Global Pathogen Analysis Platform | AI Research

MSC THESIS | FEB–JUN 2027 | REMOTE OR IN PERSON

Genome Foundation Models for Pathogen Analysis

Genome foundation models such as Evo 2, NTv3 and Carbon show strong results on many genomic tasks. We push the limits of these models for pathogen analysis tasks, for example in antimicrobial resistance surveillance and the evolution of viral genomes.

Mid-training | Post-training | SFT, DPO and RL | Data mixtures with FineDNA | Pathogen genomes | Context extension (YaRN) | FlashAttention | Flash linear attention | Kernel optimization | vLLM inference

1 **Baseline**

Evaluate Evo 2, NTv3, Carbon and other genome foundation models.

2 **Data**

Build training data from FineDNA. Build new evaluation benchmarks.

3 **Training**

Post-training and optimization, from data mixtures to kernels.

4 **Analysis**

Measure which data choices help. Release the datasets and code.

Resources

5,000+ **NVIDIA B200 GPU hours**

Plus in-house NVIDIA H100 and RTX PRO 6000 GPU compute nodes.

Potential Outputs **Top ML conference/journal paper**

Who can apply

MSc thesis: MSc students at the University of Copenhagen (KU) and the Technical University of Denmark (DTU).

Location: remote, or in person in Copenhagen, Denmark.

Research Team



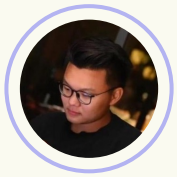
Samir Bhatt
Professor, FMedSci, MAE
Principal Investigator
University of Copenhagen



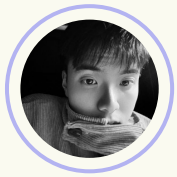
Harrison Bo Hua Zhu
Assistant Professor
Technical University of Denmark



Neil Scheidwasser
Postdoctoral Researcher
University of Copenhagen



Binghan Xiao
PhD Fellow
University of Copenhagen



Xingpeng Liu
MSc Student and Research Assistant
University of Copenhagen



Zak Rehman
Research Assistant
University of Copenhagen

How to apply

We look for strong Python and PyTorch skills and experience with LLM training or evaluation. You do not need to know about genomics or infectious diseases, but an interest in learning more about these topics is essential. Send your CV, transcript and a short research statement to Harrison Bo Hua Zhu at **habhuz@dtu.dk**. Selection is competitive.

gpap-project.github.io/join