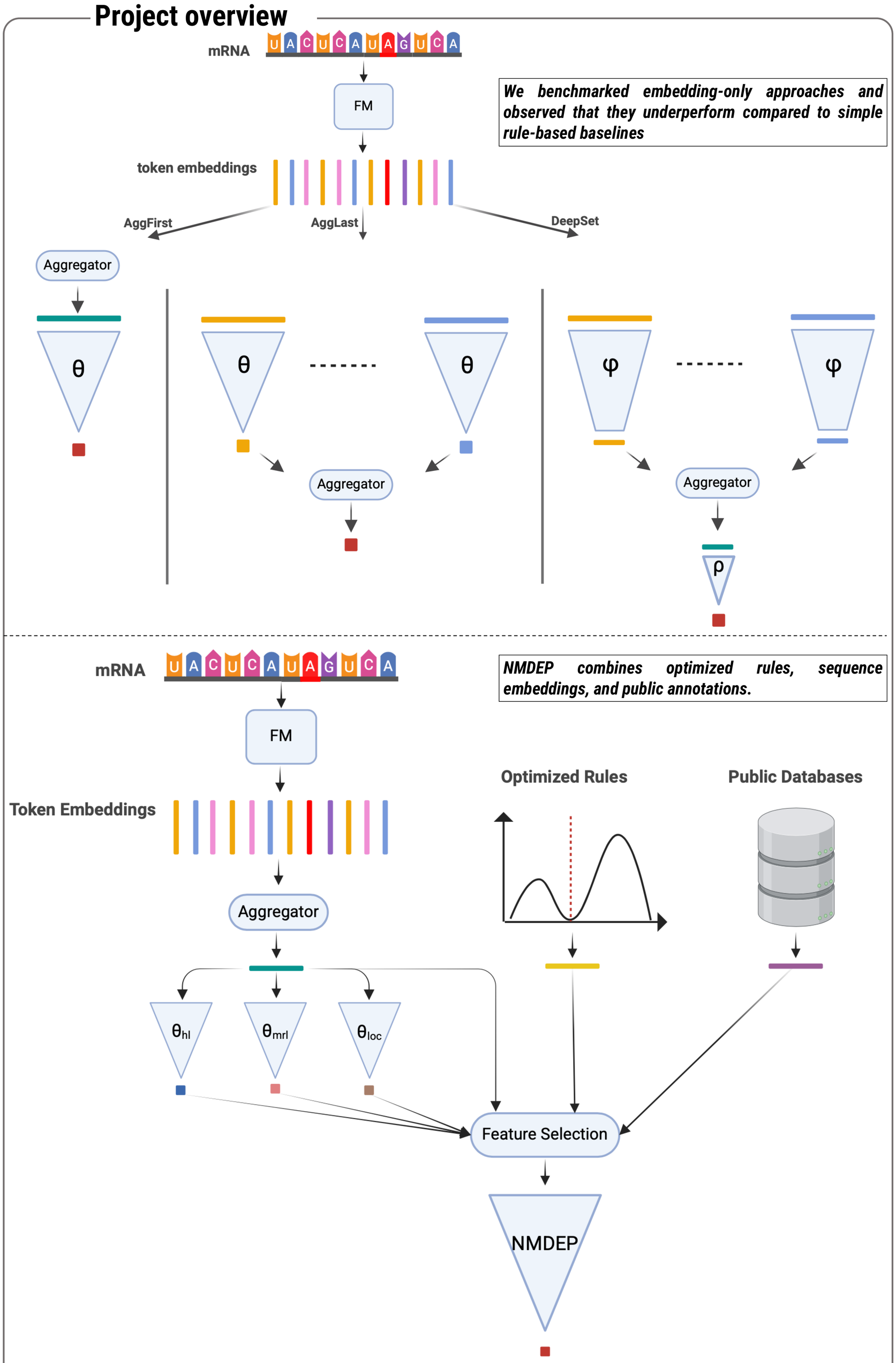
From mutation to degradation: predicting nonsense-mediate decay with NMDEP

Ali Saadat [1,2], Jacques Fellay [1,2,3]

[1] Global Health Institute, School of Life Sciences, EPFL, Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3] Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

Project overview



We benchmarked embedding-only approaches and observed that they underperform compared to simple rule-based baselines

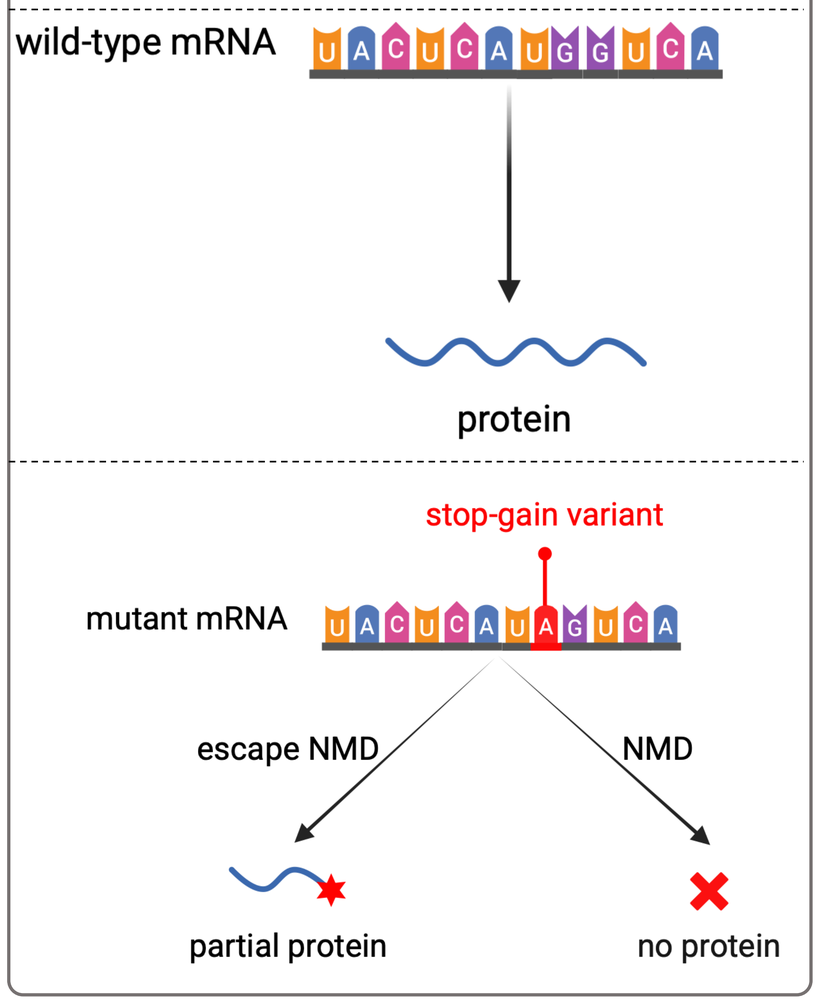
NMDEP combines optimized rules, sequence embeddings, and public annotations.

Model	MAE ↓	RMSE ↓	R^2 ↑	Spear. Corr ↑	Pear. Corr ↑
Baseline (4 rules)	0.8	1.07	0.35	0.67	0.6
4 rules optimized	0.75	0.98	0.41	0.71	0.66
Best of embedding-only models	0.89	1.16	0.18	0.48	0.45
Features from Kim et al. (2024)	0.78	1.06	0.3	0.63	0.56
NMDEP	**0.67**	**0.89**	**0.51**	**0.76**	**0.73**

NMDEP outperformed both rule-based and embedding-only models, achieving state-of-the-art accuracy in predicting NMD efficiency.

Background

Nonsense-mediated mRNA decay (NMD) is a critical post-transcriptional surveillance mechanism that degrades transcripts with premature termination codons, safeguarding transcriptome integrity and shaping disease phenotypes. However, accurately predicting NMD efficiency remains challenging, as existing models often rely on simplistic rule-based heuristics or limited feature sets, constraining their accuracy and generalizability.



Methods & Results

Data: We used paired DNA and RNA sequencing data from The Cancer Genome Atlas (TCGA) to quantify NMD efficiency for over 4,000 high-confidence stop-gain variants, calculated as the negative log-ratio of RNA to DNA variant allele frequencies.

Benchmarking: We benchmarked various token-to-sequence embedding aggregation strategies but found that embedding-only models underperformed compared to simple rule-based heuristics based on four binary features.

NMDEP: To improve the predictive performance, we optimized these heuristics via a two-step grid search and curated a comprehensive set of biological features. Building on this, we developed NMDEP, a machine learning framework that integrates optimized rules, functional annotations, and sequence embeddings.

Interpretation: Using explainable AI, we identified both known and novel determinants of NMD efficiency, enhancing model interpretability and biological insight.

Application: By applying NMDEP to 2.9 million simulated stop-gain variants, we enabled large-scale transcript degradation assessments and uncovered novel regulatory features driving NMD.

Some figures were created with BioRender.com