TOPH (True Retrieval Of Proteins Homologs): Adapting A Contrastive Question-Answering Framework for Protein Search

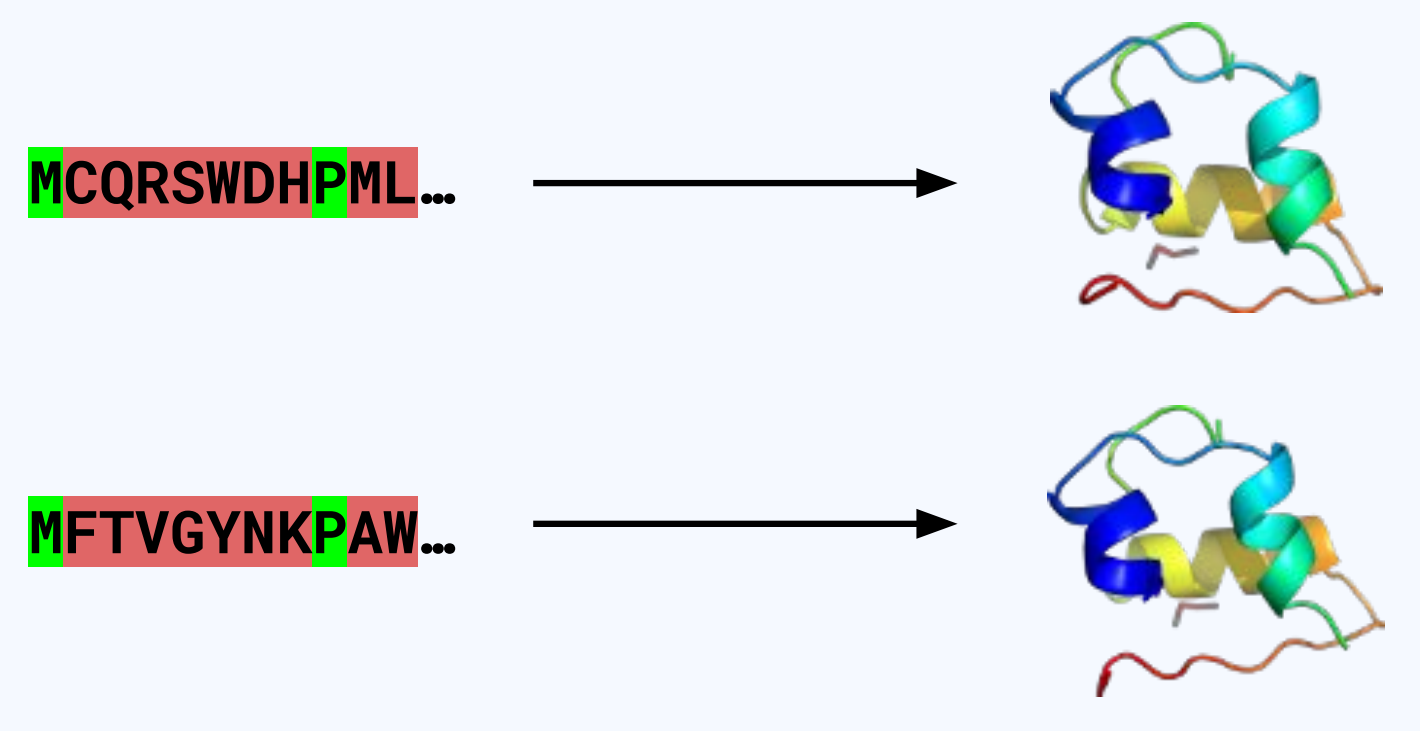
Ron Boger*[1], Amy Lu*[2], Seyone Chithrananda*[1], Kevin Yang[2], Petr Skopintsev[1], Ben Adler[1], Eric Wallace[2], Peter Yoon[1], Pieter Abbeel[2], Jennifer Doudna[1]

[1] UC Berkeley, Innovative Genomics Institute
[2] UC Berkeley, Berkeley Artificial Intelligence Research

Tl;dr: We present a **protein semantic similarity search** method for **RNA-Guided endonuclease discovery**, inspired by dense retrieval methods in open-domain question answering, and augmented by domain-specific hard negatives during training.

Motivation: Discovering New Biology Through Search

- Identification of **protein homology** (proteins which share evolutionary ancestry) is a critical tool for discovery in biology
  - E.g. metagenomic mining for **CRISPR-Cas enzymes** to harness sequences created through natural evolution for gene editing
- Homology detection provide insights into **structure and function**, but is challenging for **remote homology detection**
  - Traditional bioinformatics methods such as BLAST and HMMER relies on sequence match, which may neglect evolutionarily related sequences of bioengineering relevance, but has low sequence similarity to query
- **Structural searches** (DALI, TM-align) confer higher sensitivity, but at **infeasible speeds** for large protein datasets (1+ mo for all v all protein search)
- Searching for **semantically similar words** with **low sequence similarity** in a large natural language dataset offers an analogous challenge
- Can we adopt similar embedding-based and contrastively trained methods to find remote homologs with similar functional & structural semantics?



Proteins of identical function and structure can have little to no sequence similarity!

Dense Passage Retrieval (DPR)

- Adapts Dense Passage Retrieval (DPR), a method from open-domain question answering, to improve protein homology search.
  - Contrastively trained to distinguish a "correct pair" amongst other "incorrect pairs"
- Using a dual encoder architecture with ESM2 (Lin et al.) as the embedding method, finetunes final layers **using full proteins as the 'questions' and 'passages'** in the DPR framework.
  - Model must capture features relevant for semantic similarity, rather than sequence-level matches in traditional methods.
- Employs **hard negative sampling** and **in-batch negative sampling from misclassified proteins** during training
  - Adds domain-relevant inductive biases through data curation
- At inference, **retrieves the top k closest embeddings** to the query as the homologs.

$$sim(q,p) = E_Q^T(q)E_P(p)$$

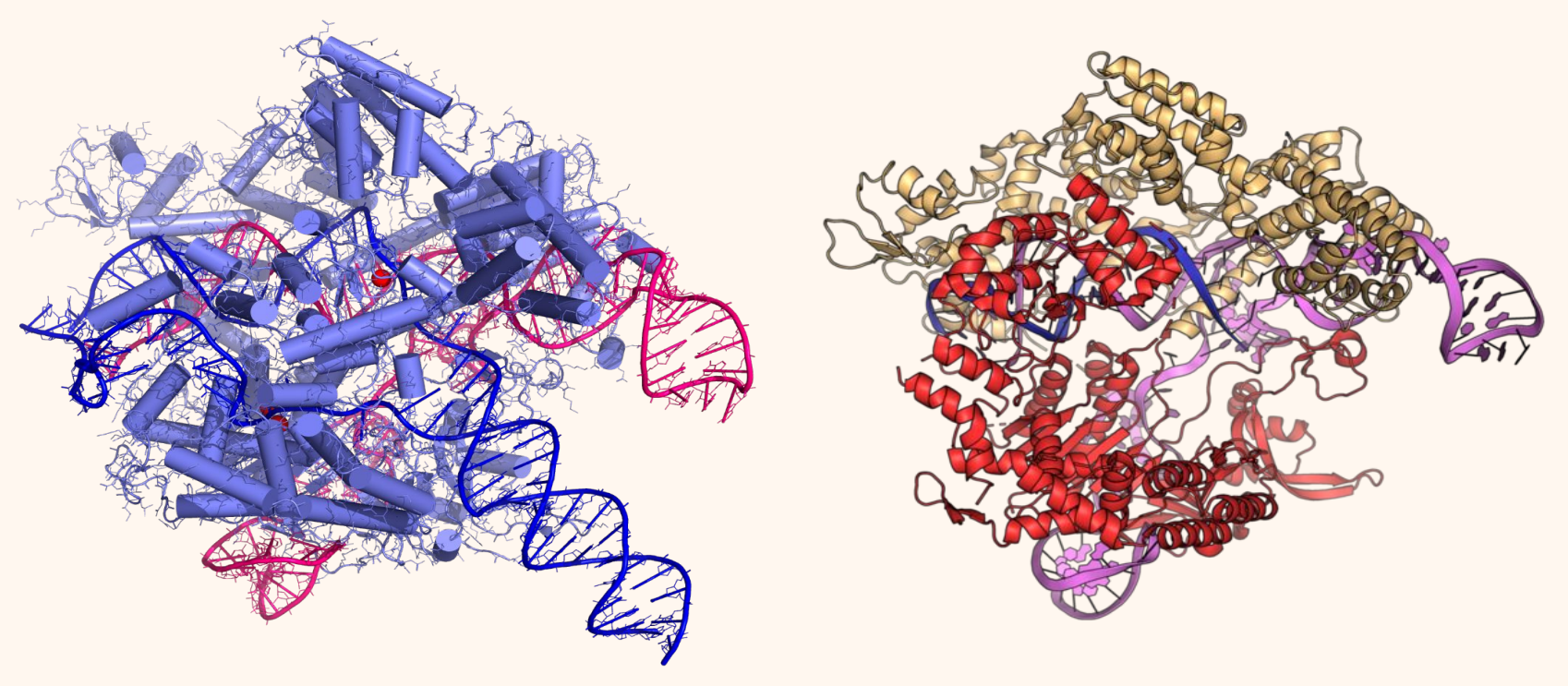
A biencoder model with dot-product similarity is fine-tuned on homologous protein sequences

$$L(q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i,p_i^+)}}{e^{\mathrm{sim}(q_i,p_i^+)} + \sum_j e^{\mathrm{sim}(q_i,p_{i,j}^-)}}$$

The model is trained using a contrastive objective function that maximizes the similarity between positive protein pairs while minimizing their similarity to negative examples.

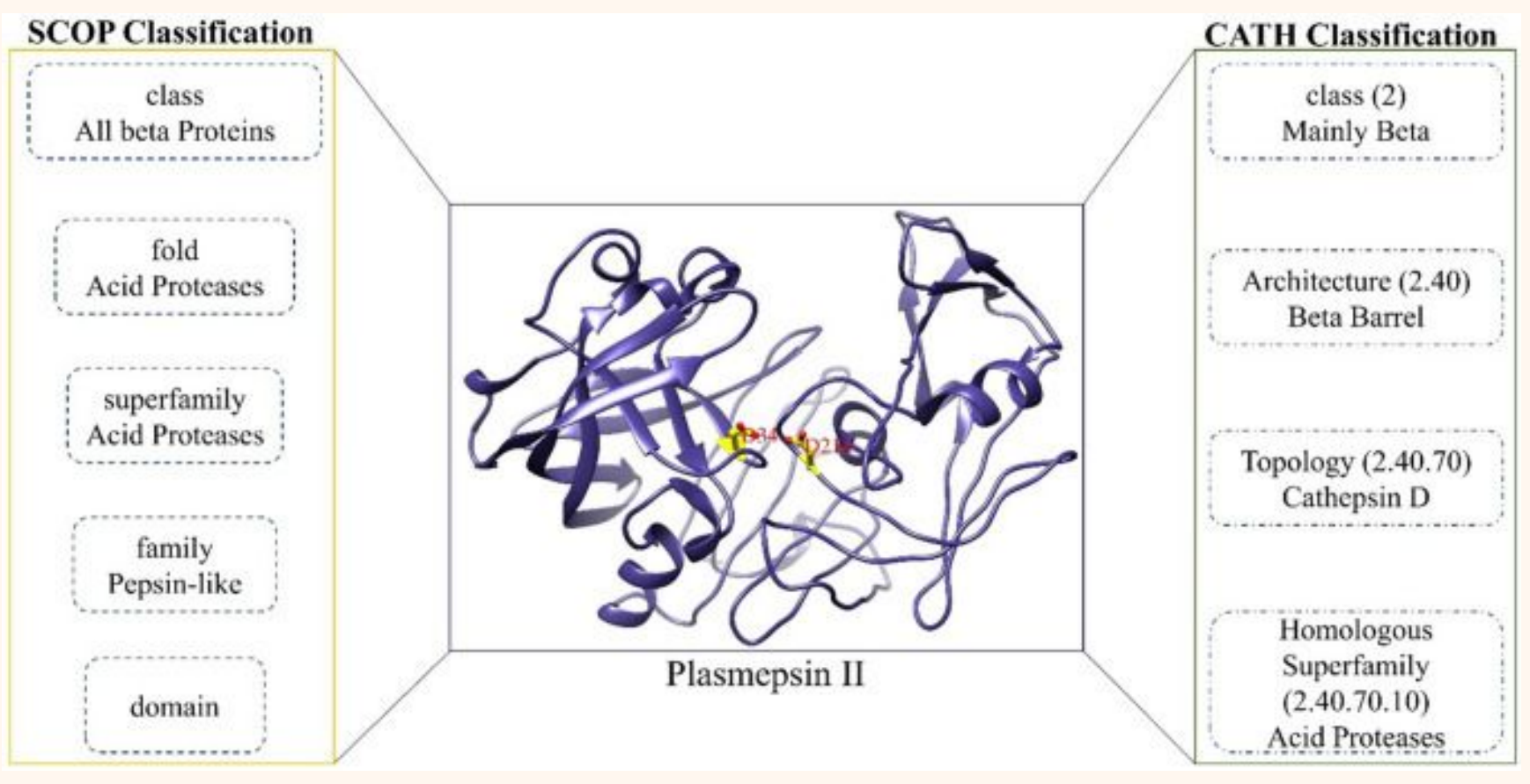
RNA-Guided Endonucleases are Remote Homologs

- We utilize a diverse CRISPR-Cas and **evolutionary related nucleases protein dataset** for remote homology detection, a key component of **bacterial defense against foreign genetic elements**.
- We introduce **2 datasets**, drawn from multiple sources and hand-curation from structural biologists, offers **verifiable remote homologs** due to the unique positioning of Cas genes upstream of CRISPR loci.
- RNA-Guided Endonucleases, such as **CRISPR-Cas9**, display incredible **diversity in structure and sequence** and may be a valuable testbed**.**
- Evidence suggests **limitations of existing models in detecting Cas proteins**, highlighting the need for improved methods.



Model Training

- Trained on **Astral Structural Classification of Proteins 2.08 (SCOPe)** clustered at **40% sequence similarity**
  - Dataset has intrinsic hierarchical structure:
    - Family: significant sequence identity
    - Superfamily: different families with structural and functional similarities
    - Fold: different superfamilies with the same topological arrangement of major secondary structures
    - Class: secondary structure composition
- **15,177 domains** in the training set across **4693 families**.
- For evaluation, we use a test set of **400 domains**, ensured to have less than 30% sequence identity to the training set proteins.
- **Two models** were trained: one fine-tuning **`esm2_t6_8M_UR50D`** and the other **`esm2_t33_650M_UR50D`** as the question and passage encoders.
- Trained on a **single NVIDIA A100 GPU**



The model is trained using a contrastive objective function that maximizes the similarity between positive protein pairs while minimizing their similarity to negative examples.

Results

	Family	Superfamily	Fold
ESM2 (8M)	0.412	0.265	0.010
ESM2 (650M)	0.314	0.134	0.010
ESM2 (3B)	0.477	0.221	0.014
MMseqs2	*0.433*	*0.165*	*0.001*
TM-Vec	0.848	0.596	0.121
TM-Align (avg)	0.868	0.619	0.163
DALI	***0.885***	***0.709***	***0.168***
Foldseek	*0.821*	*0.578*	*0.070*
Progres	*0.878*	*0.680*	*0.144*
TOPH (ESM-650M)	0.818	0.528	0.065
TOPH (ESM-8M)	0.571	0.392	0.0376

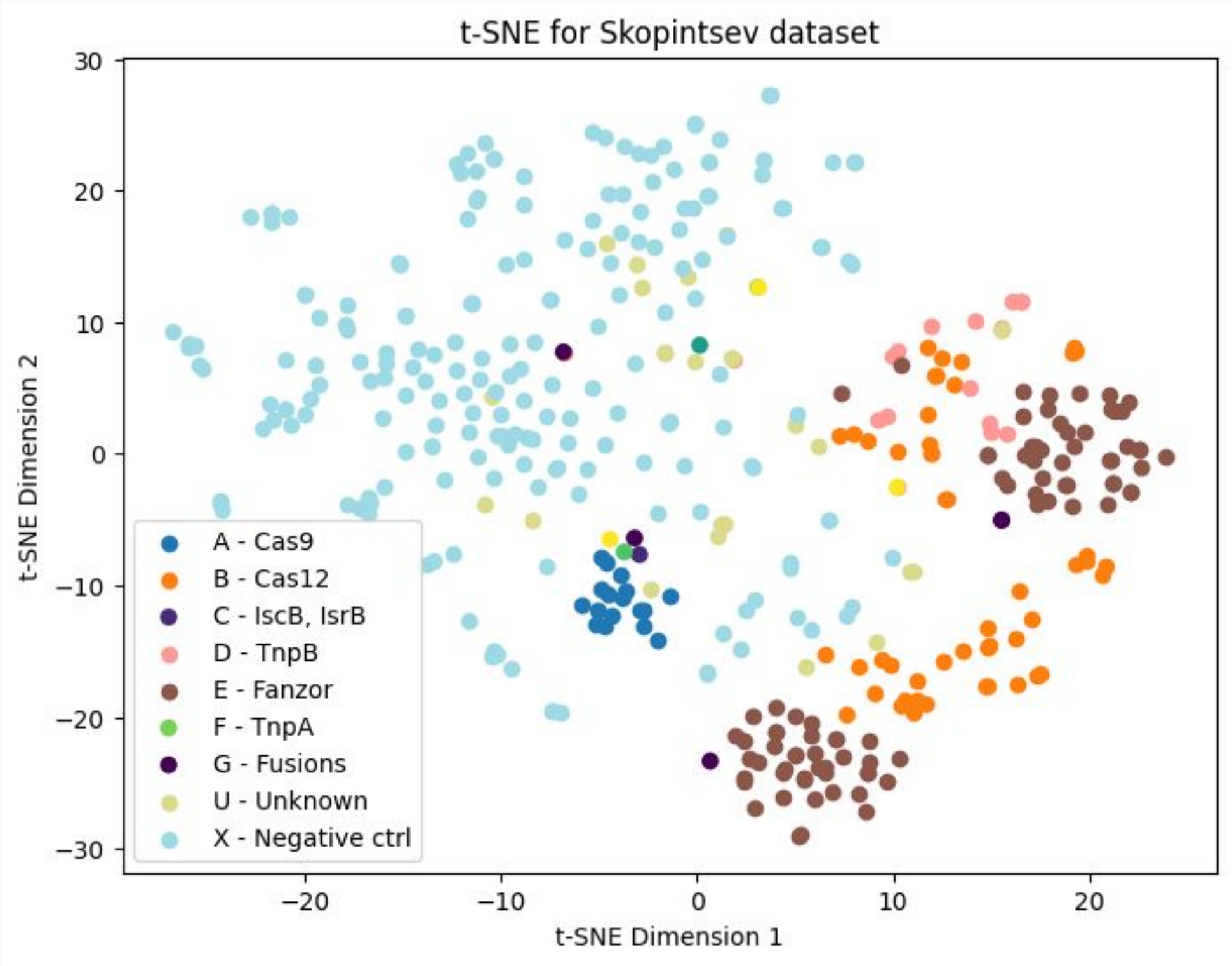
Results

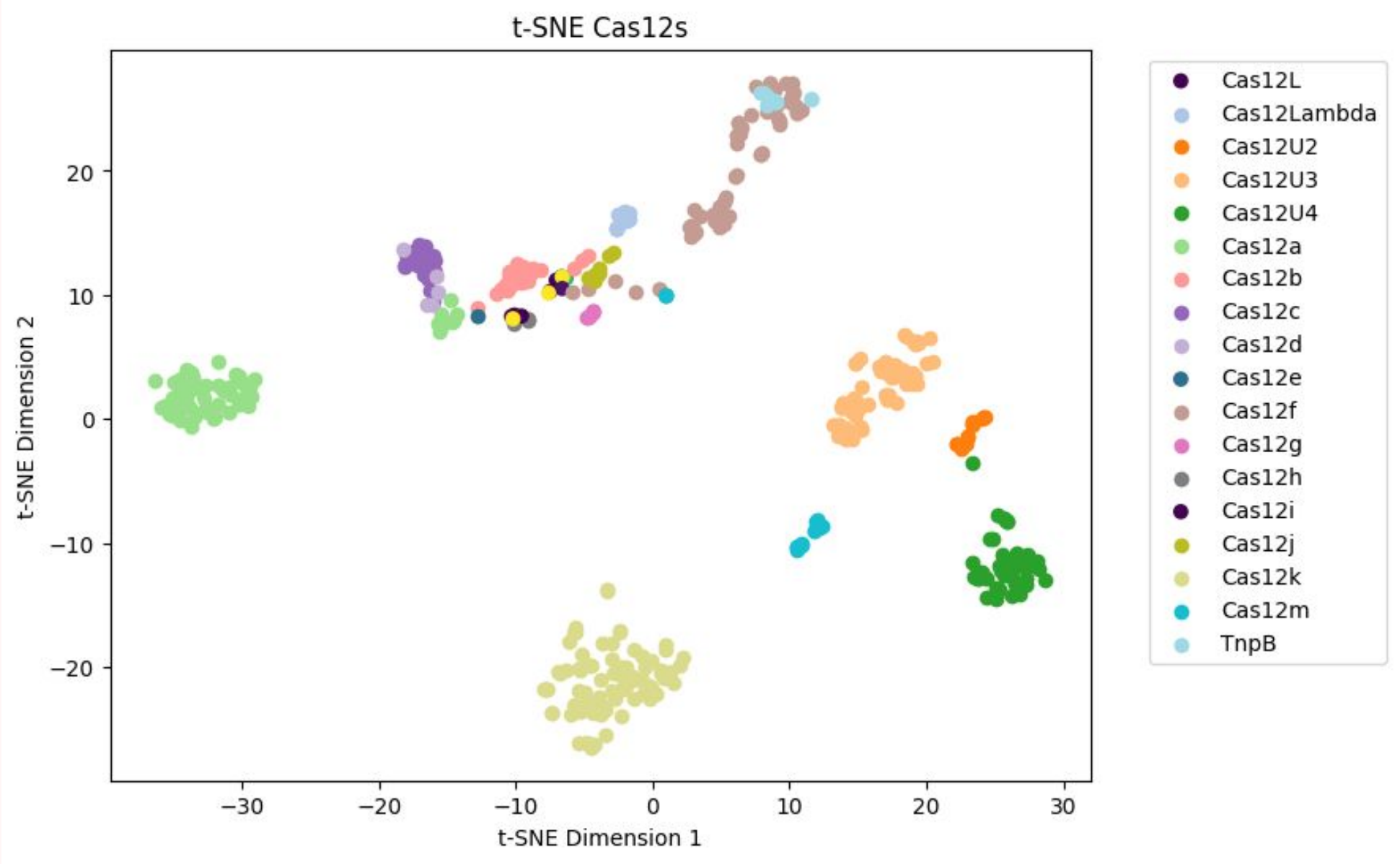
SCOPe2.08 Evaluation

- Sensitivity was measured as the **fraction of true positives (TPs) until the first incorrect fold**.
- Results were **comparable to structural methods**, but **without processing or folding.**
- Despite no hyperparameter tuning or training on multiple GPUs, **TOPH outperformed all classical sequence models and ESM models** that were not fine-tuned on the family detection task.

Cas enzyme Identification

- **Cas12 Differentiation**: Our model successfully distinguishes between different Cas12 subtypes and ancestors, with uncharacterized proteins Cas12U2, Cas12U3, and Cas12U4 emerging as distinct, hinting at unique biological roles.
- **Skopintsev Dataset**: Our model differentiated between Cas9, Cas12, and their ancestors, revealing more diversity within the Cas12 group.





Future Directions

- Enable sequence-structure search by employing a structure encoder for query sequences
- **Curriculum learning** (i.e. increasing difficulty via data curation) on family, superfamily, fold
- Improve bioinformatic usability for large-scale databases:
  - Incorporate **"reader" of protein domains**, following theme of retriever+reader in DPR
  - Incorporate high-capacity vector-based similarity search infrastructure (e.g. FAISS)
- Incorporate retrieval-augmented generation

References

1. Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. 2020
2. Evolutionary-scale prediction of atomic-level protein structure with a language model. Lin et al. 2023
3. Avatar: The Last Airbender. Michael Dante DiMartino and Bryan Konietzko 2005