## INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

# Viral Threat Detection From Proteomics Data With Machine Learning

**Christopher Puglisi** Netrias **Amy Sims** Pacific Northwest National Laboratory **Isabelle O'Bryon** Pacific Northwest National Laboratory **Kristie Oxford** Pacific Northwest National Laboratory **Mohammed Eslami** Netrias

Emerging threats posed by biological agents are of critical concern for national security. Detection of an agent’s pathogenic potential represents a complex challenge that necessitates cross-domain solutions. Given recent advancements in machine learning (ML) techniques, researchers can now use these tools to advance the current state of biological threat detection. We seek to demonstrate that a combination of distinct modalities of data can be leveraged to train ML models that yield results otherwise impossible to glean from a single omics-type alone. These multi-omic models will enhance researcher’s understanding of the bio-agents, detect threats from different sources, and support the identification of new pathogenic biomarkers.

For this effort, we use ML for threat detection from proteomics data across three objectives chosen to simulate real-world challenges. Objective 1 aims to detect the presence of threat given a known viral agent in a known host. Objective 2 aims to detect the presence of threat given a known viral agent in an unknown host. Objective 3 aims to detect the presence of threat given an unknown viral agent.

The modeling framework employs state-of-the-art bioinformatics analysis for proteomics preprocessing and ML to learn profiles indicative of threat. This modular framework will be easily extended to other omics data types.  The framework automates quality control and normalization techniques to preprocess proteomics data and implements differential abundance to select the proteins for training the models. The framework leverages explainable models with interpretable feature importances like Logistic Regression and Random Forest classifiers to detect threats given the selected features. This research analyzes CARES Act liquid chromatography–mass spectrometry-based host-response proteomics in four treatment groups: NL63, CoV2-WA, CoV2_IT, and time-matched mock-infected cultures. For each of the 3 donors and four treatment groups, researchers collected 5 replicates at 6 time points for a total of 360 samples.

For Objective 1, the models detect a threat given a known viral agent with 99% accuracy. For Objective 2, the models successfully detect an agent's pathogenicity in one of the unknown hosts with 97.6% accuracy. For Objective 3, we withhold CoV2, training on NL63 and Mock, and reach 83.7% accuracy in detection of a “novel” threat. Withholding NL63 from training, we reach 93.6% accuracy. Further, we identify a critical window of time after infection in which proteomics samples are most indicative of threat. For viral threat detection, we demonstrate that the models perform best when features are selected using traditional bioinformatics analyses. These analyses identify differentially abundant features, which enhances model explainability and provides the most critical features of focus for the model. Therefore, this workflow leverages a combination ML and bioinformatics to detect emerging threats. Although analysis and modeling of proteomics is a viable approach for threat detection, additional data sources from other omics types are required to validate the generalizability of the model. This framework is not limited to viral agents, and is currently being used to explore threats in a variety of bacterial agents. This research leverages innovative cross-domain solutions that can be used to enhance the chem-bio defense landscape.