

# Antibiotic Drug Discovery



Corey J Sinnott

# TOC

- **Overview**
- **Project Objective**
- **Analysis**
- **Findings & Recommendations**



# Understanding the problem

**1.2b** USD

The cost to produce a new antibiotic in the US.

Pharmaceutical companies do not have a cost-benefit incentive to produce new antibiotics.

**7** years

The average length of a traditional drug discovery pipeline.

Patients infected with antibiotic resistant bacteria need urgent treatment.

**50%**

More resistant strains of bacteria over the past 4 years.

Organisms are evolving new mechanisms of resistance faster than we can create new treatments.

# The Solution

- **Machine Learning**
  - Training models to predict successful antibiotics.
- **AI**
  - Using trained models to invent new antibiotics.

**Project objective:**

**Develop a drug discovery pipeline.**



---

First Step:

Develop a model

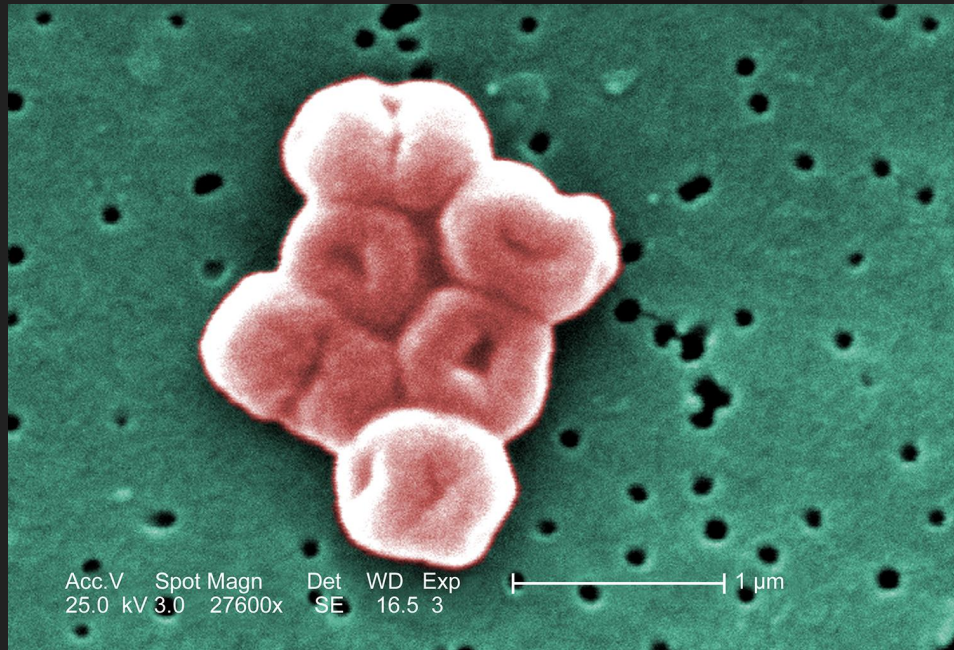


# Drug Discovery Pipeline

## Choose an Organism

- **Acinetobacter Baumannii**
  - Blood, wound, urinary tract, and lung infections
  - Becoming resistant to most antibiotics

*Pipeline optimized to start with any specified target, and be completely reproducible.*



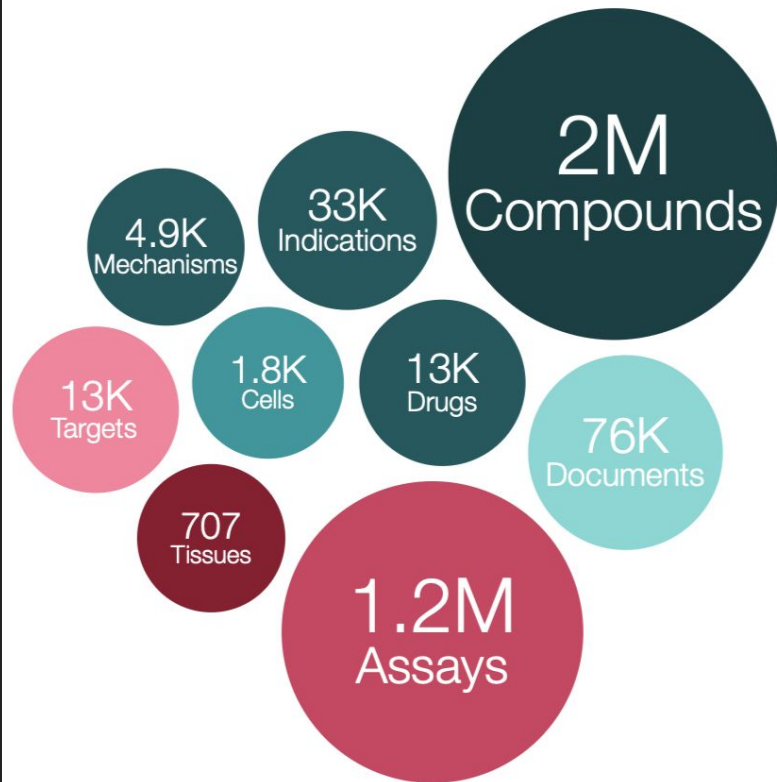


# Drug Discovery Pipeline

## Search for Classified Compounds

- ~ 5000 compounds obtained using ChEMBL web-client.
- Filtered and sorted for Minimum Inhibitory Concentration (MIC).
  - MICs ranged from <10nM (very effective) to >500,000nM (not effective).

ChEMBL

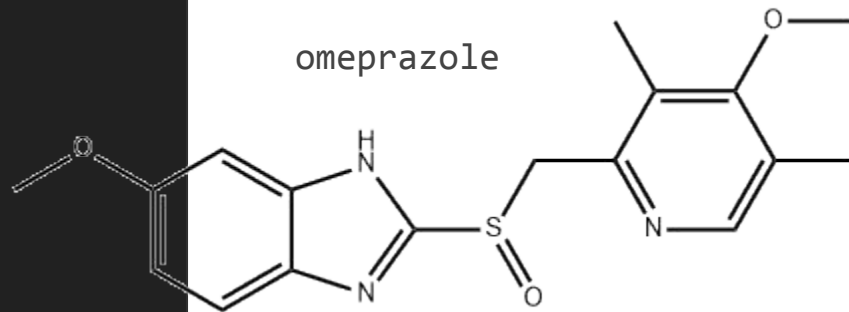




# Drug Discovery Pipeline

## Engineer Features

- Lipinski Descriptors
  - Molecular Weight
  - Log-p
    - Lipophilicity / Solubility
  - # Proton Donors
  - # Proton Acceptors
- Vectorization using Morgan's fingerprint algorithm.
- Standardized target with  $-\log_{10}$ 
  - MIC  $\Rightarrow$  pMIC



SMILES: CC1=CN=C(C(=C1OC)C)CS(=O)C2=NC3=C(N2)C=C(C=C3)OC

Mol wt =  $345.42 \text{ g} \cdot \text{mol}^{-1}$

Log-p = 2.43

# H<sup>+</sup> Donors: 1

# H<sup>+</sup> Accept: 3

# Drug Discovery Pipeline

---

## Classification Model

- HistGradient Boosting Classification
- Binary target
  - Active (<35nM MIC) vs Inactive
  - Intermediate values removed
- Standard Scaler
- Max iterations = 800



# Drug Discovery Pipeline

## HistGradient Boost Metrics

- Accuracy  $\Rightarrow$  98%
- Precision  $\Rightarrow$  96%
- Recall  $\Rightarrow$  97%
- F1 score  $\Rightarrow$  0.97
- ROC AUC  $\Rightarrow$  0.98

*Null accuracy 60%*



# Drug Discovery Pipeline

---

## Regression Model

- HistGradient Boosting Regression
- Predicting pMIC
  - $-\log_{10}$  of MIC
- Standard Scaler
- L2 regularization = 0.0001
- 1000 max iterations

*Random Forest Regression did better with outliers. Mention residuals*





# Drug Discovery Pipeline

## HistGradient Boost Metrics

- $r^2$   $\Rightarrow$  0.76
- MSE  $\Rightarrow$  0.696
- RMSE  $\Rightarrow$  0.834
- MAE  $\Rightarrow$  0.568
- Null MSE  $\Rightarrow$  3.088
- Performed 78% greater than a null model



# Drug Discovery Pipeline

---

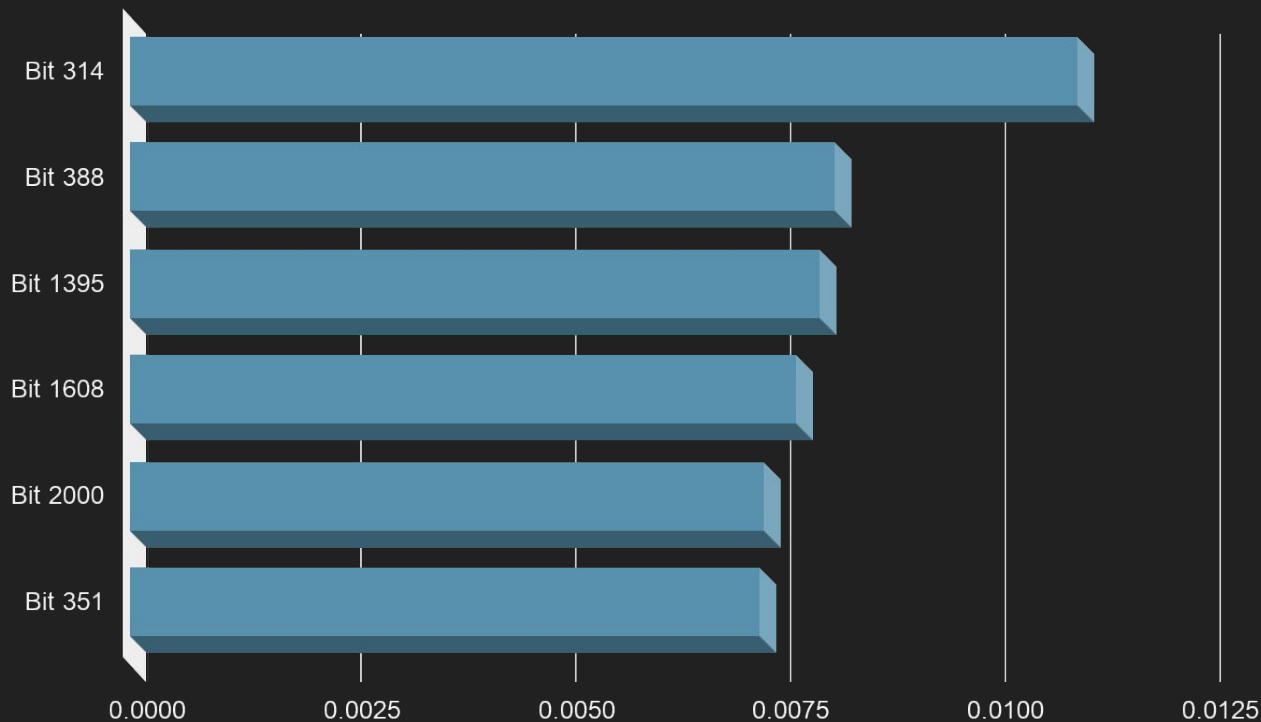
## Feature Importances

- Same molecular fragments, or “bits,” important for both classification and regression.



# Drug Discovery Pipeline

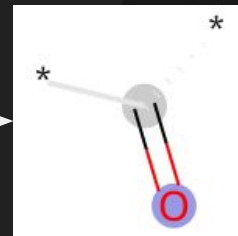
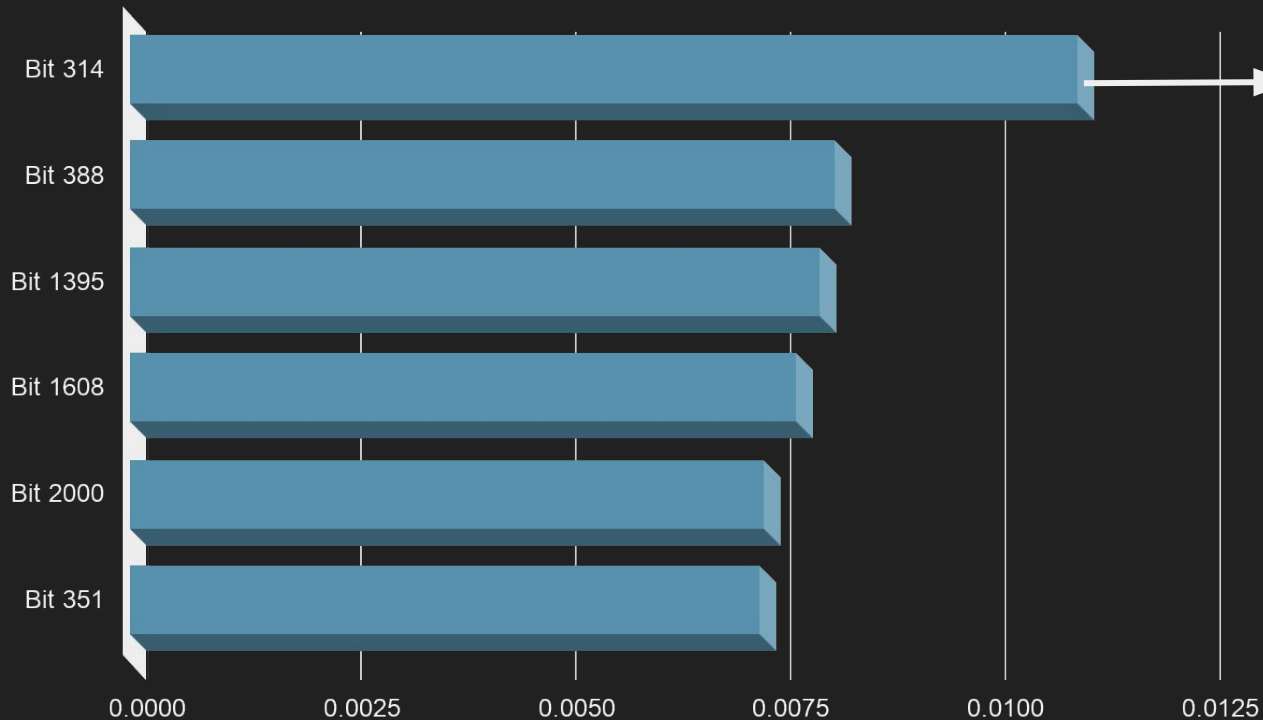
## Feature Importances





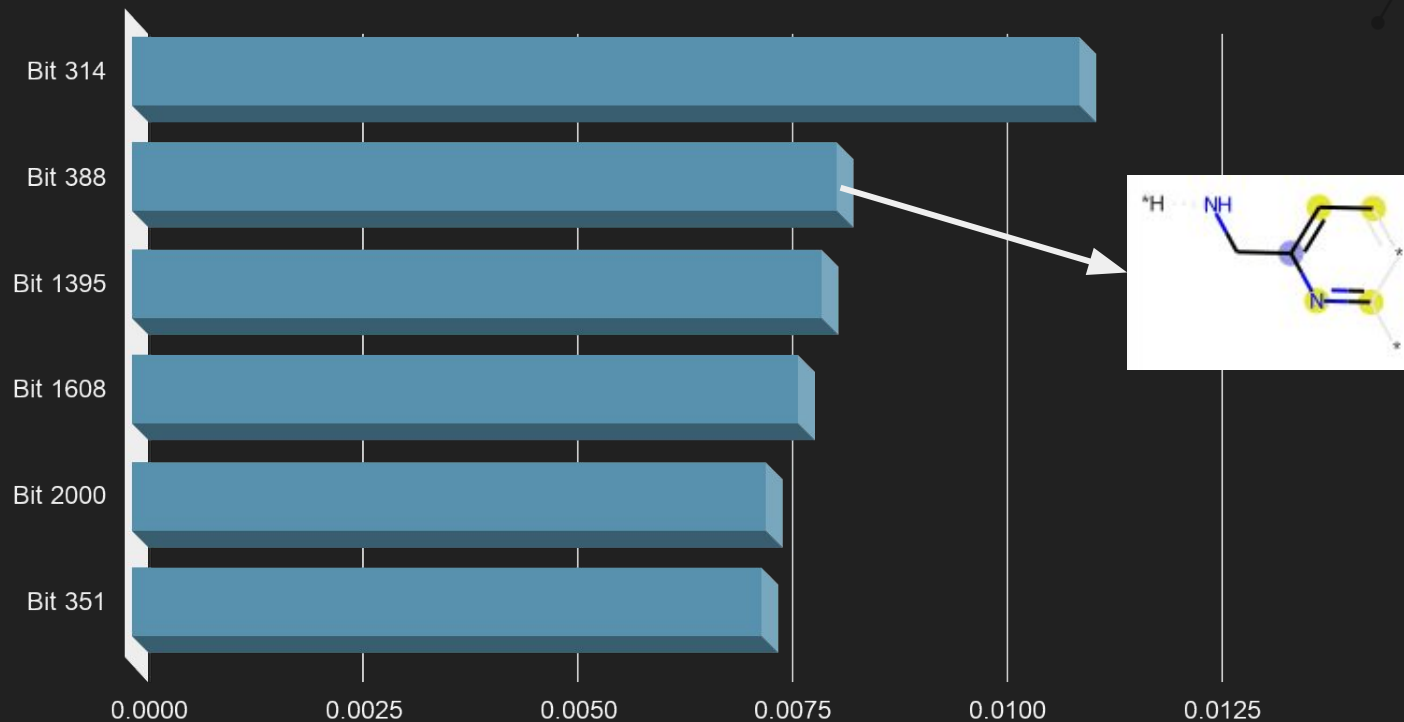
# Drug Discovery Pipeline

## Feature Importances



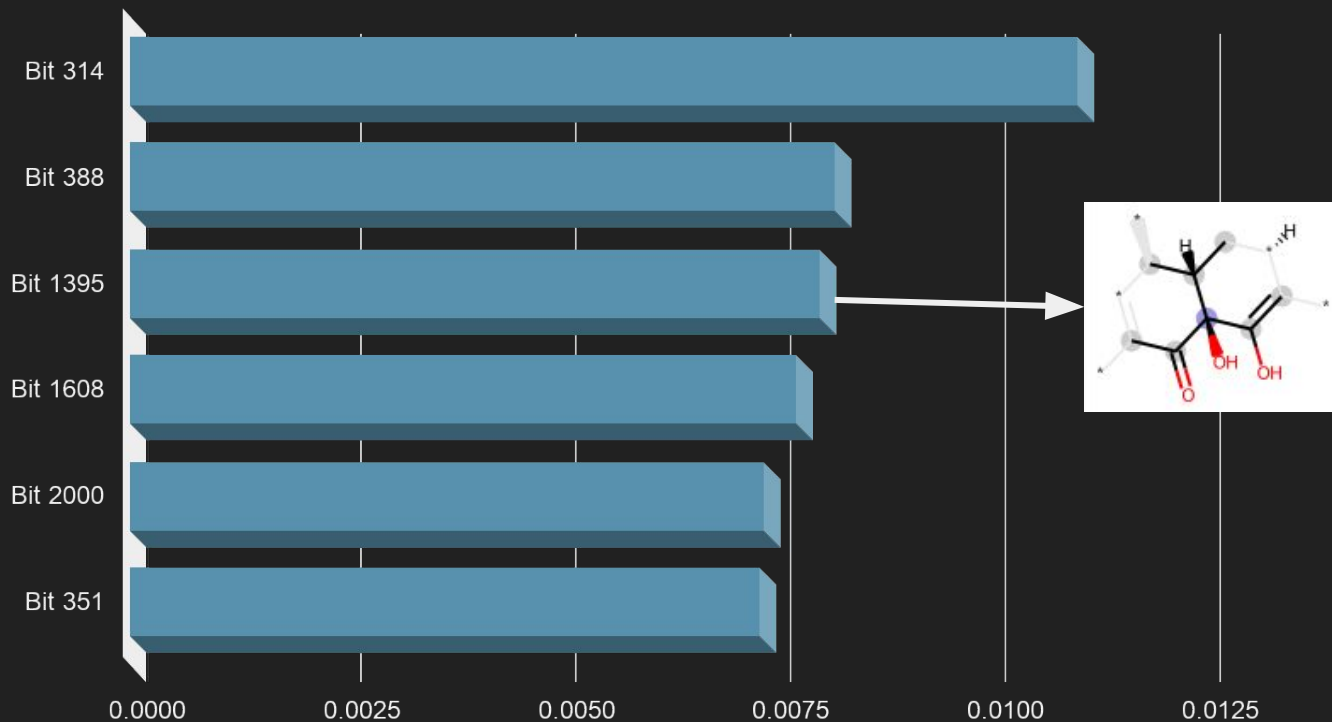
# Drug Discovery Pipeline

## Feature Importances



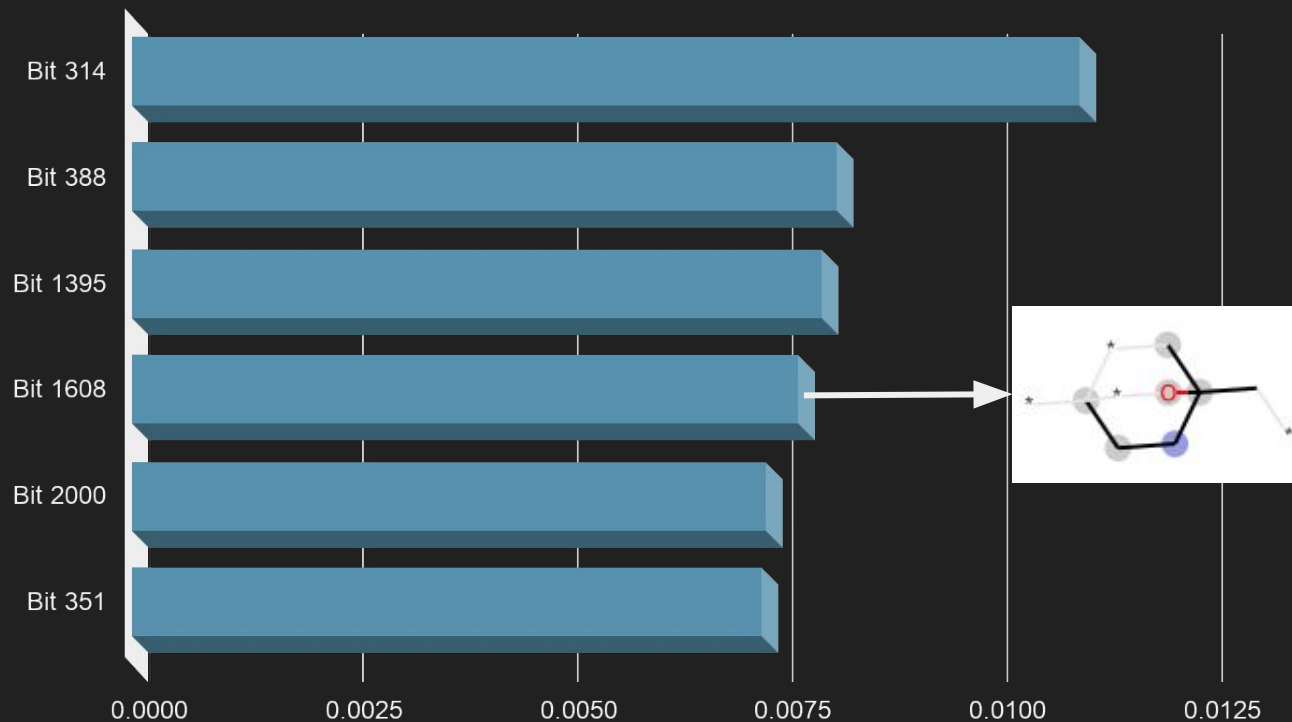
# Drug Discovery Pipeline

## Feature Importances



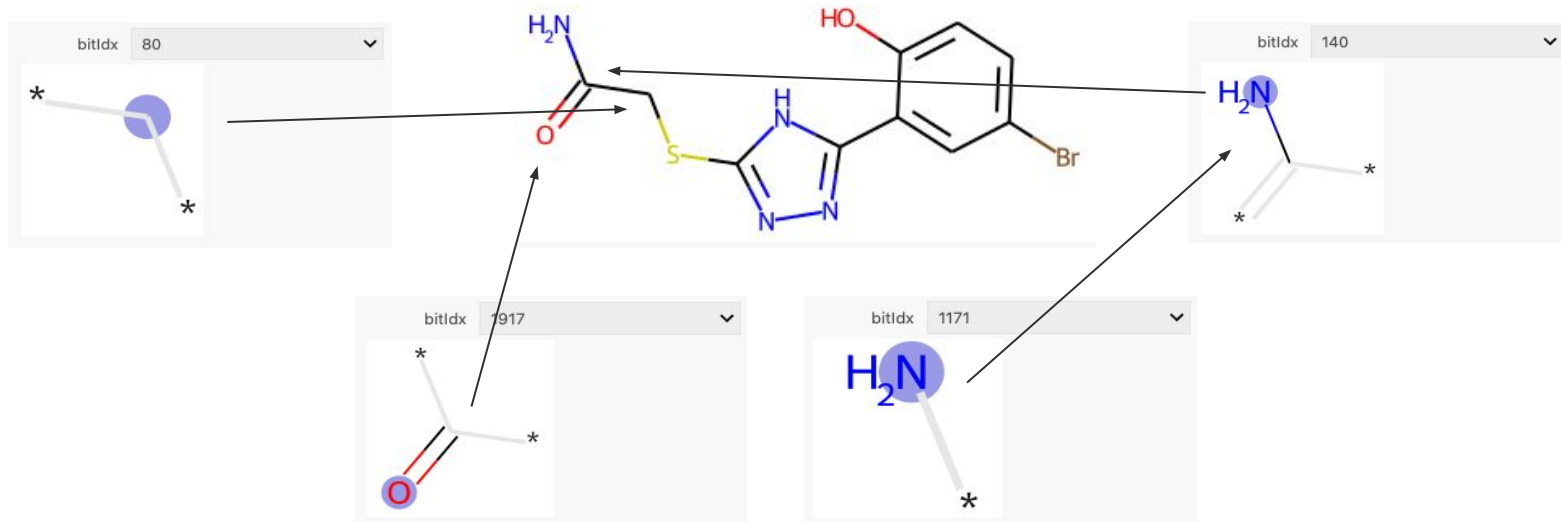
# Drug Discovery Pipeline

## Feature Importances



# Drug Discovery Pipeline

## Residuals



---

Next Step:

# Practical Application



# Drug Discovery Pipeline

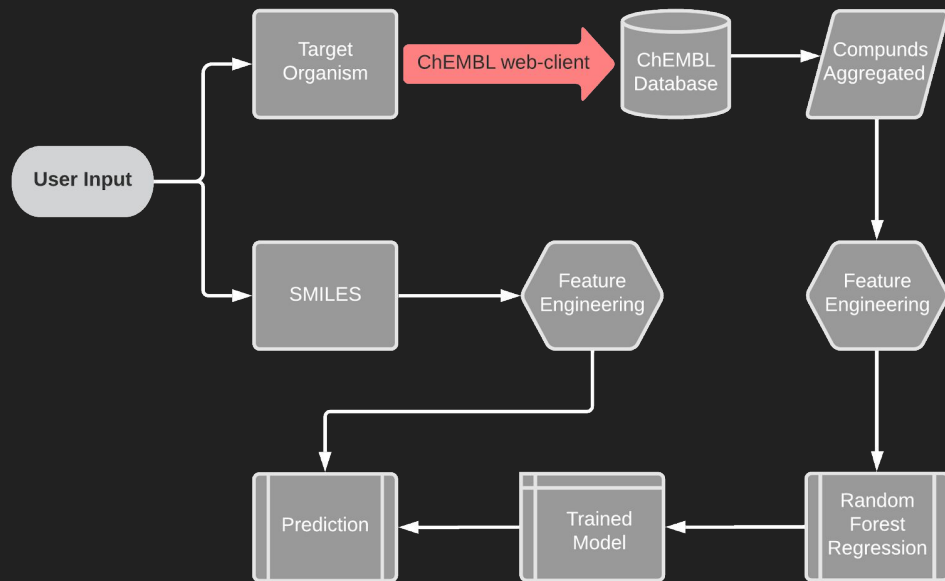
## Practical Application

- User inputs a target organism and SMILES.
- App outputs model metrics and prediction.

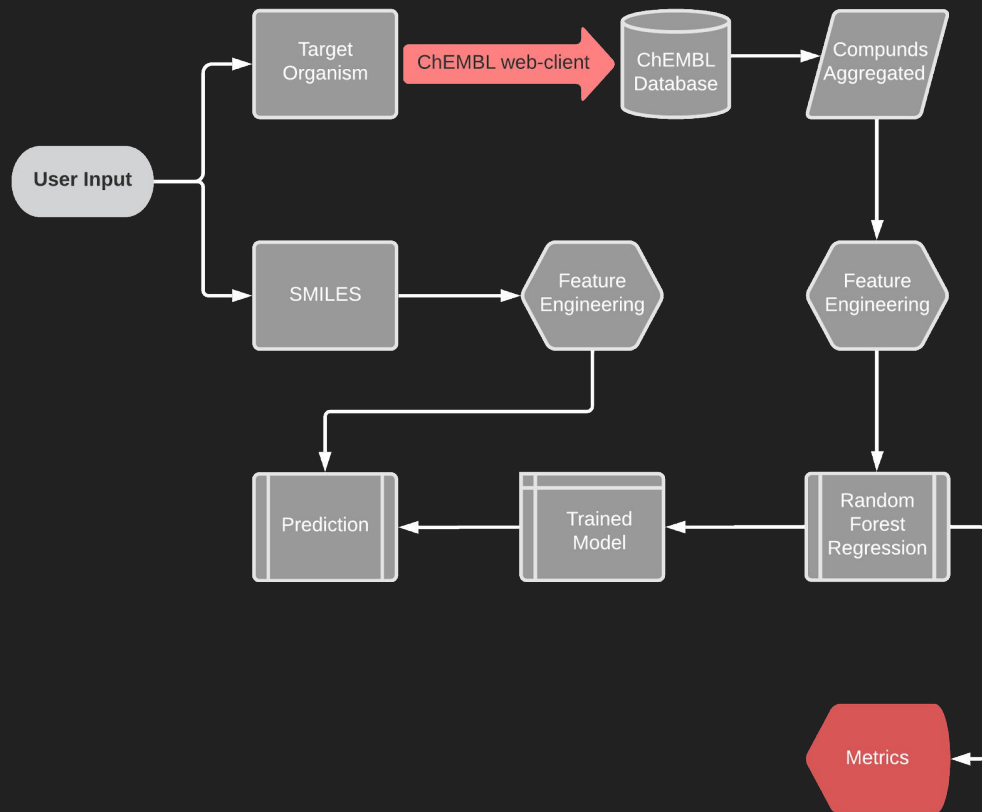




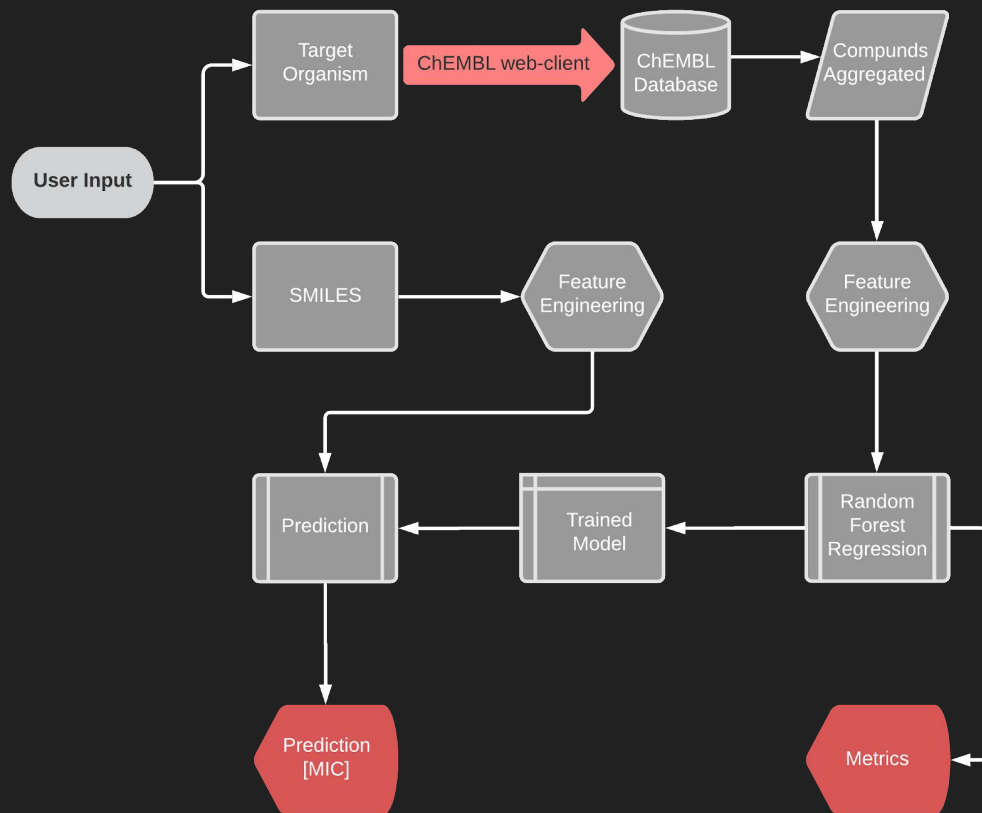
# Drug Discovery App



# Drug Discovery App



# Drug Discovery App



# App Demo

Submit

✓

Compound

✓

Target



 **Drug Discovery Pipeline**

for Chemists & Researchers

This app creates a model from every available drug in the ChEMBL database, and will return a Minimum Inhibitory Concentration (MIC) prediction for your compound in  $\mu\text{M}$ .

© 2020 David Goodrich



# Drug Discovery Pipeline

---

## What's next?

- New molecules generated by recurrent neural networks.



# Drug Discovery Pipeline

## Molecule Generator

- Tensorflow NN
- LSTM layers
- Trained on every available compound.
- **X** variable is an individual character from a molecule.
- **Y** variable is the following character.

*Adapted from Deep Learning with Python*

*Work in-progress*

## DEEP LEARNING with Python

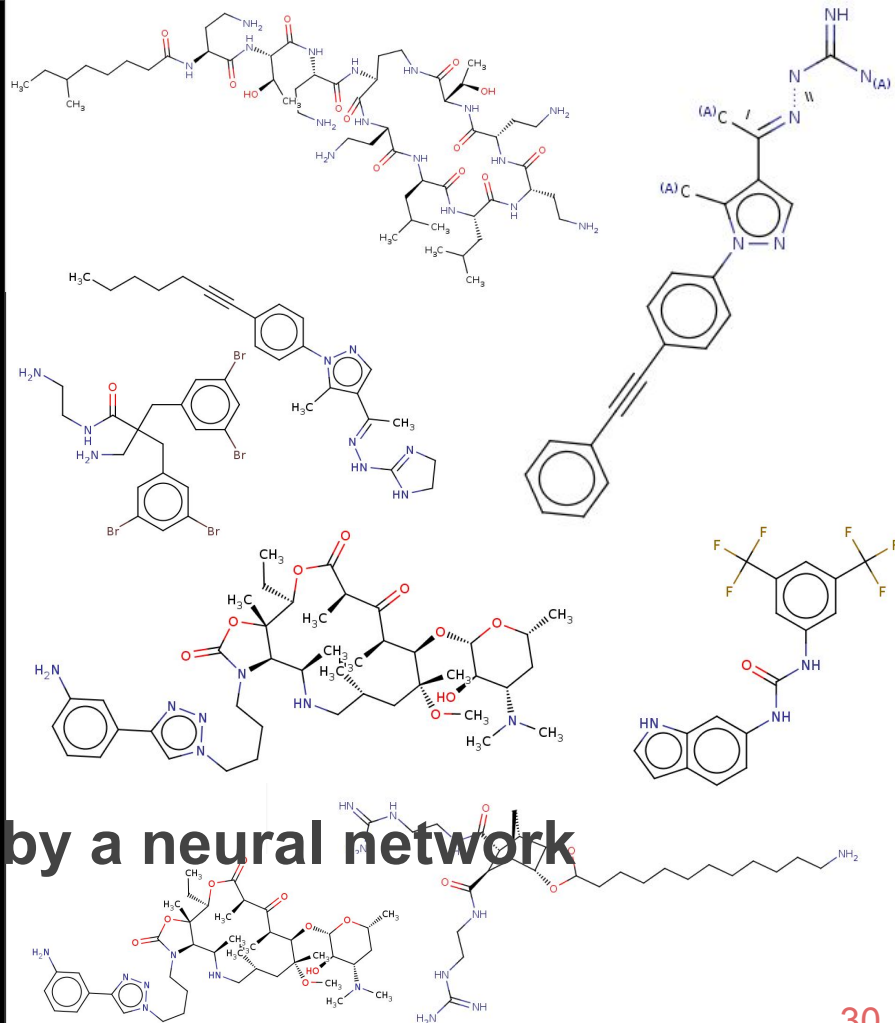
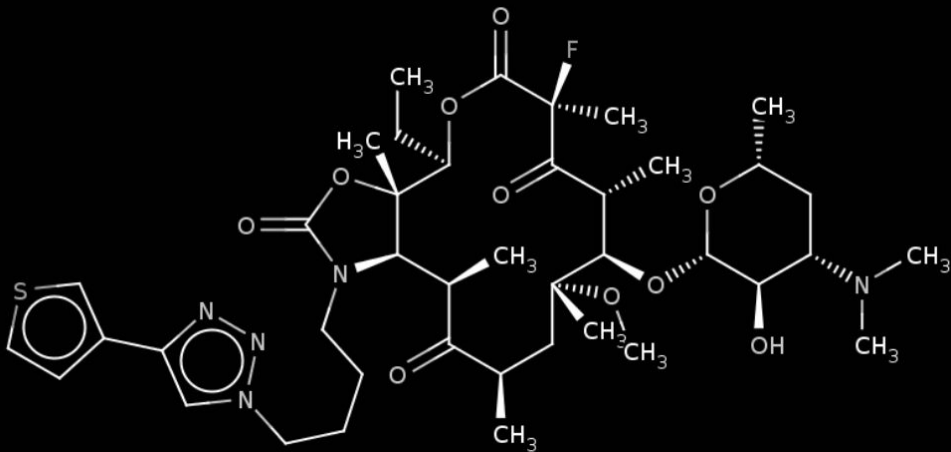
François Chollet



MANNING

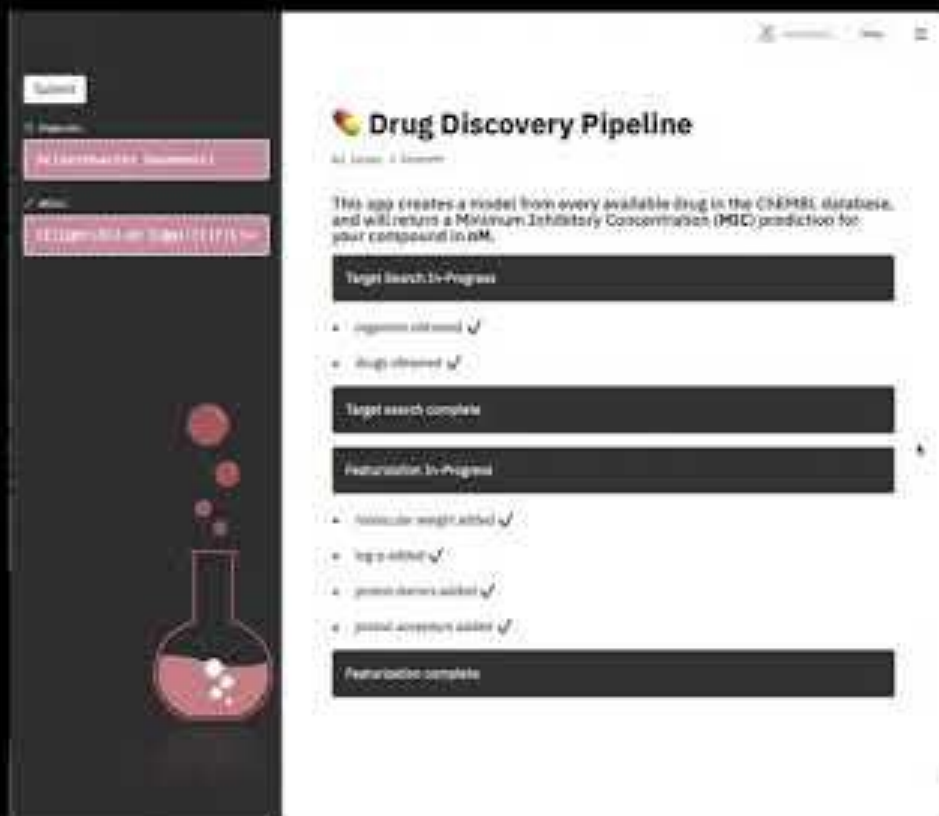


# What's next?



These molecules were generated by a neural network

# Generator Demo



The screenshot displays the 'Drug Discovery Pipeline' web application. On the left is a dark sidebar with a 'Tutorial' button at the top, followed by a 'Home' button, and two red buttons labeled 'Get Recommendations' and 'Generate New Compound'. Below these is a red illustration of a round-bottom flask with red liquid and three red bubbles rising from it. The main content area has a white background and features the application title 'Drug Discovery Pipeline' with a logo. Below the title is a description: 'This app creates a model from every available drug in the ChEMBL database, and will return a Minimum Inhibitory Concentration (MIC) prediction for your compound in  $\mu\text{M}$ .' The interface shows a progress bar for 'Target Search In Progress' with a list of items: 'Target search complete', 'Feature extraction In Progress', and 'Feature extraction complete'. Each item has a checkmark icon next to it.

**Drug Discovery Pipeline**

This app creates a model from every available drug in the ChEMBL database, and will return a Minimum Inhibitory Concentration (MIC) prediction for your compound in  $\mu\text{M}$ .

Target Search In Progress

- Target search complete ✓
- Feature extraction In Progress ✓
- Feature extraction complete ✓

# Conclusion

---

A robust, simple **drug discovery pipeline** was created.



# Drug Discovery Pipeline

A decorative graphic in the top right corner of the slide, consisting of a network of interconnected hexagons and lines, resembling a molecular structure or a honeycomb pattern, rendered in a light gray color against the dark background.

**Classify**

**Classify successful antibiotics with 98% accuracy.**

**Identify**

**Identify molecular fragments most important to model.**

**Predict**

**App can give an estimate of efficacy for a new drug in under a minute.**

**Generate**

**Ability to create new drugs with an RNN is evolving.**

# Any Questions?

---



# Thanks!



**Corey J Sinnott**

---

**Data Scientist**

# Sources

1. A Deep Learning Approach to Antibiotic Discovery:  
[https://www.cell.com/cell/fulltext/S0092-8674\(20\)30102-1?returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867420301021%3Fshowall%3Dtrue](https://www.cell.com/cell/fulltext/S0092-8674(20)30102-1?returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867420301021%3Fshowall%3Dtrue)
2. Antibiotic resistance: bioinformatics-based understanding as a functional strategy for drug design:  
<https://pubs.rsc.org/en/Content/ArticleLanding/2020/RA/D0RA01484B#!divAbstract>
3. MIC database: A collection of antimicrobial compounds from literature:  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2823385/>
4. Machine learning-powered antibiotics phenotypic drug discovery: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6428806/>
5. Helping Chemists Discover New Antibiotics: <https://pubs.acs.org/doi/10.1021/acsinfecdis.5b00044>
6. New Statistical Technique for Analyzing MIC-Based Susceptibility Data:  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3294928/>
7. Applying machine learning techniques to predict the properties of energetic materials:  
<https://www.nature.com/articles/s41598-018-27344-x>
8. QBMG: quasi-biogenic molecule generator with deep recurrent neural network:  
<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0328-9>
9. Are the physicochemical properties of antibacterial compounds really different from other drugs?:  
<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0143-5>
10. Deep Learning with Python by Francois Chollet