Project Proposal

AI-guided Protein Science

Alberto Santos — Multi-omics Network Analytics (MoNA)

Kinase-substrate Phosphorylation Prediction

Objective — Design and **train** an model to predict links (interactions) between kinases and substrates. Input —Peptide sequence / Output — Kinase that most likely phosphorylates the given peptide

Background:

Phosphorylation is a critical post-translational modification regulating numerous cellular processes. Kinases are enzymes responsible for phosphorylation, recognising specific substrate sequences or motifs. Predicting kinase-substrate relationships is key to understanding signal transduction pathways and designing therapeutic interventions.

Methods:

- Protein Language Models: Use pre-trained models (e.g., ESM, ProTrans)

- Graph Machine Learning: Model kinases and substrates as nodes in a graph, and employ GNNs to learn the relationships between kinases and substrates

- Combined approach: Integrate sequence features (from protein language models) and graph-based structural information for link prediction

Data availability:

- An atlas of substrate specificities for the human serine/threonine kinome

  ‣ Supplementary Table 3

- Using explainable machine learning to uncover the kinase–substrate interaction landscape

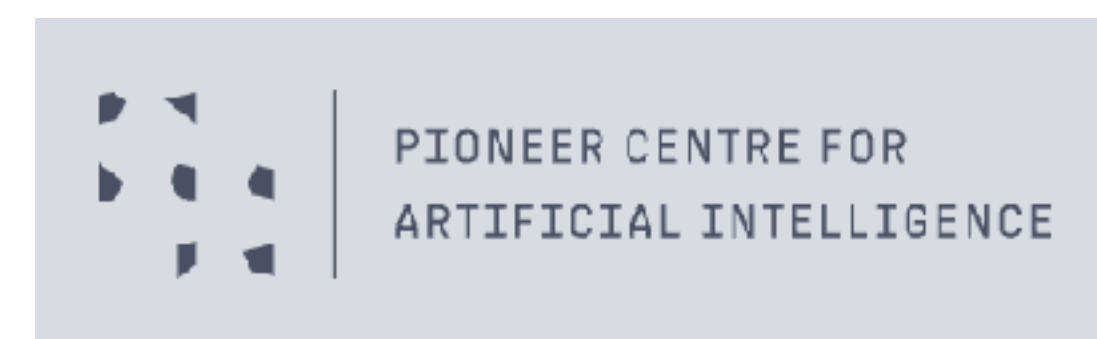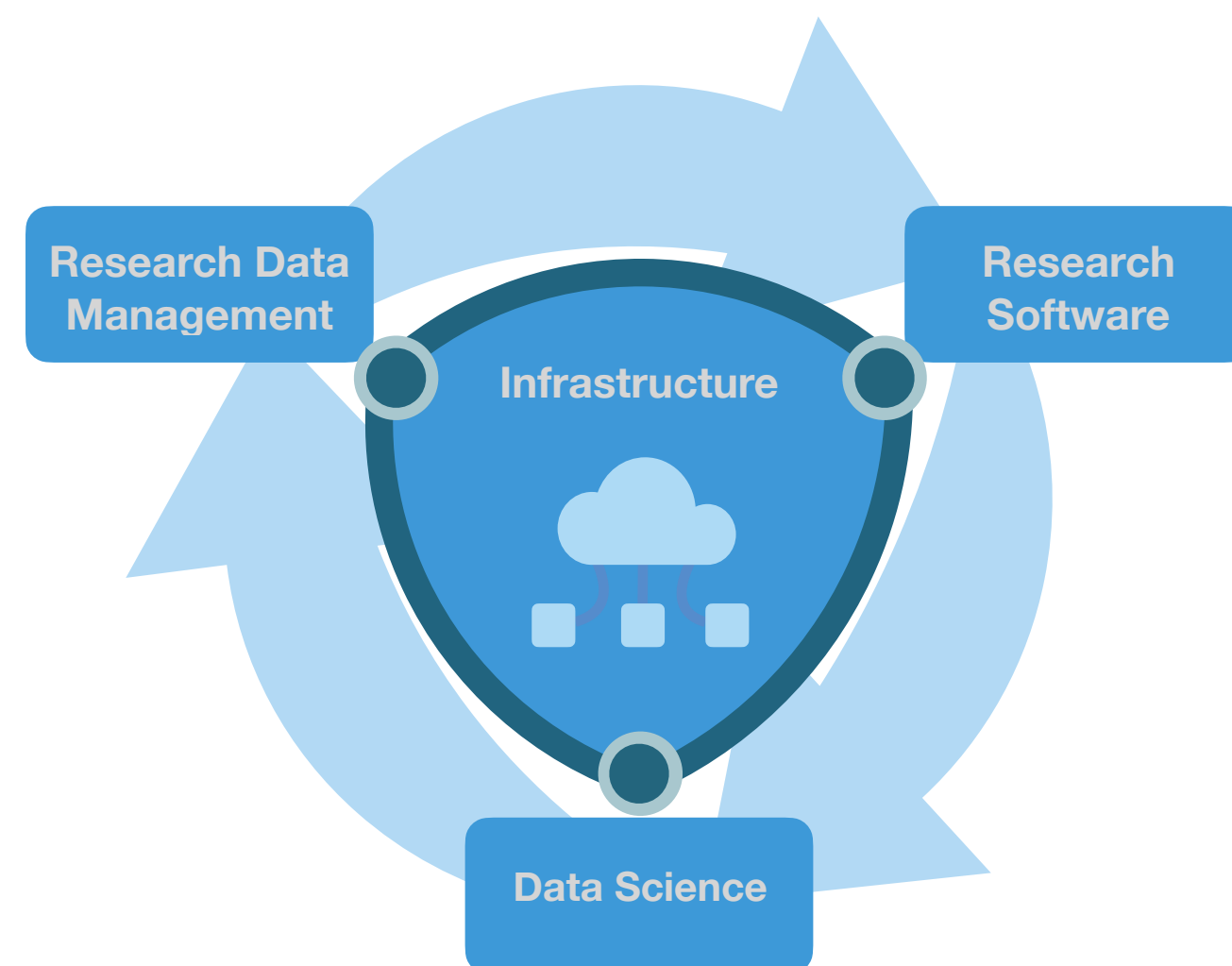 ‣ https://huggingface.co/datasets/waylandy/phosformer_curated/tree/main

Evaluation:

- Solution design

- Dataset preparation and exploration

- Implementation rational

- Proposed model improvements

Thank you

Multi-omics Network Analytics Research Group

Informatics Platform

Research Data Management

Research Software

Infrastructure

Data Science

✉ albsad@dtu.dk

🐦 @albsantosdel

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

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