TECHNISCHE UNIVERSITÄT DARMSTADT

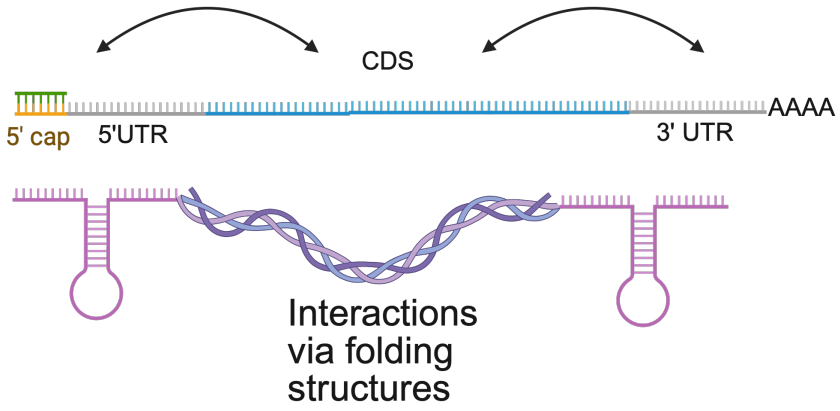
Thesis (B.Sc. / M.Sc.)
Stranger Strings: Uncovering Hidden CDS–UTR Interaction Patterns in cooperation with Merck KGaA

Department 18
Electrical Engineering and Information Technology
Self-Organizing Systems Lab

Prof. Dr. Heinz Koeppl
Head of lab

Background: Untranslated regions (UTRs) are key regulatory elements that shape RNA stability, localization, and translation. In practice, UTR design and analysis often treats the coding sequence (CDS) and UTR as largely independent despite growing evidence that sequence context and RNA structure can induce non-trivial *CDS–UTR interactions*. These interactions may influence regulatory motifs, accessibility of binding sites, and ultimately expression outcomes.

In this project, we develop deep learning methods to explicitly encode and analyze CDS–UTR interaction patterns. The goal is to move beyond independent sequence representations and learn joint, interaction-aware models that help us navigate the regulatory element design space.



Created with Biorender. RNA secondary structure can create long-range interactions between UTRs and the CDS, influencing translation efficiency and mRNA stability. We aim to learn these interaction patterns from data to guide the design of UTRs flanking a given CDS.

Thesis objective: You will design and evaluate machine learning models that capture *how* CDS and UTR sequences jointly determine regulatory behavior. The thesis combines model development with careful biological interpretation and quantitative evaluation.

Key tasks (depending on thesis level and interests):

- **Interaction-aware modeling:** Refine and perform experiments with our implemented deep learning model, which jointly represents CDS and UTR sequences to capture sequence-dependent binding and structural coupling.
- **Data pipeline & baselines:** Curate or generate paired CDS–UTR datasets, define prediction targets (e.g., expression/translation proxies), and implement strong baselines (independent encoders, k-mer models, conventional sequence models).

Philipp Froehlich
Project supervisor

S3|06 206
Merckstrasse 25
64283 Darmstadt

philipp.froehlich@tu-darmstadt.de
https://www.bcs.tu-darmstadt.de

December 25, 2025

- **Biological constraints and priors:** Integrate known mechanisms such as RNA-binding protein (RBP) motifs, predicted secondary structure, and accessibility signals to guide or validate the learned representations.
- **Exploration of the interaction space:** Use learned embeddings to map the landscape of CDS–UTR compatibility; identify recurring motifs, interaction signatures, and out-of-distribution failure modes.
- **Interpretability & validation:** Apply attribution methods, perturbation/ablation studies, and attention analyses to extract biologically meaningful hypotheses and sanity-check model behavior.

Secondary structure of original BioNTech BNT162b2 SARS-CoV2 vaccine (left) and our optimized SARS-CoV2 Vaccine generated via CDS-dependent UTR generation model (right).

Prerequisites:

- Solid foundation in machine learning / deep learning.
- Strong Python skills and experience with a DL framework (PyTorch or TensorFlow).
- Interest in bioinformatics / computational biology (prior biology knowledge is helpful but not required).
- Motivation to work in an interdisciplinary, research-oriented setting.

Nice to have (optional):

- Experience with sequence models (Transformers, CNN/RNN for sequences) or representation learning.

- Familiarity with genomics/RNA concepts (motifs, UTRs, translation, RNA structure) or willingness to learn quickly.

For further information, please contact Philipp Froehlich.