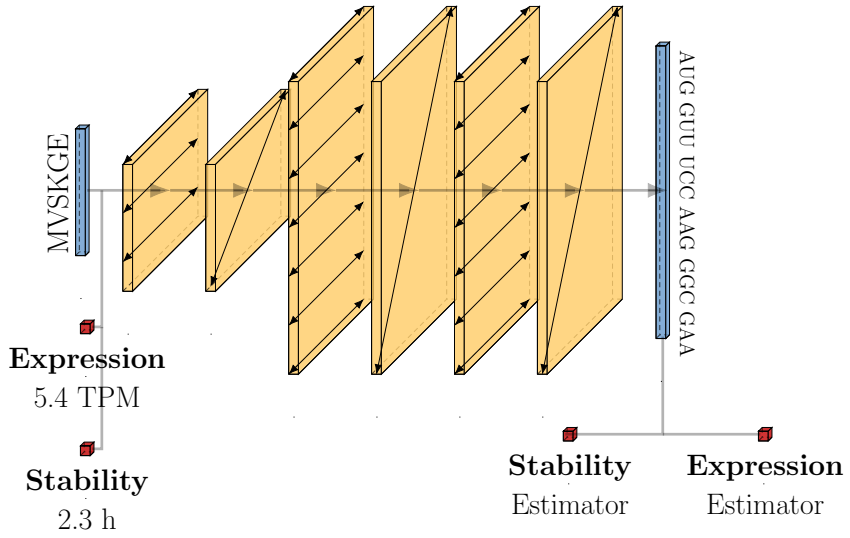
TECHNISCHE UNIVERSITÄT DARMSTADT

Thesis (B.Sc. / M.Sc.)
Surrogate-Based Sequence Design: Differentiable Optimization of mRNA Stability and Expression

Background: Rational sequence design is a core goal in synthetic biology: given a desired protein, we want to engineer a coding sequence (CDS) that yields strong expression while maintaining favorable RNA properties such as stability and accessible translation initiation regions. Two key bottlenecks are (i) evaluating sequence properties (e.g., folding energy) at scale and (ii) *optimizing* sequences under multiple, sometimes conflicting objectives.

In recent theses in our group, students have already developed *differentiable surrogate models* for two important targets: (i) a Transformer-based surrogate for RNA folding stability via Minimum Free Energy (MFE), trained to approximate RNAfold with high accuracy, and (ii) compact differentiable sequence-to-expression (S2E) models that predict protein abundance from CDS in Gram-negative bacteria. These surrogates enable fast inference *and* gradient-based optimization.

The next step is to turn these predictors into an *end-to-end design system*: a model that *proposes sequences* and uses gradients through the surrogate target models to iteratively improve them. This thesis focuses on building and evaluating such a surrogate-guided design pipeline for multi-objective sequence optimization (e.g., optimize expression while controlling MFE and sequence constraints).



Objective: Develop a differentiable sequence design framework that optimizes nucleotide sequences with respect to surrogate objectives such as RNA stability (MFE proxy) and predicted expression. You will build a design model and optimization loop that can take derivatives through the target surrogates, enabling efficient and controllable multi-objective design.

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

Prof. Dr. Heinz Koeppl
Head of lab

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

Key tasks (scope adaptable to B.Sc./M.Sc. level):

- **Formulate the design problem:** Define objective functions and constraints (e.g., maximize predicted expression; minimize/shape MFE globally or near the 5' region; enforce amino-acid identity; control GC-content; avoid forbidden motifs; restrict edit distance to a reference sequence).
- **Build a differentiable design approach:** Implement a sequence optimizer that can backpropagate through the surrogate models. Possible approaches include continuous relaxations (e.g., soft one-hot / Gumbel-Softmax), gradient-based editing, or a learnable generator trained with surrogate-guided objectives.
- **Multi-objective optimization and trade-offs:** Combine targets via weighted objectives, constrained optimization (Lagrangian methods), or Pareto-front exploration. Analyze how expression–stability trade-offs behave across genes and organisms.
- **Evaluation beyond surrogate scores:** Assess designs for biological plausibility and robustness: amino-acid preservation, codon usage statistics (e.g., CAI/tAI), motif/pathology checks (repeats, restriction sites), diversity vs. mode collapse, and (where possible) cross-checking designs with non-differentiable tools (e.g., RNAfold on a held-out subset) to detect surrogate exploitation.
- **Ablations and interpretability:** Study which objectives drive which sequence changes; identify failure modes (gradient hacking, unrealistic sequences) and implement mitigations (regularization, penalties, adversarial checks, uncertainty-aware optimization).

Expected outcomes: A reproducible design pipeline that generates optimized coding sequences under clear biological constraints, plus a systematic evaluation showing when gradient-based surrogate design yields improvements and where limitations remain (e.g., surrogate mismatch, bounded predictability of expression).

Prerequisites:

- Strong programming skills in Python; experience with PyTorch (preferred).
- Solid foundation in deep learning and optimization.
- Interest in biological sequences / synthetic biology (we provide domain onboarding).

Nice to have (optional):

- Familiarity with discrete optimization or differentiable relaxations (e.g., Gumbel-Softmax).
- Basic knowledge of codon usage, translation, and RNA secondary structure.

For further information, please contact Philipp Froehlich.