A069 · A Proteogenomic Machine Learning Approach to Evaluate Proteoform-Level Physiological Stability at Genome Scale

Presenter: Senbao Lu — Worcester Polytechnic Institute

Authors: Senbao Lu, Ziyang Gao, Oleksandr Narykov, Gloria Sheynkman, Dmitry Korkin

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Alternative splicing of mRNA precursors is a fundamental mechanism to expand the functional capabilities of the genome by generating multiple protein isoforms from a single gene, greatly diversifying protein function. Approximately 70% of alternatively spliced isoforms are predicted to induce functional or structural protein alterations, whereas others may be unstable and subject to degradation. To date, large-scale experimental characterization of isoform-level physiological stability remains limited, where we refer to the isoforms that are expressed in biologically relevant amounts at both transcript and protein levels as physiological stable. In this work, we construct a novel multi-source dataset of 3,754 paired reference and alternative isoforms with annotated physiological stability, derived from a harmonized proteogenomic integration of four experimental datasets. We then train an integrative proteogenomic machine learning framework that combines biologically derived features with foundation models-based embeddings at both transcript- and protein-levels to predict physiological stability of alternative protein isoforms and reach 83% cross-validated F1 score. We further perform orthogonal validation using proteomics experiments and wide-range temperature-series molecular dynamics (MD) simulations to assess thermal stability differences among isoforms from the same gene. Finally, we apply our model at genome scale. Analysis of the Human Protein Atlas (filtered at ≥20 TPM) identifies 7,376 stable and 8,216 unstable alternative isoforms, and analysis of GENCODE basic annotation reveals that, among 44,684 protein-coding alternative isoforms across 19,411 genes, 31,936 are predicted to be stable and 12,735 unstable, suggesting that the functional human proteome is substantially shaped by isoform-level stability constraints.

A Proteogenomic Machine Learning Approach to Evaluate Proteoform-Level Physiological Stability at Genome Scale

Senbao Lu[1,*], Ziyang Gao[1], Oleksandr Narykov, Ph.D.[2], Gloria Sheynkman, Ph.D[3,4,5,6], Dmitry Korkin, Ph.D.[1,2,7]
[1]Bioinformatics and Computational Biology Program, [2]Department of Computer Science, [7]Data Science Program, Worcester Polytechnic Institute, Worcester, MA, USA,
[3]Department of Molecular Physiology and Biological Physics, [4]Department of Biochemistry and Molecular Genetics, [5]Center for Public Health Genomics,[6]UVA Cancer Center, University of Virginia, Charlottesville, VA, USA

Alternative splicing of mRNA precursors is a fundamental mechanism to expand the functional capabilities of the genome by generating multiple protein isoforms from a single gene, greatly diversifying protein function. Approximately 70% of alternatively spliced isoforms are predicted to induce functional or structural protein alterations, whereas others may be unstable and subject to degradation. To date, large-scale experimental characterization of isoform-level physiological stability remains limited, where we refer to the isoforms that are expressed in biologically relevant amounts at both transcript and protein levels as physiological stable. In this work, we construct a novel multi-source dataset of 3,754 paired reference and alternative isoforms with annotated physiological stability, derived from a harmonized proteogenomic integration of four experimental datasets. We then train an integrative proteogenomic machine learning framework that combines biologically derived features with foundation models-based embeddings at both transcript- and protein-levels to predict physiological stability of alternative protein isoforms and reach 83% cross-validated F1 score. We further perform orthogonal validation using proteomics experiments and wide-range temperature-series molecular dynamics (MD) simulations to assess thermal stability differences among isoforms from the same gene. Finally, we apply our model at genome scale. Analysis of the Human Protein Atlas (filtered at ≥20 TPM) identifies 7,376 stable and 8,216 unstable alternative isoforms, and analysis of GENCODE basic annotation reveals that, among 44,684 protein-coding alternative isoforms across 19,411 genes, 31,936 are predicted to be stable and 12,735 unstable, suggesting that the functional human proteome is substantially shaped by isoform-level stability constraints.

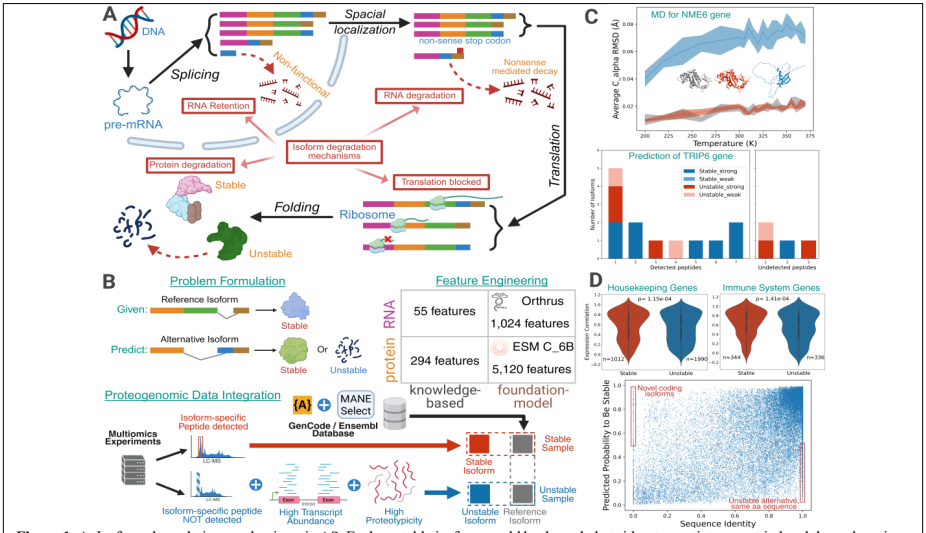

Figure 1: A. Isoform degradation mechanisms in AS. Each unstable isoform could be degraded at either transcript or protein level through various mechanisms. **B.** Top left: problem formulation of this work. Top right: Feature groups used in this work. Bottom: multi-source proteogenomic data integration pipeline. **C.** Top: MD analysis of NME6 gene shows reduced structural stability for unstable alternative isoform. Bottom: MS detects more peptides associated with stable than with unstable isoforms. **D.** Top: Stable alternative isoforms have higher transcript expression correlation with corresponding reference isoforms across 186 samples in HPA for housekeeping and immune system genes. Bottom: Analysis of GENCODE Basic reveals 1. stable alternative isoforms that have unique protein sequences (top left), and 2. unstable alternative isoforms that have identical protein sequences (bottom right).