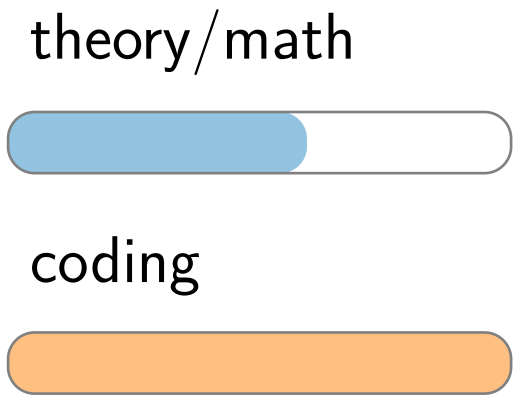
Research Internship

Large Language Models for Optimizing Bioprocesses

Topic profile



Tags

#AI

#large language models

#bioprocessing

#interdisciplinary research

Supervision

Benedikt Bollig

CNRS Researcher at ENS Paris-Saclay

Matthias Fuegger

CNRS Researcher at ENS Paris-Saclay

Thomas Nowak

Professor at ENS Paris-Saclay

Why bioprocessing?

Many industrial or biomedical products, including pharmaceuticals, biofuels, and vaccines, are produced by cultivating cells in bioreactors. Maximizing production while ensuring product quality is crucial. The current approach to optimizing bioreactor setups relies on time-consuming and costly wet-lab experiments. This internship is part of a research program that aims to reduce time and costs by developing digital twins (digital replicas of bioreactors) using machine learning.

What we are looking for

We value a curious and driven attitude. An ideal candidate is inclined to artificial intelligence, LLMs, and coding (in Python). Basic knowledge in microbiology is an advantage.

The team

You will be part of an interdisciplinary research team at ENS Paris-Saclay near Paris, working on different aspects of artificial intelligence, synthetic biology, distributed computing, and circuit design.

Research

We will leverage the capabilities of large language models (LLMs) to create digital twins, with a particular emphasis on generating biochemical reaction networks (BCRNs). These networks play an important role in digital twin development, as they encapsulate and generalize experimental data in the form of time series tracking substance concentrations within a bioreactor process. Using BCRNs, we can conduct *in silico* simulations to identify optimal bioreactor setups, bypassing additional wet-lab experiments.

LLMs excel in generating natural language as well as expressions and code from formal languages. As part of this internship, the participant will contribute to the development of an LLM that explores the space of possible (potentially novel) pathways and proposes suitable candidate BCRNs aligning with provided time-series data. A significant part of the research will involve fine-tuning base models and retrieval augmented generation based on specific domain knowledge. Although the BCRNs generated by an LLM will be qualitative in nature, their calibration is subsequently managed through parameter-estimation algorithms.

You are interested or would like to join us?

Please send us your questions or, in case you would like to apply, a short statement of interest and a CV, to Benedikt Bollig (bollig@lmf.cnrs.fr), Matthias Fuegger (mfuegger@lmf.cnrs.fr), and Thomas Nowak (thomas@thomasnowak.net). The start date of the internship is flexible, but the goal is to start in spring or summer 2024.