

UniProt: From Protein Data to Discovery



Challenges

The rapid expansion of genomic and protein sequence data has created a major challenge for understanding protein function.

Millions of newly sequenced proteins remain experimentally uncharacterised, limiting biological discovery, therapeutic development and biotechnological innovation. Furthermore, insufficient integration with other omics data prevents researchers from connecting individual findings into a coherent understanding of biological systems and generating novel hypotheses that could advance human health, sustainable agriculture and planetary health.

Keywords

Protein function # UniProt # Functional annotation # Bioinformatics # Computational biology # Protein database # Drug discovery # Machine learning datasets

Technology

Dr. Stephanie Lo and the Protein Function Content Team at EMBL-EBI have developed a comprehensive framework for the systematic and accurate annotation of proteins—the fundamental building blocks of life—within UniProt. Key components of this framework include:

- **UniProt Knowledgebase ([UniProtKB](#)):** A comprehensive, high-quality resource that integrates protein sequences and functional annotations with genetic variants, multiomics evidence and molecular interaction data across the tree of life. UniProtKB provides a particularly rich annotation of the human proteome, making it a key resource for biomedical and translational research.
- **Targeted curation projects:** Expert functional curation and integration of multi-omics, interaction and disease-relevant knowledge for proteins associated with a specific disease, with the resulting evidence integrated and summarised for analysis. Our previous Alzheimer's disease project integrated literature-derived evidence into UniProt, Gene Ontology, IMEx, and related resources, enabling improved interpretation of transcriptomic, proteomic and genetic datasets, GWAS and network analysis, and hypothesis generation.
- **Critical evaluation of AI-generated curation:** Expert curators provide critical evaluation and validation of AI-generated protein functional predictions. This is exemplified by our successful collaborations with Google DeepMind on ProtNLM and the development of our rule-based automated annotation system, ARBA, which enables functional annotation of millions of previously uncharacterised proteins at scale.
- **Standardised ontologies for analysis:** Our team is a leading contributor to the Gene Ontology (GO), developing, standardised, machine-readable functional annotations that power computational analyses at scale.

This technology enables scalable, high-quality functional annotation, ensuring the growing protein sequence data is interpretable and actionable.

Internal EMBLEM Reference

2026-046

Key Inventors

Dr. Stephanie Lo, Team Leader "Protein Function Content", EMBL-EBI, Hinxton, UK



Boxberggring 107
D-69126 Heidelberg
Germany



Tel.: +49 (0) 6221 363 22 10



info@embl-em.de



www.embl-em.de

Dr. Birgit Kerber

kerber@embl-em.de

EMBLEM TECHNOLOGY TRANSFER GMBH

UniProt: From Protein Data to Discovery



Applications

- **Drug discovery:** Target identification and biomarker discovery.
- **AI for protein function:** Curated training data for machine learning models.
- **Multi-omics analysis:** Interpretation of genomic, transcriptomic and proteomic data.
- **Precision medicine:** Disease variant interpretation.
- **Protein engineering:** Design of novel proteins and enzymes.
- **Agriculture & industrial biotech:** Discovery of enzymes for crop improvement and bioprocessing.
- **Bioinformatics platforms:** Integration into commercial databases and analysis tools.
- **Systems biology:** Pathway analysis and hypothesis generation.

Seeking

- Collaborations on large-scale protein function annotation projects, summary synthesis and hypothesis generation.
- Joint Development: Leveraging curation expertise for new computational methods and predictive tools.

Benefits

- High-quality, reliable annotations combining manual expertise with computational inference.
- Scalability to handle millions of protein sequences.
- Interoperability with other bioinformatics resources via standardized ontologies and cross-references.
- Open access enables rapid adoption in research and commercial applications.

Software

- [UniProt](#) Web Interface & APIs: Access to curated protein datasets and search tools.
- Annotation Pipelines: Computational tools for automated protein function prediction integrated with UniProt.

Further Reading

[1] <https://doi.org/10.1093/nar/gkae1010>

[2] <https://www.uniprot.org/>

[3] [Protein Function Content. Annotating protein function and properties](#)

Commercial Opportunity

- **Pharmaceuticals & Biotechnology:** Drug target identification, biomarker discovery, and protein engineering.
- **Bioinformatics & AI Tools:** Training data for machine learning or AI models in protein function prediction.
- **Agriculture & Industrial Enzymes:** Functional annotation of novel enzymes for crop improvement or industrial processes.

Intellectual Property

- UniProt and associated annotation pipelines are openly accessible under licenses that encourage research and commercial use.
- No patents are currently associated with the annotation framework itself, but the curated datasets are widely recognized as authoritative sources in life sciences.

EMBLEM TECHNOLOGY TRANSFER GMBH

Boxberggring 107
D-69126 Heidelberg
Germany

Tel.: +49 (0) 6221 363 22 10

info@embl-em.de Dr. Birgit Kerber

www.embl-em.de kerber@embl-em.de