APPLIED MATHEMATICS SEMINAR SERIES: Enhancing Machine Learning Models with Biomedical Knowledge Graph and Graph Embeddings

Mary Silva
Lawrence Livermore National Labs

Date:

3/19/2024

Time:

3:00 PM - 4:15 PM

Location:

COB2 170

About The Speaker: Mary Silva is a data scientist within the Global Security Computing Applications Division. She Joined the lab in 2020 by contributing statistics and machine learning skills towards COVID-19 related tasks, including vaccine development and identification of drug targets. In addition to being an Ambassador for Livermore Women in Data Science and the Student Outreach lead for the DSSI program, she contributes research towards the efforts of antibody developability within the Generative Unconstrained Intelligent Drug Engineering group. Her research lies at the intersection of bioinformatics and machine learning, focusing on creating pipelines to incorporate experimental data, deep sequencing data, and protein structural information into ML models.

Abstract: The integration of machine learning techniques into biological research has introduced many new challenges, particularly for tasks such as drug safety assessment, genetic interaction modeling, and a spectrum of other biological inquiries. The accuracy and effectiveness of these predictive ML models is fundamentally linked to the quality of the data collected, and the model's ability to discern underlying patterns and salient features in the data. In order to enhance predictive models, we introduce the knowledge graph known as the Scalable Precision Medicine Open Knowledge Engine (SPOKE), which was originally developed by UC San Francisco from a collection of multiple scientific databases. The SPOKE knowledge graph contains compound, protein, gene ontology pathway and biological process, and many more biologic node types. Additionally, the graph contains observed or experimental evidence to support the edge relationships. We utilize the knowledge graph to generate graphical embeddings and have additionally introduced large language models to analyze the supplementary information extracted from the graph itself. These embedding methods have proven instrumental in enhancing the precision of our prediction, notably in identifying drug compounds associated with liver toxicity. By harnessing the graphical side information, we demonstrate a marked improvement over traditional approaches that rely solely on molecular and experimentally tested data. Furthermore, the application of graphical embeddings has significantly improved our ability to accurately model genetic interactions, demonstrating the value of graph-based data augmentation in ML tasks.

For more information, contact: Maxime Theillard
mtheillard@ucmerced.edu