PRISM: A Biologically-Consistent Machine Learning Framework for Protein Function Prediction and Drug Target Discovery

Anuar, Muhammad Faiz Bin (School: Kolej PERMATA Insan)

Purpose: Antimicrobial resistance causes 1.27 million deaths annually, yet 20–30% of proteins in pathogens remain uncharacterized. In Acinetobacter baumannii, the WHO's #1 priority pathogen, this gap conceals resistance mechanisms invisible to BLAST, InterProScan, HHpred, and Foldseek, which fail below 25% sequence identity. This research developed PRISM, a resource-efficient machine learning framework operable on consumer hardware to address this gap. Procedure: PRISM integrates ESM-2 protein language model embeddings with a Logic Enforcement Layer enforcing Gene Ontology True Path Rule constraints, reducing biological hallucinations by 99.8% ($p = 2.02 \times 10^{-10}$). Requiring no GPU clusters, PRISM screened 2,265 hypothetical proteins from three clinical isolates. The top prediction, D0CFU4, a 65-amino acid protein with zero publications, was validated through 34 independent analyses including pan-genomic screening (948 genomes), evolutionary selection analysis, and structural prediction using AlphaFold3 and RoseTTAFold All-Atom. Results: D0CFU4 was detected in 46 of 948 genomes (4.9%), harboring an invariant GCN5-related N-acetyltransferase (GNAT) catalytic motif under purifying selection ($\omega = 0.5977$) across 6 independent lineages spanning 42 years. D0CFU4 replaces DinB (DNA Polymerase IV) at the UmuDAb DNA damage response locus (operator $p = 4 \times 10^{-5}$). Structural simulations revealed cofactor-induced ordering upon Acetyl-CoA binding (+4.63 pLDDT; $p = 0.0005$). Conclusions: D0CFU4 represents a novel class of minimalist, stress-regulated acetyltransferases that replaced a canonical DNA damage gene and evaded detection for four decades. PRISM demonstrates that accessible AI can uncover hidden resistance targets in priority pathogens without expensive infrastructure.

Awards Won: