

FireProt^{DB} 2.0: large-scale manually curated database of the protein stability data

Milos Musil^{1,2,3,*}, Simeon Borko^{1,3,†}, Joan Planas-Iglesias^{1,3}, David Lacko^{1,3}, Monika Rosinska³, Petr Kabourek^{1,3}, Lígia O. Martins⁴, Mateusz Tataruch⁵, Jiri Damborsky^{1,3}, Stanislav Mazurenko^{1,3,*}, David Bednar^{1,3,*}

¹Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Masaryk University, 625 00, Brno, Czech Republic

²Department of Information Systems, Faculty of Information Technology, Brno University of Technology, 612 66, Brno, Czech Republic

³International Clinical Research Centre, St. Anne's University Hospital Brno, 602 00, Brno, Czech Republic

⁴Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, 2780-157, Oeiras, Portugal

⁵Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, PL-30239, Krakow, Poland

*To whom correspondence should be addressed. Email: imusil@fit.vutbr.cz

Correspondence may also be addressed to Mazurenko. Email: mazurenko@mail.muni.cz

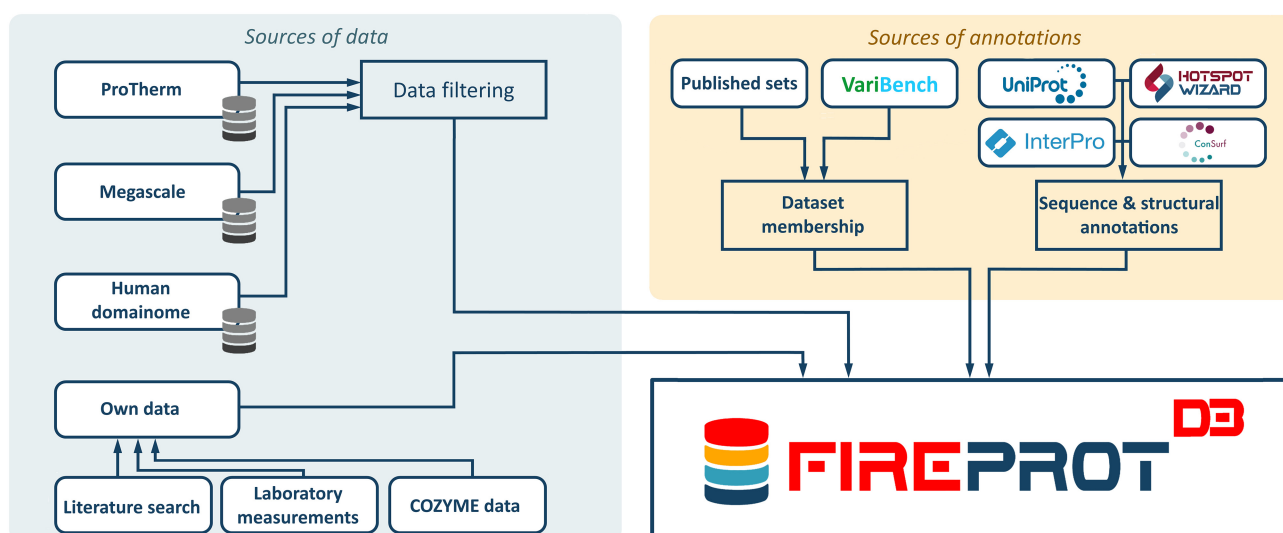
Correspondence may also be addressed to David Bednar. Email: 222755@mail.muni.cz

[†]The first two authors should be regarded as Joint First Authors

Abstract

Thermostable proteins are crucial in numerous biomedical and biotechnological applications. However, naturally occurring proteins have evolved to function in mild conditions, and laboratory experiments aiming at improving protein stability have proven laborious and expensive. Computational methods overcome this issue by providing a cheap and scalable alternative. Despite significant progress, their reliability is still hindered by the availability of high-quality data. FireProtDB 2.0 (<http://loschmidt.chemi.muni.cz/fireprotodb>) is a large-scale database aggregating stability data from multiple sources. The second version builds upon its predecessor, retaining its original functionality while introducing a new approach to data storage and maintenance. The new scheme enables the introduction of both absolute and relative data types connected with measurements of wild-types, mutants, protein domains, and *de novo* designed proteins. Furthermore, while the original database was limited to single-point mutations, more complex data such as insertions, deletions, and multiple-point mutations are now available. As a result, the inclusion of large-scale mutagenesis has increased the size of the database from 16 000 to almost 5 500 000 experiments. Moreover, the updated abstract scheme is fully expandable with any new measurements and annotations without the need for any restructuring. Finally, the tracking of history together with fixed identifiers is in accordance with the FAIR principles.

Graphical abstract



Received: September 15, 2025. Revised: October 16, 2025. Accepted: October 16, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Proteins play essential roles in a wide range of biomedical, pharmaceutical, and biotechnological applications. However, most naturally occurring proteins have evolved to function under relatively mild physicochemical conditions [1] and are therefore not inherently suited for the harsh environments typical of industrial processes. Introducing mutations that enhance protein stability enables their use under non-natural conditions, such as elevated temperatures [2] or extreme pH levels [3]. Increased protein stability is also associated with improved serum survival time [4], higher expression yields [5], and prolonged activity in the presence of denaturing agents [6]. Traditionally, the effects of mutations on protein stability are assessed through laboratory experiments, which, while reliable, are time-consuming, costly, and limited in scalability. As a result, only a small portion of the mutational space can be experimentally explored.

Over the past two decades, numerous predictive tools have been developed to reduce the vast mutational space to a more manageable subset [7]. These tools use rapid and moderately accurate computational methods to prioritize potentially stabilizing mutations for experimental validation. In general, computational approaches can be classified into three main categories: (i) force-field methods estimating protein free energy by calculating or approximating individual terms in an energy equation [8, 9]; (ii) evolution-based methods deriving insights from multiple sequence alignments and phylogenetic analyses [10, 11]; and (iii) machine learning methods extracting informative features from available data to make predictions [12–14].

While force-field methods are generally considered more robust, their advancement is limited by the current understanding of molecular interactions and the high computational cost involved. In contrast, machine learning offers a faster, more scalable alternative capable of uncovering novel, previously unknown patterns in data [15]. As a result, machine learning techniques have advanced rapidly—from simple neural networks [16, 17] to convolutional architectures [18, 19], and most recently to large-scale transformer-based language models [20–22]—all within a few decades. However, these models require large volumes of high-quality data, which can be challenging to obtain, particularly in the areas with complex biological data, such as protein stability [23, 24]. Furthermore, independent validation sets are essential for the rigorous evaluation of newly developed predictive tools [25].

To address this issue, the first version of FireProtDB [26] was released in 2021 as a suggested source of data for the training of novel predictive tools. FireProtDB addressed the previously reported errors and inaccuracies [15, 27] present in the older datasets and databases by thorough manual curation. It also enriched the data with several sequence- and structure-based annotations, all readily prepared to be utilized by researchers to develop and benchmark their tools. Despite the wide use by the community, the original FireProtDB database had a number of limitations. It was designed around single-point mutations, fixed structures, and less-scalable technologies, which did not allow for an easy expansion to new types of data and for effective searching in large datasets. It also could not store mutations in *de novo* designed protein sequences, which are becoming more abundant due to the recent progress in AI-based protein design [28, 29].

The herein presented second version of the database builds upon its predecessor by retaining its original functionality

while introducing a more abstract and flexible data structure. This new structure supports the storage of entirely new data types, annotations, and measurements, and is designed to be easily extendable for future updates. In addition, through the integration of multiple data sources [25, 30, 31] and large-scale mutagenesis studies [32], the size of the FireProtDB database has increased by two orders of magnitude—from fewer than 16 000 to over 5 million experiments, encompassing almost 13 million measurements. Combined with a fully redesigned interactive user interface enabling intuitive visualization of diverse data types and an optimized search engine capable of constructing complex queries, FireProtDB now serves as a powerful resource for various tasks. It is well-suited for both researchers aiming to build independent datasets for training, validation, and benchmarking their machine learning models, and for those simply seeking available data on their protein of interest.

Materials and methods

Data sources

The first version of the FireProtDB database provided an accessible solution for storing protein stability data and constructing training and validation datasets for machine learning applications. However, it was limited to single-point substitutions and supported only a small, predefined set of annotations and experimental measurements. The updated version supports both single- and multiple-point substitutions, as well as insertions, deletions, and measurements related to wild-type sequences. It can effectively store both absolute values (associated with the protein, such as ΔG and T_m) and relative values (associated with the mutation, such as $\Delta\Delta G$ and ΔT_m). Furthermore, the database is fully extensible as all measurements and annotations are stored as abstract key-value entities, allowing seamless integration of new data types as they become available.

In total, the second version of the FireProtDB database comprises 12 923 886 measurements, aggregated into 5 465 664 experiments across 2 762 unique proteins. It is important to note that duplicate measurement values may occur; however, these represent results obtained under different experimental conditions—such as variations in pH, method, or buffer concentration—or from independent data sources. Such cases are automatically treated as separate experiments. Currently, ~6% of mutations are present in more than one dataset, and only about 40 data points have multiple measurements of the same type but recorded under distinct conditions. The data originate from four main sources:

- i. The ProTherm database [31] was the original source of much of the experimental data stored in the older version of FireProtDB. It includes both single- and multiple-point substitutions, as well as measurements for wild-type proteins. In addition to protein stability data, ProTherm contained a variety of other experimental measurements—such as ΔC_p , ΔH , and m -values—along with detailed descriptions of experimental conditions, including pH, denaturant concentration, and measurement methods. After filtering out problematic entries (e.g. mutations with incorrect residue mapping or inconsistent data formatting), 6 691 single-point and 1 470 multiple-point substitutions were retained, along

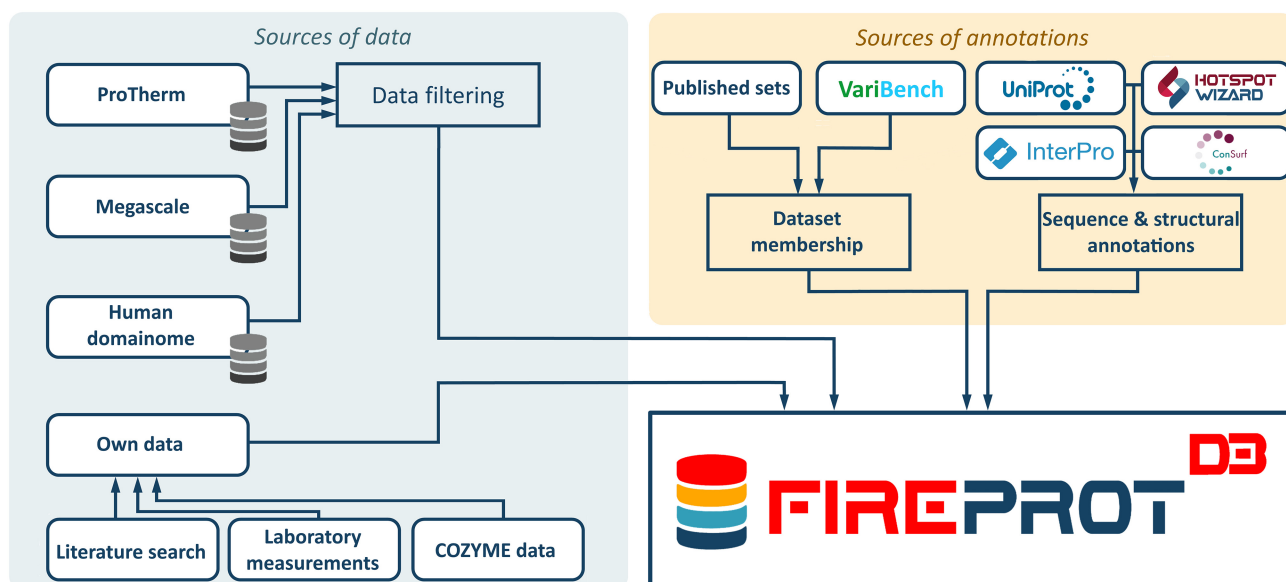


Figure 1. A schematic representation of data stored in the FireProtDB 2.0 database.

with 12 049 entries describing wild-type protein measurements.

- ii. The MegaScale dataset [30] comes from one of the largest experimental studies of protein folding stability, generated using high-throughput complementary DNA-display proteolysis. Although the dataset is limited to small protein domains and relies on protease-derived stability proxies rather than direct thermodynamic measurements, it remains a valuable source of large-scale mutational data. After applying quality filters, the dataset contributed 22 901 deletions, 47 173 insertions, 486 666 single-point, and 184 642 double-point substitutions. It also enriched the database with 148 *de novo* designed proteins.
- iii. The Human Domainome 1 dataset [32] originates from a large-scale site-saturation mutagenesis study involving over 500 human protein domains, evaluated using cellular abundance assays. Two of its subsets (Supplementary Datasets 2 and 5) were incorporated into FireProtDB due to their relevance to protein stability. This added 4 286 592 single-point substitutions, and 27 545 deletions to the database.
- iv. The remaining data were gathered from various sources, including internal datasets generated by the Loschmidt Laboratories [33] and comprehensive literature mining. Moreover, 165 additional data points [34–42] were obtained through collaboration with the members of the COZYME network [43].

In addition to experimental measurements, the database has been enriched with annotations derived from several external tools and databases. The information on protein origin was retrieved from the UniProt database [44]. InterPro [45] was used to identify overlapping protein domains and families. ConSurf [46] was employed to generate multiple sequence alignments and estimate the evolutionary conservation of individual residues. Furthermore, structural features, such as pockets, tunnels, tunnel bottlenecks, and tunnel-lining residues, were annotated using Caver [47] and FPocket2 [48] implemented in the HotSpotWizard service [49]. Finally, a benchmarking

repository of genetic variation datasets—VariBench [25]—designed to support the evaluation of computational prediction tools was utilized to annotate data with their presence in various protein stability datasets. This enables users to filter out the datasets used in the training of external computational models, which is particularly valuable for benchmarking. A full representation of data stored in the FireProtDB database is captured in Fig. 1.

Database architecture

The FireProtDB database, implemented in PostgreSQL, comprises thirty tables, which can be organized into several logical blocks, as illustrated in Fig. 2. The database is sequence-centric, meaning that structural information is not required for data to be stored. Given that the new version is designed to accommodate both absolute and relative values, as well as a wide range of annotation types, the central reference point for each database entry is the native protein sequence. This sequence is preferably retrieved from the UniProt database [44], if available, or alternatively sourced from the original publication. In cases where only the mutant sequence is available, the native sequence is reconstructed based on the provided list of mutations. The *Mutant* entity encompasses a group of tables used to track the specific insertions, deletions, and substitutions applied to generate a given mutant sequence. It also captures hierarchical relationships between designs, particularly when a mutant is derived from another mutant rather than directly from the native sequence. The *Measurement* table stores both absolute values (e.g. ΔG , T_m) linked to the native sequence, and relative values (e.g. $\Delta\Delta G$, ΔT_m) associated with the mutant. Since multiple measurements can originate from a single laboratory experiment, the *Experiment* is modeled as a distinct entity, allowing the inclusion of metadata such as pH, buffer composition, and concentration, denaturants, and other experimental conditions in the attached *Experimental annotations* table. The *Source* block contains tables describing the publications in which the data were first reported, including the list of authors. It also records the primary dataset and

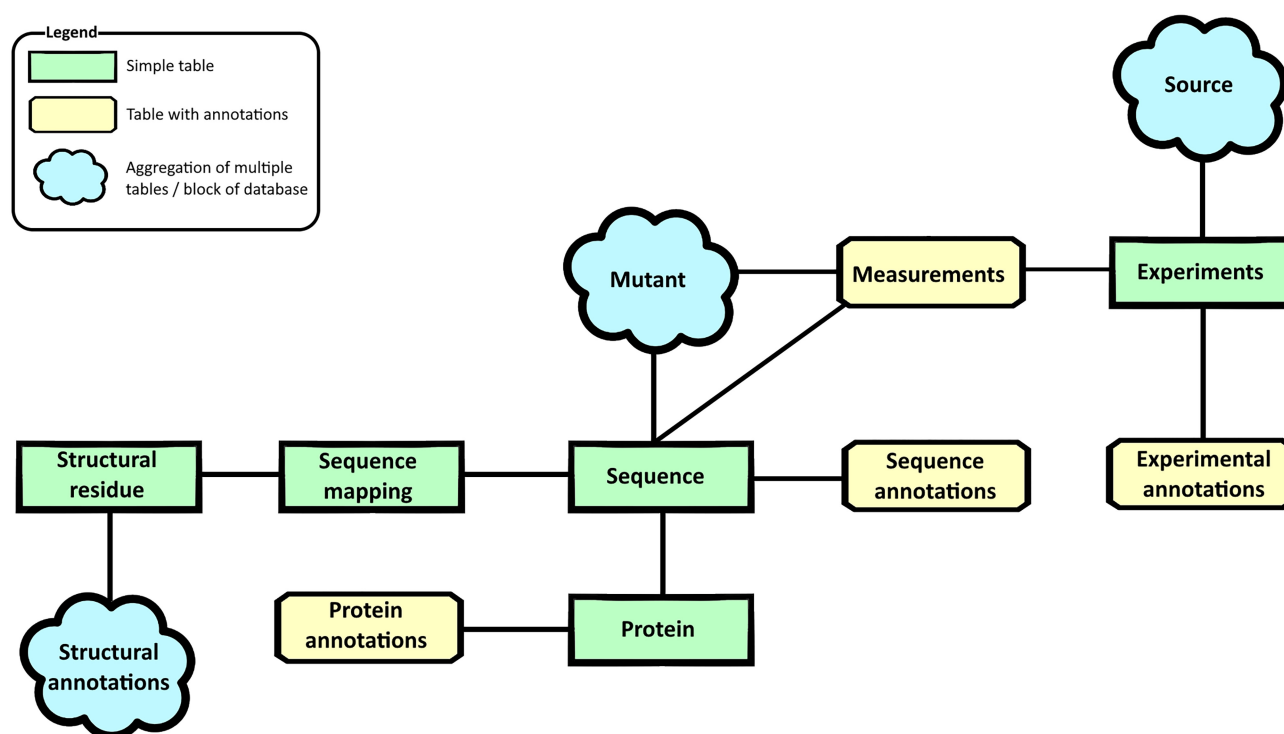


Figure 2. A schematic representation of the database architecture. Rectangular tables are the direct representation of the database entity with the chamfered rectangles being abstract key-value tables. Cloud shape represents a larger part of the database aggregated for simplicity. The Full Entity-relationship diagram can be found in the [Supplementary material 1](#).

all additional datasets in which a given experimental entry has appeared.

Each *Sequence* in the database can be associated with multiple *Sequence annotations*, computed either directly from the sequence itself or derived from multiple sequence alignments of homologous proteins. Additionally, each sequence is linked to a single entry in the *Protein* entity, which stores information about the protein's name and biological origin. Supplementary annotations—such as cross-references to external databases—are stored in a separate table, allowing for flexible integration of identifiers and metadata from various sources.

The final section of the database stores structural information, including protein tunnels, pockets, tunnel bottlenecks, and residue-specific B-factors. These data are linked not directly to the sequence, but to the corresponding tertiary structures representing the sequence. Multiple, slightly different structural models can be associated with a single sequence, each annotated independently. The central entity in this structural block is the *Structural residue*, which must be accurately mapped to the corresponding sequence residue, as structural and sequence residue numbering often differ. This mapping is performed using SIFTS [50]. The database also accommodates different biological assemblies, allowing the storage of multiple biological units derived from the same protein structure.

In addition to the main database, two materialized views have been created to enhance the efficiency of complex queries and data retrieval. The *Protein_view* aggregates information from across the entire database into a JSON format, where each entry represents a comprehensive JSON object containing all available data for a given protein. This design enables rapid loading and visualization of protein pages without the need to query multiple tables individually, significantly improving response times. The *Search_view* is specifically op-

timized to accelerate complex search queries. It consolidates arrays of searchable attributes associated with individual sequences with each column indexed using GIN indexing [51], thereby facilitating efficient filtering and retrieval. Each entry is also accommodated with a permanent unique identifier that is directly accessible in accordance with the FAIR principles [52, 53].

Results

Web interface

FireProtDB was primarily designed to enable fast and straightforward construction of protein stability datasets for training and validation, and benchmarking of novel computational tools. Nevertheless, its user interface is also well-suited for researchers outside the fields of machine learning and computational biology. A variety of data types are visually integrated along the sequence track, offering a comprehensive, in-depth view of each protein and its properties. The web server is organized into four main components: (i) a search engine for querying the database, (ii) search results displaying a summary of matching entries, (iii) a protein view presenting detailed information and visualizations for a selected protein, and (iv) a mutation view providing insights into individual variants and their associated experimental data. Alternatively, data can also be accessed via the API, documented at <https://loschmidt.chemi.muni.cz/fireprotodb/api-docs/>.

Search engine

The search engine operates on two levels of interaction to accommodate both casual users and advanced researchers. At the most basic level, users can enter a search term into the main

search bar. This initiates a standard full-text search across several data fields, including protein name, organism of origin, EC number, UniProt ID, PDB ID, InterPro ID, and publication. For more refined queries, users can click the “Advanced” button next to the search panel, enabling them to construct custom search conditions based on one or more criteria (Fig. 3A). The advanced search interface supports a total of 26 searchable fields, grouped into five distinct categories. Each field accepts one of the following input types: (i) a specific value for direct comparison, (ii) a value and a logical operator (e.g. equality, inclusion), or (iii) checkbox-based filters that define additional constraints.

Search terms can be combined using AND/OR logical operators, forming a tree-like query structure that is automatically translated into an optimized database query. This functionality makes the advanced search a powerful tool for extracting highly specific datasets. For example, users can retrieve all stabilizing mutations located in non-conserved regions of a given protein, or construct a benchmarking dataset consisting of mutations with experimentally validated $\Delta\Delta G$ values that were not used in training of the existing predictive models. The search engine, with several use cases, is further described in the [Supplementary material 2](#).

Search results

The results of a custom search are displayed in a results table (Fig. 3B), which by default includes several key measurements for each entry. Users can optionally expand the table to display up to ~40 additional annotations and measurements, providing a more comprehensive view of the data. From the results table, users can directly access both the *Protein View* and *Mutation View* for any entry. Additionally, all selected data can be exported as a .csv file directly from the results page, facilitating downstream analysis and integration into external workflows.

Protein view

The *Protein View* serves as the primary interface for exploring detailed information about a protein and its associated mutants. It is organized into several panels, each providing specific insights:

- I. Protein information: Displays key metadata including the protein’s origin, name, and relevant accession codes (Fig. 3G).
- II. Families and domains: Lists overlapping domains and families with direct links to the InterPro database for further exploration (Fig. 3H).
- III. Sequence features track: Offers an interactive visualization of various sequence- and structure-based features, mapped directly onto the protein sequence (Fig. 3C).
- IV. Sequence measurements table: Presents all available experimental measurements related to the wild-type protein, specifically absolute values such as ΔG or T_m (Fig. 3D).
- V. 3D structure visualization: Provides a graphical representation of the protein’s tertiary structure, enabling spatial interpretation of functional or mutational data (Fig. 3E).

- VI. $\Delta\Delta G$ visualization: Provides a graphical representation of the distribution of $\Delta\Delta G$ values across the protein sequence (Fig. 3J).
- VII. Mutant table: Lists all known mutants of the protein along with their associated relative measurements (e.g. $\Delta\Delta G$, ΔT_m), allowing quick comparison and analysis (Fig. 3F).

Mutant view

The Mutant View page can be accessed either directly from the search results table or via the mutant list on the corresponding *Protein View* page. It provides a more detailed analysis of a selected mutant, including all available experimental measurements related to the given variant (Fig. 3K). Additionally, the page features a sequence alignment between the wild-type and mutant sequences with all insertions, deletions, and substitutions clearly highlighted on the sequence track (Fig. 3I), allowing users to easily identify and interpret the introduced changes.

Statistics

By integrating data from multiple sources, a total of 12 923 886 measurements, aggregated into 5 465 664 experiments, were compiled into the database. Of these, 42 912 measurements represent absolute values associated with wild-type protein sequences, while the remaining 12 880 974 characterize mutant proteins through relative measurements. The dataset encompasses 47 173 insertions, 50 383 deletions, 4 778 176 single-point substitutions, and 186 194 multiple-point substitutions. A detailed breakdown of substitution frequencies for each pair of amino acids is shown in Fig. 4A, along with aggregated data on substitutions categorized by amino acid hydrophobicity (4B), charge (4C), polarity (4D), and volume (4E).

Figure 4F illustrates the distribution of mutations based on their effect on protein stability, classified into stabilizing, neutral, or destabilizing groups. The categorization is based on four criteria—experimental $\Delta\Delta G$ [54] and ΔT_m values, as well as normalized $\Delta\Delta G$ and fitness scores as defined in Human Domainome dataset [32]—applied as follows:

- i. A mutation is considered stabilizing if it meets at least one of the following thresholds: experimental $\Delta\Delta G < -0.5$ kcal/mol; normalized $\Delta\Delta G < -0.3$; $\Delta T_m > 1^\circ\text{C}$; or fitness > 0.3 .
- ii. A mutation is considered destabilizing if it meets at least one of the following conditions: experimental $\Delta\Delta G > 0.5$ kcal/mol; normalized $\Delta\Delta G > 0.3$; $\Delta T_m < -1^\circ\text{C}$; or fitness < -0.3 .
- iii. A mutation is classified as neutral if it falls within at least one of the following ranges: experimental $\Delta\Delta G \in [-0.5, 0.5]$ kcal/mol; normalized $\Delta\Delta G \in [-0.3, 0.3]$; $\Delta T_m \in [-1, 1]^\circ\text{C}$; or fitness $\in [-0.3, 0.3]$.
- iv. If more than one measurement ($\Delta\Delta G$, ΔT_m , normalized $\Delta\Delta G$, and fitness) is available for a given mutation, all have to agree; otherwise, the mutation is considered inconclusive. This can be caused by differences in experimental conditions (pH, buffer, etc.), the imperfect correlation of ~0.71 between $\Delta\Delta G$ and ΔT_m [55] (especially with mutations that are close to neutral), the average experimental measurement error of ~0.48 kcal/mol [54], or by discrepancies in the source data. In such cases, the

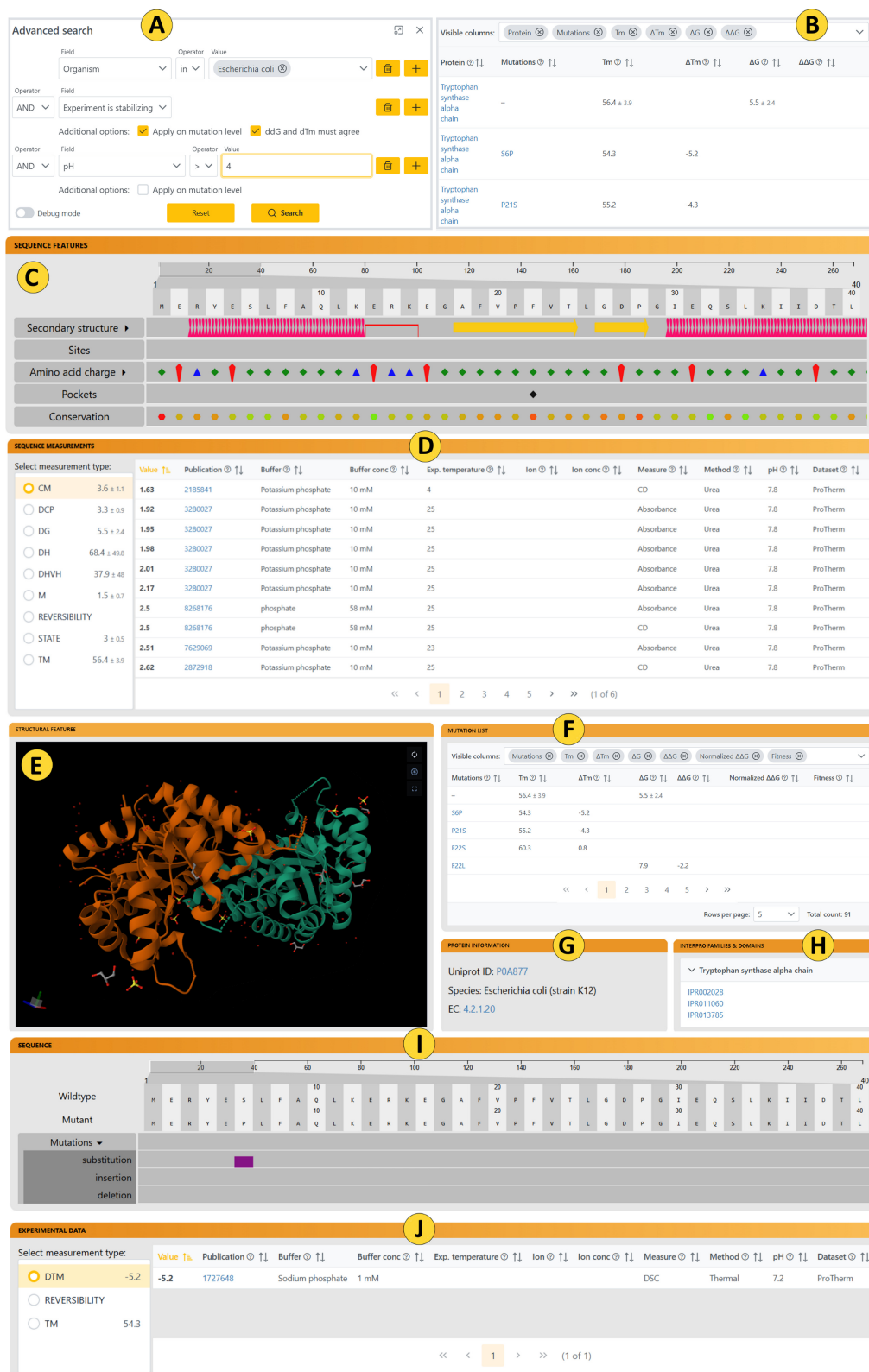


Figure 3. Graphical user interface of the FireProtDB 2.0 web server. The interface consists of several functional components: **(A)** Advanced search used for the construction of complex search queries. **(B)** A result table containing results fitting the search criteria. **(C)** Sequence track mapping various features directly on the sequence. **(D)** Table containing absolute values about the protein wild-type. **(E)** 3D structure visualization. **(F)** Mutant list showing all mutations attached to the protein. **(G)** Information about protein. **(H)** List of identified InterPro families. **(I)** Mapping of mutations on the protein sequence. **(J)** Distribution of $\Delta\Delta G$ across the protein sequence. **(K)** List of relative values connected to the selected mutant.

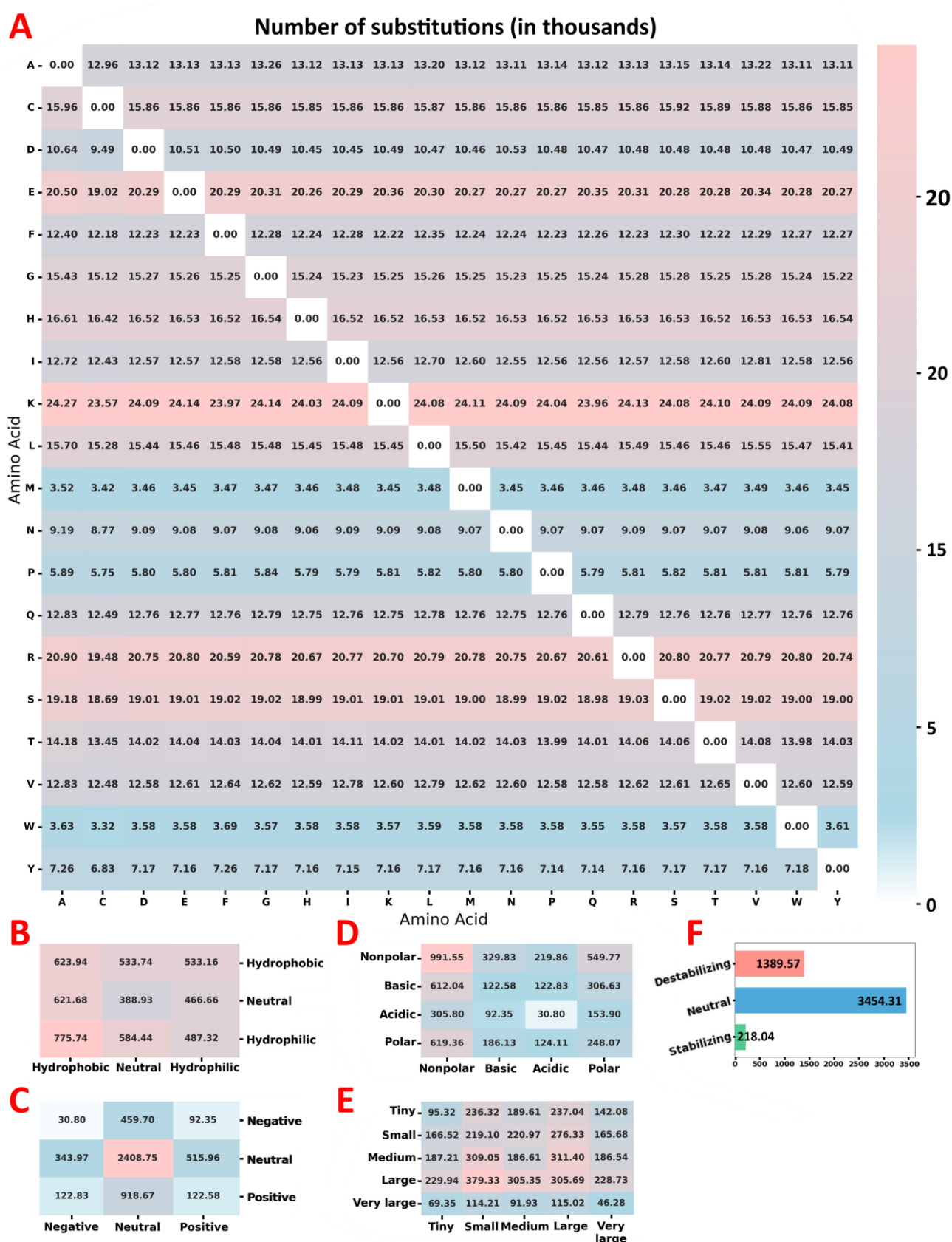


Figure 4. The statistics of the FireProtDB database, measured in thousands—the horizontal axis represents wild-type amino acids, and the vertical axis shows substitutions. (A) Amount of substitution data for each pair of amino acids; (B) substitutions based on the change in hydrophobicity; (C) substitutions considering charged amino acids; (D) substitutions causing a change in amino acid polarity; (E) substitutions based on the change in the size of the mutated amino acid; (F) a total number of stabilizing, destabilizing, and neutral mutations.

mutation will not be categorized as either stabilizing or destabilizing, unless explicitly specified by the user by limiting the search to only a selected type of measurement.

As expected, the dataset is dominated by neutral and destabilizing mutations, comprising 3 454 310 and 1 389 567 entries, respectively, compared to a smaller subset of 218 044 stabilizing mutations. Nevertheless, even this limited number of stabilizing variants represents a substantial dataset, sufficient for meaningful training, validation, and benchmarking of computational prediction tools.

Conclusions

The availability of high-quality protein stability datasets is essential for the continued advancement of predictive computational tools. While the original FireProtDB 1.0 provided researchers with a substantial collection of manually curated data, it was limited by a rigid database architecture and restricted data scope. The new version of FireProtDB 2.0 addresses these limitations through a complete redesign of the database structure, adopting a more abstract and flexible model that supports the seamless integration of new data types. In addition to relative information on single-point mutations, the database now includes data on multiple-point mutations, insertions, deletions, and absolute measurements related to wild-type proteins. Importantly, the system is fully extensible, enabling the incorporation of new annotations and experimental measurements as they become available, ensuring that FireProtDB remains a scalable and valuable resource for the scientific community.

In total, the latest release of the FireProtDB 2.0 database comprises 12 923 886 measurements, aggregated into 5 465 664 experiments conducted on 2 762 unique proteins, thus increasing the number of experiments by two orders of magnitude compared to the original set of 15 987 mutations. This comprehensive dataset is the result of integrating data from previously published datasets, large-scale mutagenesis studies, literature mining, and in-house experimental work, including collaborations with partner laboratories through the COZYME initiative (<https://cozyme.eu/>). In addition to experimental data, the database includes a wide range of automatically generated annotations, sourced from external databases such as UniProt, InterPro, and VariBench, and computed using tools such as ConSurf and HotSpotWizard.

FireProtDB 2.0 provides an accessible and comprehensive platform for the construction of high-quality training and validation datasets for predictive tools, as well as for researchers seeking to analyze proteins of interest through an intuitive user interface. This dual functionality establishes FireProtDB 2.0 as an invaluable resource for both machine learning applications and fundamental biological research. Moreover, its flexible architecture ensures that FireProtDB 2.0 can evolve to accommodate future advances while maintaining fast and efficient complex query performance through optimized use of native database functions.

Acknowledgements

Author contributions: Milos Musil (Conceptualization [equal], Data curation [equal], Software [equal], Visualization [equal], Writing—original draft [equal]), Simeon Borko

(Conceptualization [equal], Project administration [equal], Software [equal], Supervision [equal], Visualization [equal], Writing—review & editing [equal], Writing—review & editing [equal]), Joan Planas-Iglesias (Conceptualization [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal]), David Lacko (Data curation [equal], Software [equal], Validation [equal], Writing—review & editing [equal]), Monika Rosinska (Software [equal], Writing—review & editing [equal]), Petr Kabourek (Software [equal], Writing—review & editing [equal]), Lígia O. Martins (Data curation [equal], Writing—review & editing [equal]), Mateusz Tataruch (Data curation [equal], Writing—review & editing [equal]), Jiri Damborsky (Conceptualization [equal], Resources [equal], Resources [equal], Supervision [equal], Validation [equal], Writing—review & editing [equal]), Stanislav Mazurenko (Conceptualization [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal]), David Bednar (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

The authors thank the RECETOX, e-INFRA, and ELIXIR CZ Research Infrastructures (nos. LM2023069, 90254, and LM2023055), financed by the Czech Ministry of Education, Youth and Sports, and Grant Agency of the Czech Republic (no. 25-18233M). This work was also supported by the European Union's Horizon 2020 research and innovation programme (no. 857560 TEAMING), and by the European Union Centre of Excellence CLARA (no. 101136607), COST Action COZYME (no. CA21162), and Brno University of Technology (no. FIT-S-23-8209). 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 to pay the Open Access publication charges for this article was provided by European Union's Horizon 2020 research and innovation programme (no. 857560 TEAMING).

Data availability

The sequences and measurements for the Human Domainome 1 dataset were obtained from <https://zenodo.org/records/13629491> (Extended table 2 and 4), the MegaScale data are freely accessible at <https://zenodo.org/records/7992926>, and ProTherm data were downloaded from <https://web.iitm.ac.in/bioinfo2/prothermdb/>. Annotations of mutation presence in various training datasets were based on <https://structure.bmc.lu.se/VariBench/data/effectspecific/stability.php>, and protein annotations were automatically sourced from <https://www.uniprot.org/> and <https://www.ebi.ac.uk/interpro/> based on their UniProt access codes. Documentation for the Fire-

ProtDB API can be found at <https://loschmidt.chemi.muni.cz/fireprotodb/api-docs/>.

References

- Modarres HP, Mofrad MR, Sanati-Nezhad A. Protein thermostability engineering. *RSC Adv* 2016;6:115252–70. <https://doi.org/10.1039/C6RA16992A>
- Daniel RM, Dines M, Petach HH. The denaturation and degradation of stable enzymes at high temperatures. *Biochem J* 1996;317:1–11. <https://doi.org/10.1042/bj3170001>
- Jaenicke R. Protein stability and molecular adaptation to extreme conditions. *Eur J Biochem* 1991;202:715–28. <https://doi.org/10.1111/j.1432-1033.1991.tb16426.x>
- Gao D, Narasimhan DL, Macdonald J *et al*. Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol Pharmacol* 2009;75:318–23. <https://molpharm.aspetjournals.org/content/75/2/318.short>
- Ferdjani S, Ionita M, Roy B *et al*. Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnol Lett* 2011;33:1215–9. <https://doi.org/10.1007/s10529-011-0560-5>
- Polizzi KM, Bommarius AS, Broering JM *et al*. Stability of biocatalysts. *Curr Opin Chem Biol* 2007;11:220–5. <https://doi.org/10.1016/j.cbpa.2007.01.685>
- Qiu Y, Huang T, Cai YD. Review of predicting protein stability changes upon variations. *Proteomics* 2024;24:2300371. <https://doi.org/10.1002/pmic.202300371>
- Leaver-Fay A, Tyka M, Lewis SM *et al*. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 2011;487:545–74. <https://doi.org/10.1016/B978-0-12-381270-4.00019-6>
- Delgado J, Reche R, Cianferoni D *et al*. FoldX force field revisited, an improved version. *Bioinformatics* 2025;41:btaf064. <https://doi.org/10.1093/bioinformatics/btaf064>
- Lehmann M, Loch C, Middendorf A *et al*. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* 2002;15:403–11. <https://doi.org/10.1093/protein/15.5.403>
- Musil M, Khan RT, Beier A *et al*. FireProtASR: a web server for fully automated ancestral sequence reconstruction. *Brief Bioinform* 2020;22:1–11. <https://academic.oup.com/bib/article/22/4/bbaa337/6042664>
- 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. <https://doi.org/10.1038/s41467-024-49780-2>
- Zhou Y, Pan Q, Pires DEV *et al*. DDMut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Res* 2023;51:W122. <https://doi.org/10.1093/nar/gkad472>
- Chen CW, Lin MH, Liao CC *et al*. iStable 2.0: predicting protein thermal stability changes by integrating various characteristic modules. *Comput Struct Biotechnol J* 2020;18:622–30. <https://doi.org/10.1016/j.csbj.2020.02.021>
- Musil M, Konegger H, Hon J *et al*. Computational design of stable and soluble biocatalysts. *ACS Catal* 2019;9:1033–54. <https://doi.org/10.1021/acscatal.8b03613>
- Frenz CM. Neural network-based prediction of mutation-induced protein stability changes in Staphylococcal nuclease at 20 residue positions. *Proteins* 2005;59:147–51. <https://doi.org/10.1002/prot.20400>
- Yang Y, Ding X, Zhu G *et al*. ProTstab – predictor for cellular protein stability. *BMC Genomics* 2019;20:804. <https://doi.org/10.1186/s12864-019-6138-7>
- Fang X, Huang J, Zhang R *et al*. Convolution neural network-based prediction of protein thermostability. *J Chem Inf Model* 2019;59:4833–43. <https://doi.org/10.1021/acs.jcim.9b00220>
- Buyanov I, Sarycheva A, Popov P. OrgNet: orientation-gnostic protein stability assessment using convolutional neural networks. *Bioinformatics* 2025;41:i458–65. <https://doi.org/10.1093/bioinformatics/btaf252>
- Jiang F, Li M, Dong J *et al*. A general temperature-guided language model to design proteins of enhanced stability and activity. *Sci Adv* 2024;10:1–19. https://www.science.org/doi/full/10.1126/sciadv.adr2641?utm_source=chatgpt.com
- Jung F, Frey K, Zimmer D *et al*. DeepSTABp: a deep learning approach for the prediction of thermal protein stability. *Int J Mol Sci* 2023;24:7444. <https://doi.org/10.3390/ijms24087444>
- Gong J, Jiang L, Chen Y *et al*. THPLM: a sequence-based deep learning framework for protein stability changes prediction upon point variations using pretrained protein language model. *Bioinformatics* 2023;39:btad646. <https://doi.org/10.1093/bioinformatics/btad646>
- 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. <https://doi.org/10.1016/j.csbj.2020.07.011>
- Montanucci L, Martelli PL, Ben-Tal N *et al*. A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics* 2018;35:1513–17. <https://doi.org/10.48550/arXiv.1809.10389>
- Sasidharan NP, Vihinen M. VariBench: a benchmark database for variations. *Hum Mutat* 2013;34:42–9. <https://doi.org/10.1002/humu.22204>
- Stourac J, Dubrava J, Musil M *et al*. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res* 2021;49:D319–24. <https://doi.org/10.1093/nar/gkaa981>
- Yang Y, Urolagin S, Niroula A *et al*. PON-tstab: protein variant stability predictor. importance of training data quality. *Int J Mol Sci* 2018;19:1009. <https://doi.org/10.3390/ijms19041009>
- Chu AE, Lu T, Huang PS. Sparks of function by *de novo* protein design. *Nat Biotechnol* 2024;42:203–15. <https://doi.org/10.1038/s41587-024-02133-2>
- Kortemme T. *De novo* protein design—from new structures to programmable functions. *Cell* 2024;187:526–44. <https://doi.org/10.1016/j.cell.2023.12.028>
- 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. <https://doi.org/10.1038/s41586-023-06328-6>
- Nikam R, Kulandaisamy A, Harini K *et al*. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res* 2021;49:D420–4. <https://doi.org/10.1093/nar/gkaa1035>
- Beltran A, Jiang X, Shen Y *et al*. Site-saturation mutagenesis of 500 human protein domains. *Nature* 2025;637:885–94. <https://doi.org/10.1038/s41586-024-08370-4>
- Stourac J Protein engineering group. <https://loschmidt.chemi.muni.cz/peg/> (26 August 2025, date last accessed).
- Brissos V, Gonçalves N, Melo EP *et al*. Improving kinetic or thermodynamic stability of an azoreductase by directed evolution. *PLoS One* 2014;9:e87209. <https://doi.org/10.1371/journal.pone.0087209>
- Brissos V, Tavares D, Sousa AC *et al*. Engineering a bacterial DyP-Type peroxidase for enhanced oxidation of lignin-related phenolics at alkaline pH. *ACS Catal* 2017;7:3454–65. <https://doi.org/10.1021/acscatal.6b03331>
- Borges P, Silva D, Silva TFD Unveiling molecular details behind improved activity at neutral to alkaline pH of an engineered DyP-type peroxidase. *Comput Struct Biotechnol J* 2022;20:3899–910. <https://www.sciencedirect.com/science/article/pii/S2001037022003129>
- Simone MD, Alvigini L, Alonso-Cotchico L *et al*. Rationally guided improvement of NOV1 dioxygenase for the conversion of

- lignin-derived isoeugenol to vanillin. *Biochemistry* 2023;62:419–28. <https://pubs.acs.org/doi/full/10.1021/acs.biochem.2c00168>
38. De Simone M, Alonso-Cotchico L, Lucas MF *et al.* Distal mutations enhance efficiency of free and immobilized NOV1 dioxygenase for vanillin synthesis. *J Biotechnol* 2024;391:92–8. <https://doi.org/10.1016/j.jbiotec.2024.06.012>
 39. Rodrigues CF, Borges PT, Scocozza MF *et al.* Loops around the heme pocket have a critical role in the function and stability of BsDyP from *Bacillus subtilis*. *Int J Mol Sci* 2021;22:10862. <https://doi.org/10.3390/ijms221910862>
 40. Brissos V, Ferreira M, Grass G *et al.* Turning a hyperthermostable metallo-oxidase into a laccase by directed evolution. *ACS Catal* 2015;5:4932–41. <https://doi.org/10.1021/acscatal.5b00771>
 41. Brissos V, Borges PT, Núñez-Franco R *et al.* Distal mutations shape substrate-binding sites during evolution of a metallo-oxidase into a laccase. *ACS Catal* 2022;12:5022–35. <https://doi.org/10.1021/acscatal.2c00336>
 42. Brissos V, Borges PT, Sancho F *et al.* Flexible active-site loops fine-tune substrate specificity of hyperthermophilic metallo-oxidases. *J Biol Inorg Chem* 2024;29:339–51. <https://doi.org/10.1007/s00775-023-02040-y>
 43. COZYME – pan-european network on computational enzyme engineering. <https://cozyme.eu/> (26 August 2025, date last accessed).
 44. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–31. <https://doi.org/10.1093/nar/gkac1052>
 45. Hunter S, Apweiler R, Attwood TK *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009;37:D211–5. <https://doi.org/10.1093/nar/gkn785>
 46. Celniker G, Nimrot G, Ashkenazy H *et al.* ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Israel J Chem* 2013;53:199–206 <https://onlinelibrary.wiley.com/doi/full/10.1002/ijch.201200096>. (6 August 2025, date last accessed).
 47. Chovancova E, Pavelka A, Benes P *et al.* CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* 2012;8:e1002708. <https://doi.org/10.1371/journal.pcbi.1002708>
 48. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 2009;10:168. <https://doi.org/10.1186/1471-2105-10-168>
 49. Sumbalova L, Stourac J, Martinek T *et al.* HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res* 2018;46:W356–62. <https://doi.org/10.1093/nar/gky417>
 50. Velankar S, Dana JM, Jacobsen J *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 2013;41:D483–9. <https://doi.org/10.1093/nar/gks1258>
 51. 64.4. GIN Indexes. PostgreSQL documentation. <https://www.postgresql.org/docs/17/gin.html> (6 August 2025, date last accessed)
 52. Wilkinson MD, Dumontier M, IjJ A *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>
 53. Clifford D, Ausloos J. Data protection and the role of fairness. *Yearbook Eur Law* 2018;37:130–87. <https://doi.org/10.1093/yel/yey004>
 54. Khatun J, Khare SD, Dokholyan NV. Can contact potentials reliably predict stability of proteins? *J Mol Biol* 2004;336:1223–38. <https://doi.org/10.1016/j.jmb.2004.01.002>
 55. Pucci F, Bourgeois R, Rooman M. Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoTMuSiC. *Sci Rep* 2016;6:23257. <https://doi.org/10.1038/srep23257>