Predicting Drug-Drug Interactions

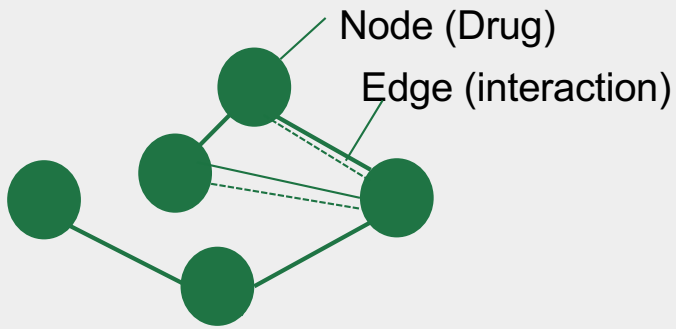
Jay Russell

Problem Statement

The discovery of Drug to Drug interactions (DDIs) are dependent on clinical trials and resourcefully intensive research. The surveillance of drug safety profiles is difficult and time consuming, leaving patients prone to risk of an adverse event if not addressed/ discovered in time.

In this project, I explore computational methods to predict unknown DDIs using Deep Learning Neural Networks and other machine learning models such as XGboost and Logistic Regression.

What is link prediction?



Methodology

Data Preparation

Hosting SQL server, cleaning data, achieving comparability

Preprocessing

Neural networks require certain formatting

Optimization

Adjusting hyperparameters to get best results

EDA

Insight on the types of DDIs, chemical structures...

Modeling

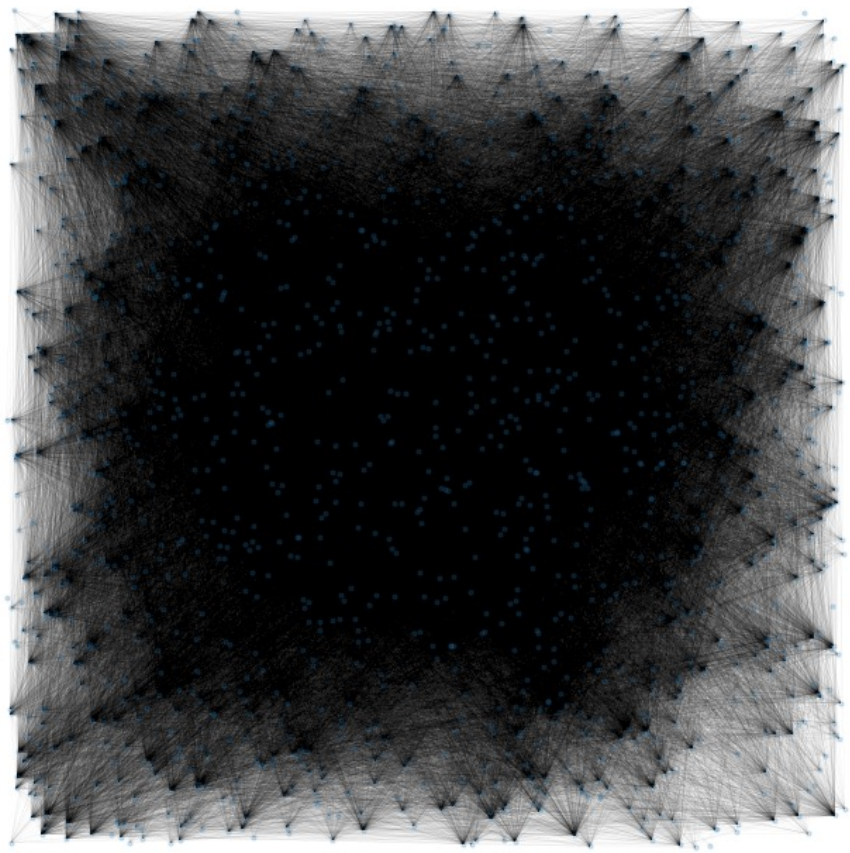
Neural Network, XGBoost, Logistic Regression

Overview of dataset

This

drug_a	drug_b
Naltrexone	Abacavir
Abacavir	Orlistat
Aluminum hydroxide	Dolutegravir
Aprepitant	Dolutegravir
Attapulgite	Dolutegravir

Turns into this



Formulating the problem

Train

All known interactions – 20,000

T → T1

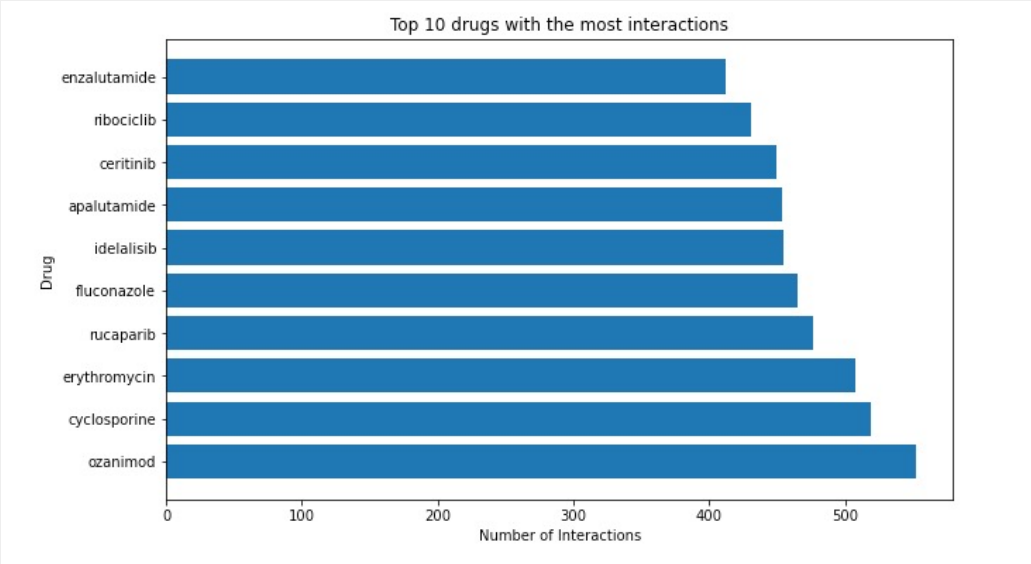
Test

All known interractions

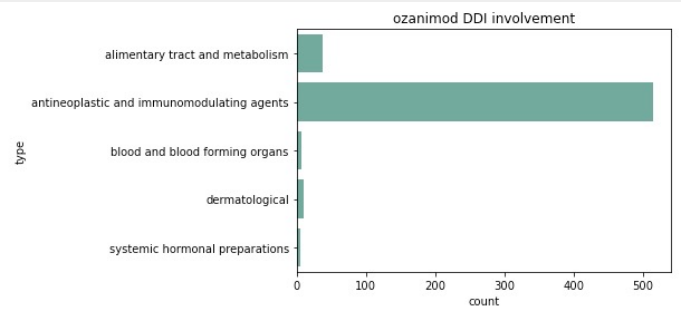
Neural Network

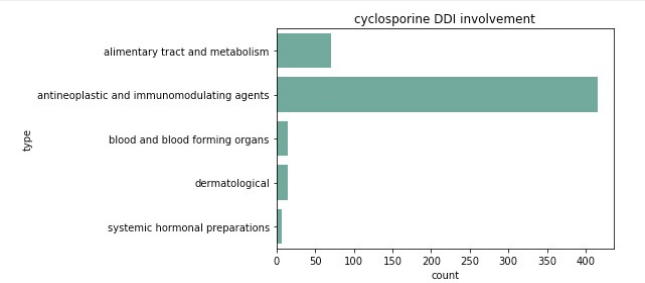
Exploratory Data Analysis

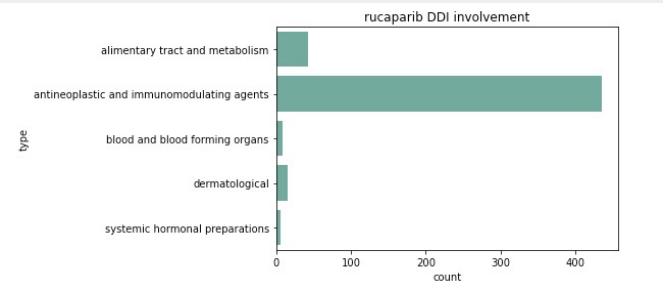
Most interactive drugs

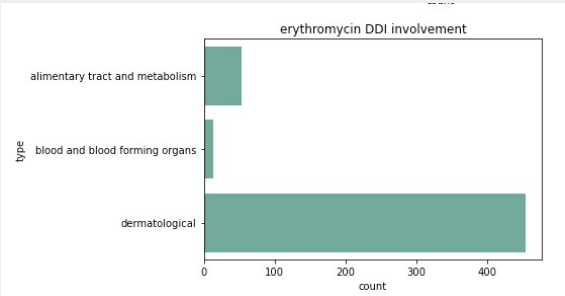


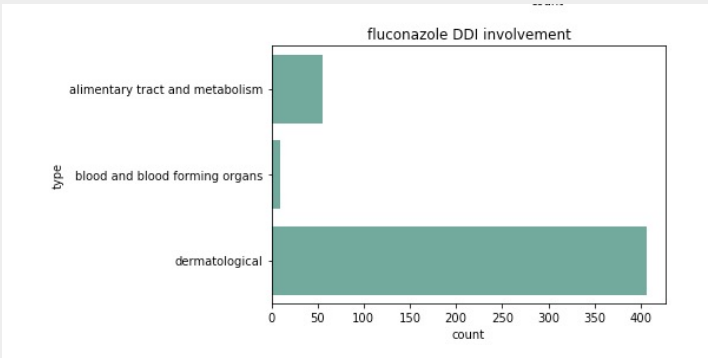
Profiles of most interactive drugs

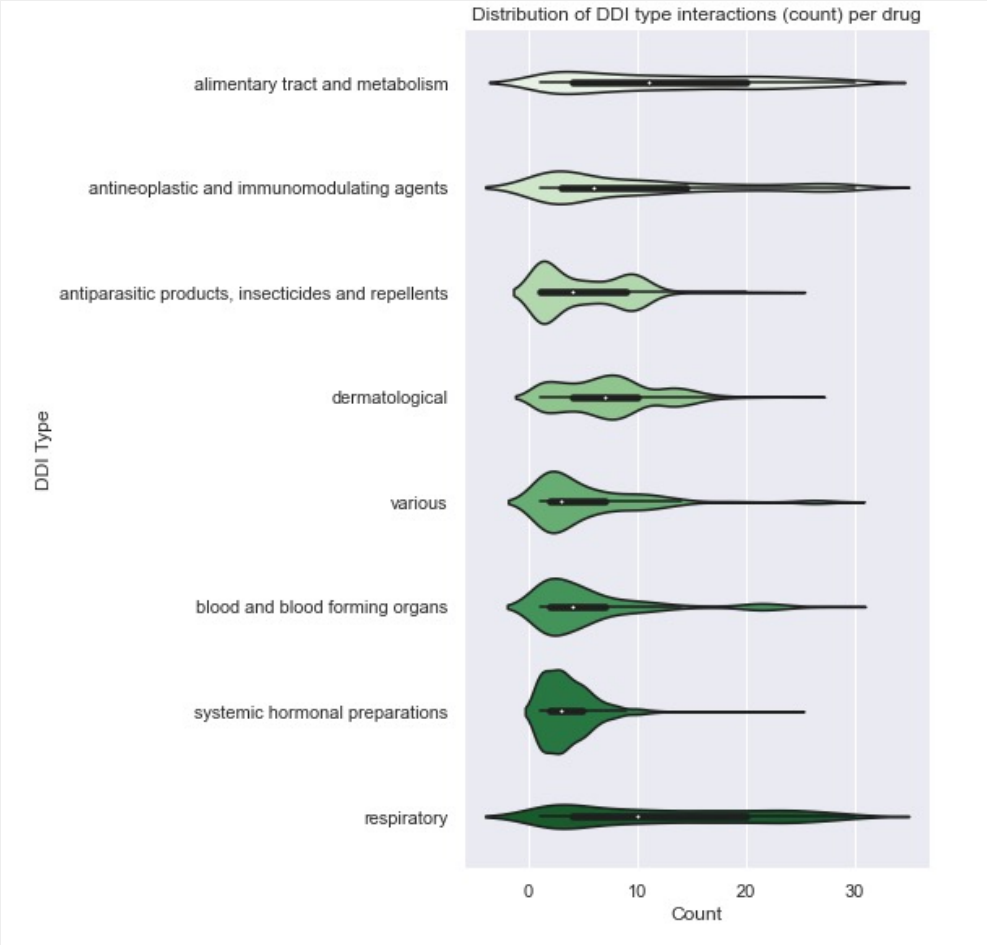











Distribution of DDI type interactions (count) per drug
alimentary tract and metabolism
antineoplastic and immunomodulating agents
antiparasitic products, insecticides and repellents
dermatological
various
blood and blood forming organs
systemic hormonal preparations
respiratory
DDI Type
0
10
20
30
Count

Types of Drug interactions

type	interactions_sum
antineoplastic and immunomodulating agents	30168
alimentary tract and metabolism	28964
respiratory	9836
dermatological	7157
blood and blood forming organs	4962
various	3242
antiparasitic products, insecticides and repellents	2378
systemic hormonal preparations	2301

Feature Engineering

Chemical Similarities

01

Data Extraction

ChEMBL

02

Achieving comparability

Reduced drug portfolio

03

Tokenization

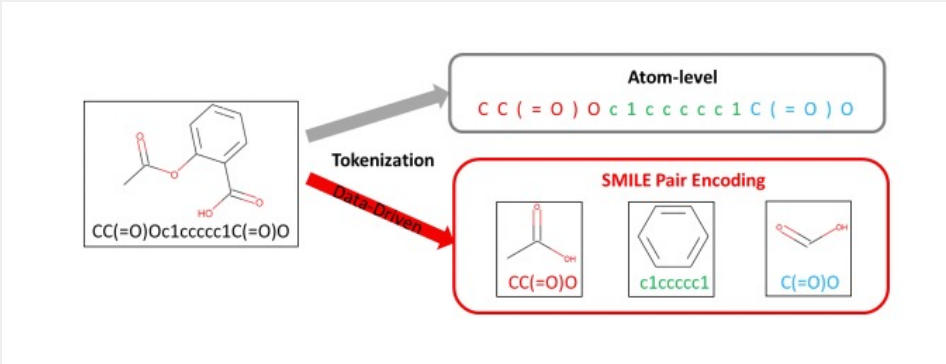
You can describe the topic of the section here

04

Feature engineering

Pairwise distance

SMILES tokenisation



Graph features

Jacard

Acadamic Adar

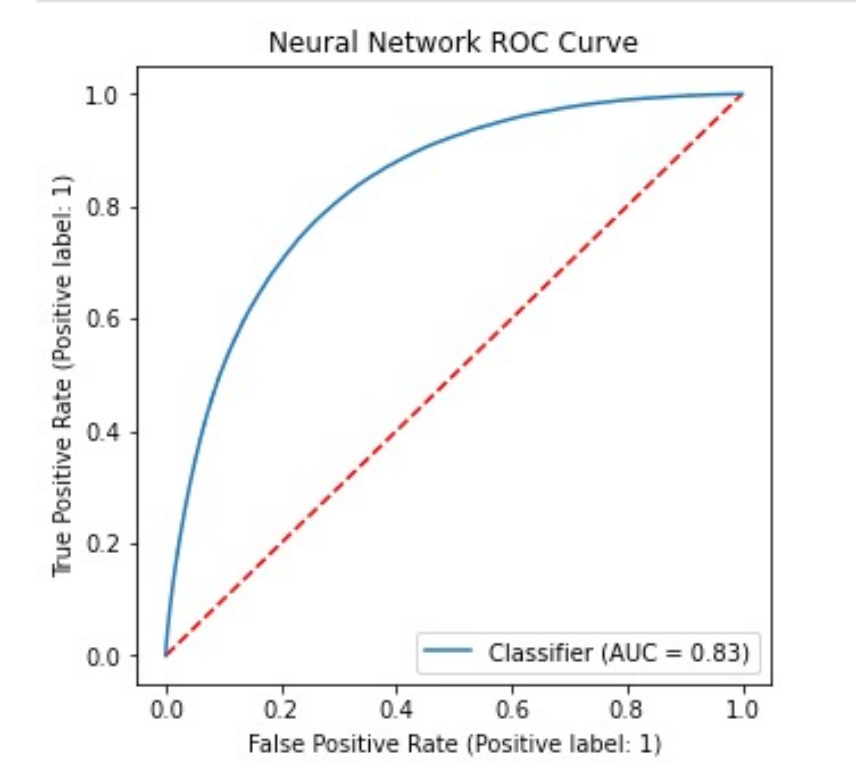
Pairwise Distance

Results

Results

	Train Accuracy	Test Accuracy	Train Loss	Test Loss
XGBoost	0.63	0.77	0.53	0.42
NN	0.77	0.76	0.66	0.65
Logistic Regression	0.76	0.82	0.41	0.40
Baseline	0.75	0.64	0.48	0.52

Area Under Curve (AUC)



AUC: 0.85

CONCLUSIONS

1

Exploring other methodologies

2

Applying to completely novel drugs