Name: FireProtDB: database of manually curated protein stability data

Cathegory: Social relevance

Jan Štourač, Juraj Dubrava, Miloš Musil, Jana Horáčková, Jiří Damborský, Stanislav Mazurenko, David Bednář

FireProtDB: database of manually curated protein stability data

Software, 2021

Cathegory: Social relevance

Summary

One of the major challenges in applying proteins to biotechnological processes is their limited stability at elevated temperatures or in the presence of denaturing agents. Traditionally, improving stability required extensive and costly experimental screening. In recent decades, this effort has been partially replaced by computational approaches, often based on machine learning. While these methods are faster and more economical, their progress is constrained by the lack of high-quality stability data for training and validation. As a result, most predictive tools are trained on noisy or overlapping datasets, which complicates reliable benchmarking.

FireProt-DB was created to address these limitations. It integrates published datasets, manually curated data from recent literature, and measurements generated at Loschmidt Laboratories and by collaborating groups. In addition, all entries are automatically annotated using multiple bioinformatics tools and databases. The interactive user interface is designed to accommodate a wide range of users, from computational biologists to experimental researchers. Currently, FireProt-DB contains nearly 16,000 stability measurements across 242 unique proteins.

Since its launch in December 2020, FireProt-DB has attracted nearly 20,000 users and has been employed in the development of several protein stability predictors. The database is supported by a publication in Nucleic Acids Research (Impact Factor 2021: 19.16, Q1 in Biochemistry and Molecular Biology). Since its publication in January 2021, the article has been cited 76 times in Web of Science, 81 times in Scopus, and 138 times in Google Scholar.

Finally, FireProt-DB was developed in collaboration with Loschmidt Laboratories at Masaryk University and the International Clinical Research Center of St. Anne's University Hospital.

Results

FireProt-DB integrates data from multiple sources, including protein stability datasets, protein databases, and results generated by computational tools (Figure **1**).

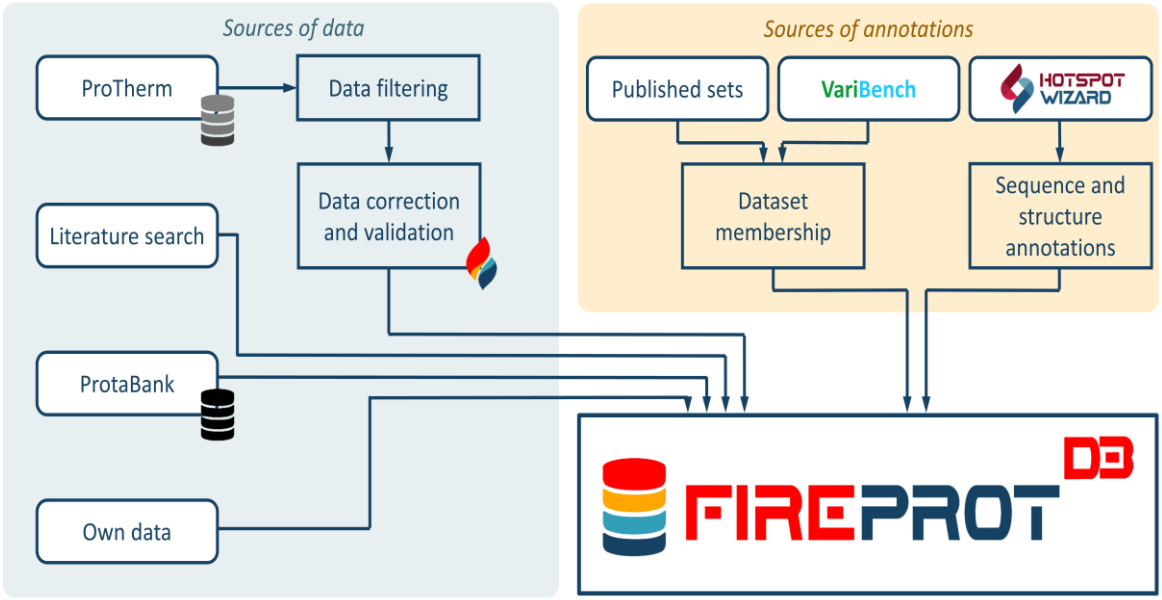


Figure 1: *A schematic representation of data contained in the FireProt-DB database.*

ProTherm served as the primary data source for FireProt-DB. At the time of release, it was the largest available collection of protein stability measurements, however, its usefulness was limited by substantial noise and inaccuracies. To address this, we applied a thorough manual curation process that included: (i) retaining only entries reporting ΔΔG or ΔT_m values; (ii) mapping entries to valid SwissProt accession codes and/or PDB identifiers; (iii) verifying the consistency between sequence and structural indices; (iv) remapping to higher-resolution PDB structures where available; and (v) cross-checking all entries against the original publications to identify and correct misinterpretations.

In addition to ProTherm, FireProt-DB was enriched with stability data from ProtaBank, manually extracted information from literature searches, and experimental measurements generated at Loschmidt Laboratories and through collaborations. To complement these experimental data, we integrated sequence and structural annotations from multiple sources. Core protein information was obtained from UniProt, while over forty VariBench datasets were incorporated to indicate whether specific entries were included in benchmarking or training sets of predictive tools. Furthermore, the HotSpotWizard workflow was applied to every protein in the database to provide additional sequence- and structure-derived features, including: (i) conservation scores, (ii) amino acid correlations, (iii) secondary structure, (iv) solvent-accessible surface area, and (v) localization within pockets, tunnels, or tunnel bottlenecks. Basic physicochemical properties were also extracted from AAIndex.

In total, FireProt-DB offers nearly 16,000 curated stability measurements across 242 unique proteins, each enriched with a comprehensive set of sequence and structural annotations. At the time of its publication, this made it the largest resource of protein stability data available.

To maximize usability, FireProt-DB is equipped with an interactive user interface that supports both simple full-text search and the construction of complex queries without requiring knowledge of database query languages (https://loschmidt.chemi.muni.cz/fireprotdb/). The advanced search functionality enables users to create highly specific subsets of data for training and benchmarking of computational tools. In addition, integrated visualization features, such as a 3D structure viewer and a sequence track, allow users to analyze proteins directly within the interface.

The rationale of excellence

FireProt-DB was developed in collaboration with Loschmidt Laboratories at Masaryk University and the International Clinical Research Center of St. Anne's University Hospital. The project is also part of the ELIXIR Czech Republic initiative, which focuses on the organization, storage, sharing, and interoperability of life science data.

The project has provided the scientific community with two key outputs: (i) a peer-reviewed article published in Nucleic Acids Research and (ii) a freely accessible web server containing nearly 16,000 protein stability measurements. With its interactive user interface and the ability to construct highly specialized data subsets, FireProt-DB serves as a valuable resource for both computational scientists developing or benchmarking predictive tools, and experimental researchers seeking comprehensive information about their protein of interest.

The FireProt-DB web server was released at the end of 2020, followed by the publication in Nucleic Acids Research in early 2021. At that time, the journal had an impact factor of 19.16 (ranked 8th out of 296 in Biochemistry and Molecular Biology) and a Journal Citation Indicator of 3.12 (ranked 5th out of 321 in the same field). Since publication, FireProt-DB has been cited 76 times in Web of Science, 81 times in Scopus, and 138 times in Google Scholar. The web server has also been accessed by close to 20,000 users worldwide (see Figure **2** for distribution).

Data from FireProt-DB have already been used in the development of several protein stability predictors, including DDMut, StabilityOracle, MutCompute, ProStage, and SCONES. In addition, the database has been integrated into PDBe-KB database and employed in the construction of benchmarking datasets such as BenchStab.

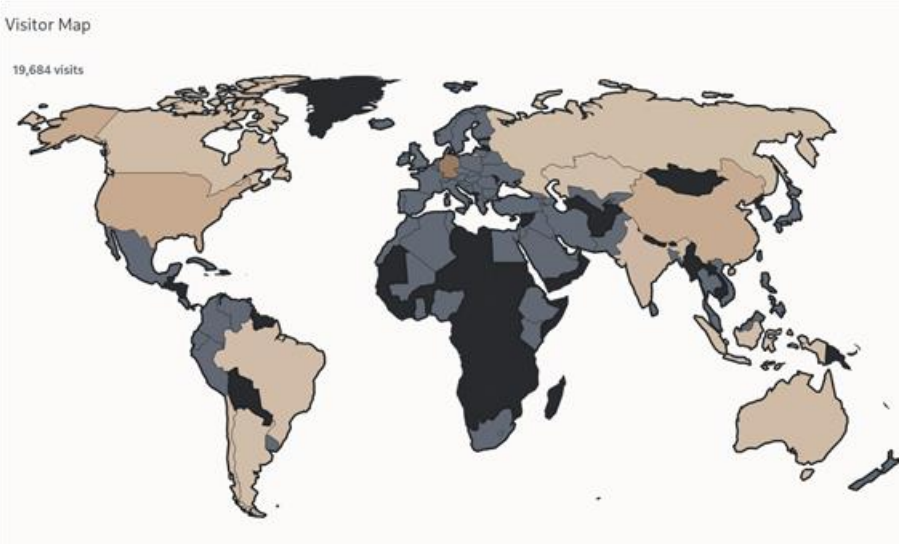


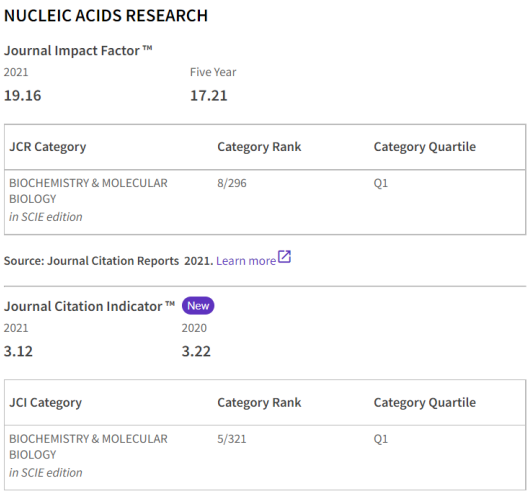
Figure 2: *Distribution of users of the FireProt-DB web server.*

The above-mentioned points demonstrate that FireProt-DB represents a world-class achievement in terms of both originality and scientific impact. This is evidenced by its thousands of active users, its publication in a highly influential journal, and 81 citations according to Scopus. Moreover, FireProt-DB has become a cornerstone for further research in protein stability, particularly in the development of machine learning tools and in independent benchmarking. Until now, such benchmarking has been hindered by extensive overlaps between the training datasets of individual tools, a limitation that FireProt-DB helps to overcome.

Citation analysis

The publication was released in *Nucleic Acids Research* in January 2021 (Figure **3**).

Citations WoS: 76, citations Scopus: 81, citations Scholar: 138, Q1 in Biochemistry and Molecular Biology. Scopus FWCI 4.18, 96 percentile.



NUCLEIC ACIDS RESEARCH

Journal Impact Factor™

2021	Five Year
19.16	17.21

JCR Category	Category Rank	Category Quartile
BIOCHEMISTRY & MOLECULAR BIOLOGY *in SCIE edition*	8/296	Q1

Source: Journal Citation Reports 2021. Learn more

Journal Citation Indicator™ New

2021	2020
3.12	3.22

JCI Category	Category Rank	Category Quartile
BIOCHEMISTRY & MOLECULAR BIOLOGY *in SCIE edition*	5/321	Q1

Figure 3: *Journal statistics according to WoS.*

Selected citations (from works that are themselves well cited):

- Machine learning-guided protein engineering (ACS Catalysis, 128 citations)
- Proteingym: Large-scale benchmarks for protein fitness prediction and design (NeurIPS proceedings, 270 citations)
- Transfer learning to leverage larger datasets for improved prediction of protein stability changes (PNAS, 92 citations)
- Enzymes, In Vivo Biocatalysis, and Metabolic Engineering for Enabling a Circular Economy and Sustainability (Chemical Reviews, 207 citations)
- Thermal stability enhancement: Fundamental concepts of protein engineering strategies to manipulate the flexible structure (International Journal of Biological Macromolecules, 107 citations)
- The 2022 Nucleic Acids Research database issue and the online molecular biology database collection (Nucleic Acids Research, 195 citations)

Selected citations (utilized FireProt-DB for training or benchmarking):

- Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations (Nature, 57 citations)
- ProSTAGE: Predicting Effects of Mutations on Protein Stability by Using Protein Embeddings and Graph Convolutional Networks (Journal of Chemical Information and Modeling, 20 citations)
- SCONES: Self-Consistent Neural Network for Protein Stability Prediction Upon Mutation (The Journal of Physical Chemistry, 32 citations)
- VenusMutHub: A systematic evaluation of protein mutation effect predictors on small-scale experimental data (Acta Pharmaceutica, 3 citations)
- VariBench, new variation benchmark categories and data sets (Frontiers in Bioinformatics, 3 citations)
- DDMut: predicting effects of mutations on protein stability using deep learning (Nucleic Acids Research, 161 citations)

Authors contribution

The second and third authors are current or former members of Brno University of Technology. Juraj Dúbrava was primarily responsible for the development of a large part of the FireProt-DB webserver, while Miloš Musil contributed to the database architecture, data collection, refinement, storage, and the calculation of sequence and structural features, as well as to the writing of the manuscript. The first author, Jan Štourač, was involved in all aspects of the development and data collection. Jana Horáčková assisted with data collection, while Jiří Damborský, Stanislav Mazurenko, and David Bednář served as consultants and contributed to the refinement of the dataset.