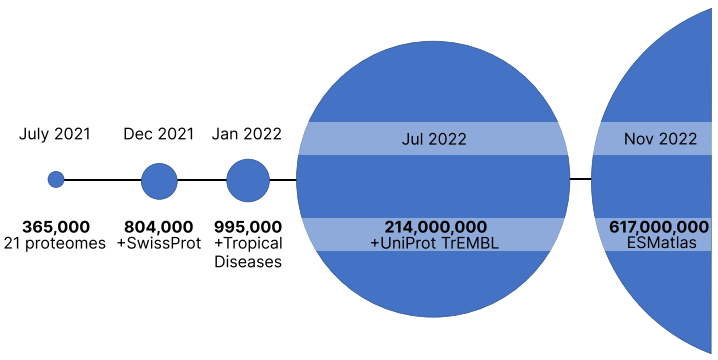
Metagenome-scale structural homology detection with Foldseek-ProstT5

Milot Mirdita[1], Victor Mihaila[1,2], George Bouras[3], Michael Heinzinger[4], Martin Steinegger[1,2,5,6]

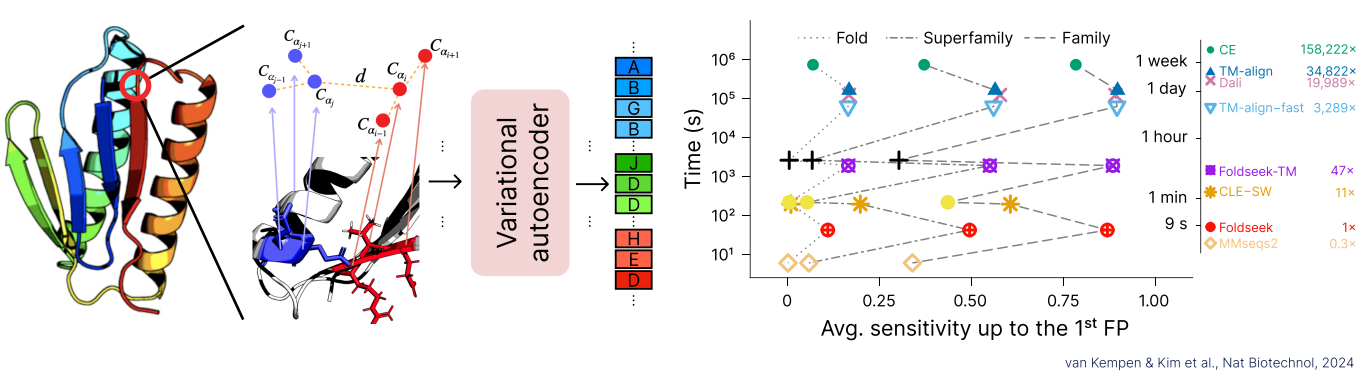
[1]School of Biological Sciences & [2]Interdisciplinary Program in Bioinformatics & [5]Institute of Molecular Biology and Genetics & [6]Artificial Intelligence Institute, Seoul National University, Korea. [3]Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Australia. [4]School of Computation, Information, and Technology, Department of Informatics, Bioinformatics & Computational Biology, Technical University of Munich, Germany.

Abstract Foldseek enables fast and sensitive homology detection of protein structures. Together with predicted structures from ColabFold—a ~100x accelerated version of AlphaFold2—Foldseek facilitated sensitive annotation of dark parts of a sponge proteome (Ruperti & Papadopoulos, et al. Genome Biology, 2023), identifying an additional 50% beyond sequence-based methods. However, structure prediction of entire metagenomes remains cost-prohibitive. Here, we present **Foldseek-ProstT5**, an extension to Foldseek to enable sequence-based structural metagenomics and sensitive annotation of previously dark proteins. Utilizing the ProstT5 protein language model, we replace costly structure prediction with >3500x accelerated translation of amino-acid sequences directly to structural interaction (3Di) tokens. On the Foldseek sensitivity benchmark, ProstT5's 3Di sequences improve sensitivity for Fold, Superfamily, and Family recognition by 4.3%, 12.8%, and 23.1% respectively without backbone (Cα) coordinates, and by 3.4%, 10.5%, and 18.2% respectively if Cα backbone coordinates are available.

An avalanche of predicted protein structures

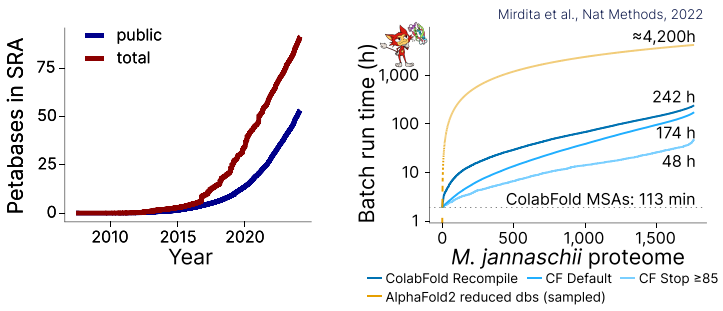


Foldseek is a fast and sensitive aligner that discritizes protein structures to 3Di seq

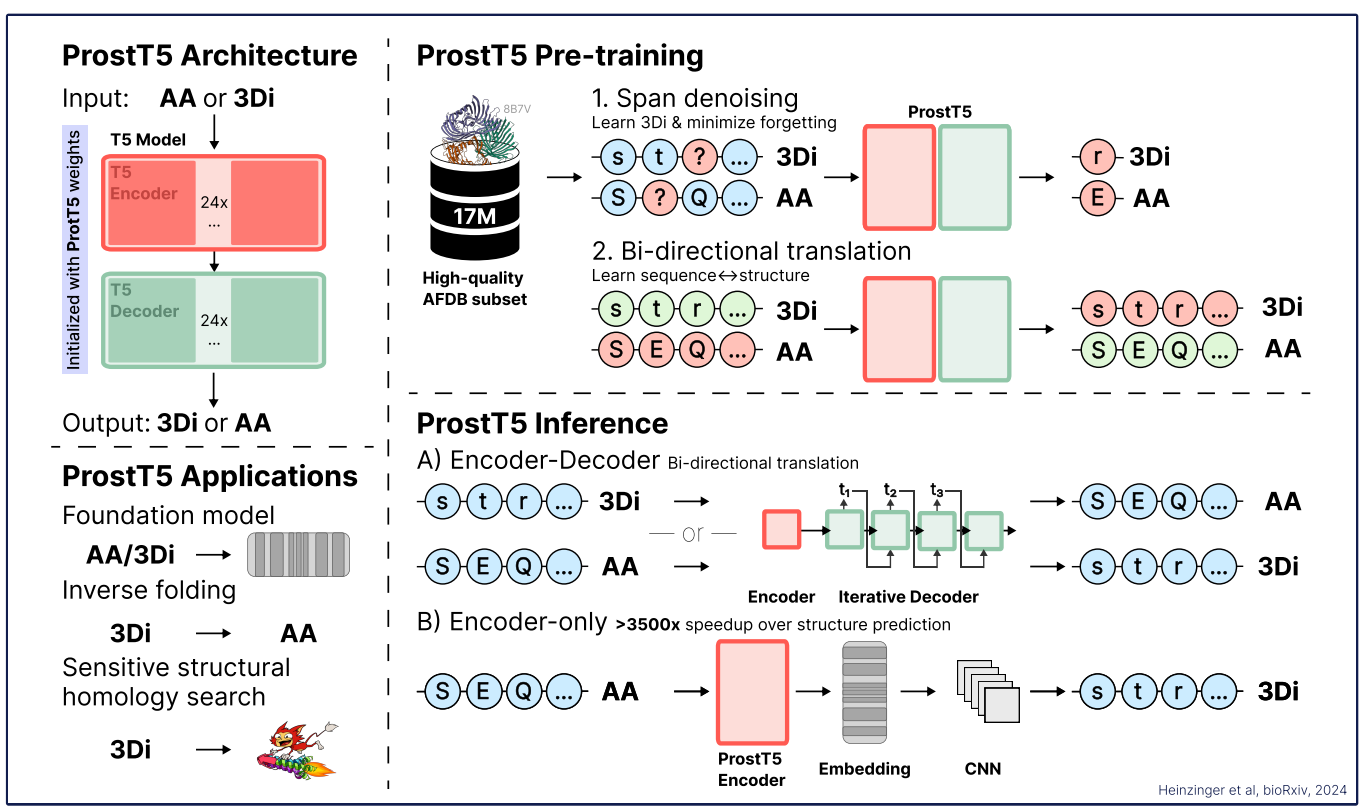


van Kempen & Kim et al., Nat Biotechnol, 2024

Expensive metagenomic structure prediction

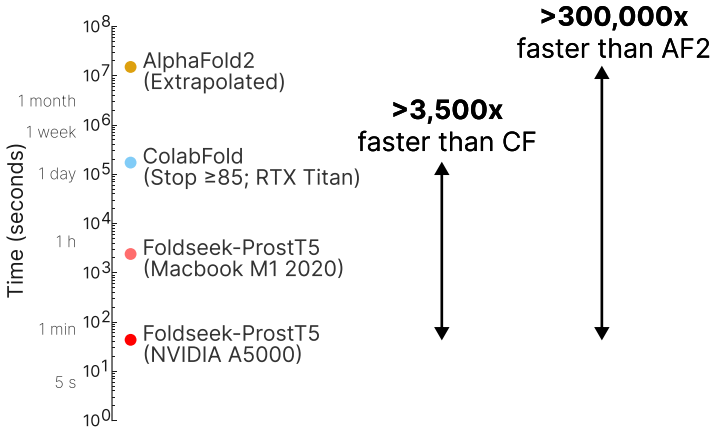


ProstT5 is a protein language model train bilingual on sequence and structure

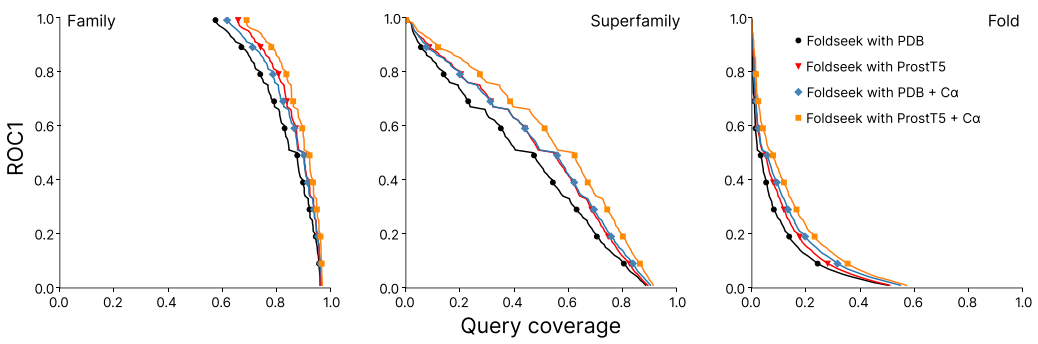


Heinzinger et al, bioRxiv, 2024

Accelerated 3Di with FS-ProstT5



High sensitivity with FS-ProstT5 3Di and exceeding FS's with Cα



Outlook

Model distillation and optimization
We aim to achieve +10-100x speed-up. Particularly for CPU-only
Fine tuning on out-of-distribution viral and metagenomic proteins
Easy installation: No Python or CUDA libraries
Integration into our whole suite of methods
ColabFold, FoldMason (Gilchrist et al, bioRxiv, 2024), Spacedust (Zhang et al., bioRxiv, 2024), Pharroka (Bouras et al., Bioinformatics, 2023), ...

Availability Foldseek-ProstT5 is free and open source software available for Linux and macOS at **foldseek.com** and as a webserver at **search.foldseek.com**.

```
conda install -c conda-forge -c bioconda foldseek
```

RECOMB Seoul | 서울 | 2025
Visit us at RECOMB 2025 in Seoul

Poster **Server** **GitHub**

Acknowledgments

NRF