**Title**: Generating understanding and interpretation of multi-omics data with an automated and generalizable pipeline.

**Team**:

1. Samantha Erwin (AI Research)
2. Lisa Bramer (Domain Expert)
3. Daniel Claborne (AI Engineer)
4. Javier Flores (AI Research)

**Domain & Relevant Sector(s**): *Predictive Phenomics and ‘omics Discovery*

1. Karl Mueller - Basic Energy Sciences: Chem, Geo, and Biosciences

2. David Wunschel - National Security: Chem/Bio

**DC Reviewer:** Lauren Charles

**Specific Aims:** In this work, we will develop a protype of an automatic pipeline for prediction of leading mechanisms driving disease pathogenesis in a rapid, scalable approach using multi-omics data. The aim of this work is to demonstrate feasibility using generative AI (Gen AI) for discovery of host biological mechanisms from key features identified during multi-omics data harmonization.

**Domain Background and Impact**: Advances in instrumentation have led to increased and rapid collection of multi-omics data, paving the way for a more holistic understanding of biological systems and detection of mechanisms leading to pathogenesis. Increasing amounts of data drive a crucial need for leveraging recent advances in data harmonization methods and generative AI techniques to glean useful information rapidly. This current project team has already developed a scalable multi-omics data harmonization deep learning model from an adapted variational information bottleneck approach to identify key features from multi-omics data sets2. Furthermore, a generative AI-informed Bayesian model for phenomic prediction has been developed but requires seamless interpretation of predictive features. We will leverage the already developed pipelines and pre-existing datasets from lethal human viral infections (proteomics, lipidomics, transcriptomics, and metabolomics) as a proof-of-concept workflow to use generative AI to understand the key mechanisms and reactions driving the host biological response. While there are currently hand curated methods for elucidating mechanisms from multi-omics1, this work would demonstrate an automated and operationalizable pipeline that could be rapidly leveraged on a multitude of previously collected data for discovery of biological mechanisms. By demonstrating the feasibility of using Gen AI in this use case, we open the door for rapid analysis of multi-omics datasets, which are currently being collected across a multitude of domains.

**Generative AI Approach** (model type and how it produces output): To discover mechanisms leading to pathogenesis, we will first use our already developed multi-omics data harmonization deep learning model that identifies key features driving a phenotypic outcome. After identifying a subset of features from the multi-omics dataset, we will use Llama 3 and Retrieval Augmented Fine-Tuning (RAFT) to understand underlying ‘omics mechanisms leading to pathogenesis4. Specifically, we will use RAFT to fine tune Llama 3 to the textual corpora of 'omics-related literature harvested from publicly available abstracts in PubMed, Medline, and other sources. The resulting, updated Llama 3 model will then accept queries about relationships between the key features from our harmonization model as input and produce as output, descriptions of any potential relationships/mechanisms learned through the RAFT fine tuning on 'omics literature. Recent studies indicate that RAFT is an ideal fine-tuning method for domain specific problems and would be leveraged to elucidate in-host mechanisms across omics that cannot be traditional identified with classical statistical and mathematical techniques.

**Description of your data source**: Multi-omics data (proteomics, lipidomics, transcriptomics, and metabolomics) will be used from the PNNL work, *A compendium of multi-omics data illuminating host responses to lethal human virus infections,* Eisfeld *et. al.*, 2024. This data captures host response to infection generated from 45 individual experiments involving human viruses (*Orthomyxoviridae*, *Filoviridae*, *Flaviviridae*, and *Coronaviridae* families). In addition to the omics data, Llama 3 will be fine-tuned on ‘omics-related literature from publicly available sources as well as Wikipathways, which includes genes, proteins, and metabolites3.

**Metrics for success and mid-point go/no-go gate**: 30 day go/no-go: Fine tune Llama 3 with RAFT using ‘omics-related literature and pathways data. 60 day go/no-go: Apply the RAFT-tuned model to successfully demonstrate the ability of the model to yield (known) biological mechanisms.

**Funding Request**: $60k

**POC for Proposal**: [samantha.erwin@pnnl.gov](mailto:samantha.erwin@pnnl.gov)

1. Erwin, S et al. 2024; doi:10.1101/2023.09.06.556597
2. Lee, C & van der Schaar, M. 2021; doi:10.48550/2102.03014
3. Slenter DN et al. 2018; doi:10.1093/nar/gkx1064
4. Zhang, T et al. 2024; doi:10.48550/2403.10131