INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

AI-driven Harmonization

Alexander Verbitsky Netrias **Patrick Boutet** Netrias **Christopher Puglisi** Netrias **Mohammed Eslami** Netrias **Mark Weston** Netrias

The emerging potential of artificial intelligence (AI) to revolutionize threat detection through multi-omics data is currently hampered by siloed and inconsistent annotations in training datasets. These annotations are present in filenames, separate spreadsheets, or within the dataset itself (e.g. sample IDs or column headers). Current techniques use manual, one-off scripts to parse this information and standardize it to generate a training corpus. Data is now being collected at such a scale that these manual solutions are inadequate to supply models with new and updated training data in a timely manner. This limitation impedes detection of novel or engineered biological threats and development of effective countermeasures. This calls for a new capability to eliminate the biodefense bottleneck caused by unharmonized multi-omics data. By developing AI-driven harmonization techniques, we aim to streamline integration and analysis of diverse multi-omics datasets, thus accelerating the ability to generate new models for threat detection, biomarker discovery for pathogenicity and resistance, and enhancing diagnostic and therapeutic capacities against novel threats.

Our methods employ generative AI for metadata standardization as well as data annotation and integration. Key components include the use of language models for automatic metadata segmentation, annotation, and standardization. The approach resolves inconsistencies across diverse datasets and includes algorithms for the systematic annotation and integration of multi-omics data. First, if multiple entities are combined (e.g. a bacteria name and replicate ID), they are detected and segmented. Once segmented, the model infers the entity's type, and, finally, standardizes its representation. This creates a comprehensive data curation workflow that enriches datasets for more accessible and effective use in threat detection and analysis. Initial developments have demonstrated significant progress in standardizing metadata for multi-omics datasets, achieving over 90% accuracy, with successful applications to both bacterial and viral pathogens.

The impact of our research aligns directly with the JSTO mission to counter weapons of mass destruction and emerging threats. Through the rapid standardization and integration of multi-omics data, we enhance the Joint Force's capability for threat analysis, situational awareness, and the development of advanced countermeasures. This ensures operational readiness and protects the health of service members, sustaining the U.S. military's technological superiority in biodefense.