WORKSHOP

ML For Biotech And Pharmas

Etienne Goffinet, PhD
Senior Researcher
Technology Innovation Institute

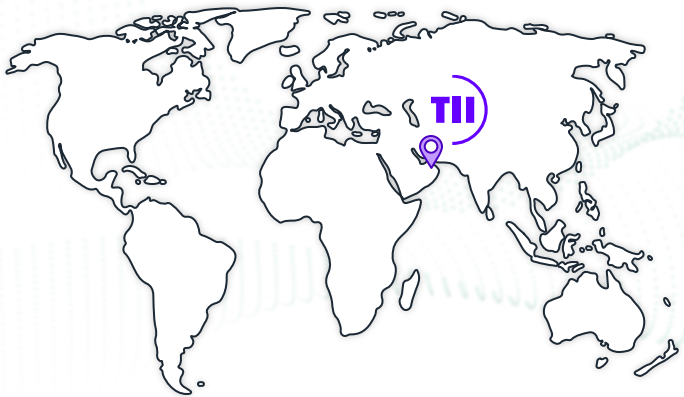
Introduction to Protein Language Models for Synthetic Biology

ODSC EAST 2024

BOSTON
APRIL 23–25

THE LEADING
AI TRAINING CONFERENCE

Technology Innovation Institute



Biotechnology Research Center

Bio-medicine

Bio-robotics and nano

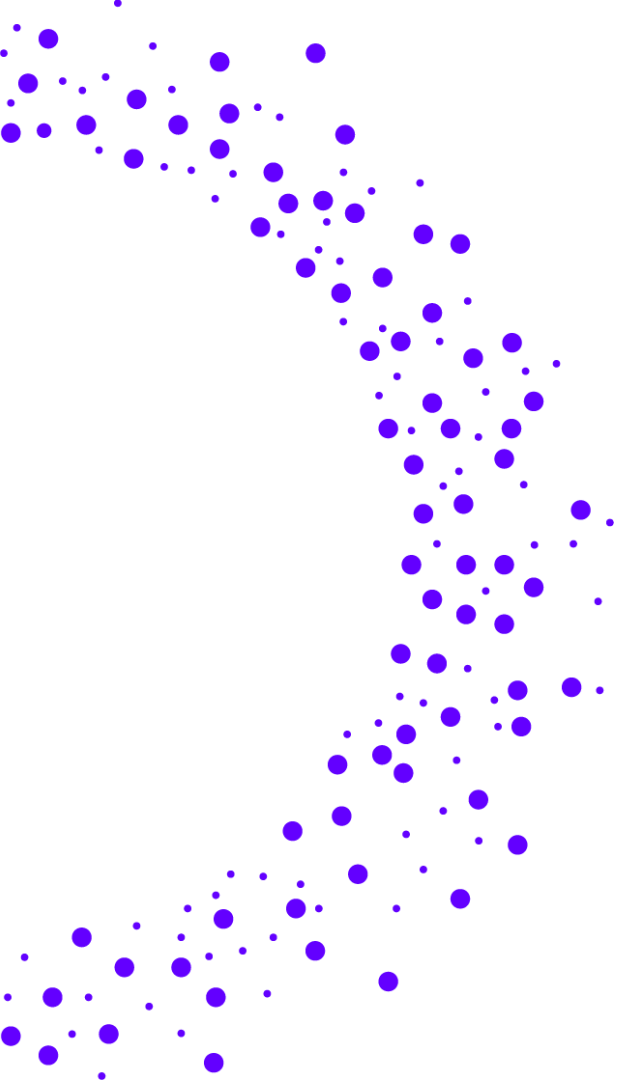
Environmental biotech

Molecular biotech and genomics

Bio-informatics

Requirements

1. Functional computer with internet access
2. Google Colab account

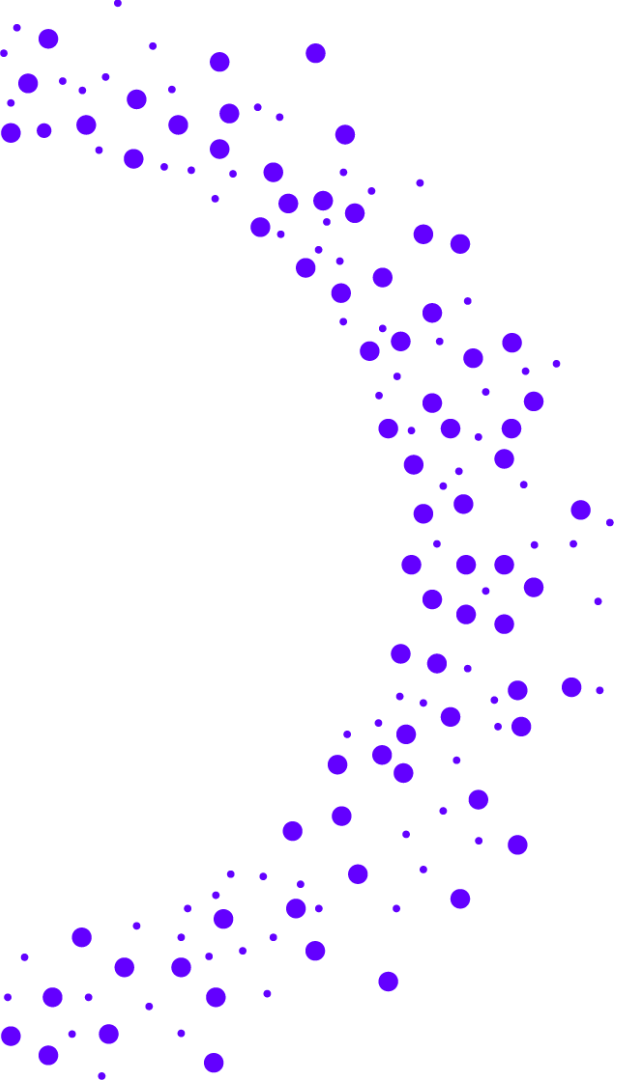


Agenda

1. Protein as a Language (20 min)
    1. From LLM to PLM
    2. Overview of the PLM ecosystem

2. Application to Protein Function Prediction (30 min)
    1. Task and Dataset
    2. Method and Implementation -> Notebook

Q&A (10 min)



Protein as a Language

Protein as a Language

Proteins: roles, representation

They are Everywhere

Immune system (Antibodies)

Digestion (Catalytic enzymes)

Growth (Hormones)

Blood (Hemoglobin, Myoglobin)

Bones, tendons, cartilage (Collagen)

Skin, hair, nails (Keratin)

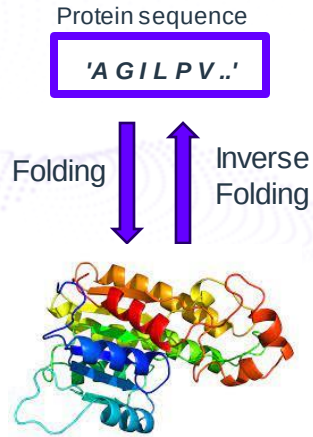
Source of energy, body repair & healing, ...

3D Structure --> Binding capacity -> Function & properties

Primary structure : sequence of amino-acids

Length between 12 and 2000

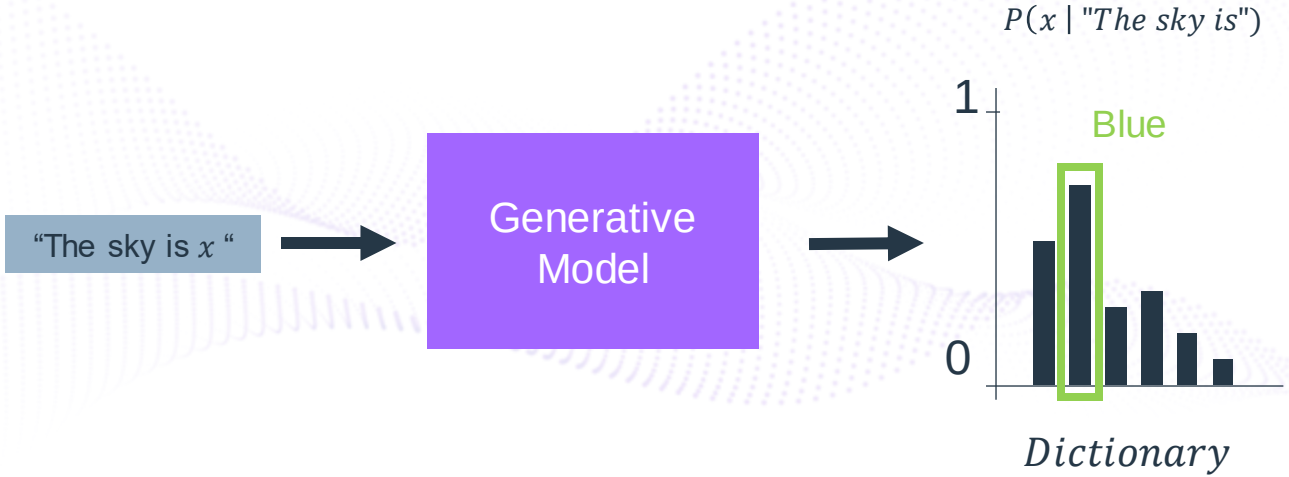
from an alphabet of size 20



Tertiary structure: 3D structure

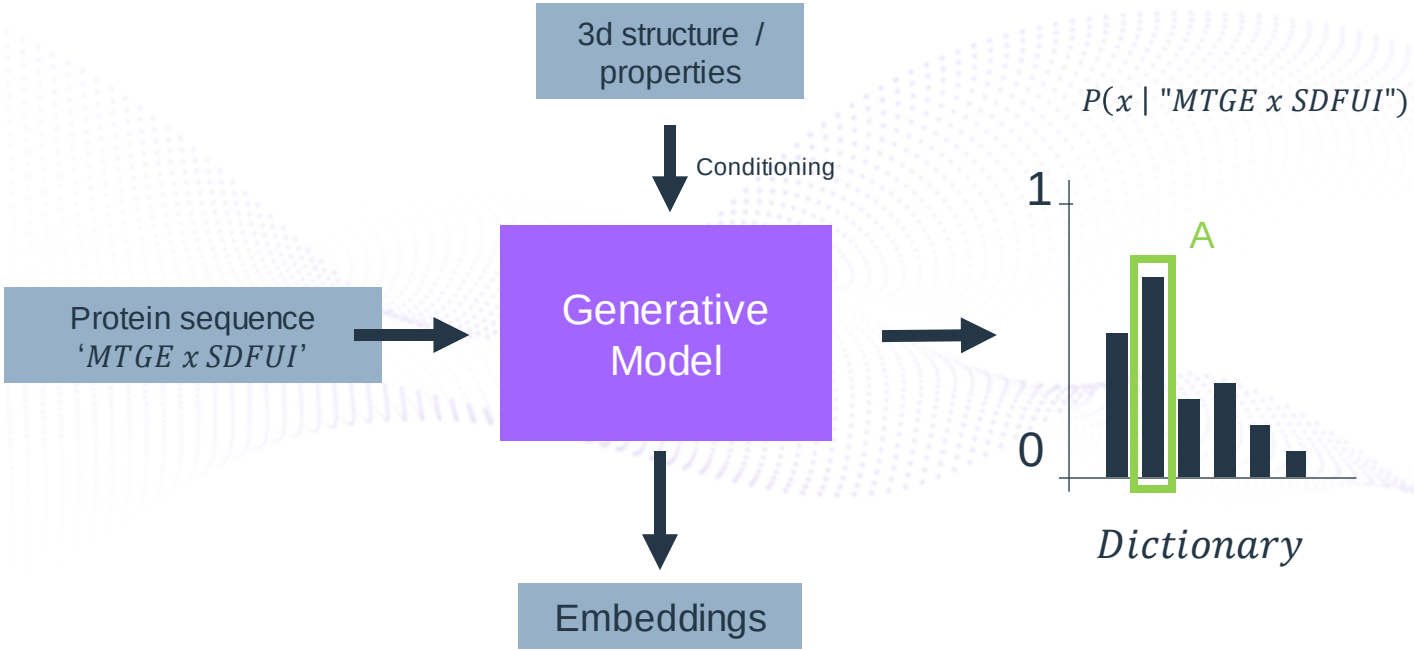
Protein as a Language

From LLM to PLM



Protein as a Language

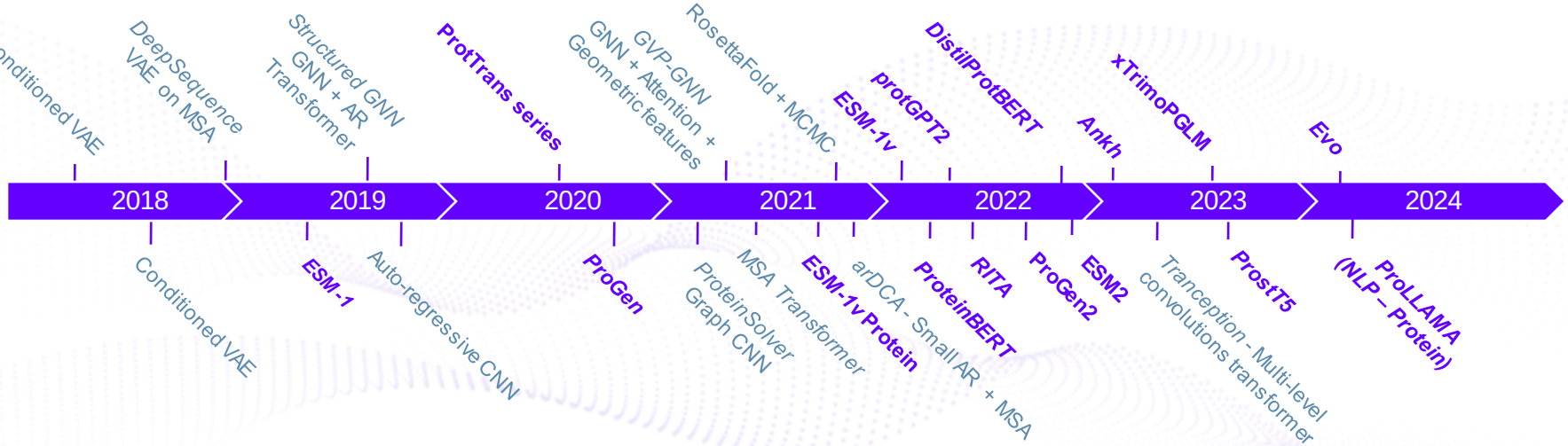
From LLM to PLM



Exploration / Annotation / Property estimation /
Interaction prediction / Conditioned generation

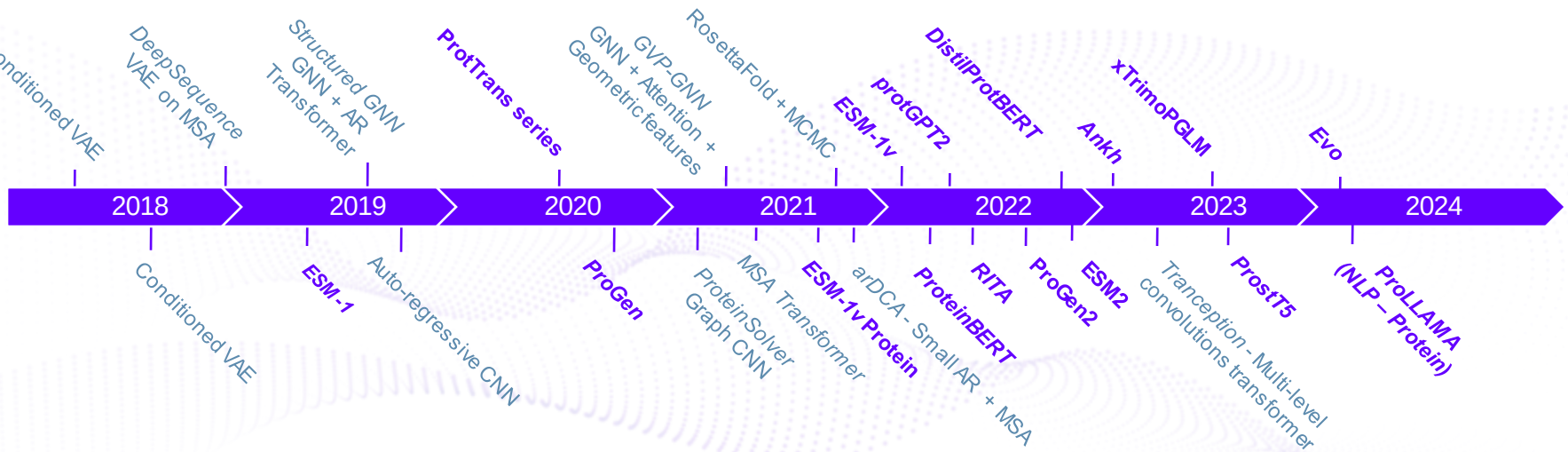
Protein as a Language

Overview of the PLM ecosystem



Protein as a Language

Overview of the PLM ecosystem



Powered by:

UniProt

- Uniref100: ~390M sequences
- Uniref90: ~180M sequences
- Uniref50: ~63M sequences

MC BFD

BFD: 2.1 Billions sequences from Uniprot/TrEMBL+Swissprot, Metaclust and Soil Reference Catalog

EMBL-EBI MGnify

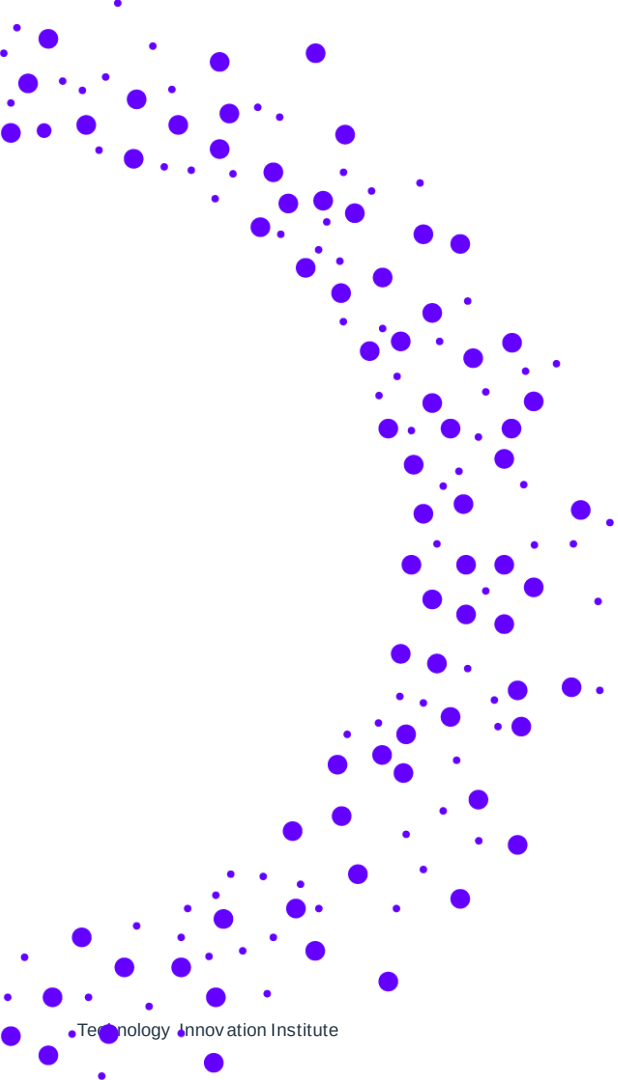
2.4 billion non-redundant sequences predicted from metagenomic assemblies

Protein as a Language

Tasks and evaluation

Category	Task	Task	Source	Data size
Structure Prediction	Contact	Classification (AA-wise)	trRosetta	15k samples
	Folding	Classification	Scope 2009	15k
	Secondary Structure	Classification	CASP12 and CASP14	11k
Properties	Solubility	Classification (binary)	DeepSol	70k
	Stability	Regression	Rocklin et al (2017)	68k
	Optimal Temperature	Regression	DeepET	1.8k
	Temperature Stability	Classification (binary)	TemStaPro	410k
Interaction	Metal ion Binding	Classification (binary)	Cheng et al (2023) from PDB	7.3k
	Enzyme Catalytic Efficiency	Regression	EcGEMs (Li et al 2022)	17k
	Peptide-HLA Affinity	Classification	Ccbhla	900k
	TCR-pMHC Affinity	Classification	EpiTCR (from VDJdb)	24k
Function	Antibiotic Resistance	Classification (MultiLabel)	CARD	3.3k
	Fluorescence	Regression	TAPE	54k
	Fitness	Regression	FLIP	8.5k
	Localization	Classification	DeepLoc (from UniProt)	8.4k

..And many others, including TAPE, PEER, ProteinGlue, ProteinGym, FLIP benchmarks..



Application to Protein Function Prediction

Application to Protein Function Prediction

Task Description: Protein Molecular Function Prediction [1]

Protein Molecular Function (MF)

'The enduring potential of a [protein] to perform actions, such as catalysis or binding, on the molecular level of granularity' [2]

Task Type

Proteins can have several functions

↓

Multi-Label Classification

Dataset [3]

Training / Validation / Test sizes
1k / 0.5k / 0.5k

677 possible functions

Strong **imbalance** of occurrence

Dependencies between functions

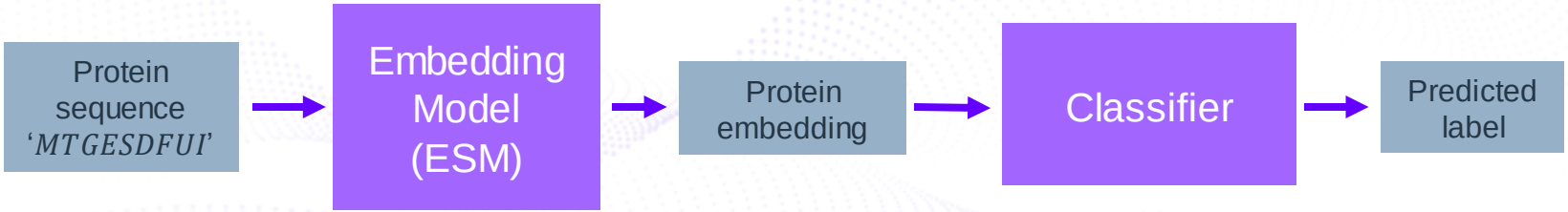
[1] Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., ... & Salakoski, T. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome biology, 20, 1-23.

[2] Hill, D. P., Smith, B., McAndrews-Hill, M. S., & Blake, J. A. (2008). Gene Ontology annotations: what they mean and where they come from. BMC bioinformatics

[3] Oliveira, G. B., Pedrini, H., & Dias, Z. (2023). TEMPROT: protein function annotation using transformers embeddings and homology search. BMC bioinformatics

Application to Protein Function Prediction

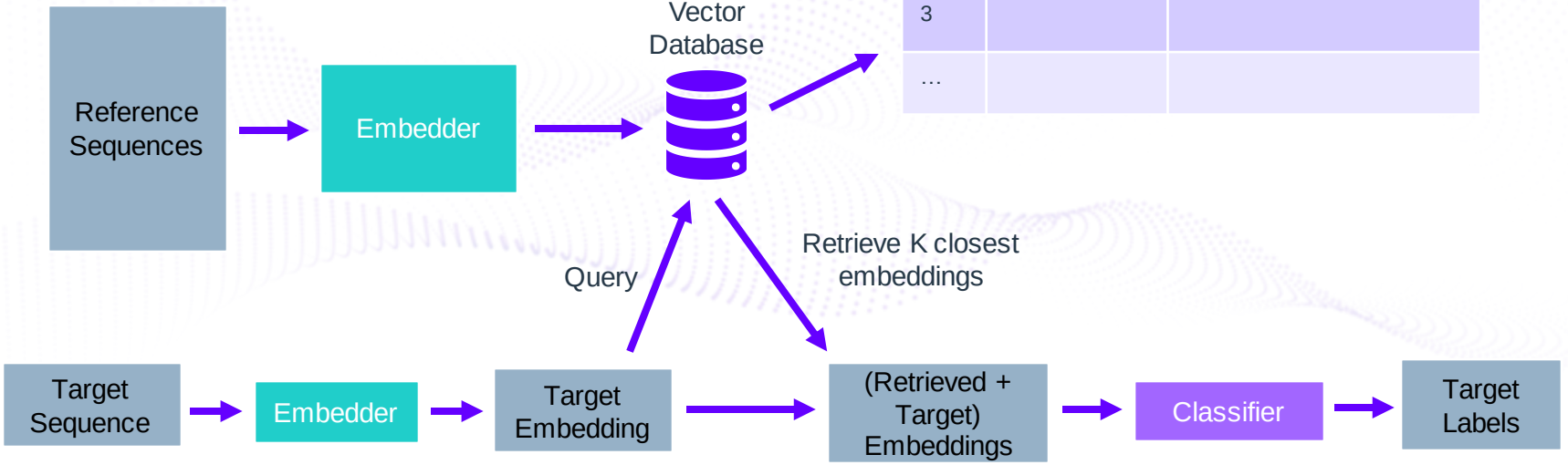
Approach 1: Finetuning



Application to Protein Function Prediction

Approach 2: Retrieval-Augmented Classification

Id	Embeddings	Labels
1	[0.12, -.234, ..]	[‘antifungi’, ‘antimicrobial’, ..]
2	…	…
3		
…		



Protein as a Language

Notebook

Q&A