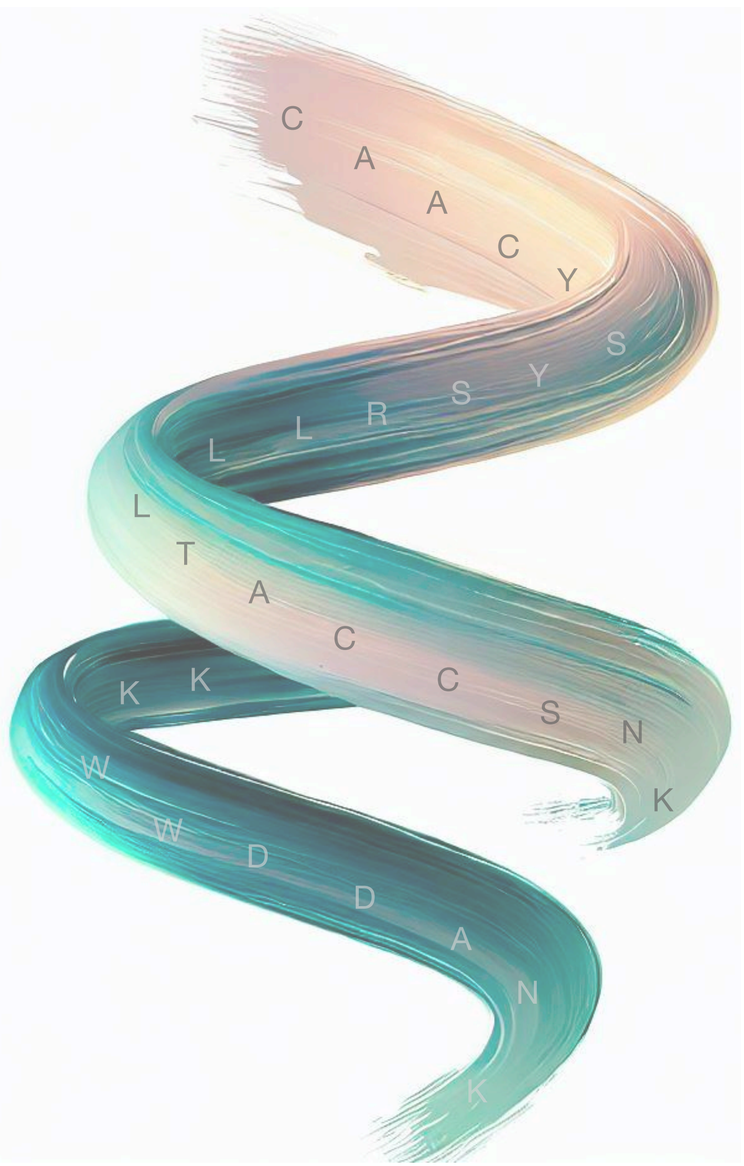
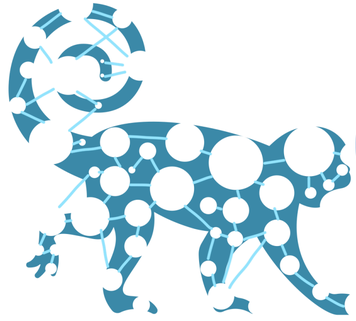


Protein Language Modeling

Applied Deep Learning in Bioinformatics Summer School

Alberto Santos — Multi-omics Network Analytics Group

13.08.2024

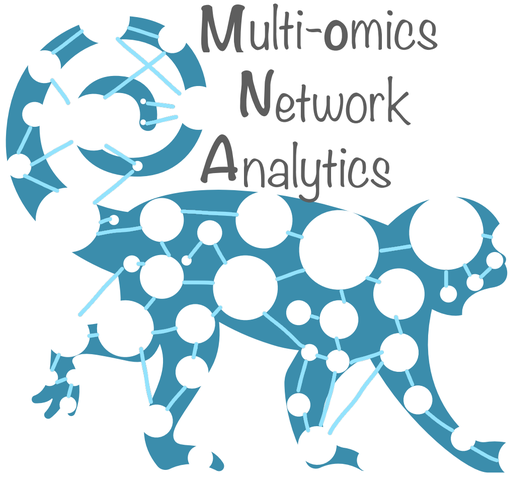


The Novo Nordisk Foundation
Center for Biosustainability

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: DTU Danmarks Tekniske Universitet — The Novo Nordisk Foundation Center for Biosustainability

Multi-omics

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

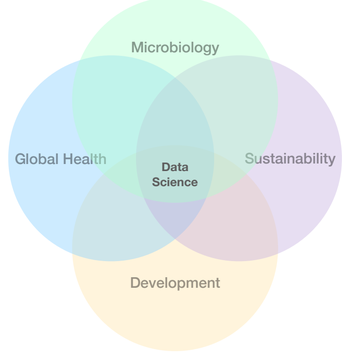
Network

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

Analytics

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

Data Science



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

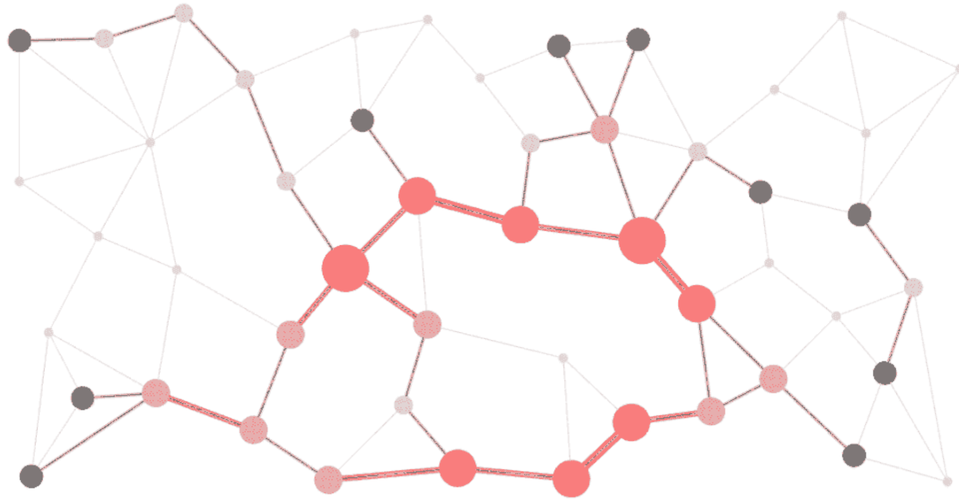
Education

Data Science democratisation

Open science

FAIR data

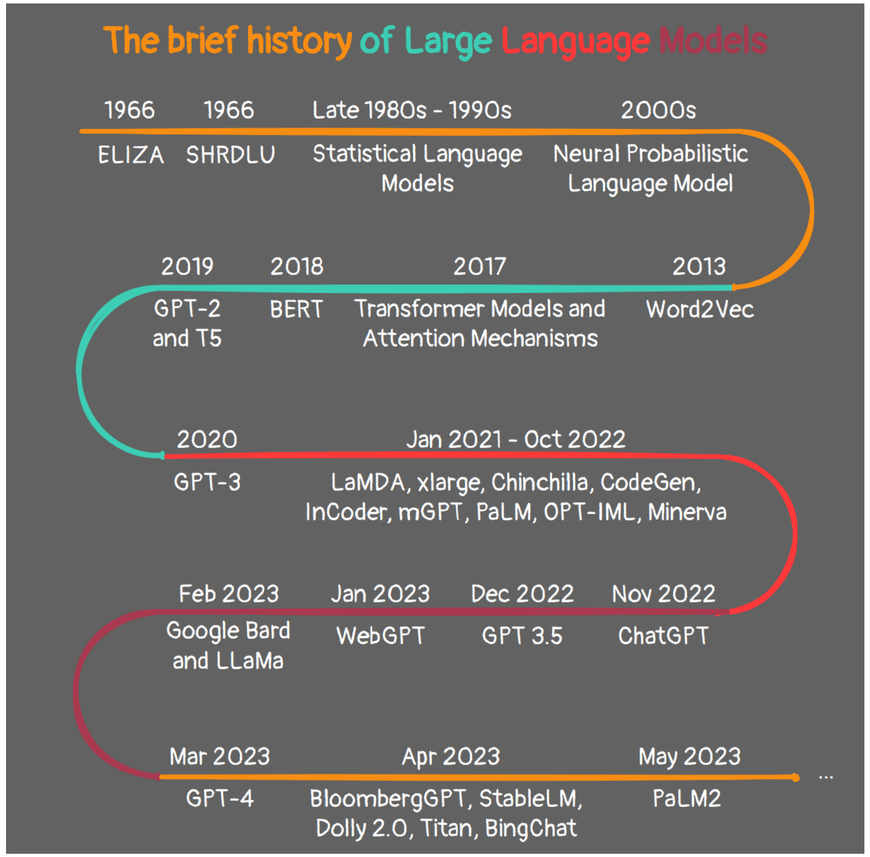
Graphs

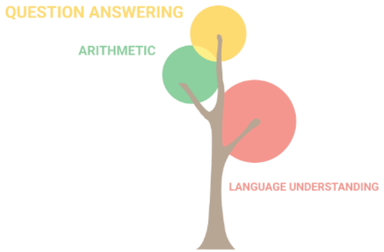


Basic Concepts

Language Models (LMs)

- **Language models (LMs)** enable analysing **patterns in language** by predicting words
- **Different tasks**: sentiment analysis, machine translation and text summarisation
- **Best performing models** are based on **Transformer architectures** and using **Attention mechanisms** (LLMs use a decoder-only variant of this architecture)
- 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:

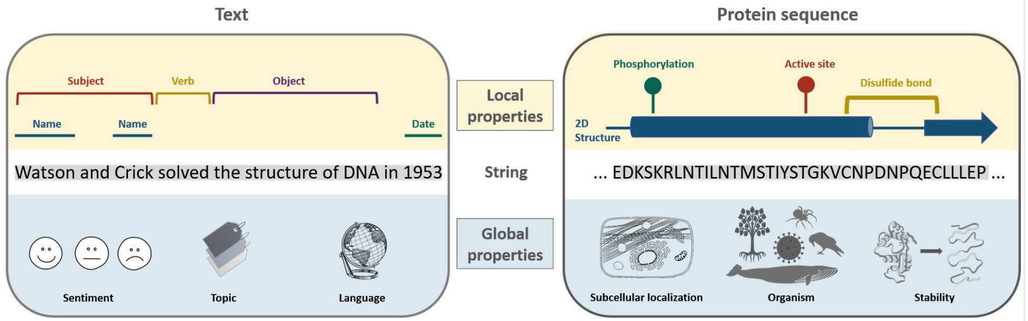




Attention Is All You Need
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv:1706.03762 [cs.CL] https://doi.org/10.48550/arXiv.1706.03762
https://techblog.ezra.com/an-overview-of-different-transformer-based-language-models-c9d3adafead8
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** - words
- Protein **motifs and domains** are the functional building blocks of proteins — 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



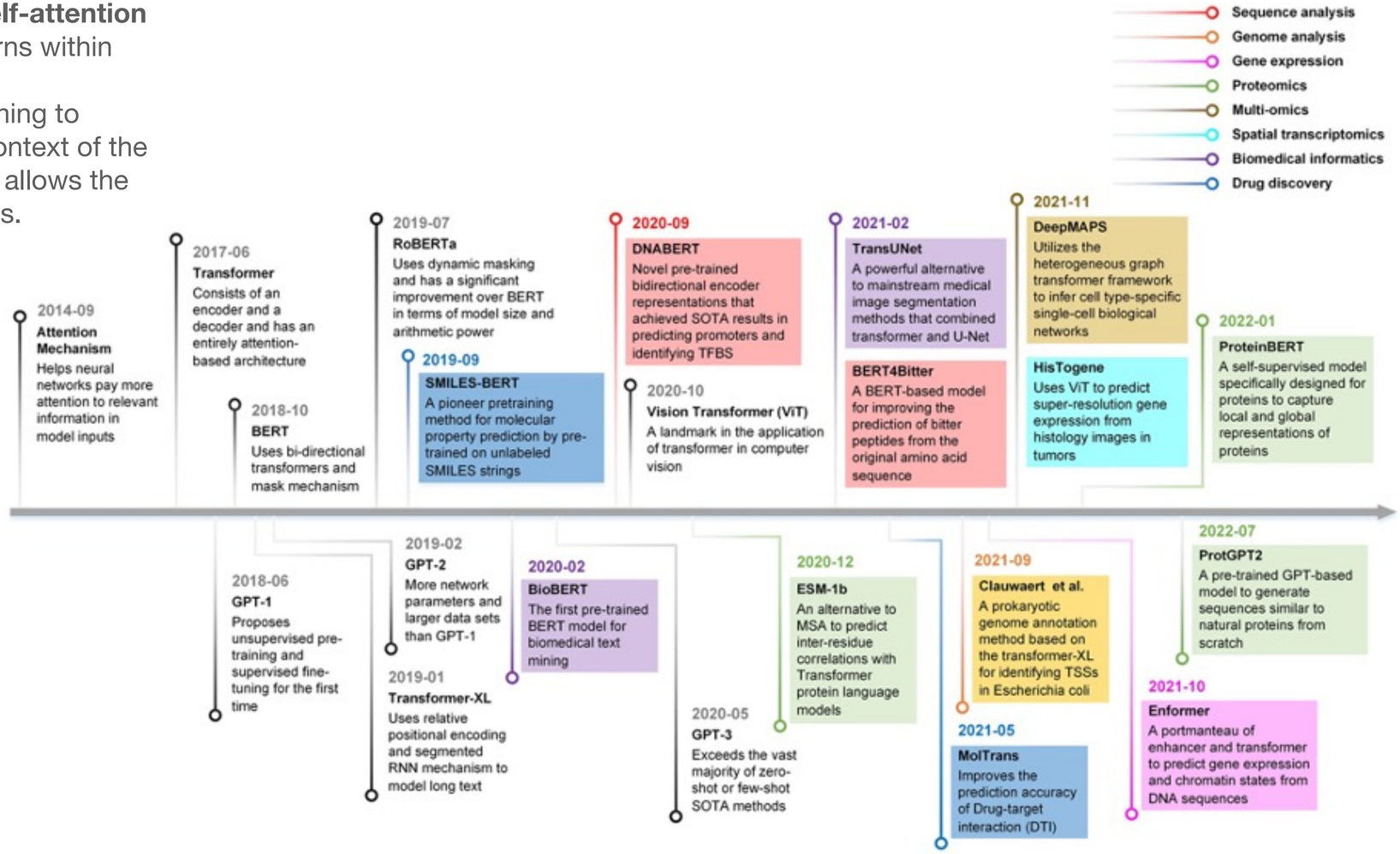
The language of proteins: NLP, machine learning & protein sequences

Dan Ofer, Nadav Brandes, Michal Linial.

Comput Struct Biotechnol J. 2021; 19: 1750–1758, 2021. https://doi.org/10.1016/j.csbj.2021.03.022

Protein Language Models — pLMs

Pre-training use t**ransformer-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 transformer-based language models in bioinformatics: a survey
Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, Wanwen Zeng.
Bioinformatics Advances, Volume 3, Issue 1, 2023, vbad001, https://doi.org/10.1093/bioadv/vbad001

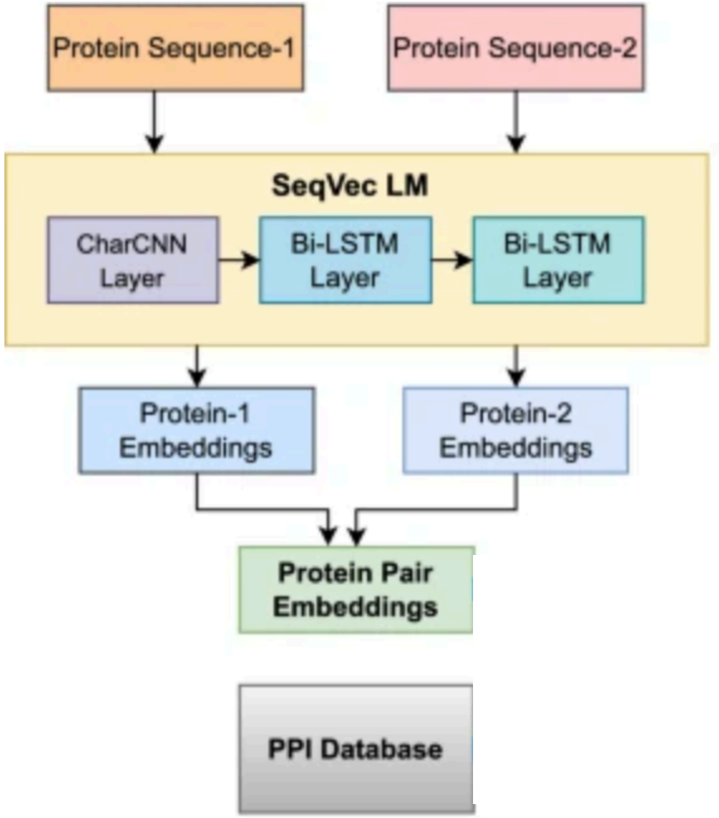
Applications of pLMs

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**



Graph-BERT and language model-based framework for protein–protein interaction identification
Kanchan Jha, Sourav Karmakar & Sriparna Saha. Scientific Reports volume 13, Article number: 5663 (2023)

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.

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, …

Protein sequence generation

```
GCYVQAFGMAGCKINALNQYVINSLNYVTNVP
YNLVGGKGYIYVWVLCGGNGGNGGNGAD
AIPGGADGGGGAGDPGAVLMSFYAYYPGATVT
IKVGNAGAAGGNGTNGGSGGTSTTVVNG
TTYTFTGGTSGKSGGTGVSGSSGLPGVDSSGT
TPGGGNGANGITTVFWS
```

```
MTQDDRKRYRQLSMGQELAQMCAECGARQS
LTVGTSSQRFCSEQCAYAARNRSHKAATSR
SQGLDPSRRQYLAVKRENNQFCVLCRAEGP
PSHYDRPLAVDDHDHDHESGPPRRLLCNRC
NLGPLGKFDDPALLRDAVAHLEGMAGRA
```

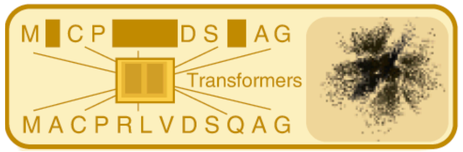
ProtGPT2 is a deep unsupervised language model for protein design
Noelia Ferruz, Steffen Schmidt & Birte Höcker. Nature Communications volume 13, Article number: 4348 (2022)

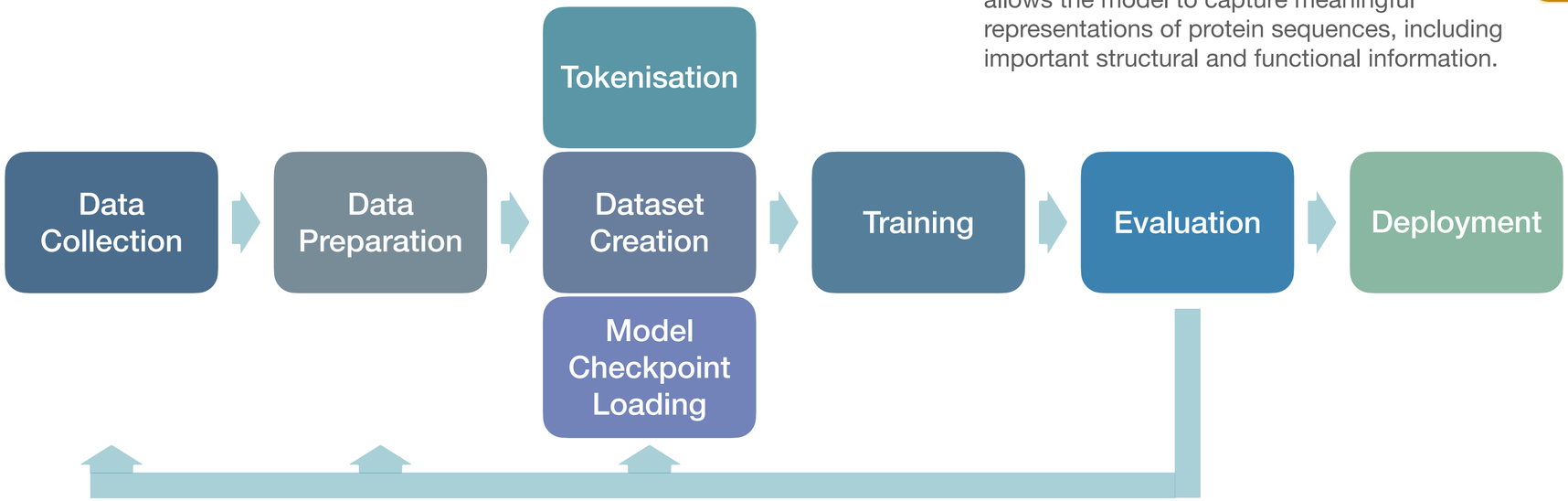
Fine-tuning

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.

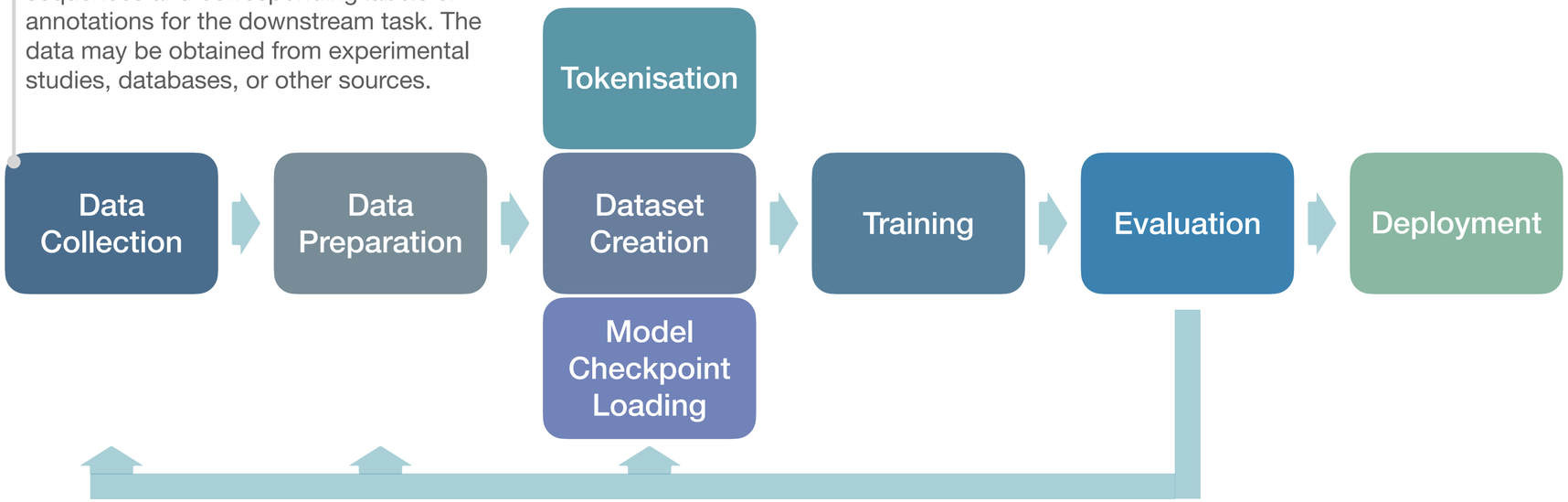




Single-sequence protein structure prediction using a language model and deep learning
Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M. Church, Peter K. Sorger & Mohammed AlQuraishi.
Nature Biotechnology (2022)

Fine-tuning pLMs

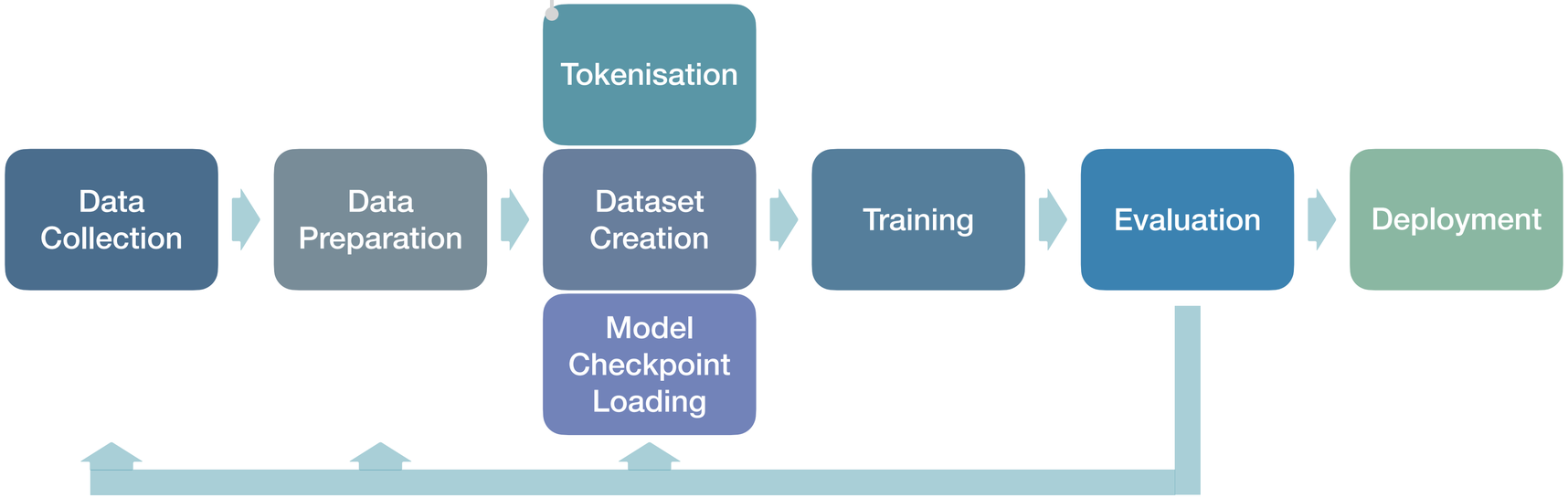
Data Collection and Preparation
task-specific labeled data is collected or created. This data typically includes protein sequences and corresponding labels or annotations for the downstream task. The data may be obtained from experimental studies, databases, or other sources.



Fine-tuning pLMs

Tokenisation is a required task in NLP that divides text into smaller units called tokens. In NL, tokens can be either words, characters, or subwords. pLMs typically tokenise at character level (i.e., residue), so no need to train a custom tokeniser.

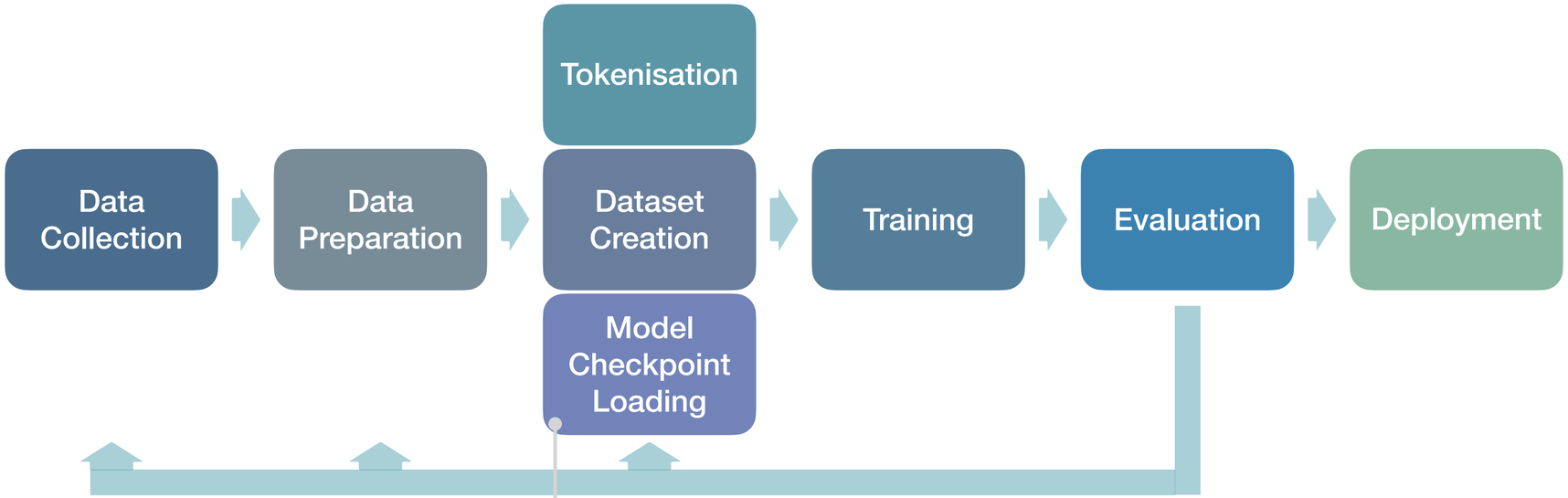
String:	The cat sat on the mat	MSTIYSTGKVCNP...
Possible tokenizations:	[*start*] [T] [h] [e] [*space*] [c] [a] [t] ...	[*start*] [M] [S] [T] [I] [Y] [S] [T] [G] ...
	[*start*] [The] [cat] [sat] [on] [the] [mat]	[*start*] [MS] [TI] [YS] [TG] ...
	[*start* The] [cat sat on] [the] [mat]	[*start* M] [STI] [YST] [GK] [VCN] ...



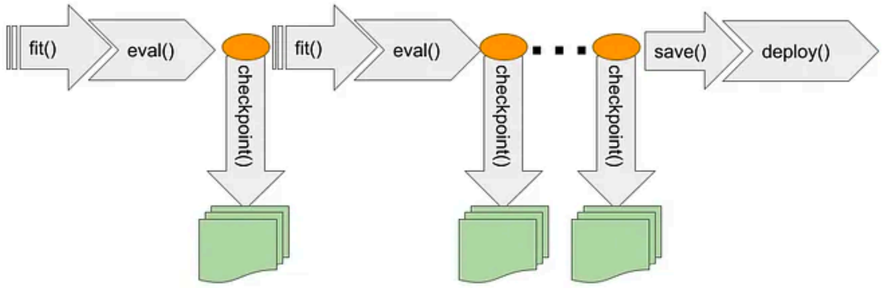
The language of proteins: NLP, machine learning & protein sequences

Dan Ofer, Nadav Brandes, Michal Linial.
Comput Struct Biotechnol J. 2021; 19: 1750–1758, 2021. https://doi.org/10.1016/j.csbj.2021.03.022

Fine-tuning pLMs



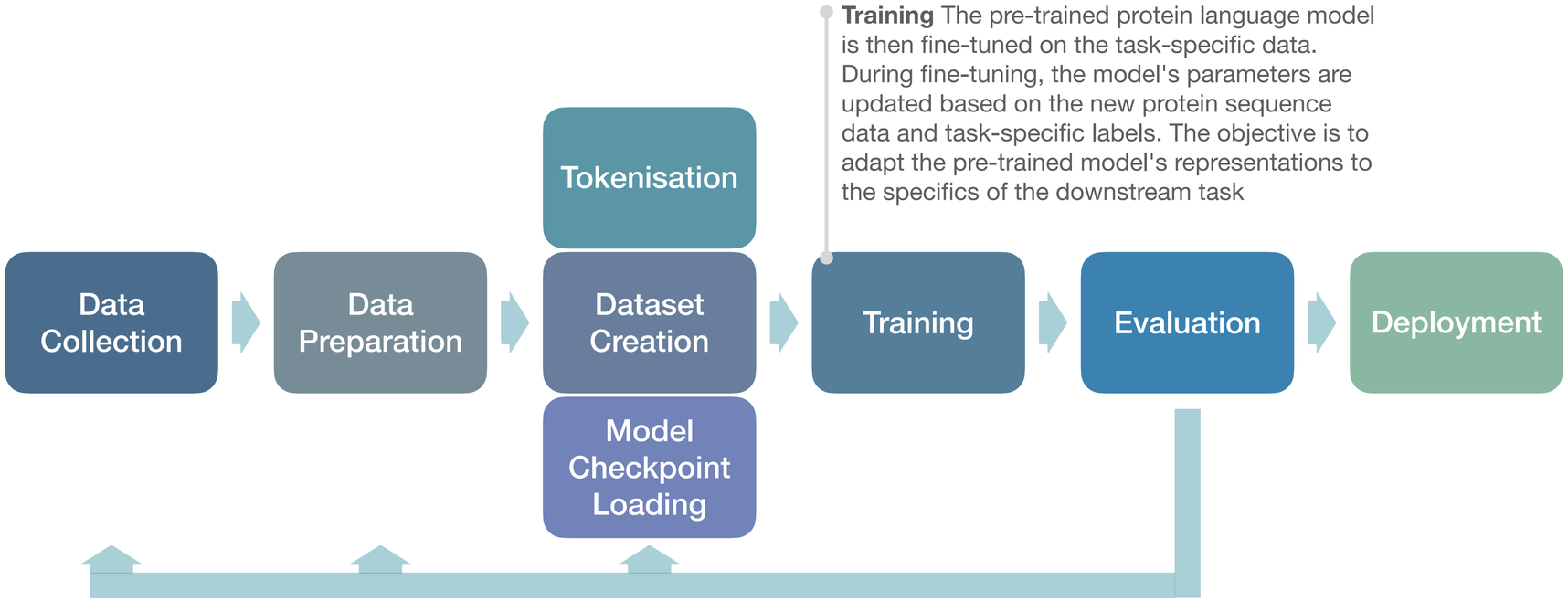
Checkpoint is an intermediate dump of a model's internal state — weights, current learning rate, parameters, etc. This allows to resume the training from that point. Many existing pLMs provide checkpoints that are ready to initialise (warm start) your own models and in many cases get better performance than a baseline model trained from scratch.



Exploiting pretrained biochemical language models for targeted drug design.
Gökçe Uludoğan, Elif Ozkirimli, Kutlu O Ulgen, Nilgün Karalı, Arzucan Özgür
Bioinformatics, September 2022, https://doi.org/10.1093/bioinformatics/btac482

https://towardsdatascience.com/ml-design-pattern-2-checkpoints-e6ca25a4c5fe

Fine-tuning pLMs



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

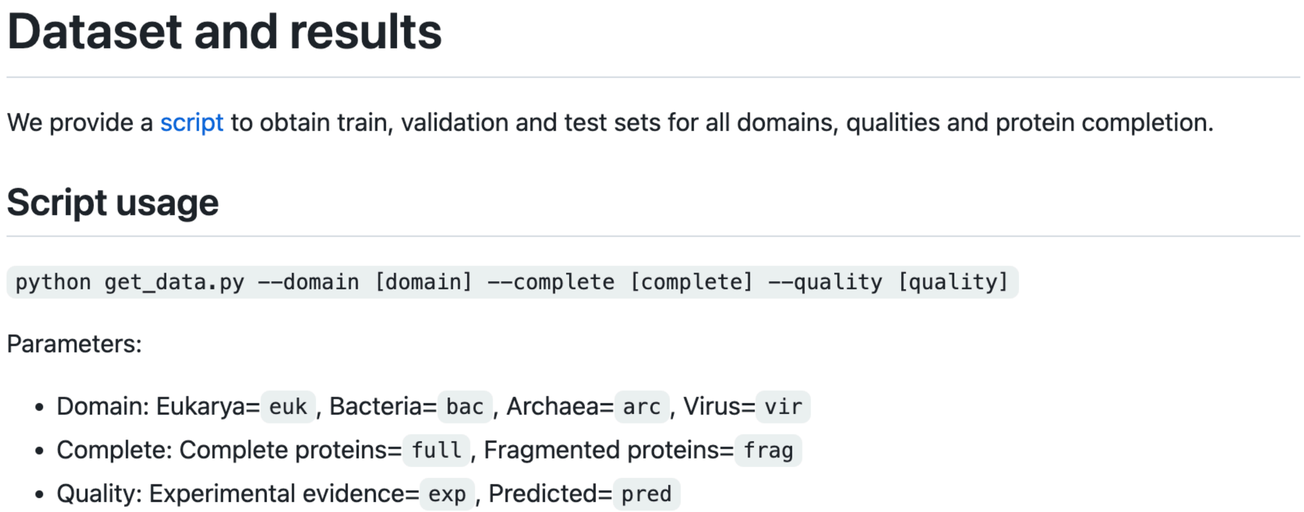
Functional analysis of proteins by **classifying them into families and predicting domains and important sites.**

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`

Language modelling for biological sequences – curated datasets and baselines
Jose Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, Henrik Nielsen
doi: https://doi.org/10.1101/2020.03.09.983585

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.

https://benchmark.protein.properties/

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

Practical

https://github.com/Multiomics-Analytics-Group/course_protein_language_modeling

SCAN ME

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	*De novo* protein design and engineering using a LLM	ProtGPT2, ESM

Thank you

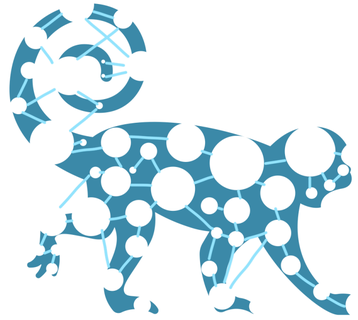
novo
nordisk
fonden

albsad@dtu.dk

@albsantosdel

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

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



The Novo Nordisk Foundation
Center for Biosustainability