

## **Breakout Session 1: Track B**

# **UniProt Knowledgebase to Enable AI/ML Readiness and Applications**

Dr. Cathy Wu

*Professor and Director, Data Science Institute, University of Delaware*

# UniProt AI Readiness

<https://www.uniprot.org/>

NIH ODSS AI Supplement Program PI Meeting

NIH FY22 AI-Readiness program (3U24HG007822-09S2)

*UniProt - Protein sequence and function embeddings for AI/ML readiness*

MPIs: Alex Bateman & Cathy Wu

March 27, 2024

Cathy Wu

University of Delaware

# Project Goals

1. Organize UniProt data to be AI-ready
  - Enable the community to harness AI/ML using UniProt data
2. Work with the community to advance AI-readiness and applications
  - Focus on useful problems through setting challenges
3. AI-driven innovation for UniProt resource development
  - Identify opportunities and new partners for collaborative development

Lack of **well-curated** data  
is a major barrier for AI Research

# AI/ML-Readiness, Engagement & Innovation

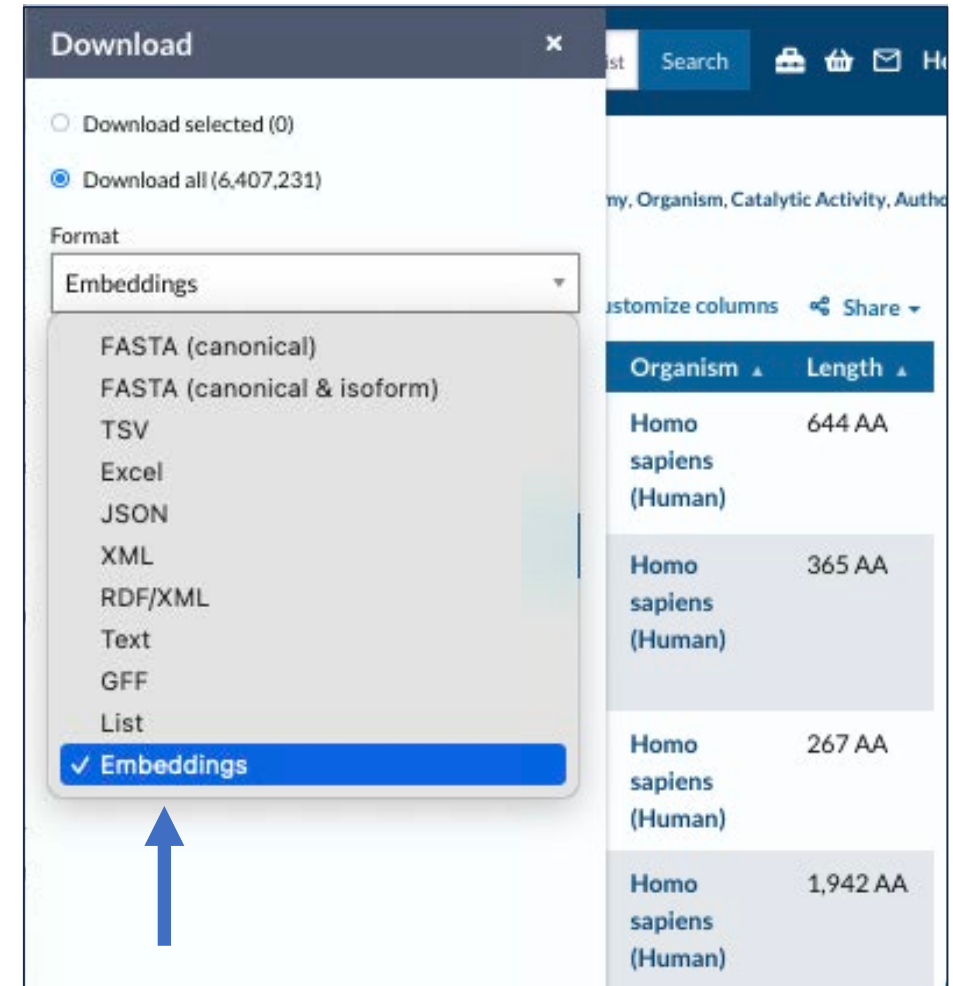
1. Dissemination of AI/ML models/data (collaborator-generated)
  - AlphaFold 2.0
  - ProtT5 sequence embeddings
2. AI/ML community engagement
  - Challenge evaluation – CAFA-UniProt metal binding challenge
  - Community workshops to discuss AI/ML readiness and applications
3. Collaborative development with AI communities
  - ProtNLM language model for protein name and function prediction
  - OntoGPT for extracting structured information from text with Large Language Model (LLM)
  - Text mining and LLM for UniProt annotation

# AI/ML Dissemination – AlphaFold

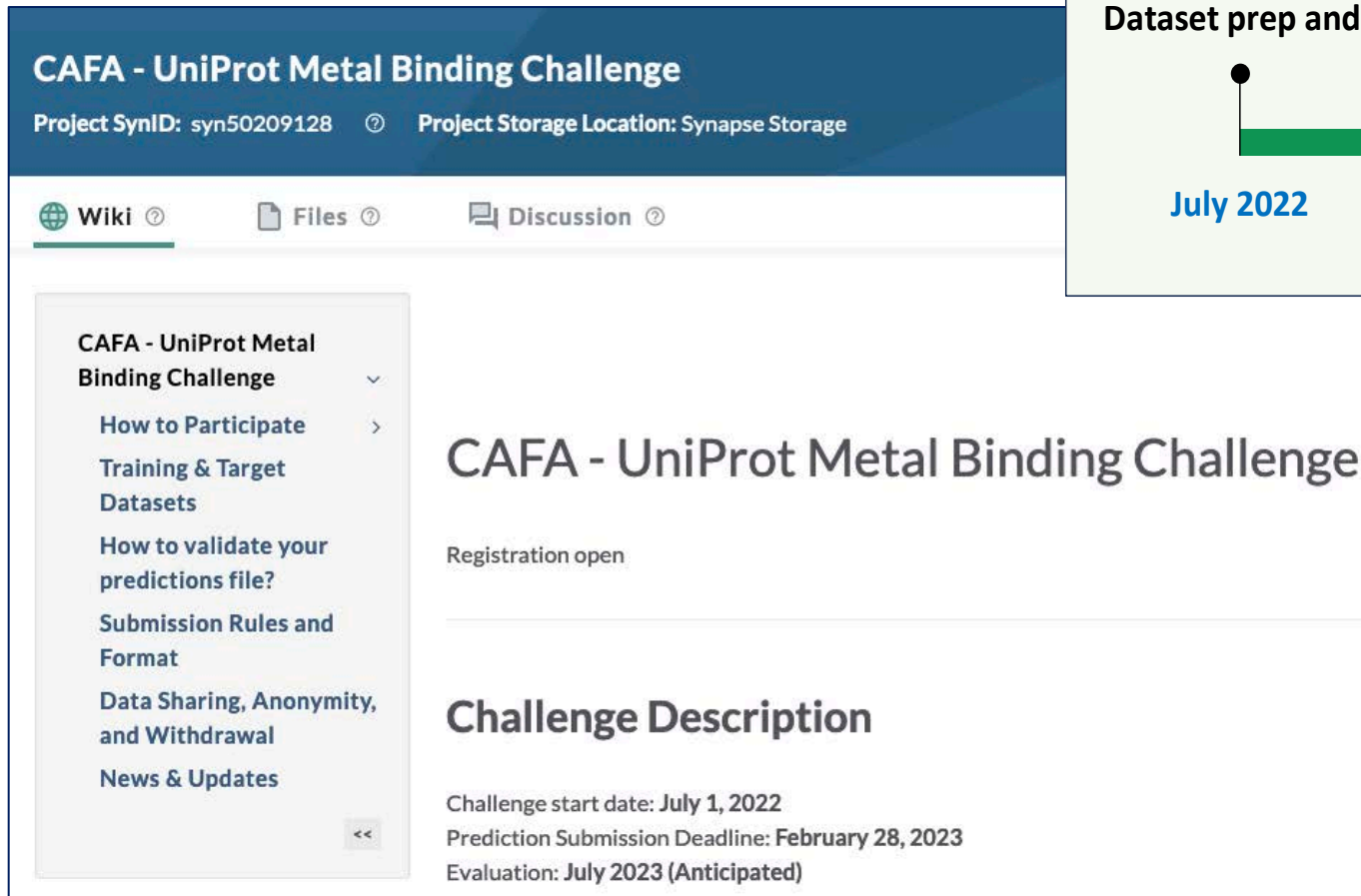
- Deep-learning based **AlphaFold 2.0** demonstrates atomic accuracy
- EMBL-EBI collaborated with **DeepMind** to release AlphaFold models and launched the AlphaFold Protein Structure Database in 2021
- UniProt has developed a process for making AlphaFold structures available from UniProt when the structures are released in the AlphaFold database
- Currently over 188 million UniProtKB proteins have AlphaFold 2.0 predictions

# AI/ML Dissemination – Sequence Embedding

- Protein **sequence embedding**: Encode functional and structural properties of a protein from its sequence in a vector representation
- UniProt **data ready for AI/ML** via sequence embedding: Save community compute and enable the community to harness AI/ML
- Provide a generic framework for a wide range of AI/ML tasks
- Precomputed UniProtKB datasets
- User-tailored datasets through website (e.g., all proteins with structure from PDB, all bacteria)



# Community Engagement – CAFA-UniProt Challenge



**CAFA - UniProt Metal Binding Challenge**

Project SynID: syn50209128 Project Storage Location: Synapse Storage

Wiki Files Discussion

CAFA - UniProt Metal Binding Challenge

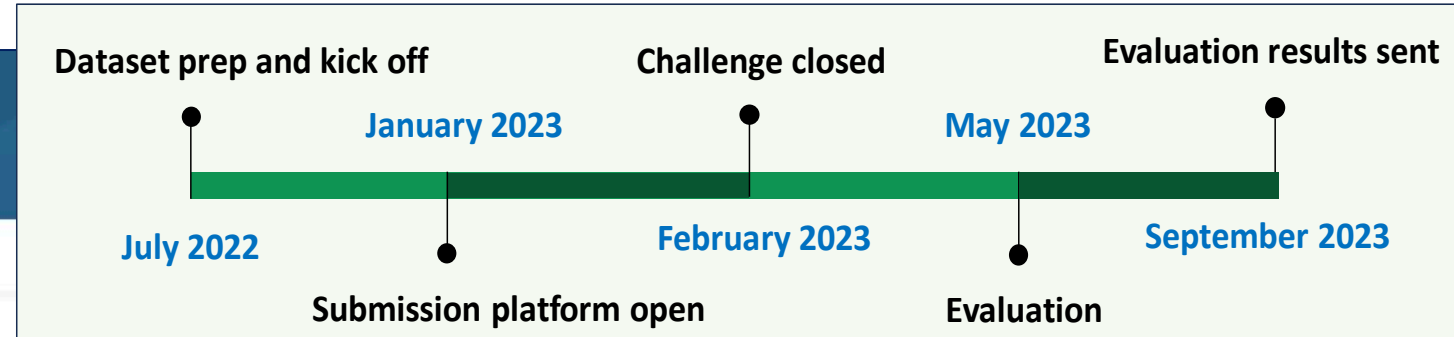
- How to Participate
- Training & Target Datasets
- How to validate your predictions file?
- Submission Rules and Format
- Data Sharing, Anonymity, and Withdrawal
- News & Updates

**CAFA - UniProt Metal Binding Challenge**

Registration open

**Challenge Description**

Challenge start date: July 1, 2022  
 Prediction Submission Deadline: February 28, 2023  
 Evaluation: July 2023 (Anticipated)



## Evaluation

- Metal Binding Challenge hosted on **Synapse**
- 17 registered participants
- High false positives in the submissions

## Next steps

- Host challenge on **Kaggle** to attract more participants and data scientists
- Leverage CAFA expertise and frameworks

# Community Engagement - UniProt AI Workshop

Provider of useful AI-ready datasets and embeddings to the community



User of AI methods that improve aspects of its functioning, e.g., functional annotation

## Community Engagement Workshop, 2023

- Working with the community to advance **AI-readiness and applications**
  - Sharing advances of UniProt AI development
  - Learning ongoing AI development to be leveraged in UniProt
  - Understanding community needs through solicitation of use cases



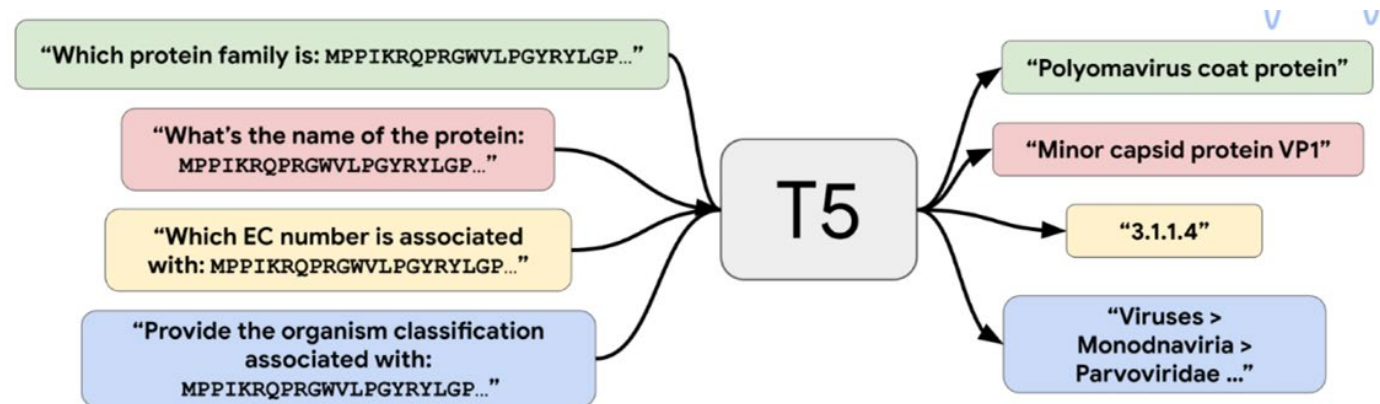
# ProtNLM Natural Language Models for Annotation

## ProtNLM (Name predictions to uncharacterized proteins)

- Fine-tuned algorithm and improved naming in training data set – better prediction quality
- 28,972,944 uncharacterized proteins have names from ProtNLM in UniProt 2024\_01

## ProtNLM (full UniProt predictions) - in progress

- Selection of prediction types (EC numbers, Function, etc.)
- Preliminary assessment of annotations
- Evidence strategy (e.g., compare with other annotations in entry and pHMMER alignments)



Collaboration with Lucy Colwell and Max Bileschi groups at Google Research (DeepMind)

<https://ebrevdo.github.io/publication/gane-2022-az/>

# Future Work & Pilots: LLM in UniProt



Identification and  
review of relevant  
literature

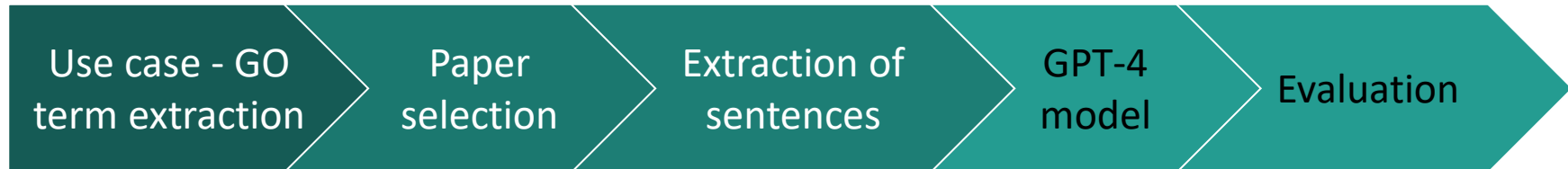
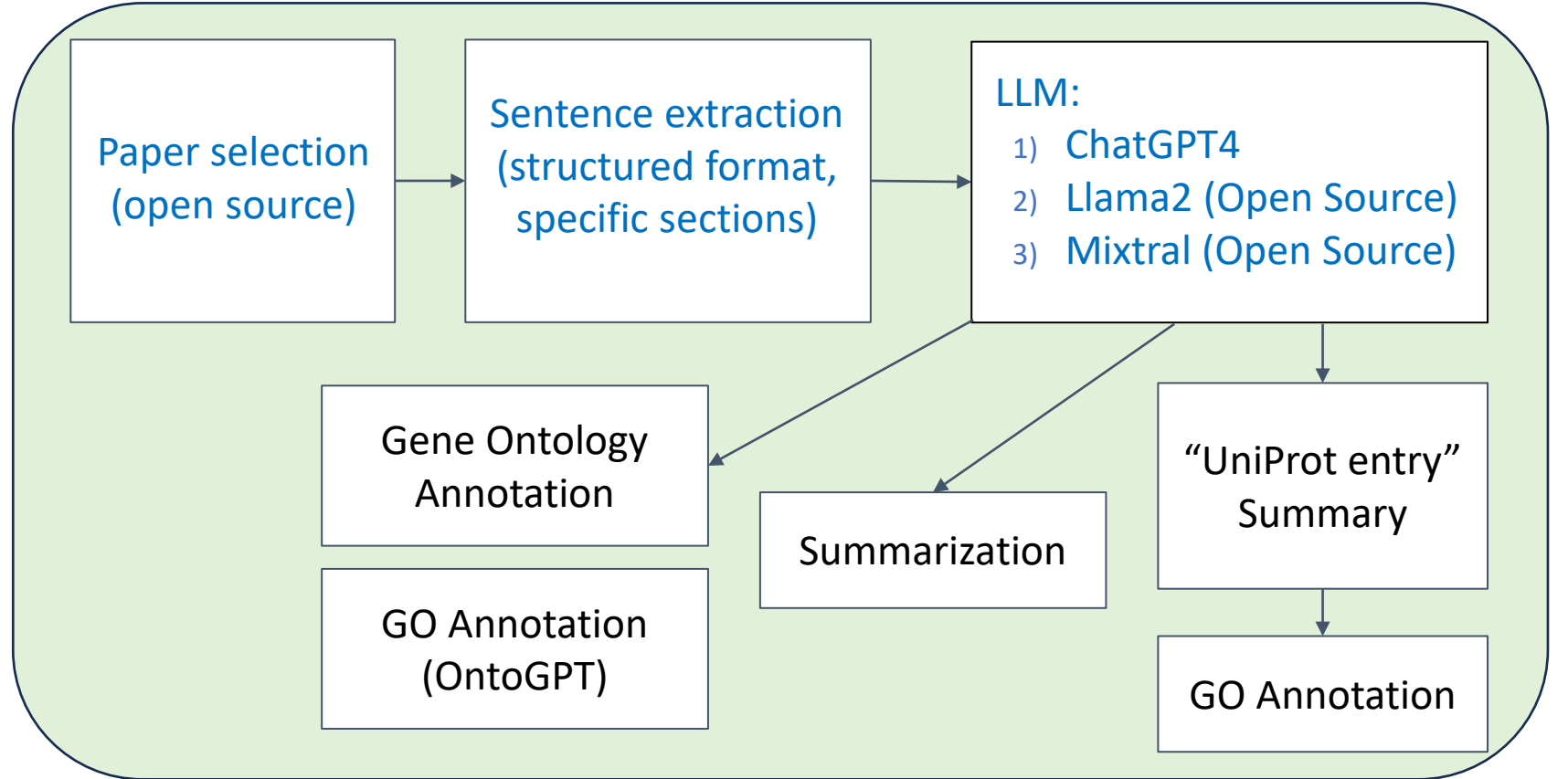
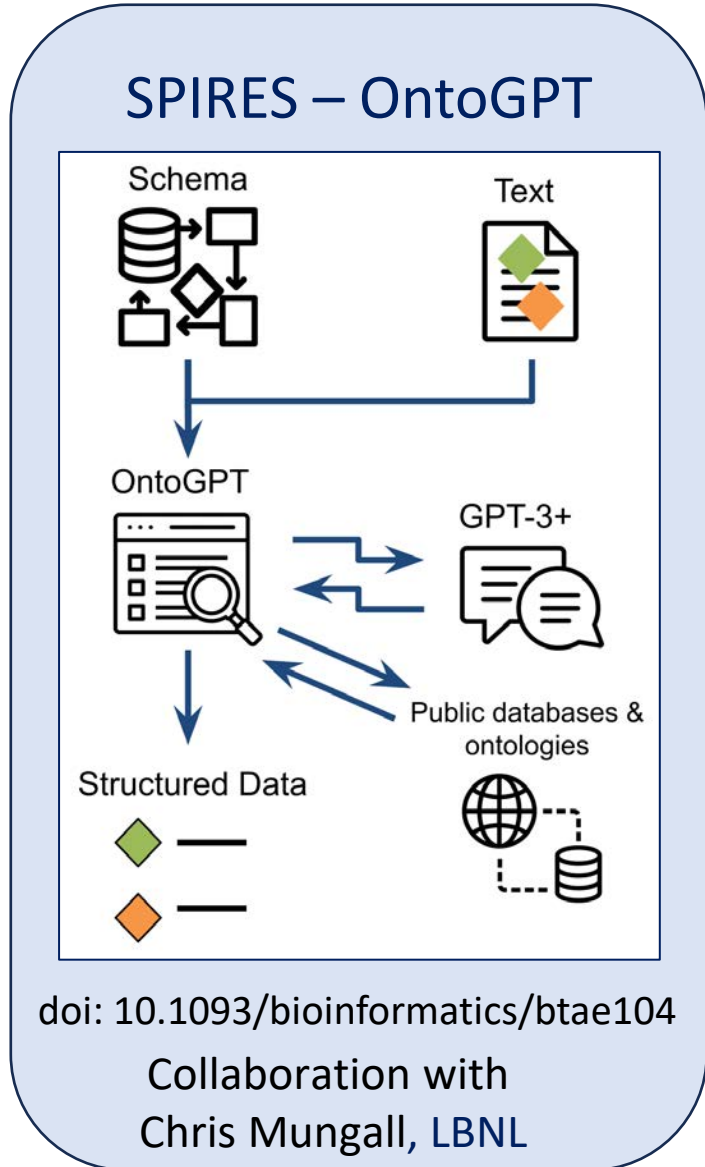
## Manual Curation

- To scan/summarize published papers with a focus on the gene of **function** interest
- To automatically identified which ones are adding “new information”
- To be used a co-pilot for curators writing scientific summaries (integrating information from multiple papers)

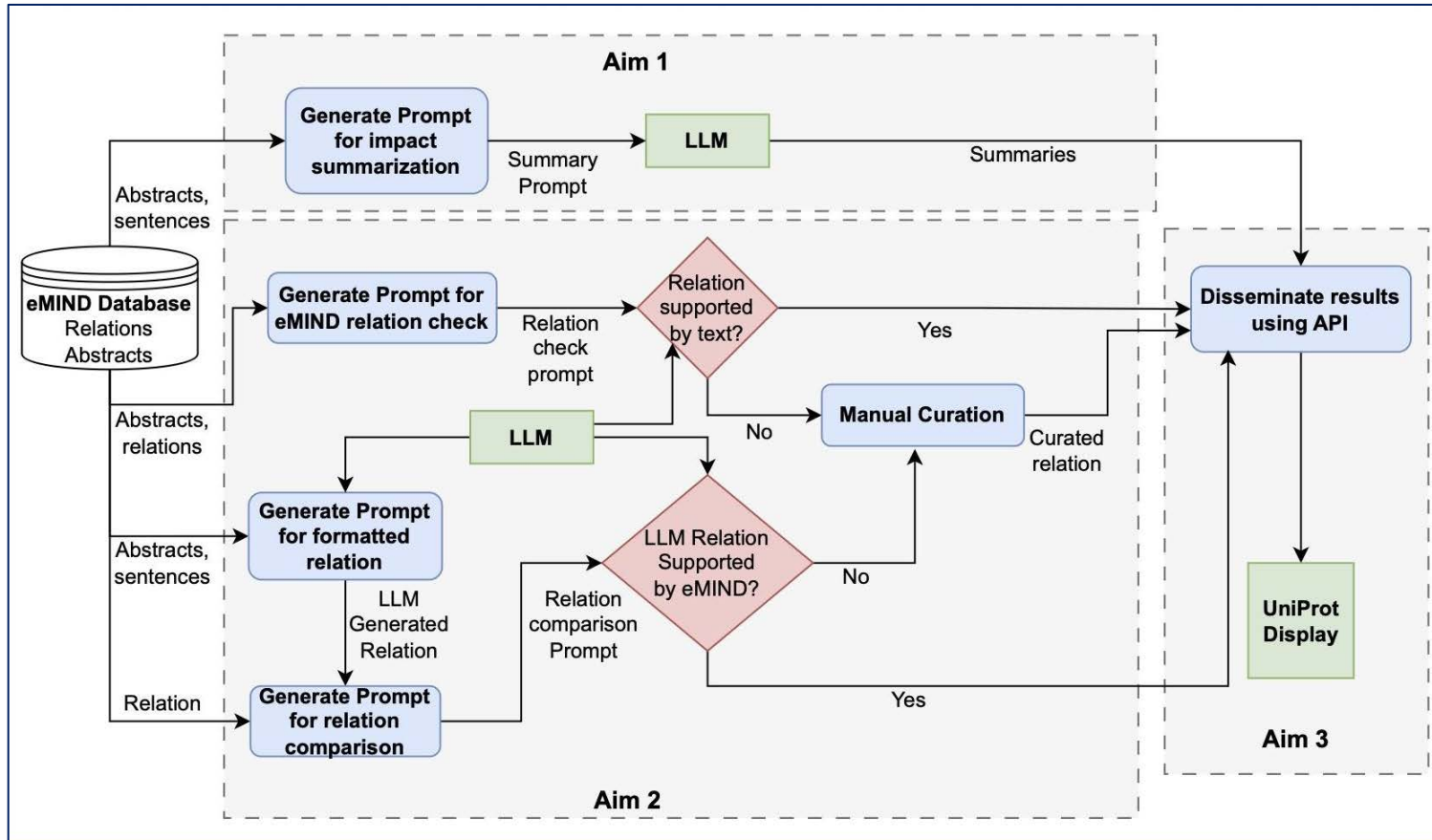
## Automatic Annotation

- To summarize literature
- To “extract” relevant information to create annotation types (e.g., GO terms)

# LLM for Summarization and Term Extraction



# Text Mining & LLMs for UniProt Annotation



- **eMIND text mining** pipeline to process all PubMed abstracts and extracts information on functional consequences of variants in Alzheimer’s Disease (AD)
- Combine eMIND output with ChatGTP to **write summaries** and **extract relations** about the impact of AD-associated variants

doi: 10.1101/2023.09.07.556602  
 Collaboration with Vijay Shanker  
 University of Delaware

# Display of Results – Confidence Scoring

## Disease & Variants<sup>1</sup>

### Involvement in disease<sup>1</sup>

**Alzheimer disease 1 (AD1)**

30 Publications

**Note** | The disease is caused by variants affecting the gene represented in this entry

**Description** | A form of Alzheimer disease, a neurodegenerative disorder characterized by progressive dementia, loss of cognitive abilities, and deposition of fibrillar amyloid proteins as intraneuronal neurofibrillary tangles, extracellular amyloid plaques and vascular amyloid deposits. The major constituents of these plaques are neurotoxic amyloid-beta protein 40 and amyloid-beta protein 42, that are produced by the proteolysis of the transmembrane APP protein. The cytotoxic C-terminal fragments (CTFs) and the caspase-cleaved products, such as C31, are also implicated in neuronal death. It can be associated with cerebral amyloid angiopathy. Alzheimer disease can be associated with cerebral amyloid angiopathy.

**See also** | [MIM:104300](#)

UniProt Annotations

AI-powered Summaries

| MUTATION     | IMPACT   | SOURCE (PMID)                     |
|--------------|--|-----------------------------------|
| <b>E693G</b> | causes decreased Abeta42 and Abeta40 levels in plasma  | <a href="#">11528419</a>          |
|              | leads to dementia with clinical features similar to Alzheimer's disease                              | <a href="#">19329229</a>          |
|              | is sufficient to cause amyloid deposition and cognitive dysfunction                                  | <a href="#">19329229</a>          |
|              | leads to dementia with clinical features similar to Alzheimer's disease                              | <a href="#">21880397</a>          |
|              | favors proamyloidogenic APP processing by increased beta-secretase cleavage                          | <a href="#">17448150</a>          |
|              | exhibits a purely cognitive phenotype that is typical of Alzheimer's disease                         | <a href="#">28890319</a>          |
|              | decreases cell viability in human neuroblastoma cells and enhances sensitivity to toxic stress       | <a href="#">12052536</a>          |
| <b>A673T</b> | facilitates amyloid-β protofibril formation and generates clinical symptoms of Alzheimer's disease   | <a href="#">22118948</a>          |
|              | displays intracellular amyloid deposits but not plaques and has a relatively mild epilepsy phenotype | <a href="#">26825094</a>          |
|              | attenuates APP-BACE-1 interactions   | <a href="#">26642089</a>          |
|              | not relevant in Asian population   | <a href="#">23652020~24126161</a> |
|              | reduces the risk for Alzheimer's disease   | <a href="#">28003277</a>          |
|              | reduces the incidence rate of Alzheimer's disease  | <a href="#">24646423</a>          |

**SUMMARY** ✕

**LLM summary for the role of E693G mutation in APP**

- The E693G mutation in the APP gene, also known as the Arctic mutation, is associated with Alzheimer's disease and leads to enhanced formation of amyloid-beta (Aβeta) protofibrils ([11528419](#), [21880397](#)).
- Carriers of the Arctic mutation exhibit decreased levels of Aβeta42 and Aβeta40 in plasma ([11528419](#)).
- The Arctic mutation causes amyloid deposition and cognitive dysfunction, similar to Alzheimer's disease, and can be studied using the TgAPP<sup>Arctic</sup> mouse model ([19329229](#)).
- The Arctic mutation favors proamyloidogenic APP processing by increased beta-secretase cleavage, resulting in increased levels of Arctic Aβeta, particularly at intracellular locations ([17448150](#)).
- Mice with the Arctic mutation display intracellular amyloid deposits but not extracellular amyloid plaques ([26825094](#)).
- The Arctic mutation enhances sensitivity to toxic stress, contributing to the pathogenic pathway leading to Alzheimer's disease ([12052536](#)).
- The Arctic mutation leads to distinct plaques and accumulation of truncated forms of amyloid-beta ([22118948](#)).
- Tracers for amyloid-beta fibrils show reduced binding affinity toward the Arctic mutation compared to the native type ([28898051](#)).

In conclusion, the E693G mutation in the APP gene, known as the Arctic mutation, plays a significant role in the pathology and neurodegeneration associated with Alzheimer's disease. It leads to enhanced formation of amyloid-beta protofibrils, amyloid deposition, cognitive dysfunction, and increased sensitivity to toxic stress. The mutation also affects APP processing and the accumulation of truncated forms of amyloid-beta.

*AI summary powered by OpenAI*

**Publications**

PMID: [28898051](#)  
Effect of Alzheimer Familial Chromosomal Mutations on the Amyloid Fibril Interaction with Different PET Tracers: Insight from Molecular Modeling Studies  
Balamurugan K, Murugan NA, Långström B, Nordberg A, Ågren H.ACS Chem Neurosci. 2017 Dec 20 8(12):2655-2666.  
doi: 10.1021/acscchemneuro.7b00215. Epub 2017 Oct 3.

PMID: [26825094](#)  
Increased Epileptiform EEG Activity and Decreased Seizure Threshold in

Assign confidence score to output of text mining tools

Using eMIND as a test tool to:

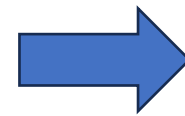
- Collect eMIND positive abstracts
- Ask LLM to answer what impact of mutation is
- Check with overlap with eMIND output

Scalable framework for other text mining/relation extraction tools

# UniProt Vision for AI

- Transformative Impact of AI: Enable the user community to harness AI using UniProt data
- AI approaches are being applied to many aspects of UniProt: Close collaboration with the AI research communities to innovate new approaches and solutions
- Scaling up protein functional annotation: AI-assisted literature information extraction and automated functional annotation
- Organizing and sustaining the growing sequence space: AI-enabled sequence clustering and similarity search

Lack of **well-curated** data  
is a major barrier for AI Research



UniProt **well-curated** and  
**AI/ML-ready** data