

Protein Language Modeling

Applied Deep Learning in Bioinformatics

Summer School

Alberto Santos and Angel L. Phanthanourak –
Multi-omics Network Analytics Group

13.08.2024



Outline

- **Introduction**
- **Basic Concepts**
 - Language Models (LM)
 - Proteins (p)
 - Protein Language Models — pLMs
- **Applications of pMLs**
 - Protein Structure
 - Protein-protein Interactions
 - Protein Engineering and Design
- **Fine-tuning**
- **Data Resources**
- **Limitations**
- **Practical**

Introduction



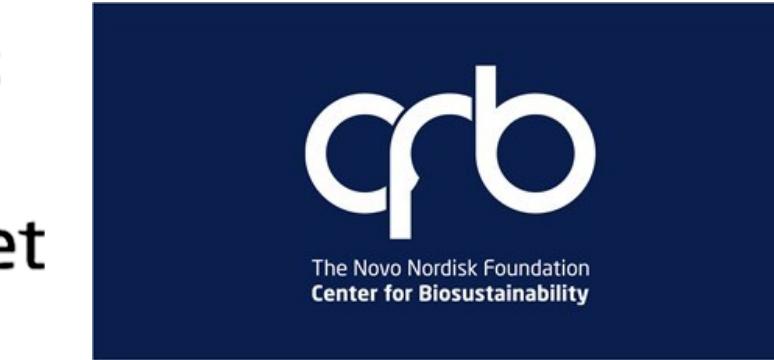
Mision

Getting a holistic view of **biological systems and their context** to understand their complexity and provide new insights and applications that can **benefit human and environmental health**

Located at:



Danmarks
Tekniske
Universitet



Multi-omics

Using **multimodal data** to have a comprehensive view on **(micro) biology problems**

Network

Exploiting **graphs** to **structure, represent, integrate and analyse** data

Analytics

Applying **Machine Learning** to answer **complex biological questions**

Data Science

Me
WANTS
DATA



Data

Internal — Biofoundry, platforms, iModulon, images, etc.

External — databases such as MicroPheno, MiMeDB, Exposome, MGnify, FooDB, GNPS, etc.

Meta-omics (metagenomics, metaproteomics)

Proteomics and metabolomics

Protein-protein Interactions (PPI)

Technology

Knowledge Graphs

Data integration

Natural Language Processing

Graph Machine Learning

Databases

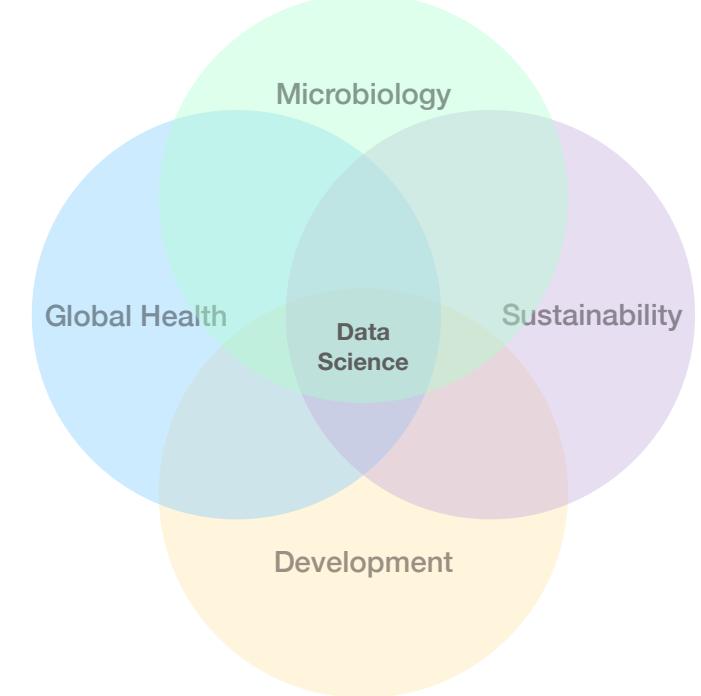
Education

Data Science democratisation

Open science

FAIR data

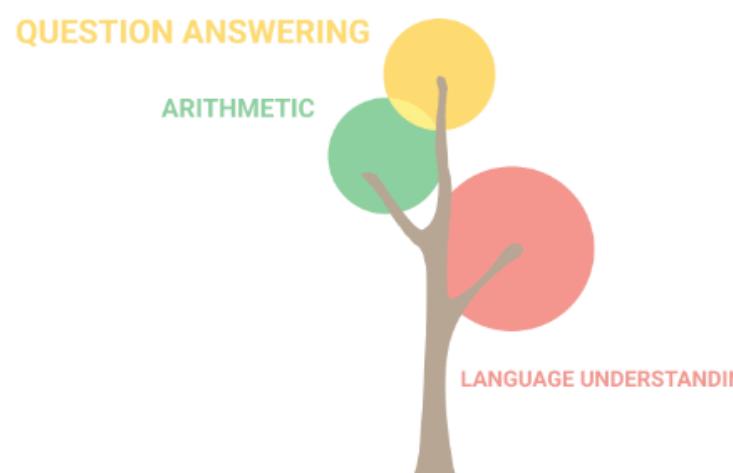
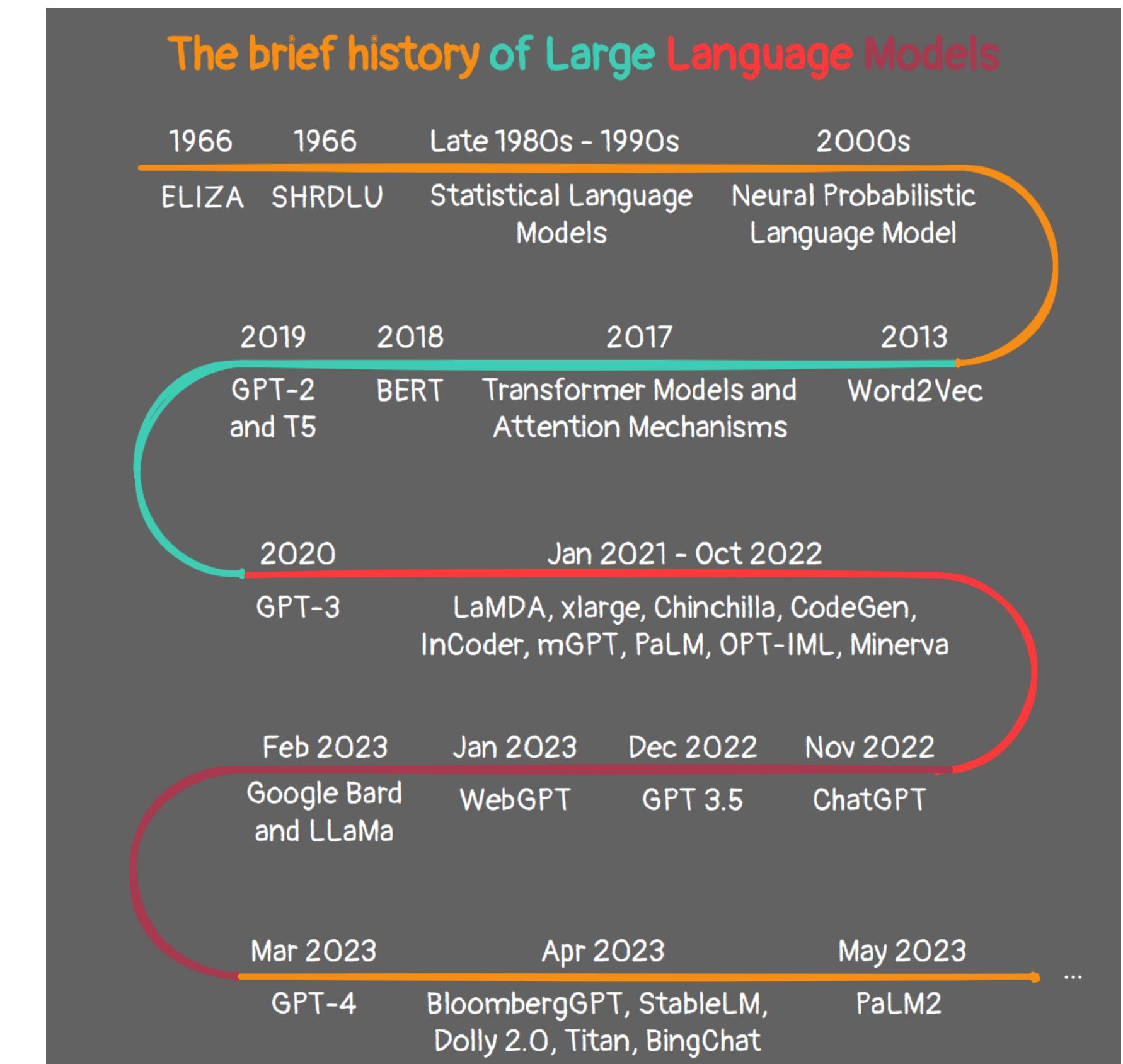
Graphs



Basic Concepts

Language Models (LMs)

- **Language models (LMs)** enable analysing **patterns in language** by predicting words
- **Best performing models** are based on **Transformer architectures** and using **Attention mechanisms** (LLMs use a decoder-only variant of this architecture) — masked language model, next sentence
- LMs are trained using **self-supervised learning**:
 - **Masked self-attention**: looks at other tokens in the sequence (i..e, those that precede the current token).
 - **Feed-forward transformation**: transforms each token representation individually.
- Applications:



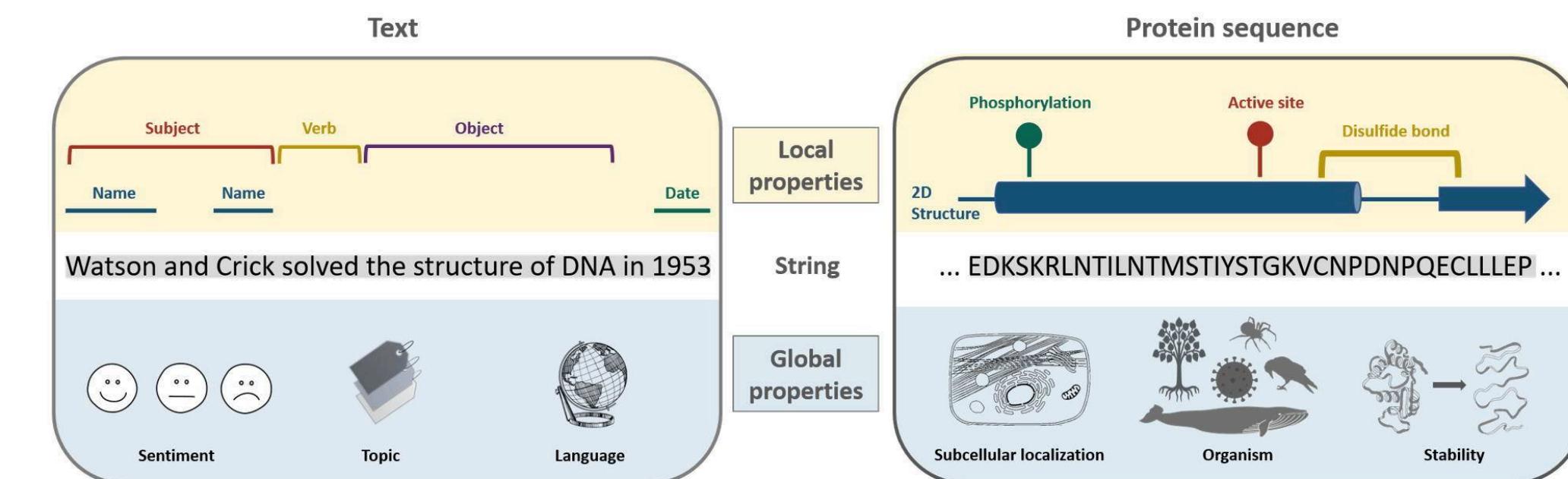
8 billion parameters

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin
arXiv:1706.03762 [cs.CL] <https://doi.org/10.48550/arXiv.1706.03762>
<https://techblog.ezra.com/an-overview-of-different-transformer-based-language-models-c9d3adafead8>
<https://tinyurl.com/LMtimeline>
<https://github.com/Hannibal046/Awesome-LLM/tree/main>

Proteins

- They play a **crucial role** in the **structure, function, and regulation of cells** and **tissues** within living organisms — Building blocks
- Protein sequences can be represented as **strings of amino acid letters** — fit for NLP
- The protein alphabet consists of **20 amino acids (AAs)** - letters
- These amino acids form **secondary structural elements** (e.g., motifs and domains)- words
- **Full sequences** — sentences
- **Structure, function and context** - meaning
- Differences with NL:
 - No punctuation no differentiation of words, sentences and paragraphs
 - High variability in length
 - NL have fewer distant interactions — proteins 3D structure
 - Letters of proteins can be modified (e.g., post-translational modifications)

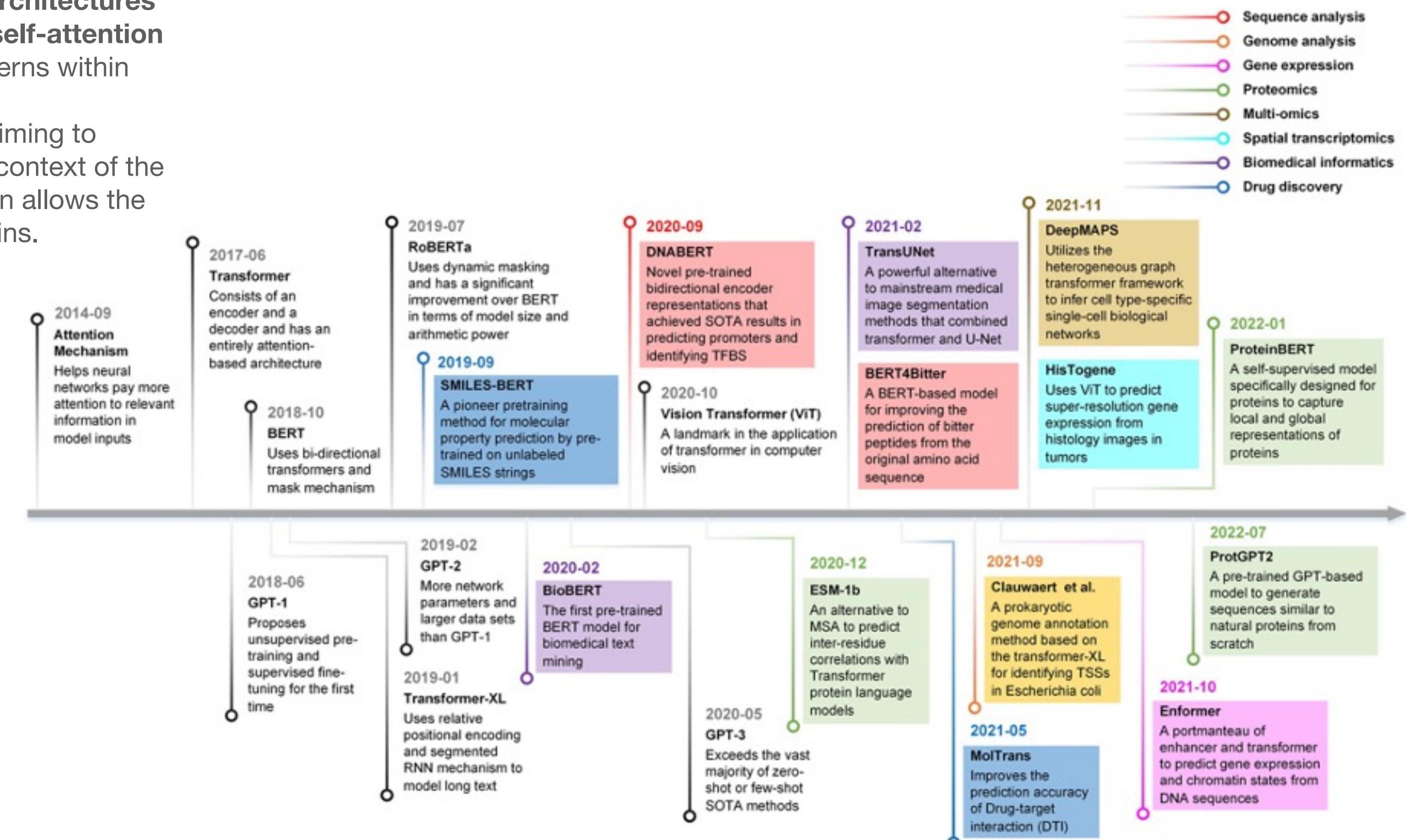


Protein Language Models – pLMs

Pre-training

use transformer-based architectures similar to NLP models. These architectures leverage **self-attention mechanisms** to learn complex relationships and patterns within protein sequences.

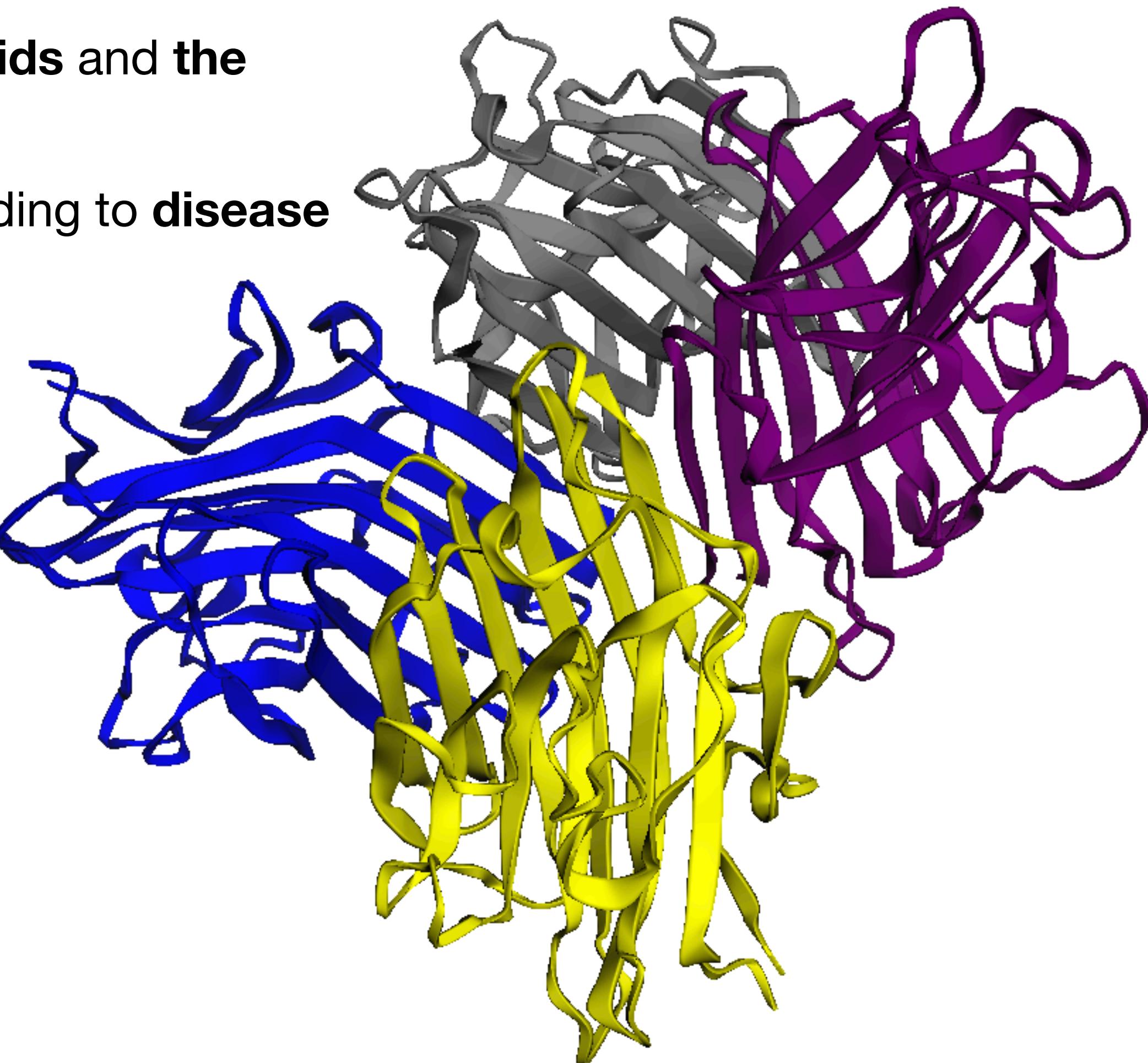
The model is trained using self-supervised learning, aiming to predict missing parts of protein sequences given the context of the surrounding sequences. This context-based prediction allows the model to capture meaningful representations of proteins.



Applications of pLMs

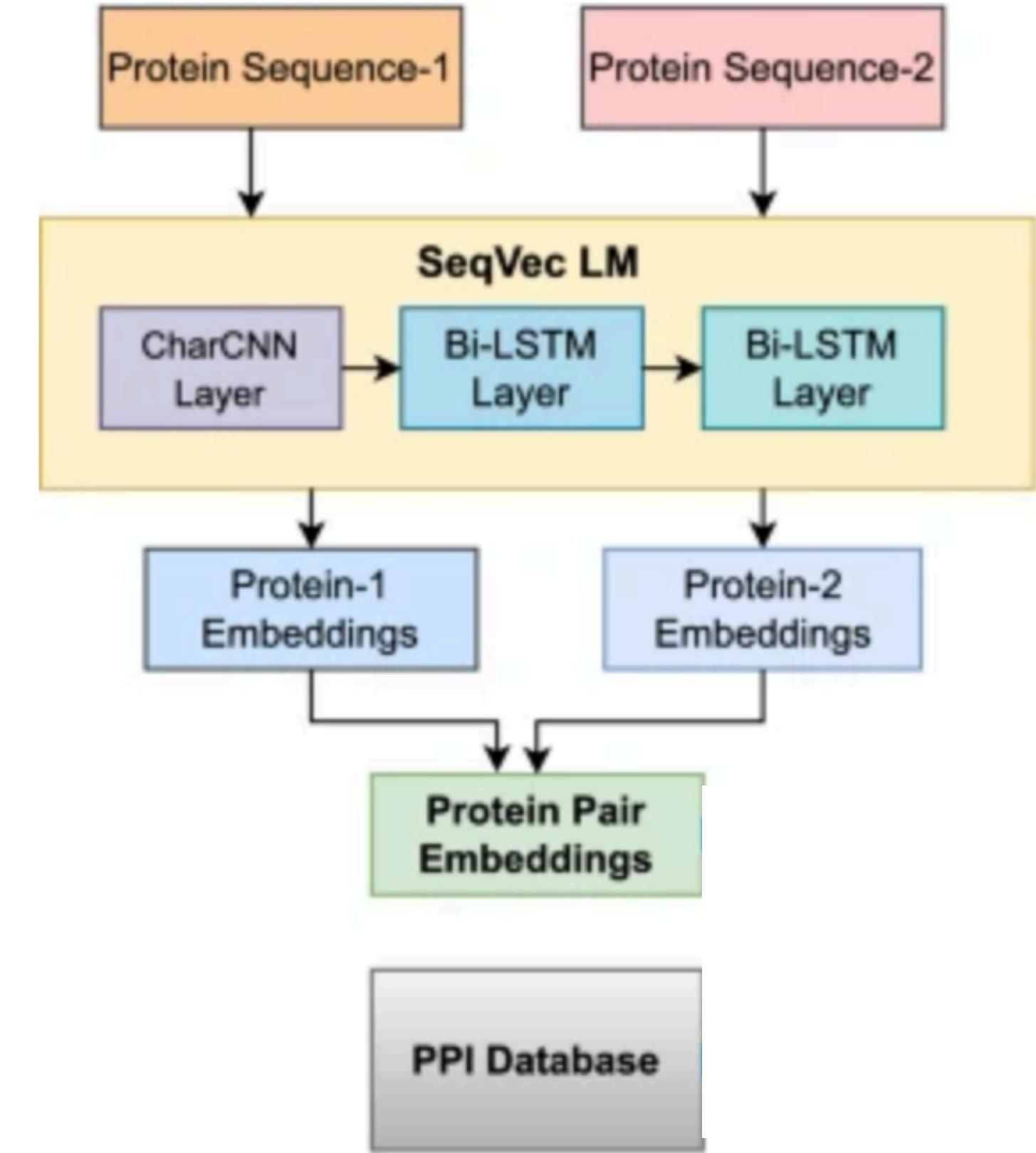
Protein Structure and Function

- The **function of a protein** depends on its **3D structure**
- The structure is determined by the **specific sequence of amino acids** and **the interactions between them**
- Any **changes** in the sequence can alter the shape and function, leading to **disease**
- **Remarkable success:**
 - Evolution-based: ESM
 - Single-sequence-based: RGN2
- Applications:
 - Function prediction
 - Drug design
 - Mutation impact



Protein-protein Interactions

- Protein–protein interactions (PPI) is a **critical problem in biology**
 - Some **protein functions** in a cell depend on having **physical contact** with other proteins present
 - PPIs identified through **experiments cover only a small percentage** of PPI networks
 - The entire PPI network cannot be explored experimentally due to **costs and time requirements**



Protein Engineering and Design

- Pre-trained transformer models enabled **text generation** including texts **with specific properties** such as style
- Fine-tuning pre-trained models on **protein families** will similarly enable the generation of **novel potentially functional sequences**.
- **Additional information** such as cellular compartment or function can aid the **control over specific properties**

Text generation

When I was living in Lisbon I was so shocked at the lack of diversity **that everyone chose to go back home. No matter there background, it was always the same things.** I was shocked at the amount of food options... **it was like one-dish-and-you're-on-your-face food with almost no meat options.**

Protein sequence generation

GCYVQAFGMAGCKINALNQYVINSLNVTNVP
YNLVGGKGYIYVWVLCGGNGGGNGAD
AIPGGADGGGGAGDPGAVLMSFYAYYPGATVT
IKVGNAGAAGGNGTNGSGGTSTTVNG
TTYTFTGGTSGKSGGTGVSGSSGLPGVDSSGT
TPGGGNGANGITTVFWS

Lisbon is a world-renowned destination for great culture, world-class dining, and an unparalleled opportunity for relaxation. It is also one of the safest cities in Europe, with an official crime rate that is half that of Paris or New York, ...

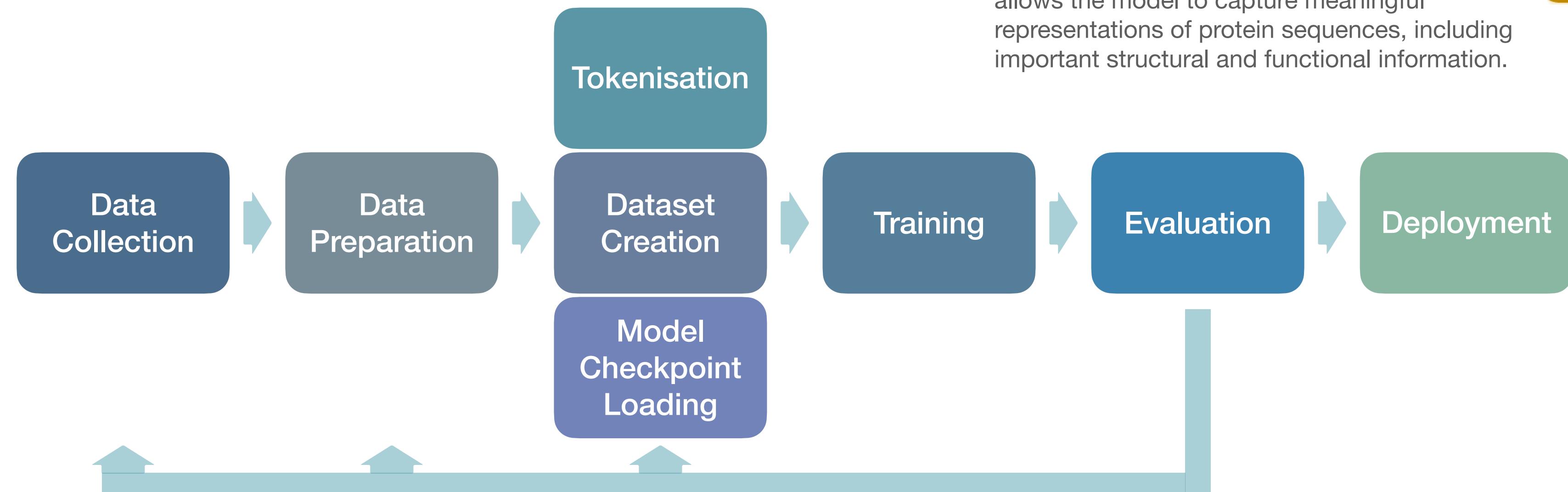
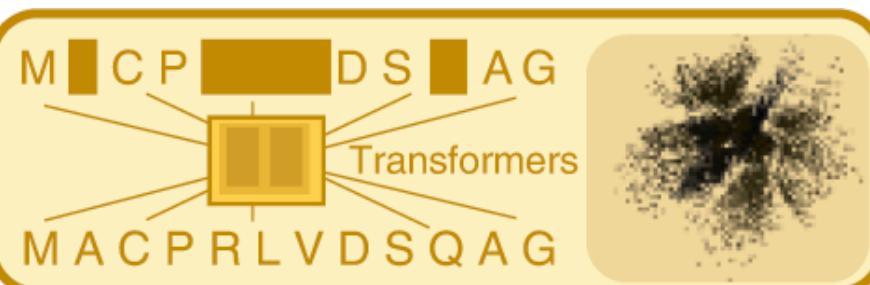
MTQDDRKRYRQLSMGQELAQMCACGARQS
LTVGTSSQRFCSEQCAYAARNRSHKAATSR
SQGLDPSRRQYLAVKRENNQFCVLCRAEGP
PSHYDRPLAVDDHDHDHESGPPRLLCNRC
NLGPLGKFDDPALLRDAVAHLEG MAGRA

Fine-tuning

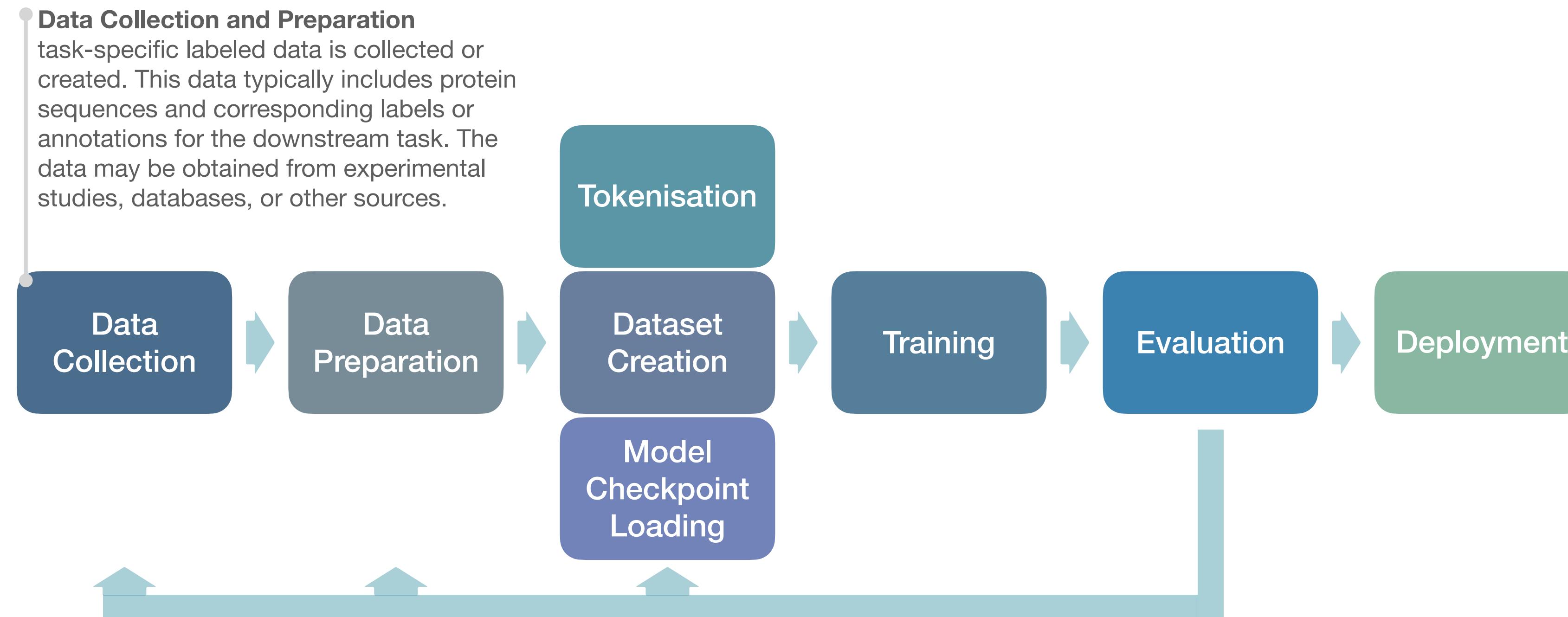
Fine-tuning pLMs

Fine-tuning refers to a learning technique where a **pre-trained language model** is further trained on a specific downstream task using task-specific data. The goal is to adapt the knowledge captured by the pre-trained model to the specific characteristics of the target task, thereby improving the model's performance on that task.

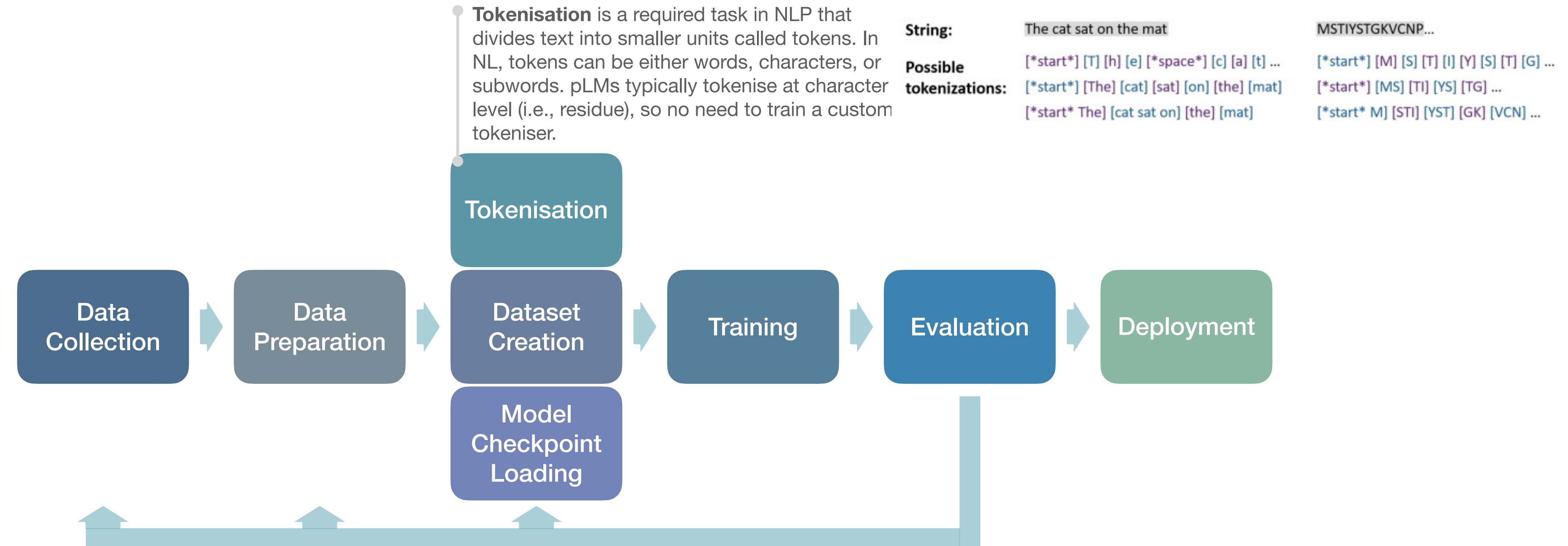
Pre-trained pLM is a model pre-trained on massive protein sequence databases. During pre-training, the models learn to predict missing parts of protein sequences based on the context provided by the surrounding sequences. This allows the model to capture meaningful representations of protein sequences, including important structural and functional information.



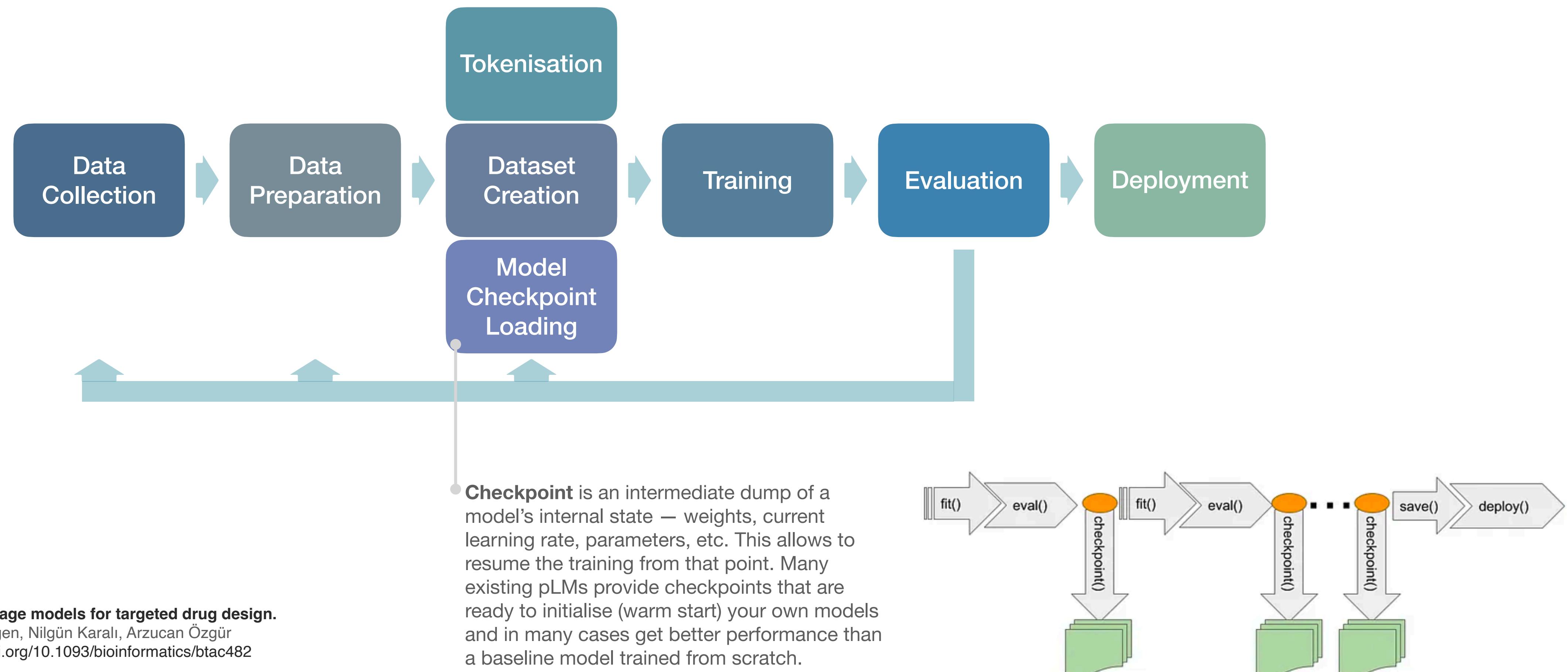
Fine-tuning pLMs



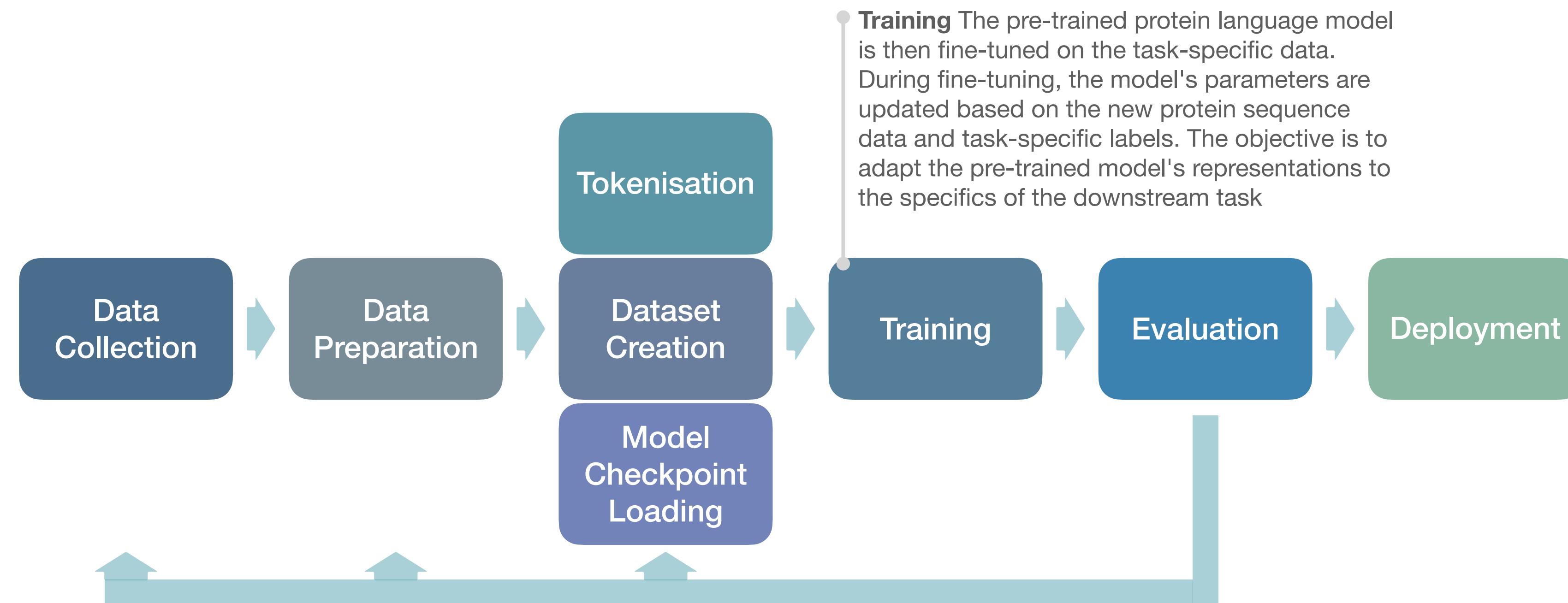
Fine-tuning pLMs



Fine-tuning pLMs



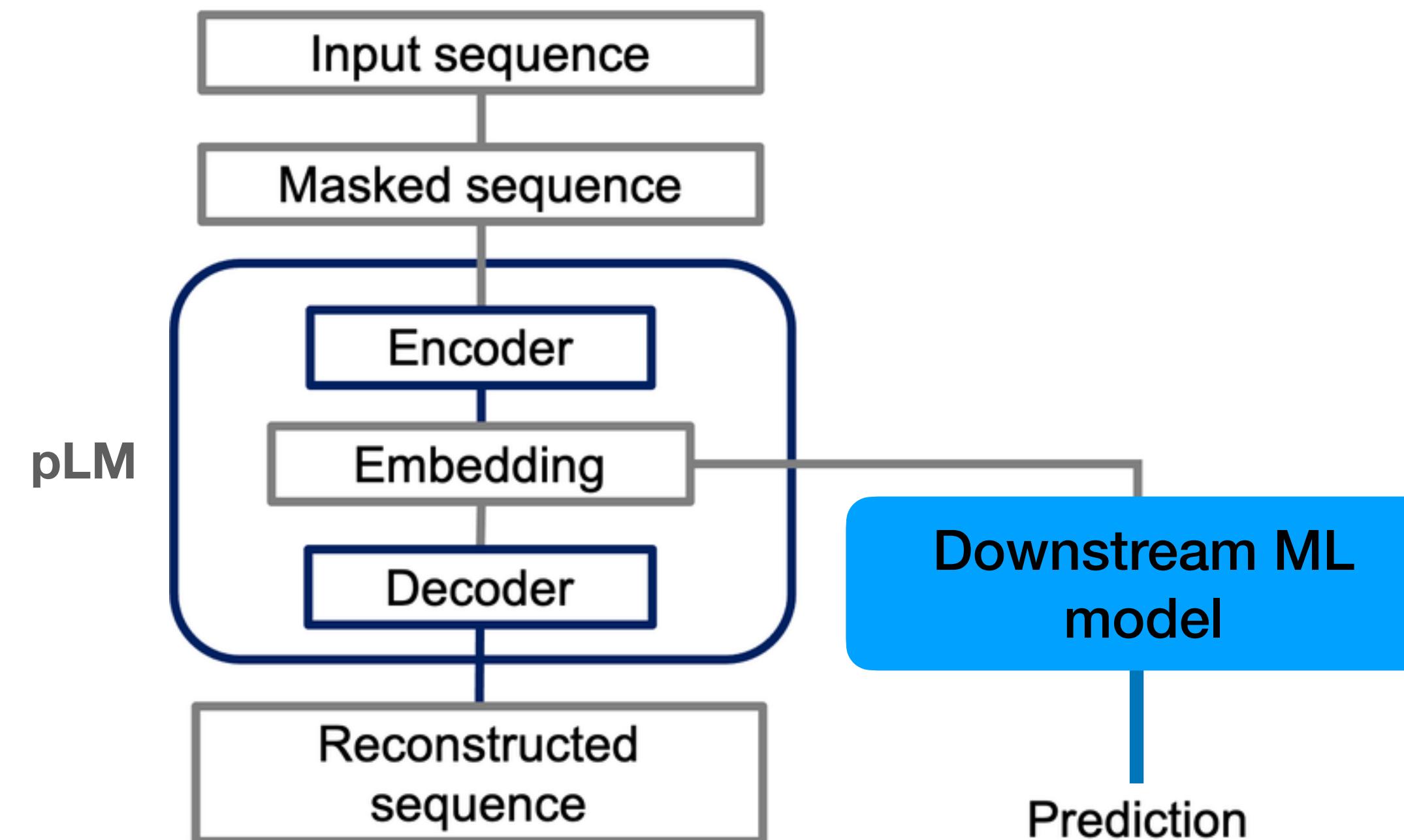
Fine-tuning pLMs



Protein Embeddings

Protein Embeddings

- **Represent** functional and structural properties of a protein in a machine-friendly format – **vector representation**
- Once encoded, they can be used in **multiple tasks** (e.g., sequence similarity, prediction, classification)



Data Resources

Databases



High-quality, comprehensive and freely accessible resource of **protein sequence and functional information**



RCSB Protein Data Bank (**RCSB PDB**) provides access and tools for exploration, visualisation, and analysis of:
Experimentally-determined 3D structures
Computed Structure Models (CSM) from AlphaFold DB and ModelArchive



Functional analysis of proteins by **classifying them into families and predicting domains and important sites.**



A database of **orthology** relationships and **functional annotation** based on **5090 organisms** and **2502 viruses**



IntAct provides a free, open source database and analysis tools for **molecular interaction data**. It also includes **negative datasets**

UniLanguage Datasets

- Analysis of the entire UniProt database
- Define different properties that can bias the performance of LMs such as homology, domain of origin, quality of the data, and completeness of the sequence
- This is the first protein dataset with an emphasis on language modelling

Dataset and results

We provide a [script](#) to obtain train, validation and test sets for all domains, qualities and protein completion.

Script usage

```
python get_data.py --domain [domain] --complete [complete] --quality [quality]
```

Parameters:

- Domain: Eukarya= `euk` , Bacteria= `bac` , Archaea= `arc` , Virus= `vir`
- Complete: Complete proteins= `full` , Fragmented proteins= `frag`
- Quality: Experimental evidence= `exp` , Predicted= `pred`

Benchmarking

Bio-Benchmarks

Code to process data, compute baselines and download splits can be found on <https://github.com/J-SNACKKB/FLIP>

Similar to **CASP** or **CAFA** used to assess **protein structure and function predictions**

Fitness Landscape Inference for Proteins (FLIP)

Provides a benchmark for function prediction to encourage rapid scoring of representation learning for protein engineering.

Limitations

- Lack of protein information
- Incomplete training data — negative data
- Biological complexity
- Lack of biological interpretation
- Model complexity and resource requirements
- Complex biological context and limited contextual understanding

Hands-on

<https://tinyurl.com/2u4v8sme>

Course Structure

Theory		Link	Data	Models
	Topics	slides		
Hands-on				
	Sequence Analysis	Seq. Analysis Notebook  Open in Colab	Exploring protein sequence and structure data	
	Fine-tuning a model	Model Training Notebook  Open in Colab	Taking an existing model and tuning it for other prediction/classification tasks	Evolutionary Scale Modeling (ESM)
	Working with Embeddings	Embeddings Notebook  Open in Colab	Accessing Protein representations (embeddings) generated by existing LMs	ProTrans
	Predictions	pML Predictions Notebook  Open in Colab	Using embeddings for predicting features or classifying sequences	ProTrans
	Protein Design	Protein Design Notebook  Open in Colab	<i>De novo</i> protein design and engineering using a LLM	ProtGPT2, ESM



Hands-on

<https://tinyurl.com/2u4v8sme>

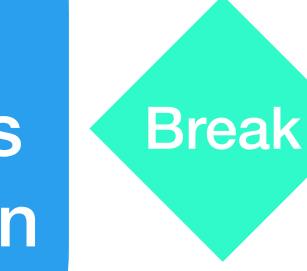
Data
collection and
preparation

Fine-tuning a
model for
classification

Generating
embeddings

Reusing
embeddings
for prediction

Applying for
protein design

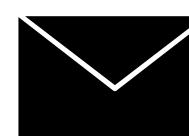


15'

15'

Thank you

novo
nordisk
fonden



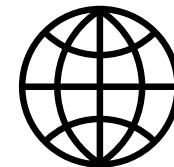
albsad@dtu.dk



@albsantosdel



<https://github.com/Multiomics-Analytics-Group>



<https://multiomics-analytics-group.github.io/>



crb

The Novo Nordisk Foundation
Center for Biosustainability