Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Differentiable Protein Expression Models for Enhanced mRNA Translation

Project Manager
Philipp Fröhlich

Researchers
Farzad Dehghani Eshratabad

Principal Investigator
Prof. Dr. Heinz Koeppl

Project Term
2025 - 2025

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

Institute
Self-Organizing Systems

University
Technische Universität Darmstadt

Introduction

Designing coding sequences that produce high protein levels is a major challenge in biotechnology, synthetic biology, and microbial engineering. Many synonymous DNA sequences encode the same protein, but they can differ strongly in translation efficiency and final protein abundance. This project investigated whether differentiable machine learning models can be used to guide this design problem: first by predicting protein abundance directly from coding DNA, and ultimately by serving as surrogate objectives for optimizing new gene sequences. The long-term goal is to support computational gene design before experimental testing. Such models could help propose coding sequences for recombinant protein production, enzyme engineering, synthetic gene circuits, and optimized microbial production strains, reducing costly trial-and-error in the laboratory.

Methods

We developed lightweight neural predictors that take nucleotide sequences, and where available organism or strain information, as input. Starting from *E. coli*, the project was extended to several related bacterial species to study both shared and host-specific sequence determinants of expression. The modelling strategy moved from an initial low/medium/high classification setup to regression, which better preserved the quantitative structure needed for design. The core models were compact convolutional neural networks, chosen for their ability to capture

local sequence and codon-level patterns while remaining efficient enough for future optimization pipelines. We compared these models with attention-based alternatives, tested learned species representations, and evaluated different training objectives and input variants.

Results

High Performance Computing on the Lichtenberg cluster was essential for this work. It enabled GPU-accelerated training, large-scale hyperparameter optimization, cross-validation, ablation studies, and repeated comparisons across model families. This made it possible to identify models that were not only accurate, but also compact and robust enough to be useful as differentiable surrogates. The best regression models captured a meaningful part of the variation in protein abundance from coding sequence alone. Multi-species training improved performance over single-species baselines, and learned species embeddings further increased predictive quality while keeping the models small.

Discussion

Overall, the project provides a first step towards differentiable predictor-guided mRNA and coding-sequence design for improved protein expression. Future work will combine expression prediction with additional design objectives such as codon adaptation, RNA stability, GC content, and host-specific biological constraints to generate experimentally plausible sequences with improved translation properties.

Last Update: 2026-06-09 15:41