Edwin Ventura, Jase Straka, Bruce Nguyen, Michael Niescier, Diane Lagare, Sedem Kakrada, Ashley Garcia

Elanco Symposium '24

PURDUE UNIVERSITY | The Data Mine

Elanco™

Project Overview

Elanco Animal Health is a global animal health company that develops products and knowledge services to prevent and treat disease in food animals and pets in more than 90 countries. Elanco provides comprehensive products and services to improve animal health and food-animal production in more than 70 species.

Our project is to utilize **machine learning, knowledge graphs, Python programming, and DuckDB database management** to effectively improve the identification, prioritization and validation of drug targets. The main focus of the project is to leverage systematic human biomedical data and orthologous relationship between human and animal genes, more specifically cats and dogs, to identify and prioritize disease targets.

To achieve our goal, the project the following steps were taken:

- Data Organization:

The team organized diverse biomedical data entities (molecules, diseases, drugs) into a coherent and interconnected graph structure. This structure representation allowed for the visualization of the discoveries and analysis made.

- Relationship Mapping:

Utilizing Python and DuckDB, the team mapped out the intricate relationships between entities for comprehensive analysis. This includes the establishment of associations between drugs and diseases, and identifying potential drug targets which can further allow for valuable insights into the therapeutic pathways and treatment modalities.

With this goal in mind, our team set out to analyze the various datasets and make relations through them, then implementing and portraying these discoveries into knowledge graphs. Through the utilization of our knowledge of knowledge graphs and advanced database management techniques, our project aims to make the process of identifying and validating drug targets more effective. Thus, our vision is to enhance the effectiveness and efficiency of drug discovery by bridging the gap between human and animal data, particularly focusing on cats and dogs. Through our efforts, we seek to not only make the process more efficient but also make groundbreaking analyses.

Project Scope

- **Knowledge Graphs for Biomedical Discovery:**

Definition: Organize data from multiple sources, and capture information about entities of interest in a given domain or task (like people, places or events), and forge connections between them.
Usefulness: Allows for the identification of complex patterns and insights within the data, Allowing for new discoveries to be made as analysis is done.

- **Project:**

Multi-model biochemical data offers opportunities for disease target identification and drug discovery to further the understanding of the interconnectedness of the biological data and explore the biomedical relationship between humans and animals.

Research

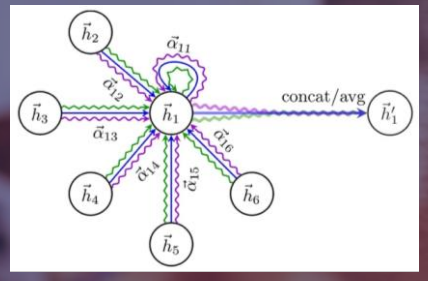
Graphical Neural Networks

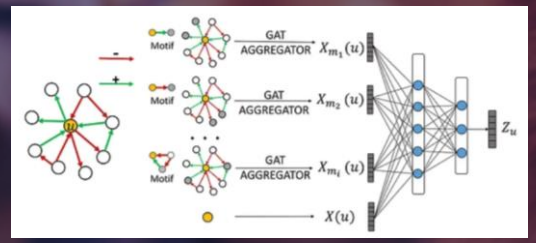
- Graph Attention Networks (GATs):

Incorporates attention mechanisms to capture important relationships between nodes in a graph. Assign attention weights to the neighboring nodes of a target node, allowing the network to focus on the most relevant neighbors during the information propagation process.

- Graph Convolutional Network (GCN):

A specific type of GNN that uses convolutional operations to propagate information between nodes in a graph. Convolutional operation is commonly used in image processing adapted to the graph domain.





Datasets

Datasets

- Molecules (Drugs): Contains information about chemical compounds under investigation or identified as potential drugs.
- Targets (Genes): Contains information about specific biological targets (biotypes, homologues, proteins, pathways) associated with diseases in humans and animals.
- Diseases: Contains information about various diseases affecting humans and animals, including their characteristics, symptoms, and therapeutic areas.

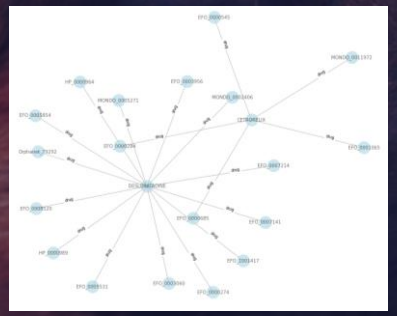
The various datasets serve as a foundation for the implementation of knowledge graphs thus to make things more efficient the steps taken were:

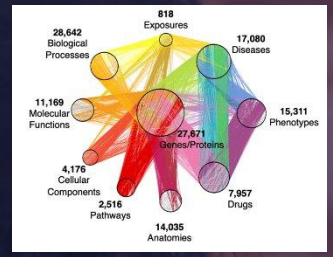
- Data Merging:

Data from the datasets were stored in separate Parquet files, a total of 200 Parquet files in each. The Parquet files were merged into three separate CSV files named 'drugs.csv', 'targets.csv', and 'diseases.csv'.

- Relating Datasets:

Python and DuckDB were implemented to store and query the merged datasets. Data was loaded into database tables which could then be analyzed through the execution of SQL.





Conclusion

Our goal for our data was to discover connections between the various datasets, all while taking a closer look at the drugs that target specific genes and discovering which genes are affected by certain diseases. Through this analysis, we could then use this data to prescribe drugs to animals suffering from said diseases. Using the diseases datasets, we queried the desired diseases, such as diabetes and osteoarthritis, finding their linked diseases, ancestors, descendants, and everything else we could find. To make things more efficient through the combining of the drugs and target datasets we hoped to attain a broader picture of the drugs that can treat it and how these connections can be made with the diseases on animals. Through relationship mapping, we establish associations between drugs, diseases, and potential drug targets, utilizing techniques like Graph Attention Networks (GATs) and Graph Convolutional Networks (GCNs). All in all, our approach involves merging datasets on molecules, targets, and diseases, storing them in a unified database for efficient querying and analysis, and ultimately aiming to bridge the gap between human and animal data for groundbreaking drug discovery and treatment modalities.

Future Plans/Research

- Develop and optimize Graph Convolutional Network (GCN) and Graph Attention Network (GAT) models to predict animal disease treatments, validating predictions by comparing them with existing literature and clinical trial data.
- Enhance knowledge graphs with integrated datasets on drug interactions, pathways, gene expression, and protein-protein interactions.
- Create a user-friendly interface for exploring and visualizing potential treatments through the incorporation of interactive visualization tools.

References

- Importing data. DuckDB. (2024, April 3). https://duckdb.org/docs/data/overview
- Kumar, A. (2024, February 20). Graph attention networks. Medium. https://medium.com/@ashish28082002.ak/graph-attention-networks-52f03591b3cc
- Learning pytorch with examples. Learning PyTorch with Examples - PyTorch Tutorials 2.2.1+cu121 documentation. (n.d.). https://pytorch.org/tutorials/beginner/pytorch_with_examples.html

Acknowledgements

- **The Data Mine**

David Glass
Kevin Amstutz
Cai Chen
Suyash Johari
Victor Penedo

- **Elanco**

Shrinivas Mane, PhD
Matthew Mazalouskas, PhD