# Proposal 1: Virus Classification

**Motivation:**

* Some pathogens, particularly viruses without reference genomes, are hard to identify using sequence alignment techniques.
* Machine learning can help predict viral sequences from patient data that cannot be mapped to known human or pathogen genomes.
* There is a need for software specifically designed to predict these viral sequences accurately.

**Objective:**

The model classifies sequences into one of three categories:

* **Infectious viruses**: Infectious viruses are pathogens that can invade host cells, replicate, and cause diseases. These include well-known viruses like influenza, HIV, and SARS-CoV-2.
* **Endogenous retroviruses (ERVs)**: ERVs are ancient viral sequences that have integrated into the human genome over millions of years. They originated from retroviruses, which are a type of virus that can insert their genetic material into the host’s DNA.
* **Non-ERV human sequences**: Non-ERV human sequences refer to the standard parts of the human genome that do not originate from retroviruses. This includes all the DNA that makes up genes, regulatory elements, and other non-viral genetic material in human cells.

**Dataset Construction**:

* The viral sequences came from two sources:
  + **In-house viral reference genome database**: 377,009 sequences.
  + **NCBI Viral Genomes database**: 13,273 complete genome sequences.
  + Combined, this totaled **390,282 viral sequences**.
* **ERV sequences** were extracted from the human genome reference (GRCh38.p13), and they collected **614,316 ERV sequences**.
* **Non-ERV human sequences** came from:
  + The GRCh38.p13 human genome with ERV regions masked.
  + The NCBI Consensus Coding Sequence (CCDS) database, which included **33,408 sequences**.

**Sequence Feature Extraction**:

* Each DNA sequence was transformed into a **numerical vector** using the **K-tuple relative abundance** method, which extracts specific patterns and frequencies from the sequence.
* These features will be fed into the ML/NN model for classification.

**Performance Evaluation**:

* Model performance was evaluated using common metrics:
  + **Accuracy**: Proportion of correct predictions.
  + **Recall**: Sensitivity or the ability to correctly identify positive instances.
  + **Precision**: The ratio of correctly predicted positive observations.
  + **F1-score**: The harmonic mean of precision and recall.

Virus Classification offers a fast, accurate, and interpretable tool for viral sequence discovery, particularly in cases where the origin of a disease is unclear. Its ability to handle uncharacterized viral sequences, including ERVs, makes it a valuable asset for research into viral disease etiologies and broader species classification tasks.

# Proposal 2: Pathogen/Virus Classification using Multi-omics Integration

**Introduction**

The rise of multi-omics data has revolutionized our understanding of diseases, including viral pathogens. Initial attempts at virus prediction primarily leveraged classical machine learning algorithms, which, while effective for simpler datasets, fall short in capturing the intricate relationships between diverse omics data. To address this challenge, the integration of multi-omics data—spanning genomics, transcriptomics, proteomics, and metabolomics—offers a more comprehensive perspective on viral characteristics. However, traditional machine learning models lack the capacity to fully exploit the complex interactions inherent in such data.

Recent advances in **graph-based neural networks** have shown significant promise in modeling the intricate relationships within multi-modal data. Graph-based approaches, such as **Graph Convolutional Networks (GCNs)** and **Graph Attention Networks (GATs)**, offer a structured means to integrate multi-omics data, capturing both global and local topological patterns. Despite their potential, there is a lack of comprehensive benchmarking specifically assessing graph-based models for pathogen or virus-type classification, particularly in the context of multi-omics data.

**Motivation**

The complexity and heterogeneity of multi-omics data necessitate advanced modeling techniques that can capture both the individual and collective contributions of various biological layers. Motivated by the gap in graph-based methods for virus classification, our project aims to establish a comprehensive benchmarking framework. This effort will systematically evaluate different graph representation learning techniques for virus prediction across a wide array of omics datasets.

Our work seeks to bridge this gap by assessing how **graph neural networks (GNNs)**, such as GCNs, **GraphSAGE**, **GAT**, and **Graph Isomorphism Networks (GIN)**, perform when applied to multi-omics data. This approach will not only help improve virus classification methods but also advance our understanding of how different biological layers contribute to disease outcomes.

**Significance**

By developing a comprehensive benchmarking framework for virus classification using graph-based models and multi-omics data integration, this project aims to provide insights that will:

* Advance the state-of-the-art in virus prediction, moving beyond traditional machine learning approaches.
* Highlight the potential of **multi-omics integration** in disease research, with direct applications to diagnostics and therapeutic development.
* Offer a robust comparison of graph neural network architectures, enabling better model selection for similar tasks in healthcare and biomedical research.

**Methodology**

* 1. **Feature Selection**Effective feature selection is critical when dealing with high-dimensional multi-omics data. We will explore a hybrid approach that combines machine learning and statistical techniques for feature extraction and dimensionality reduction. Specific methods include:
* Recursive Feature Elimination (RFE): A machine learning technique that recursively removes the least important features, narrowing down the dataset to the most predictive features.
* SelectFromModel: Leverages feature importance scores from models such as random forests or support vector machines to select relevant features.
* Boruta: A wrapper algorithm designed for feature selection based on random forests.
* Lasso (Least Absolute Shrinkage and Selection Operator): A linear model that uses L1 regularization to encourage sparse solutions, effectively selecting a subset of the most important features.
  1. **Network Generation**Graph construction plays a crucial role in translating multi-omics data into a format suitable for graph-based models. We will use cosine similarity to calculate similarity scores between features, creating an adjacency matrix that defines the connections between nodes (features). To ensure the graph structure reflects meaningful biological relationships, we will apply a similarity score cutoff to prune weaker connections. The resulting graph will represent the multi-modal interactions among different biological entities (e.g., genes, proteins, metabolites).
  2. **Graph-based Modeling**To explore the potential of graph-based models in virus classification, we will implement several state-of-the-art graph neural networks (GNNs):
* Graph Convolutional Networks (GCNs): These models use convolutional layers to aggregate and update node embeddings, enabling the network to capture local graph structure and enhance virus classification.
* GraphSAGE (Graph Sample and Aggregation): By sampling neighboring nodes and aggregating their features, GraphSAGE can handle large-scale graphs and provide a broader view of the data.
* Graph Attention Networks (GATs): These models apply attention mechanisms to weigh the importance of neighboring nodes, allowing for more precise information aggregation from relevant nodes in the graph.
* Graph Isomorphism Networks (GINs): Known for their ability to distinguish between non-isomorphic graphs, GINs capture complex topological features and are particularly well-suited for modeling subtle structural differences in multi-omics networks.
* Topology Adaptive Graph Convolution Networks (TAGCNs): These networks adapt their convolutional kernels to the graph's topology, ensuring that the learned features are sensitive to the structure of the underlying data.
  1. **Message Passing and Node Aggregation**Message passing between nodes is at the core of GNNs. During this process, information from neighboring nodes is aggregated to update the representation of each node. The quality of this aggregation process directly affects the model's ability to learn meaningful patterns from the graph.
  2. **Loss Function and Training**The primary objective will be multi-class virus-type classification, for which we will use cross-entropy loss. This loss function is well-suited for tasks where the output consists of multiple classes, allowing for a probabilistic interpretation of the predictions.

**Expected Outcomes**This project aims to deliver a set of best practices and recommendations for using GNNs in virus classification tasks, with broader implications for other disease research. By establishing a comprehensive benchmark, we hope to shed light on the trade-offs between different GNN architectures, the role of feature selection, and the impact of network construction methods. The final outcome will provide a roadmap for researchers and clinicians aiming to leverage multi-omics data in pathogen discovery and virus classification.

A screenshot of a computer

Description automatically generated