# A general machine-learning workflow identifies data requirements for functional insights from promoter libraries.

Strain engineering in biotechnology modifies metabolic pathways in microorganisms to overproduce target metabolites. To alter metabolic pathway activity, gene expression is an effective and easy manipulated target, specifically the promoter sequence that is recognized by sigma factors. Promoter libraries are generated to scan different promoter sequences for their activity and to identify predictive sequence positions for activity. To maximize the information retrieval, a well-designed experimental setup is required. We present a workflow to predict gene expression and to identify important sequence features for fine control of expression activity. The workflow is based on a Python Jupyter Notebook and covers (i) statistical sequence analysis, (ii) sequence-input to expression-output machine learning, (iii) estimator performance evaluation, and (iv) new sequence prediction with defined activity. The workflow can process various scenarios: multiple promoter libraries across species or reporter proteins and classification or regression. We tested the workflow on six promoter libraries and found that the strongest predictions were achieved when the promoters in the library were recognized by a single sigma factor. To maximize mechanistic interpretability, a high diversity of nucleotides tested at each position proved essential. In addition to recognizing established consensus sequences, we discovered unknown sequence positions to fine-tune expression. The workflow is open source and can be easily adapted to include various ML-strategies and to process sequence libraries from other expression related problems. Ultimately, this tool will support biotechnology with more efficient strain engineering.