INFO290 - Spring 2025
Final Project Proposal

A knowledge graph framework for decoding microbial social networks

Microbial communities underpin critical biological processes in human health, environmental sustainability, and industrial biotechnology. However, existing tools for analyzing interspecies interactions lack the capacity to model multi-step metabolic dependencies or generate interpretable hypotheses for community engineering. Current approaches often overlook the complex network effects that define microbial ecosystems, while large language models (LLMs) struggle to incorporate domain-specific biochemical knowledge. To address these gaps, we propose the use of retrieval-augmented generation (RAG) within a Neo4j knowledge graph architecture. This project aims to transform how researchers analyze and engineer microbial consortia by providing a unified platform for context-aware reasoning and hypothesis generation.

The methodology follows a three-phase workflow. First, raw data from metabolic models and literature will be preprocessed into a standardized format and loaded into Neo4j, creating a queryable knowledge graph. Second, a hybrid embedding engine will generate unified representations of nodes, enabling similarity searches and subgraph retrieval. The framework will map nodes (microbes and metabolites) and edges (e.g., PRODUCES, CONSUMES) to represent metabolic networks, enhanced by hybrid embeddings combining structural graph features (via Node2Vec) and semantic text embeddings of biochemical metadata. Finally, a RAG pipeline will translate natural language questions (e.g., "Which microbes support butyrate production?") into Cypher queries, retrieve relevant subgraphs, and synthesize explanations using a domain-tuned LLM.

Biologically, the project will identify cross-feeding pathways and quantify metabolic trade-offs critical for microbial community networks. By bridging graph knowledge with systems biology, this work offers a paradigm shift in microbiome analysis, enabling researchers to navigate microbial networks with the nuance of ecological theory and the precision of AI-driven insights.