Hessisches Kompetenzzentrum
für Hochleistungsrechnen

mRNA Sequence Optimization via Neural Surrogate Modeling of Translation Initiation



Project Manager
Prof. Dr. Heinz Koeppl

Principal Investigator
Philipp Fröhlich

Project Term
2025 - 2025

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

Additional Software
SciKit-Learn, Biopython

Institute
Self-Organizing Systems

University
Technische Universität Darmstadt

Introduction

In this Bachelor's thesis, a neural surrogate model was developed to predict the Minimum Free Energy (MFE) of mRNA sequences. The MFE is a crucial parameter for forecasting mRNA secondary structures, which in turn strongly influence protein expression rates. Accurate estimation of secondary structures is an essential step in many areas of modern bioinformatics, including gene expression analysis, vaccine development, and the design of synthetic RNA molecules. The main motivation for developing this surrogate network was to enable optimization and conditioning of MFE values using gradient-based methods. By creating a differentiable model that can predict MFE from mRNA sequences, we open up the possibility for the model's predictions to be used as part of loss functions in future machine learning workflows. This allows gradients to be backpropagated through the MFE prediction step, facilitating tasks such as mRNA sequence optimization to achieve desired MFE values or secondary structure properties. Having a fast and differentiable surrogate for RNA folding energy enables exciting new research avenues. High Performance Computing (HPC) resources were needed to train the model due to the high computational requirements of the transformer-based architectures and the large size of the training dataset.

Methods

The model was designed around the RNAfold algorithm from the ViennaRNA package, which served as a reference to generate

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg

training data consisting of mRNA sequences and their corresponding MFE values. An encoder-only transformer architecture was implemented, incorporating a FlashAttention kernel to accelerate training and reduce GPU memory usage. Training and evaluation were carried out on HPC infrastructure provided by TU Darmstadt as part of NHR4CES, enabling the use of multiple high-performance GPUs to efficiently process millions of sequence-MFE pairs. Hyperparameters such as learning rate, batch size, and number of layers were optimized to achieve stable convergence.

Results

After training, the model demonstrated strong predictive performance for MFE values, achieving low error rates compared to the RNAfold-generated ground truth. It exhibited good adaptability to variable sequence lengths, making it suitable for a wide range of applications. The inference speed was several orders of magnitude faster than the traditional RNAfold computation for large batches of sequences, which highlights the potential of this surrogate approach for integration into high-throughput pipelines.

Discussion

These results demonstrate that transformer-based surrogate models can accurately and efficiently predict RNA properties like MFE. The model's differentiable nature enables optimization and conditioning of MFE values using gradient-based methods. By allowing gradients to be backpropagated through the MFE prediction step, this approach opens up possibilities for optimizing mRNA sequences to achieve target MFE values or desired secondary structure properties. This could drive advancements in rational RNA design and related fields

Last Update: 2025-11-11 10:29