔΔG from − 1 kcal/mol to 1 kcal/mol), the remaining 27% were destabilising (ΔΔG greater than 1 kcal/mol). The thresholds for stabilising, neutral and destabilising categories were taken from the FireProt^{DB}.

From the 14 mutations predicted by SSP, six were stabilising. The other eight mutations had ΔΔG values

Table 3 Effects of mutations generated by SSP on the thermostability of cold shock protein validated against the collated FireProt^{DB} dataset [38]

Mutation by SSP	Mean ΔΔG (kcal/mol)	△Tm (°C)[a]	Prediction agreement type
L2R	+ 0.4	N/A	Site agreeing
E3K	− 2.48	+ 16.6	Mutation agreeing
E3Q	− 1.09	+ 7.3	Mutation agreeing
E3R	− 1.65	+ 16.0	Site agreeing
E3V	− 1.8	N/A	Site agreeing
D24N	+ 0.66	− 6.9	Site agreeing
A46E	+ 0.07	− 5.0	Site agreeing
A46K	− 1.41	+ 8.4	Site agreeing
A46L	− 0.8	N/A	Site agreeing
E50K	+ 0.33	− 5.6	Site agreeing
N55D	− 0.46	+ 3.9	Site agreeing
N55K	0	+ 0.8	Mutation agreeing
N55S	+ 0.2	N/A	Site agreeing
E66K	− 2.17	+ 12.9	Site agreeing

[a] *N/A* data not available

Table 4 Effects of mutations generated by SSP on the solubility of levoglucosan kinase [34]

Mutation by SSP	Effect on solubility	Prediction agreement type
I3L	Neutral	Mutation agreeing
I3F	Neutral	Mutation agreeing
D9G	Positive	Site agreeing
K38Q	Positive	Site agreeing
V200A	Negative	Site agreeing

between − 1 kcal/mol and 1 kcal/mol and can thus be classified as neutral. Seven out of ten mutations also increased the melting temperature (T_m) of the protein, and three were destabilising (Table 3).

Engineering solubility

SSP predictions for levoglucosan kinase were compared to experimental data from Klesmith et al. [34] available in the SoluProtMutDB [37]. This comparison assessed how well the SSP predictions matched the known effects of mutations on protein solubility.The mutations I3L and I3F (supported by two different indices) had a neutral effect on solubility. Both mutations predicted by SSP, D9G and K38Q, are known to have a slightly enhancing effect on solubility. Only V200A showed a slightly negative effect on solubility in *E. coli* (Table 4). This suggests that the expressed mutants produced via SSP do not compromise their solubility.

Engineering activity

Aminoglycoside 3′-phosphotransferase is a protein that confers resistance to aminoglycosides, such as kanamycin, neomycin, paromomycin, ribostamycin, butirosin, and gentamicin B. Melnokov et al. [36] conducted site saturation mutagenesis on this protein, transformed variants into cells and exposed them to six different antibiotics at up to four different concentrations. The amino acid enrichment (the number of identified variants with the particular mutation) was then noted in each case. A value of ~ 1 applies to wild types, while a higher value means more resistance and hence more significant enrichment of that mutant, and vice versa for a value below 1 (Fig. 4).

The average of all enrichment values across antibiotics and their concentrations in the complete dataset (AAC value) was 0.82. This means that a random, single-amino-acid variant is less likely to be resistant than the wild type, and, therefore, will have lower activity. The AAC value of 0.82 may be assumed as a proxy value for random

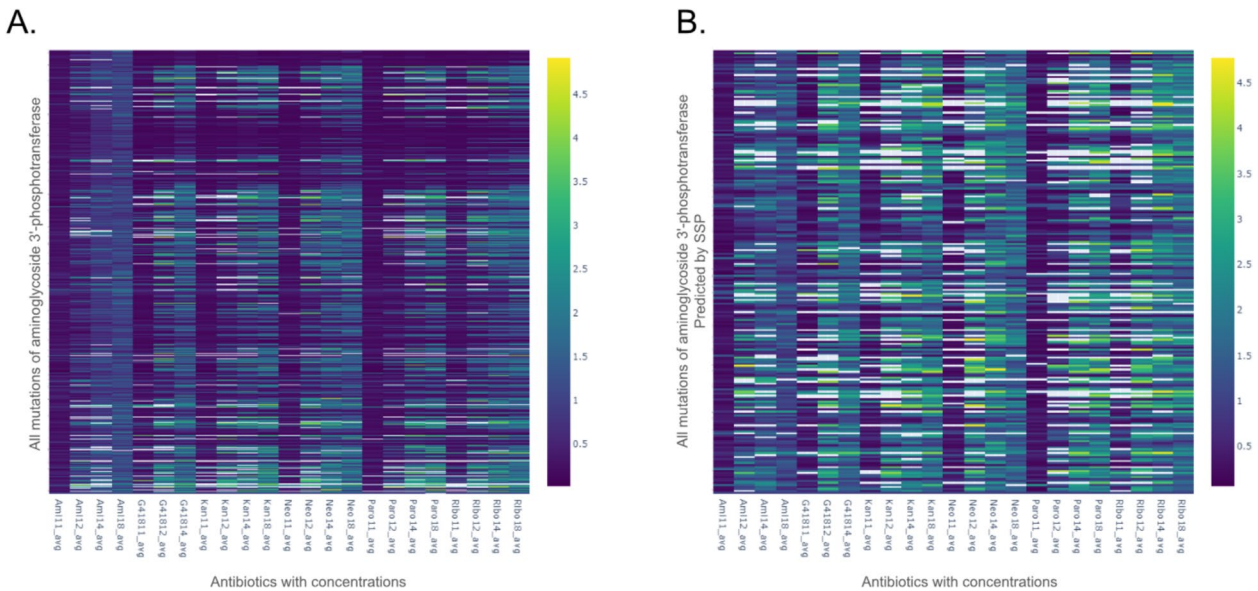


Fig. 4 Heatmap visualisations comparing the enrichment values for mutations of aminoglycoside 3′-phosphotransferase. **A** A heatmap representing the entire mutational space of aminoglycoside 3′-phosphotransferase. **B** A heatmap representing only the mutations of aminoglycoside 3′-phosphotransferase that were predicted by the SSP. The X-axis represents the antibiotics and their tested concentrations, while the Y-axis represents the relevant mutations of aminoglycoside 3′-phosphotransferase. Details of antibiotic concentrations and individual enrichment values can be found in SI 1. Each rectangle on the plot indicates the enrichment value for a mutation when exposed to the effects of the specific antibiotic concentration. The Viridis colour map is used to maintain perceptual uniformity. A value of 1 (dark blue) represents no change in enrichment from the wild type, anything below 1 (purple) represents a negative effect on enrichment, while anything above 1 (light blue to yellow) represents a positive enriching effect of the mutation. This figure contrasts the effects of random mutations on the activity of aminoglycoside 3′-phosphotransferase, against the effect of SSP suggested mutations for the same protein. The perceptual increase in 'brightness' of **B** over **A** illustrates an increase in the positive impact of mutations on the activity of aminoglycoside 3′-phosphotransferase

Table 5 Cross-matching positively selected site data of ADP-ribosylarginine hydrolase from SIPS with SSP predictions [19]

Sites selected by SSP	Adjusted p-value	Prediction agreement type
K72	0.0848	Mutation agreeing
P74	0.0473	Mutation agreeing
T77	0.1731	Mutation agreeing
Q78	0.1101	Mutation agreeing
Q109	0.0796	Mutation agreeing
H145	0.05	Non agreeing
L189	0.0002	Mutation agreeing
I355	0.0128	*Not predicted*

mutations, while 1 is the default value for the wild type. Thus, random single-point mutations are likely to reduce the protein's fitness. SSP generated 221 predictions, all with experimental validation points available from this large-scale site saturation mutagenesis study (Fig. 4). For mutants generated by SSP, the AAC value is 1.36, showing a preferable selection of enriched (more active) variants, thus an increase in fitness if a mutation is selected from SSP's output. Moreover, 61 of the 221 mutations were predicted at the MAP level, and their AAC value is 1.4. The remaining 160 predictions were made at the SAP level, and their AAC value is 1.34. As the AAC value for MAP level predictions is slightly higher than that for SAP level (1.4 *versus* 1.34), it hints at the possibility that MAPs may be slightly more reliable. This is summarised in Fig. 3C. The comparison of experimentally determined and predicted values are available in the supplementary table SI 1.

Evolutionary selection

Structure Integrated with Positive Selection (SIPS) is an online resource with positively selected sites mapped onto protein structures from an evolutionary perspective [19]. ADP-ribosylarginine hydrolase, which is one of the example cases of SIPS, has eight positively selected sites with an adjusted p-value threshold of 0.2 or higher. SSP predictions were made for ADP-ribosylarginine hydrolase to see how many of the predictions could be made for positively selected sites. Here, the emphasis was on sites and not the mutation itself, as SIPS only lists sites of evolutionary interest and not what they would mutate into. Out of the eight sites, seven were predicted by SSP, and six were MAPs, implying that SSP can selectively make predictions for sites with evolutionary significance (Table 5).

Discussion and conclusions

SSP is a protein design method that employs the prediction of the evolution of amino acids in a protein sequence. It builds a statistical, ancestral sequence reconstruction-guided evolutionary history of a protein sequence [39], which is utilised to extrapolate the possible *future* substitution at a given position. SSP makes these predictions in the context of AAindex scoring [12] applied to the reconstructed evolutionary history for each position in a protein sequence. The AAindices used for SSP have been manually selected to reflect a variety of possibly relevant physiochemical descriptors. The selected set of AAindices can be easily adjusted based on the physicochemical properties expected to be involved in shaping the evolution of a particular protein.

It should be noted that while SSP utilises ASR, they are both *fundamentally different* techniques with distinct goals. ASR aims to 'look back' into the evolutionary history of a protein sequence, while SSP is designed to extrapolate into the potential future of a protein sequence. ASR is generally used for evolutionary analysis [40] and protein engineering [39]. While ancestral proteins are more robust and with unique substrate specificities [32, 41, 42], the engineering scope of ASR is generally along the lines of improving the thermostability of a protein and its expression yield. This is because ancestral proteins, when resurrected, tend to be more robust [43]. SSP can map out potential future evolutionary trajectories of a protein, and it can also be used to engineer proteins.

Proseeker is another tool that simulates natural selection and thus mimics evolution in silico. It uses physicochemical characteristics and structural information to pick sequences that have higher probabilities of evolving a desired function [11]. However, the technique was designed specifically for binding proteins and lacks general applicability. Instead of complete protein sequences, it uses small peptide sequences (13 amino acids), and it also does not filter or select specific AAindices, rather it uses all available AAindices [12]. The selection of relevant indices and then estimating their utility for any tool in this domain is crucial as many indices are redundant, e.g., nine indices for the hydrophobicity: ARGP820101, GOLD730101, JOND750101, PRAM900101, ZIMJ680101, PONP930101, WOLR790101, ENGD860101, and FASG890101 [12]. This can lead to index weighting issues, where a certain physiochemical descriptor may have an exaggerated effect on the outcome. Furthermore, many indices are context-specific, such as hydrophobicity coefficients in specific solutions—from WILM950101 to WILM950104, and weights for alpha-helix at specific window positions—from QIAN880101 to QIAN880139 [12]. Thus a careful

selection of indices is a necessary step, SSP used manually curated non-correlated indices (Table 1 and Fig. 1). While the direct comparison between SSP and Proseeker could have been useful, it is hard to achieve as Proseeker works with shortened peptides (13 AA long) instead of the whole protein sequence. Moreover, it specifically requires binding affinity data to score every iteration of in silico evolution, thus making the technique specific to nucleic acid binding peptides. SSP is not limited in terms of the nature or the length of the target protein sequence.

SSP was validated using the datasets from different sources to test for the performance of various properties. In the case of thermostability, SSP made 14 predictions for the cold shock protein CspB, eight of which had a stabilising effect on the protein ($\Delta\Delta G < 0$), while the remaining six were neutral with $\Delta\Delta G$ values between 0 kcal/mol and 1 kcal/mol. Seven of the predicted mutations also had positive experimentally determined changes in melting temperatures $\triangle$Tm (°C), including the highest increase in melting temperature of + 16.6 °C, and only three mutation had a negative $\triangle$Tm (°C) value with the lowest being − 6.9 °C.

SSP was also used to make predictions for aminoglycoside 3'-phosphotransferase [36]. Aminoglycoside 3'-phosphotransferase is an enzyme that confers resistance to aminoglycosides with antibiotic properties. Thus an enhancement of enzyme's activity can increase the antibiotic resistance of a bacteria that codes for it. SSP made 221 predictions for Aminoglycoside 3'-phosphotransferase with an AAC value of 1.4 at the MAP level, and 1.36 at the SAP level (1 being the value for the wild type, and 0.82 being the average value for random mutagenesis), thus demonstrating predictive prowess in the context of enhancing enzymatic activity, being significantly better than random mutation, while conferring an improvement over the wild type itself.

Validation of mutations predicted from the solubility dataset showed a higher likelihood of a positive or neutral effect on the solubility of the protein, despite the sparseness of the dataset. Furthermore, evolutionary selectivity data for ADP-ribosylarginine hydrolase [19] taken from SIPS and SSP made predictions for 7 of 8 evolutionary selected sites with an adjusted p-value upper threshold of 0.2. This result suggests that SSP is selective in making predictions for sites that tend to evolve under positive selection, thus making a strong case for SSP's selectivity. However, it should be noted that the size of the dataset is quite small, and more work is required to validate this aspect of the predictor.

Analyzing and validating methods like SSP presents significant challenges. Extracting meaningful insights from diverse datasets with varying experimental standards can be complex due to limited overlap between experimentally observed mutations and the mutations predicted by SSP (Fig. 3). Finding datasets that are not only extensive but also contain experimental data for mutations that coincide with SSP predictions—enabling their validation—proved to be a substantial hurdle. This scarcity necessitated the use of all available validation sets, despite their inherent differences in physicochemical properties. Substantially more mutational data would be needed to have evenly distributed dataset for each protein property.

This study shows that the SSP approach enhances specialised proteins by predicting mutations that improve desired properties, such as thermostability, activity, and solubility. Crucially, it also shows that SSP does not make predictions for sites randomly, but picks sites that are known to evolve under positive selection. In general, SSP method will work better with the proteins under stronger selection evolutionary pressure. Further validation of the predictor with diverse protein structures is desirable to define applicability for protein engineering applications. It should also be noted that the technique has a limitation; it is dependent on the size and quality of the homolog set. The technique will not work if the protein has no or very few homologs. For our pipeline, we suggest having at least 10 trees of 150 protein sequences each, per analysis. However the exact numbers need further exploration.

As the service to the community, we are now integrating SSP as a new module into the easy-to-use web server FireProtASR (https://loschmidt.chemi.muni.cz/fireprotasr/), which will make predictions accessible to non-experts, jointly with related strategies Ancestral Sequence Reconstruction (ASR) and generation of sequences using Variational Autoencoder (VAE) [44].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00971-z.

Supplementary Material 1: Comparison of experimentally determined and predicted values for Aminoglycoside 3'-phosphotransferase, as well as the complete enrichment data.

Supplementary Material 2: A text README file on how to locally run the Successor Sequence Predictor.

Supplementary Material 3: Example case query sequence in FASTA format.

Supplementary Material 4: Example case homolog database in FASTA format.

Supplementary Material 5: Pre-run Ancestral Sequence Reconstruction output.

Supplementary Material 6.

Acknowledgements

This work was supported by Operational Programme Research, Development and Education—"Project Internal Grant Agency of Masaryk University" (No. CZ.02.2.69/0.0/0.0/19_073/0016943) and Brno University of Technology

[FIT-S-23-8209]. Authors thanks to the RECETOX Research Infrastructure (No. LM2023069) financed by the Ministry of Education, Youth and Sports. Computational resources were provided by the e-INFRA CZ and ELIXIR-CZ projects (ID: LM2018140 and LM2023055), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102)—Funded by the European Union—Next Generation EU. This project was also supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement CETOCOEN Excellence (No. 857560) and TEAMING CLARA (No. 101136607), and by the Czech Ministry of Education, Youth and Sports, and the Operational Programme Research, Development and Education (the RECETOX RI project No. LM2023069). Pavel Kohout is holder of the Brno Ph.D. Talent scholarship funded by the Brno City Municipality and the JCMM. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Author contribution

Rayyan Tariq Khan—Conceptualization, Data curation, Investigation, Methodology, Writing—original draft, Writing—review & editing Pavel Kohout—Conceptualization, Software, Investigation, Methodology, Writing—original draft, Writing—review & editing Milos Musil—Conceptualization, Software, Supervision, Writing—original draft, Writing—review & editing Monika Rosinska—Software, Methodology, Writing—review & editing Jiri Damborsky—Supervision, Funding acquisition, Writing—review & editing Stanislav Mazurenko—Conceptualization, Supervision, Writing—review & editing David Bednar—Conceptualization, Supervision, Project administration, Funding acquisition, Writing—review & editing.

Data availability

We provide scripts as a command line application written in Python 3.8, which can be found on GitHub https://github.com/loschmidt/successor-sequence-predictor. We use the LinearRegression module from the scikit-learn library to predict the evolutionary trend in the phylogenetic tree.

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic. [2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic. [3]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic.

Received: 18 September 2024 Accepted: 9 February 2025
Published online: 21 March 2025

References

1. Hall BK, Hallgrimsson B (2014) Strickberger's evolution. Jones And Bartlett.
2. Gillespie JH (1994) The causes of molecular evolution. Oxford University Press
3. Kimura M (1985) The neutral theory of molecular evolution. Cambridge University Press
4. Nosil P, Flaxman SM, Feder JL, Gompert Z (2020) Increasing our ability to predict contemporary evolution. Nat Commun 11(1) https://doi.org/10.1038/s41467-020-19437-x
5. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8(8):610–618. https://doi.org/10.1038/nrg2146
6. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7(5):337–348. https://doi.org/10.1038/nrg1838
7. Cano AV, Gitschlag BL, Rozhoňová H, Stoltzfus A, McCandlish DM, Payne JL (2023) Mutation bias and the predictability of evolution. Philos Trans R Soc B 378(1877):20220055. https://doi.org/10.1098/rstb.2022.0055
8. Yang KK, Wu Z, Arnold FH (2019) Machine-learning-guided directed evolution for protein engineering. Nat Methods 16:687–694. https://doi.org/10.1038/s41592-019-0496-6
9. Arnold FH (2019) Innovation by evolution: bringing new chemistry to life (nobel lecture). Angew Chem 58(41):14420–14426. https://doi.org/10.1002/anie.201907729
10. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413. https://doi.org/10.1093/genetics/155.3.1405
11. Raven SA, Payne B, Mitchell AF, Rackham O (2022) In silico evolution of nucleic acid-binding proteins from a nonfunctional scaffold. Nat Chem Biol 18(4):403–411. https://doi.org/10.1038/s41589-022-00967-y
12. Kawashima S (2000) AAindex: amino acid index database. Nucleic Acids Res 28(1):374–374. https://doi.org/10.1093/nar/28.1.374
13. Livada J, Vargas AM, Martinez CA, Lewis RD (2023) Ancestral sequence reconstruction enhances gene mining efforts for industrial ene reductases by expanding enzyme panels with thermostable catalysts. ACS Catal 13(4):2576–2585. https://doi.org/10.1021/acscatal.2c03859
14. Musil M, Khan RT, Beier A, Stourac J, Konegger H, Damborsky J, Bednar D (2020) FireProtASR: a web server for fully automated ancestral sequence reconstruction. Brief Bioinform. https://doi.org/10.1093/bib/bbaa337
15. Prakinee K, Phaisan S, Kongjaroon S, Chaiyen P (2024) Ancestral sequence reconstruction for designing biocatalysts and investigating their functional mechanisms. JACS Au 4(12):4571–4591. https://doi.org/10.1021/jacsau.4c00653
16. Gumulya Y, Huang W, D'Cunha SA, Richards KE, Thomson RE, Hunter DJ, Baek JM, Harris KL, Boden M, De Voss JJ, Hayes MA (2019) Engineering thermostable CYP2D enzymes for biocatalysis using combinatorial libraries of ancestors for directed evolution (CLADE). ChemCatChem 11(2):841–850. https://doi.org/10.1002/cctc.201801644
17. Gumulya Y, Baek JM, Wun SJ, Thomson RE, Harris KL, Hunter DJ, Behrendorff JB, Kulig J, Zheng S, Wu X, Wu B (2018) Engineering highly functional thermostable proteins using ancestral sequence reconstruction. Nat Catal 1(11):878–888. https://doi.org/10.1038/s41929-018-0159-5
18. Brennan CK, Livada J, Martinez CA, Lewis RD (2024) Ancestral sequence reconstruction meets machine learning: ene reductase thermostabilization yields enzymes with improved reactivity profiles. ACS Catal 14(23):17893–17900. https://doi.org/10.1021/acscatal.4c03738
19. Slodkowicz G, Goldman N (2020) Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. Proc Natl Acad Sci 117(11):5977–5986. https://doi.org/10.1073/pnas.1916786117
20. Fasman GD (1989) Practical handbook of biochemistry and molecular biology. CRC Press, New York. https://doi.org/10.1201/9781351072427/handbook-biochemistry-gerald-fasman
21. Goldsack DE, Chalifoux RC (1973) Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. J Theor Biol 39(3):645–651. https://doi.org/10.1016/0022-5193(73)90075-1
22. Wolfenden RV, Cullis PM, Southgate CCF (1979) Water, protein folding, and the genetic code. Science 206(4418):575–577. https://doi.org/10.1126/science.493962
23. Bhaskran R, Ponnuswamy PK (1988) Positional flexibilities of amino acid residues in globular proteins. Int J Pept Protein Res 32(4):241–255. https://doi.org/10.1111/j.1399-3011.1988.tb01258.x
24. Bull HB, Breese K (1974) Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. Arch Biochem Biophys 161(2):665–670. https://doi.org/10.1016/0003-9861(74)90352-x
25. Fauchere J-L, Charton M, Kier LB, Verloop A, Pliska V (2009) Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 32(4):269–278. https://doi.org/10.1111/j.1399-3011.1988.tb01261.x
26. Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21(2):170–201. https://doi.org/10.1016/0022-5193(68)90069-6
27. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(1):5–9. https://doi.org/10.1093/nar/gkn201
28. Liu Y, Hayes DN, Nobel A, Marron JS (2008) Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc 103(483):1281–1293

29. Sievers F, Higgins DG (2014) Clustal omega. Curr Protoc Bioinformat. 48(1) https://doi.org/10.1002/0471250953.bi0313s48
30. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033
31. Tria F, Landan G, Dagan T (2017) Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol 1:0193. https://doi.org/10.1038/s41559-017-0193
32. Hanson-Smith V, Kolaczkowski B, Thornton JW (2010) Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol Biol Evol 27(9):1988–1999. https://doi.org/10.1093/molbev/msq081
33. sklearn.linear_model.LinearRegression (2023) Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
34. Klesmith JR, Bacik J-P, Wrenbeck EE, Michalczyk R, Whitehead TA (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. Proc Natl Acad Sci 114(9):2265–2270. https://doi.org/10.1073/pnas.1614437114
35. Gribenko AV, Makhatadze GI (2007) Role of the charge-charge interactions in defining stability and halophilicity of the CspB proteins. J Mol Biol 366(3):842–856. https://doi.org/10.1016/j.jmb.2006.11.061
36. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS (2014) Comprehensive mutational scanning of a kinase *in vivo* reveals substrate-dependent fitness landscapes. Nucleic Acids Res 42(14):e112–e112. https://doi.org/10.1093/nar/gku511
37. Velecký J, Hamsikova M, Stourac J, Musil M, Damborsky J, Bednar D, Mazurenko S (2022) SoluProtMutDB: a manually curated database of protein solubility changes upon mutations. Comput Struct Biotechnol J 20:6339–6347. https://doi.org/10.1016/j.csbj.2022.11.009
38. Stourac J, Dubrava J, Musil M, Horackova J, Damborsky J, Mazurenko S, Bednar D (2020) FireProtDB: database of manually curated protein stability data. Nucleic Acids Res 49(D1):D319–D324. https://doi.org/10.1093/nar/gkaa981
39. Spence MA, Kaczmarski JA, Saunders JW, Jackson CJ (2021) Ancestral sequence reconstruction for protein engineers. Curr Opin Struct Biol 69:131–141. https://doi.org/10.1016/j.sbi.2021.04.001
40. Cai W, Pei J, Grishin NV (2004) Reconstruction of ancestral protein sequences and its applications. BMC Evol Biol 4:1–23. https://doi.org/10.1186/1471-2148-4-33
41. Babkova P, Sebestova E, Brezovsky J, Chaloupkova R, Damborsky J (2017) Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. ChemBioChem 18(14):1448–1456. https://doi.org/10.1002/cbic.201700197
42. Risso VA, Sanchez-Ruiz JM (2017) Resurrected ancestral proteins as scaffolds for protein engineering. In: Directed enzyme evolution: advances and applications. pp. 229–255. https://doi.org/10.1007/978-3-319-50413-1_9
43. Thomson RE, Carrera-Pacheco SE, Gillam EM (2022) Engineering functional thermostable proteins using ancestral sequence reconstruction. J Biol Chem 298(10):102435. https://doi.org/10.1016/j.jbc.2022.102435
44. Kohout P, Vasina M, Majerova M, Novakova V, Damborsky J, Bednar D, Marek M, Prokop Z, Mazurenko S (2025). Engineering Dehalogenase Enzymes Using Variational Autoencoder-Generated Latent Spaces and Microfluidics. JACS Au. https://doi.org/10.1021/jacsau.4c01101.

Publisher's Note