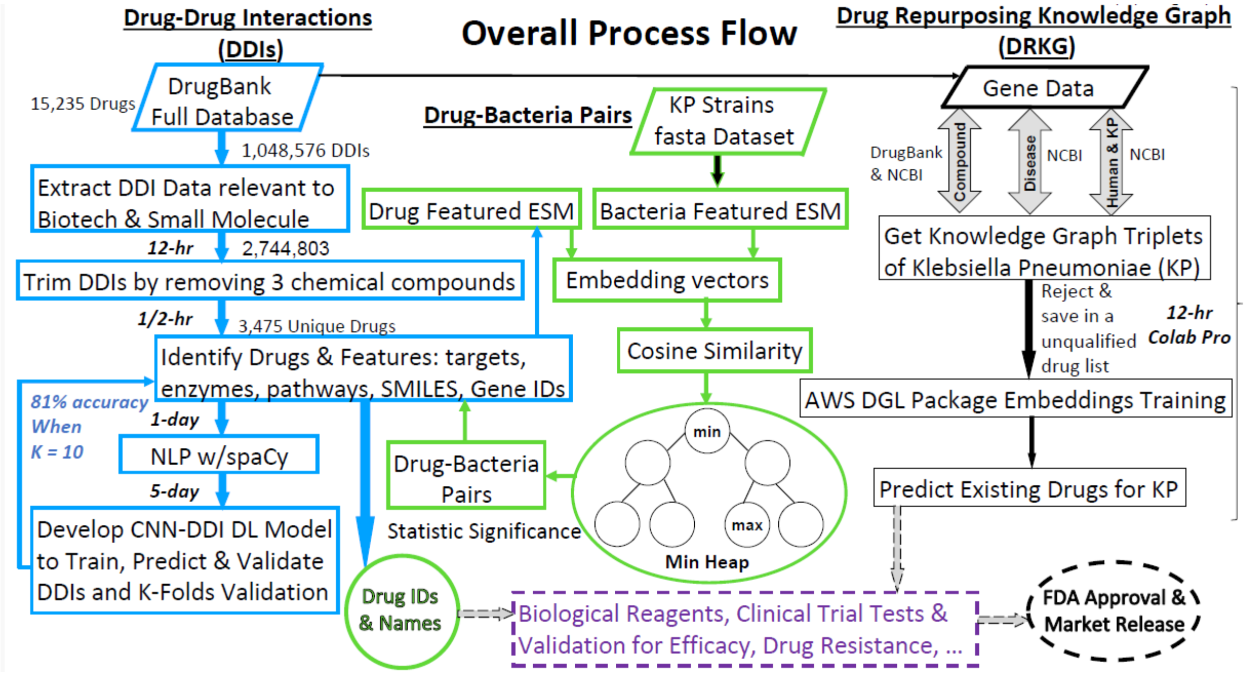
Overview:

This project employs deep learning techniques to accelerate drug discovery for combating *Klebsiella pneumoniae* (KP) infections. KP, a highly antibiotic-resistant pathogen, poses a significant global health challenge. Our approach integrates convolutional neural networks (CNN) and evolutionary scale modeling (ESM) to identify promising drug candidates from a dataset of biotech and small-molecule drugs extracted from DrugBank.

The figure below illustrates the overall process flow, including DDIs and unique drug extraction, NLP, CNN modeling, ESM with drug-bacteria similarity analysis, and the Drug Repurposing Knowledge Graph (DRKG), used in the deep learning framework to identify potential drug candidates against Klebsiella pneumoniae (KP).


Drug-Drug Interactions (DDIs)
Overall Process Flow
Drug Repurposing Knowledge Graph (DRKG)
15,235 Drugs
DrugBank Full Database
1,048,576 DDIs
Extract DDI Data relevant to Biotech & Small Molecule
12-hr
2,744,803
Trim DDIs by removing 3 chemical compounds
1/2-hr
3,475 Unique Drugs
Identify Drugs & Features: targets, enzymes, pathways, SMILES, Gene IDs
81% accuracy When K = 10
1-day
NLP w/spaCy
5-day
Develop CNN-DDI DL Model to Train, Predict & Validate DDIs and K-Folds Validation
Drug-Bacteria Pairs
KP Strains fasta Dataset
Drug Featured ESM
Bacteria Featured ESM
Embedding vectors
Cosine Similarity
min
max
Min Heap
Drug-Bacteria Pairs
Statistic Significance
Drug IDs & Names
Gene Data
DrugBank & NCBI
Compound
Disease
NCBI
Human & KP
NCBI
Get Knowledge Graph Triplets of Klebsiella Pneumoniae (KP)
Reject & save in a unqualified drug list
12-hr Colab Pro
AWS DGL Package Embeddings Training
Predict Existing Drugs for KP
Biological Reagents, Clinical Trial Tests & Validation for Efficacy, Drug Resistance, ...
FDA Approval & Market Release

Dataset

1. DrugBank 5.1.11 version
   https://go.drugbank.com/releases/latest
2. Klebsiella Pneumonia (KP) strain fasta dataset
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9769640/pdf/spectrum.02306-22.pdf
3. National Center for Biotechnology Information (NCBI) databases

Key Features

- Data-Driven Drug Discovery: Utilizes 3,475 biotech and small-molecule drugs from DrugBank, focusing on potential candidates for KP treatment.
- Deep Learning Models:
    - Convolutional Neural Networks (CNN): Applied for classification and interaction prediction.
    - Evolutionary Scale Modeling (ESM): Used for similarity analysis of drug molecules to KP strains.
- Drug-Drug Interaction Analysis: Incorporates drug-drug interaction (DDI) data for improved predictions.
- Similarity Metrics: Employs cosine similarity to identify candidates with >85% molecular similarity to KP strains.

Results

1. Accuracy: Achieved approximately 72% validation accuracy for the CNN model.
2. Candidate Identification: Five promising drugs were identified with molecular similarities exceeding 85% to KP strains.
3. Efficiency: Reduced the drug discovery timeline from years to months.
4. Robustness: PairRE outperformed other models across evaluation metrics for predictive ranking.

Significance

This project provides a scalable framework for drug discovery, offering a faster and more efficient alternative to traditional laboratory-based approaches. It holds potential for addressing bacterial pathogens like KP, aiding in effective treatment and minimizing side effects.

Repository Contents

- **/data/**: Contains DrugBank-derived dataset used for training and evaluation.
- **/models/**: Includes implementations of CNN and ESM models.
- **/scripts/**: Python scripts for data preprocessing, training, and evaluation.
- **/results/**: Output files, including accuracy metrics and identified drug candidates.
- **README.md**: This file, describing the project.