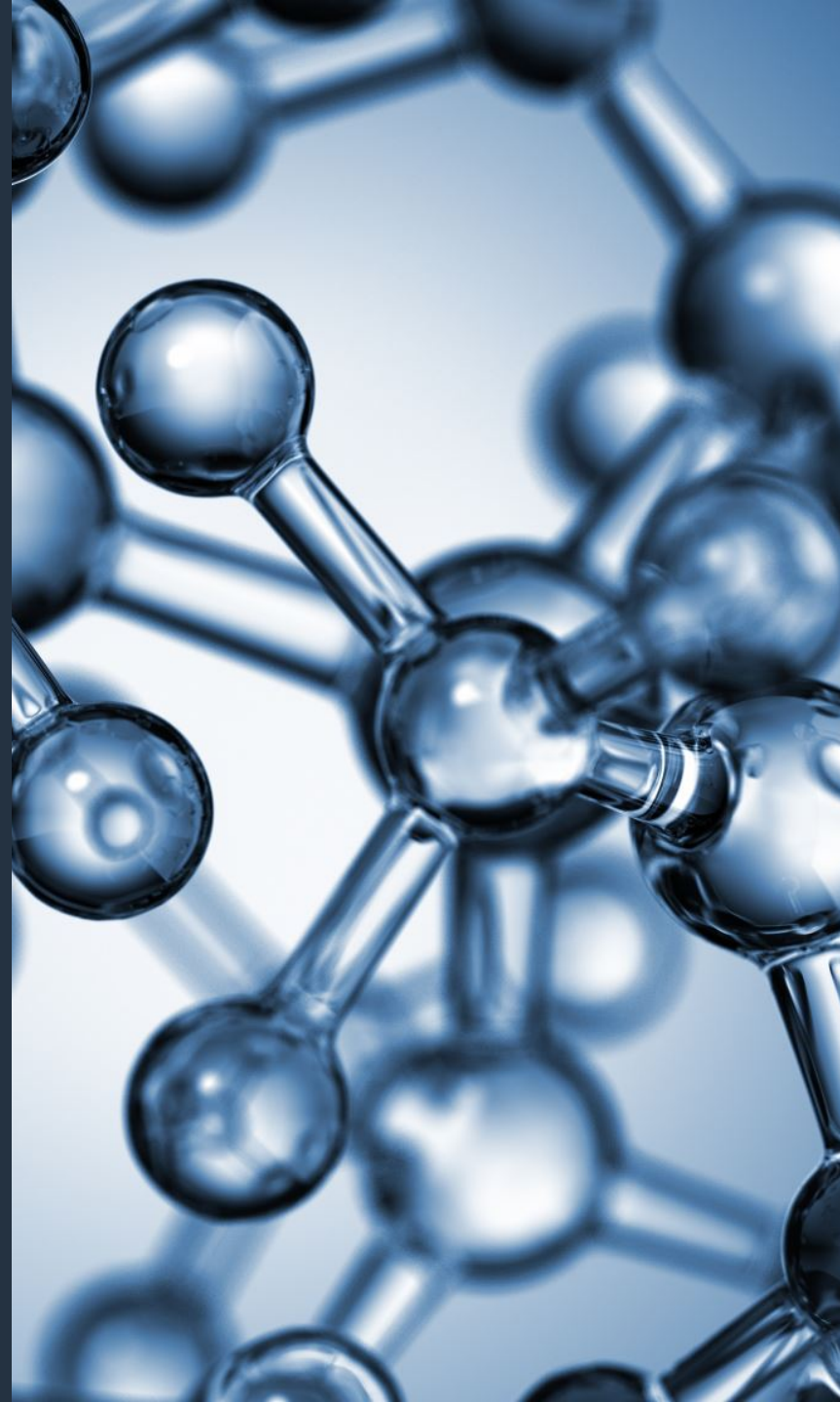
PREDICTION OF MOLECULE INTERACTIONS

GURUNADH PARINANDI

OCTOBER 28, 2020



MOTIVATION

- Pharmaceutical Industry had an economic output of $1.3 Trillion Dollars in 2015.
  - Molecule interactions are used to predict new drugs.
  - Drug Repurposing.
- Understanding molecule interactions is useful in the field of bioenergy.
- Molecule Interactions have given way to new computer algorithms and techniques in Data Science (high throughput sequencing).

OBJECTIVE

- Given it's chemical properties, predict if a molecule binds to a known target site using known Machine Learning Algorithms.
- Understand what features are most important in predicting if a molecule binds to a target site.

METHODOLOGY

- Utilized ChEMBL database of known bioactive molecules (https://www.ebi.ac.uk/chembl).
  - 6,900 Compounds
  - 9,800 Activities.
  - Database size is 1.5 GB
  - Library of Academic Papers that utilized database.
- Created personal PostgreSQL database to store data.
  - Had to utilize SQL schema to create working dataset.
  - Data curation took more time than I originally anticipated.
- Logistic Regression used for both prediction and inference.

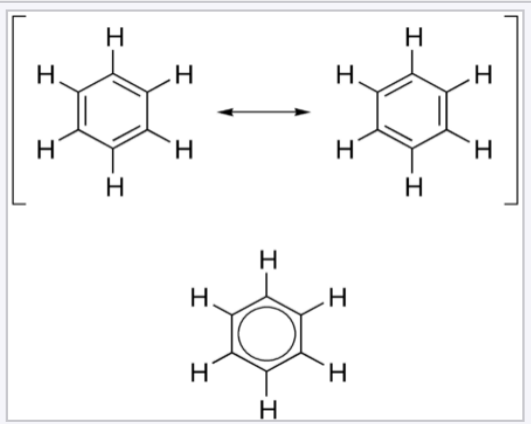
INFORMATION ON FEATURES USED IN ANALYSIS

- 19 features total were used in analysis.
- Drug Interaction is Target Feature:
  - 1 if molecule is known to bind to target site. 0 if molecule does not bind to target site, or molecule interaction is not known.
- Molregno is the unique identifier given to each molecule.
- Some of the features used :
  - Aromatic Rings - Number of Aromatic Rings.
  - PSA – Polar Surface Area
  - Full Mwt – Full Molecular Weight including salts.
  - HBA_Lipinkski – Number of Hydrogen Bond Acceptors according to Lipinski Rules.
  - HBD Donors – Number of Hydrogen Bond Donors according to Lipinski Rules.

Full Dictionary located at:
http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_23/schema_documentation.html

INFERENCE FROM LOGISTIC REGRESSION MODEL

- Aromatic Rings feature with largest norm in molecule prediction with target:
  - Cyclic property of molecule to increase molecular stability
    - Attr to Pi-BONDS.
- Other important features include HBA/HBP LIPINSKI



Two different resonance forms of benzene (top) combine to produce an average structure (bottom)

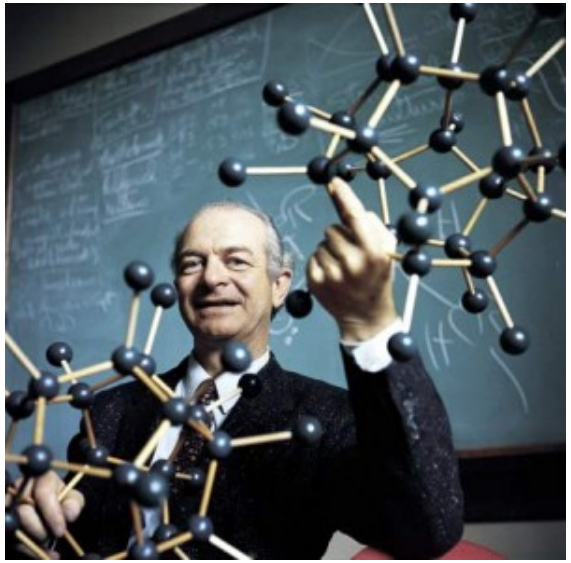
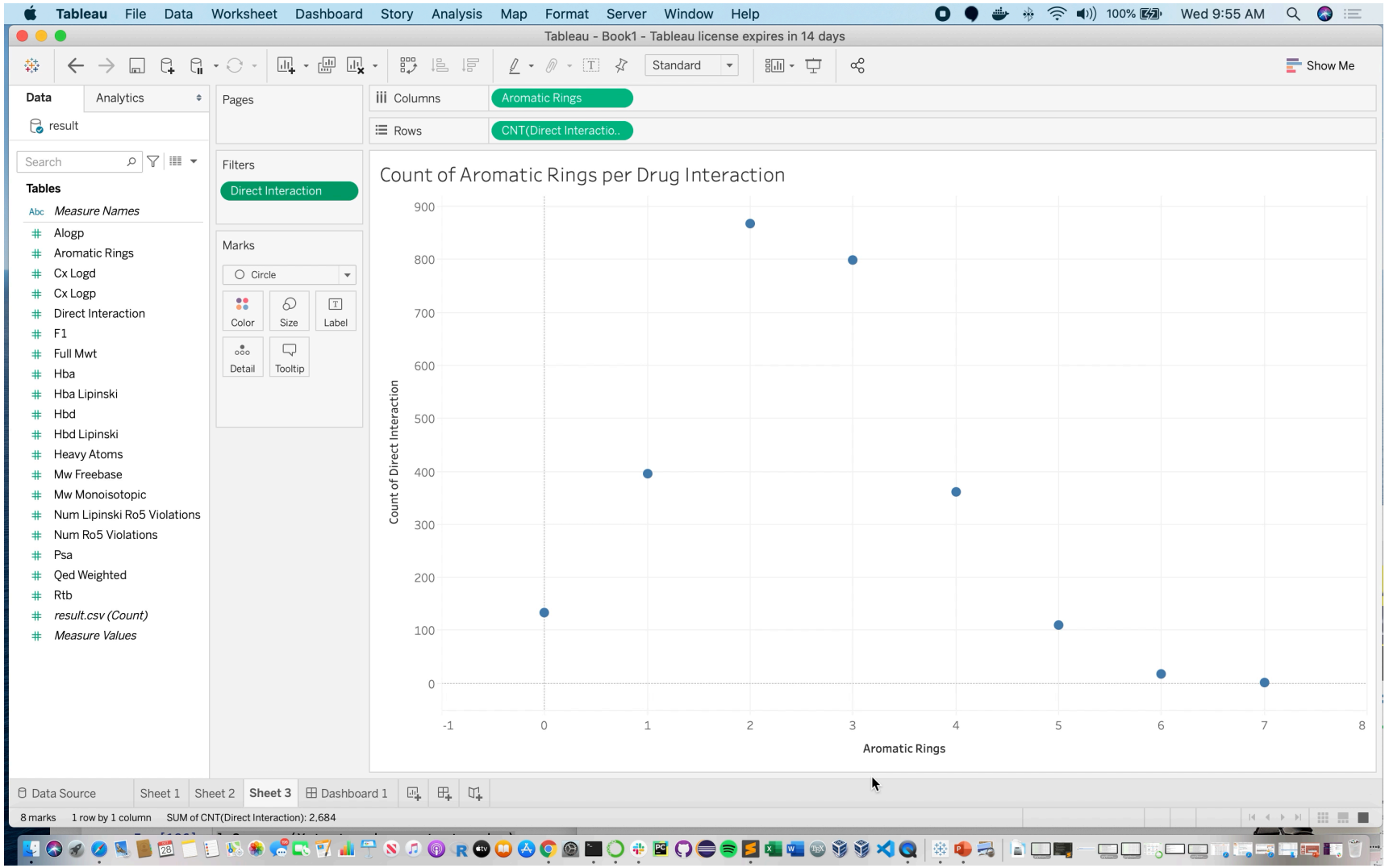


TABLEAU-# OF DRUG INTERACTIONS VS. # AROMATIC RINGS.



NEXT STEPS

- Finish modeling of Drug Interactions using Test/Train/Validate method (will be done for Code Review Deadline).
- Finish Logistic Regression vs. RandomForestClassier Compaison (will be done for Code Review Deadline).
- Predict what molecules bind to which target sites (Proteins, lipids, etc……)

APPENDICES:

CLASS IMBALANCE

0 class	1 class
1892785	2684

- 1 class is 0.014% of total data set.
- Utilized random under-sampling from larger class.
- Both classes had 2684 unique molecules to analyze.

LOGISTIC REGRESSION RESULTS :

Training Error	Testing Error
0.680	0.688

CONFUSION MATRIX:

ACTUAL CLASS

PREDICTED CLASS

	0	**1**
0	1667	470
1	901	1256