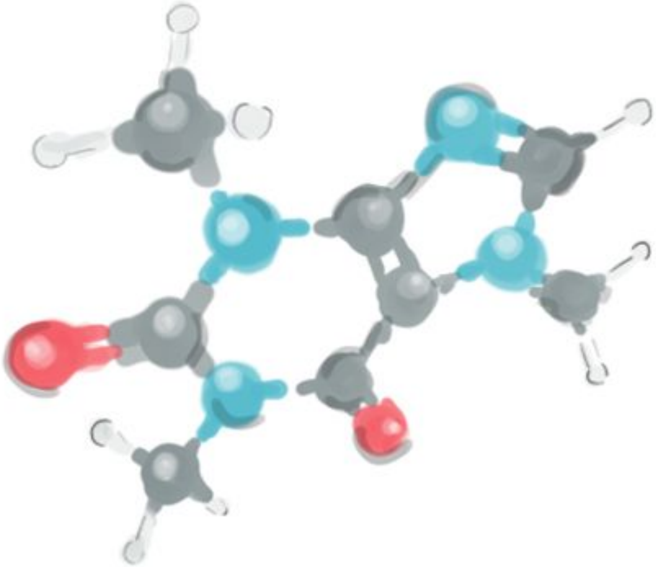
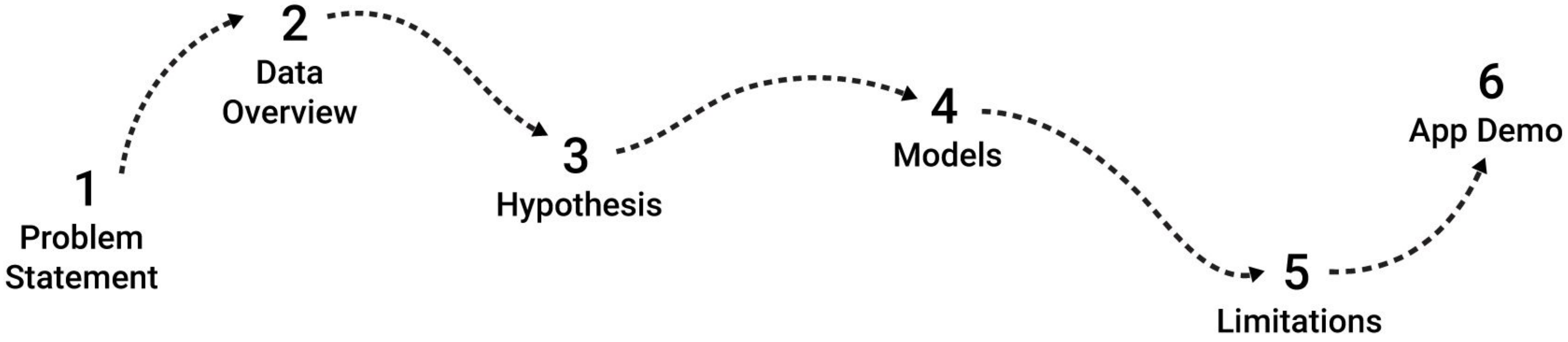
Multi Class Drug Classification Using Molecular Structures

Vivian Peng

https://github.com/veeps/molecular_classification



Overview



1 Problem Statement

2 Data Overview

3 Hypothesis

4 Models

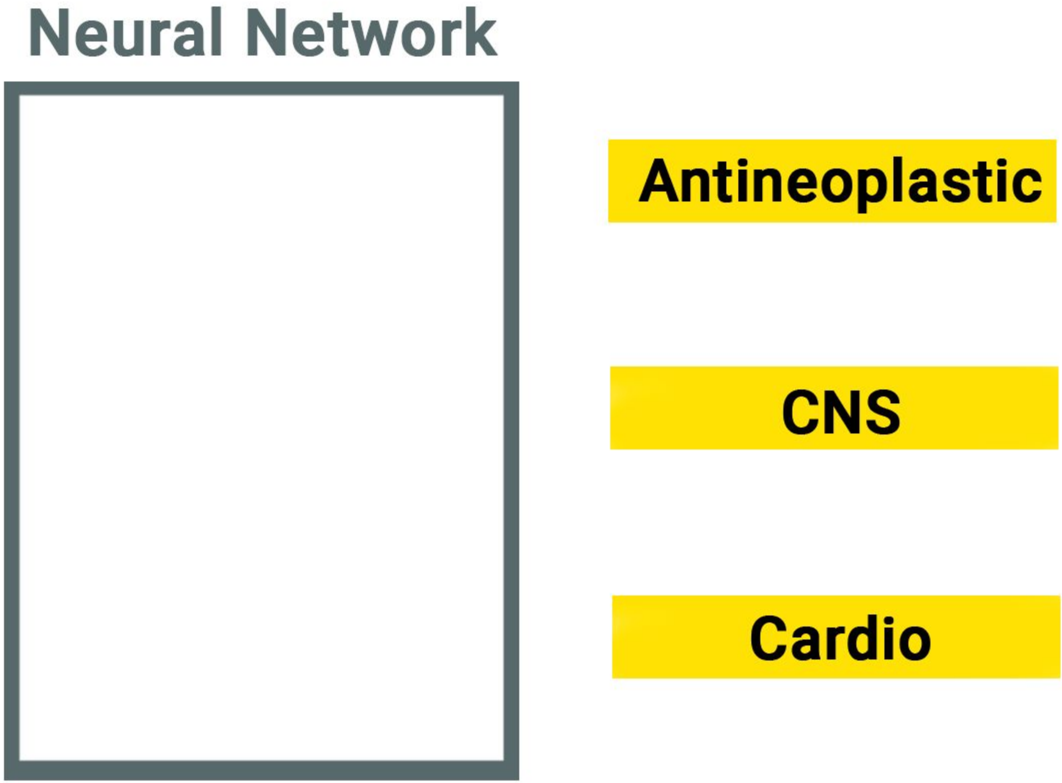
5 Limitations

6 App Demo

Problem Statement

**Current drug R&D is long and costly…
And it doesn’t incentivize for research into treatments for neglected diseases.**

What if we could improve the R&D process by reducing redundancy, and screen for multiple therapeutic uses in parallel?



Neural Network

Antineoplastic

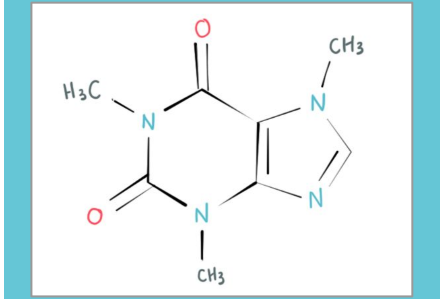
CNS

Cardio

Here’s what my data looked like:

CSV file w/ chemical properties

2D Images of molecular structures



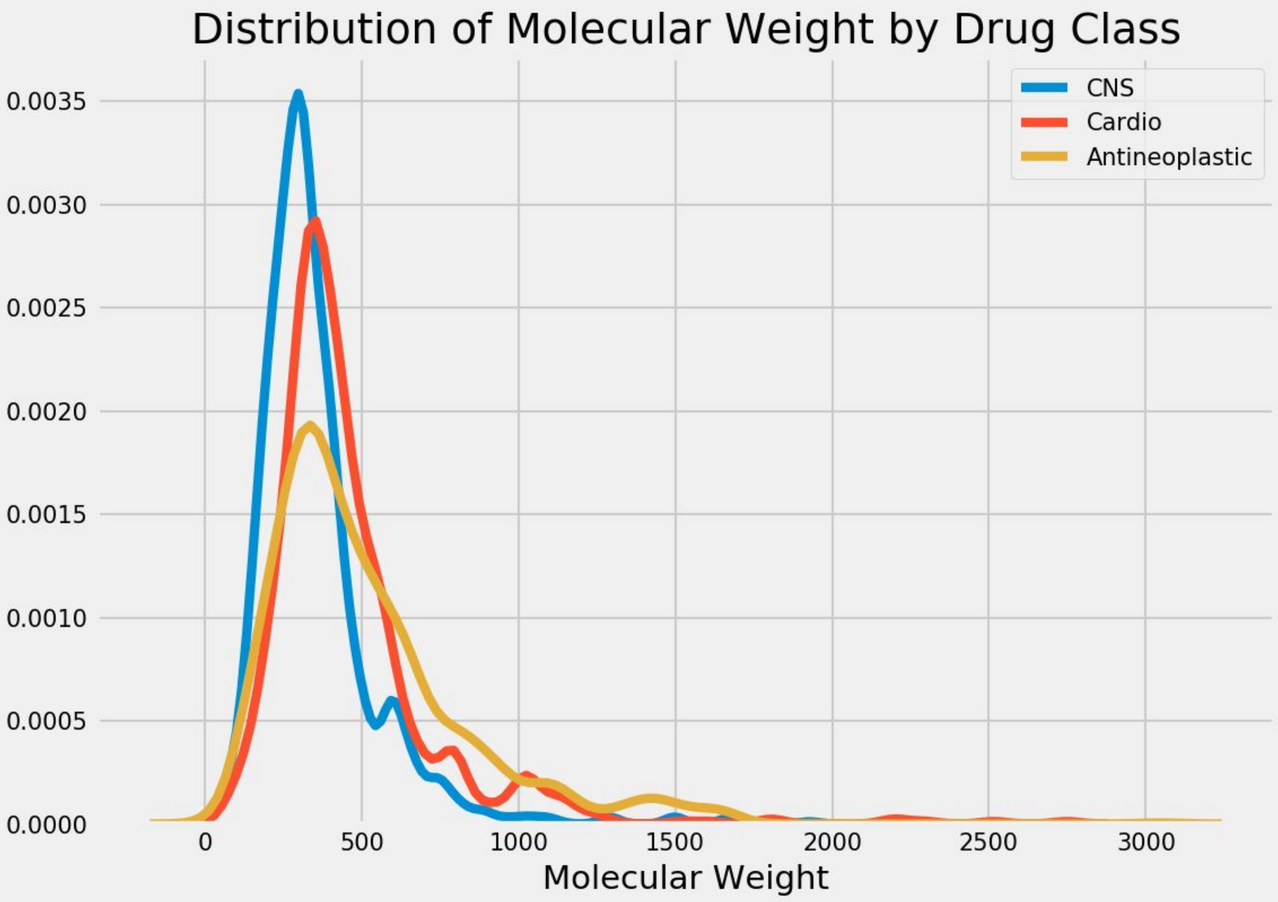
SMILES of molecular structures

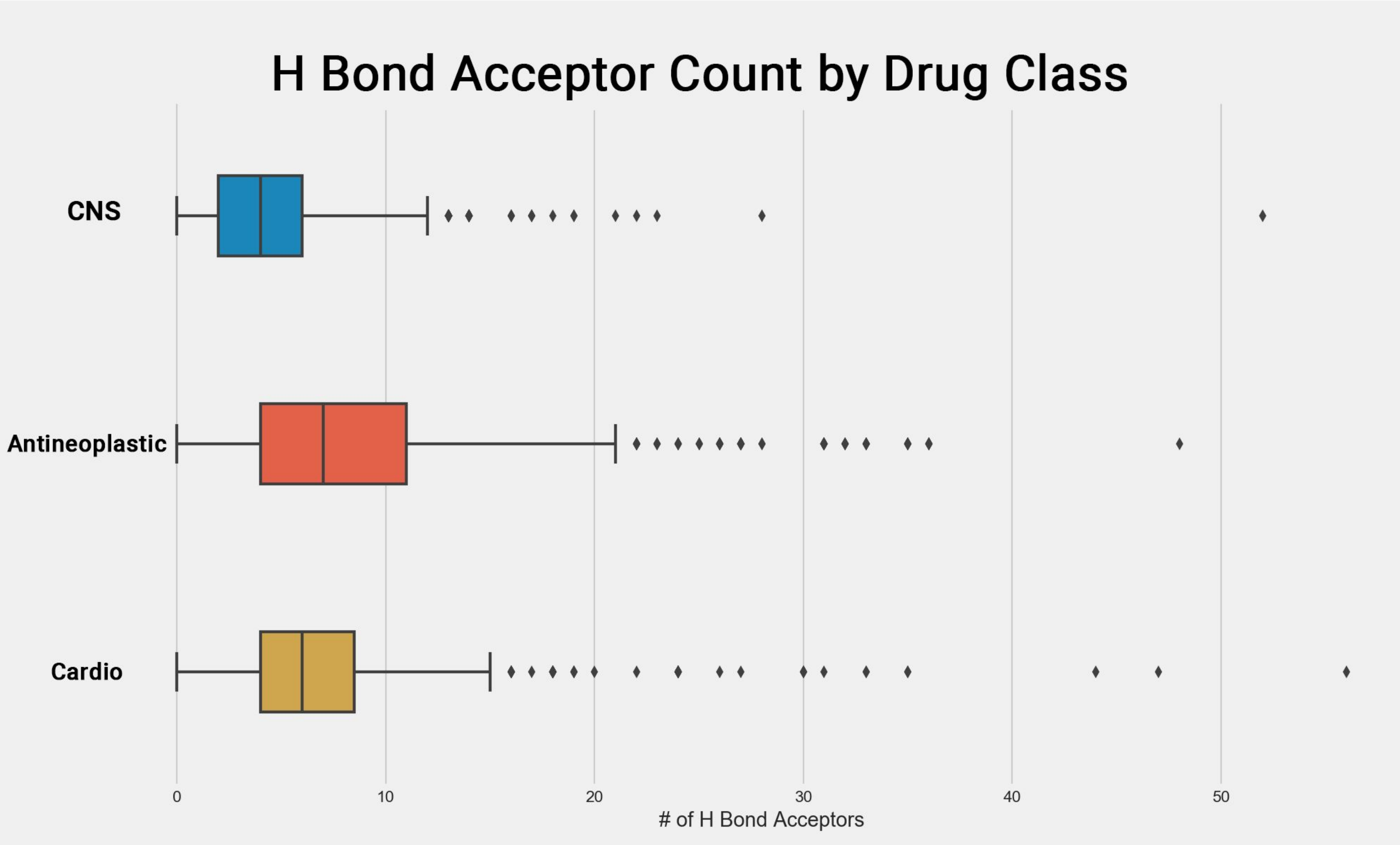
SMILE

CN1C(=O)N(C)C2NCN(C)C2C1=O

SMILE

CN1C(=O)N(C)C2NCN(C)C2C1=O

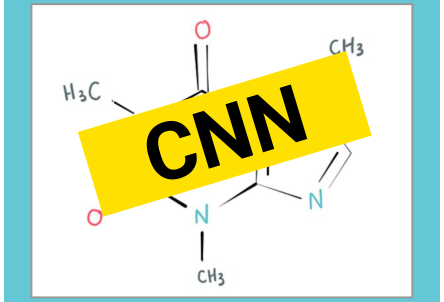

Distribution of Molecular Weight by Drug Class
CNS
Cardio
Antineoplastic
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
500
1000
1500
2000
2500
3000
Molecular Weight


H Bond Acceptor Count by Drug Class
CNS
Antineoplastic
Cardio
0
10
20
30
40
50
# of H Bond Acceptors

Ran 3 types of models:

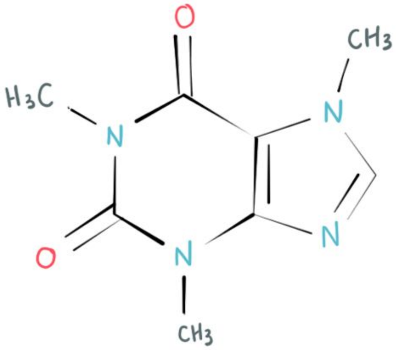
CSV file w/ chemical properties

2D Images of molecular structures



SMILES of molecular structures

How do these neural networks compare in classifying drug classes using molecular structure data?

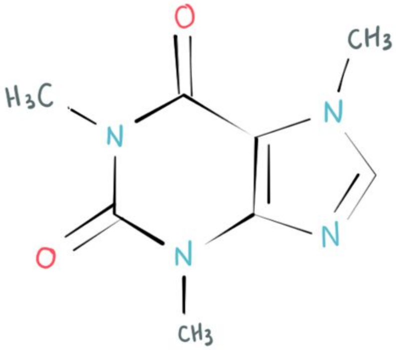


CNN + IMAGE DATA

SMILE

CN1C(=O)N(C)C2NCN(C)C2C1=O

RNN + TEXT DATA



CNN + IMAGE DATA

My hypothesis was that a CNN model using image data would be better at classifying than an RNN model using SMILES.

Models

SVC

Support Vector Classification

FEATURES:

- Molecular Weight
- Hydrogen Bond Acceptors
- Hydrogen Bond Donors
- XlogP (Solubility)

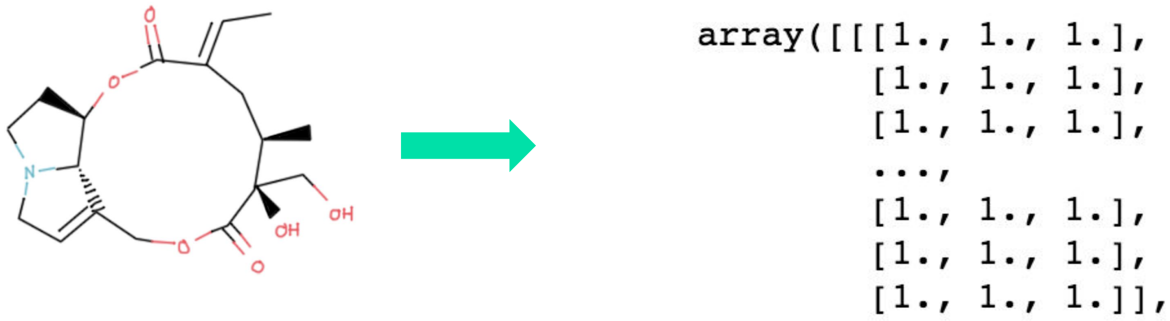
HYPERPARAMETERS:

- Ridge penalty
- Linear kernel
- Gamma = “scale”

CNN

Convolutional
Neural Network

DATA PROCESSING:



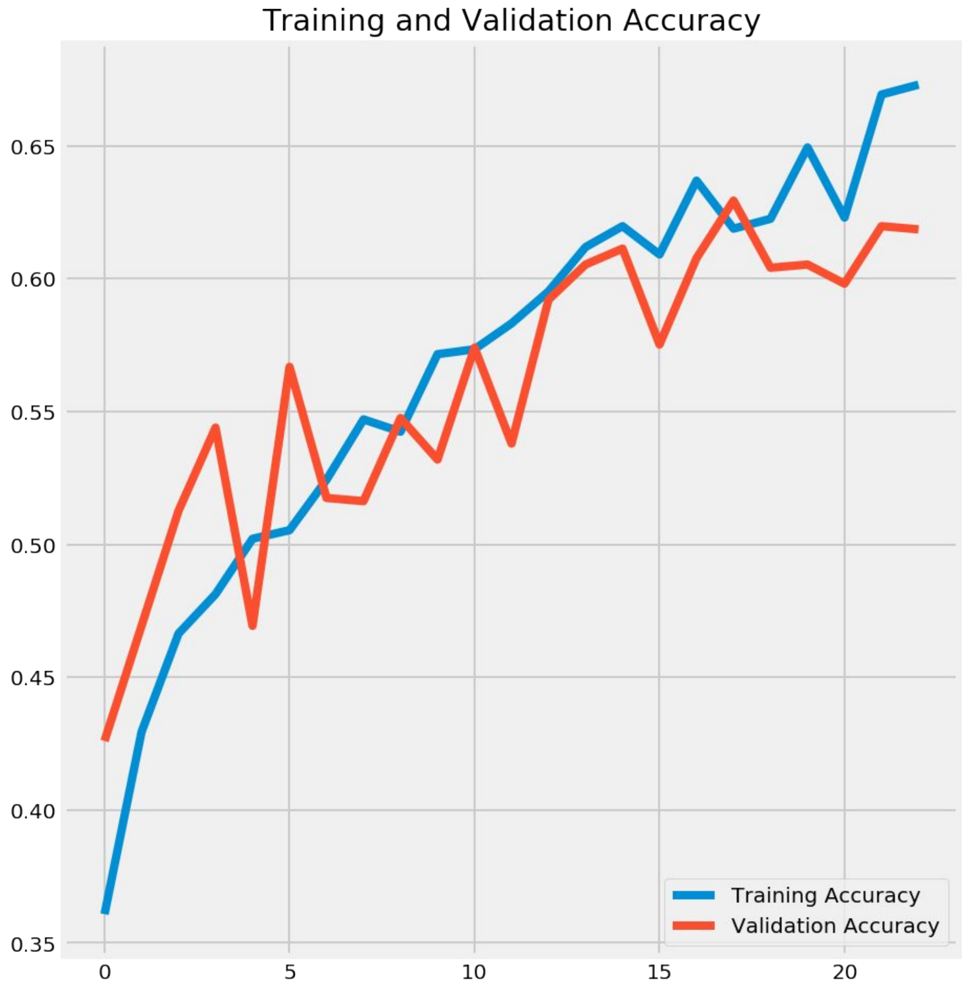
```
array([[[1., 1., 1.],
        [1., 1., 1.],
        [1., 1., 1.],
        ...,
        [1., 1., 1.],
        [1., 1., 1.],
        [1., 1., 1.]],
```

MODELS:

- **Custom convolutional neural net**
- **Pre-trained VGG16 model**
  - 2 hidden layers (256 and 128 neurons)
  - Adam optimizer

CNN

Convolutional
Neural Network

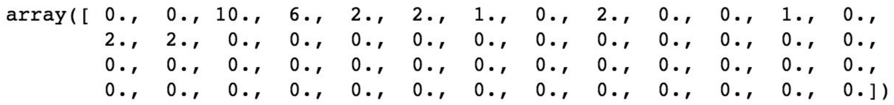


RNN

Recurrent Neural Network

DATA PROCESSING:

Brc1c(NC2=NCCN2)ccc2nccnc12



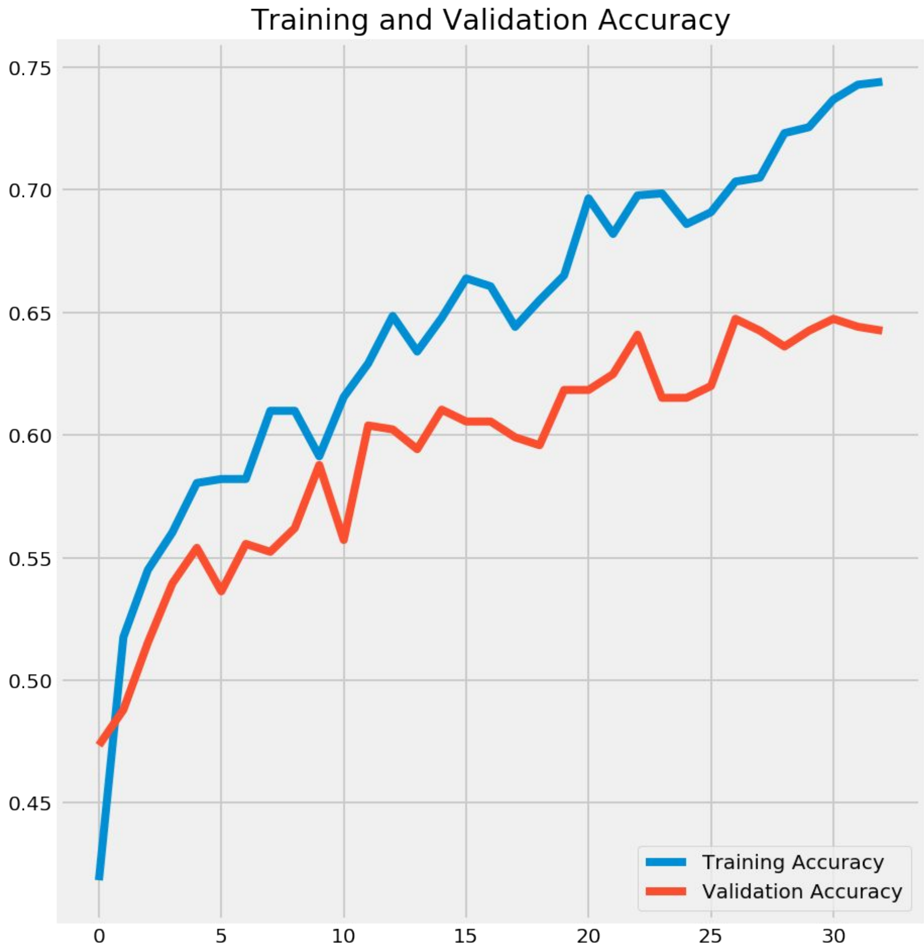
```
array([ 0.,  0., 10.,  6.,  2.,  2.,  1.,  0.,  2.,  0.,  0.,  1.,  0.,
        2.,  2.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
        0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
        0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.])
```

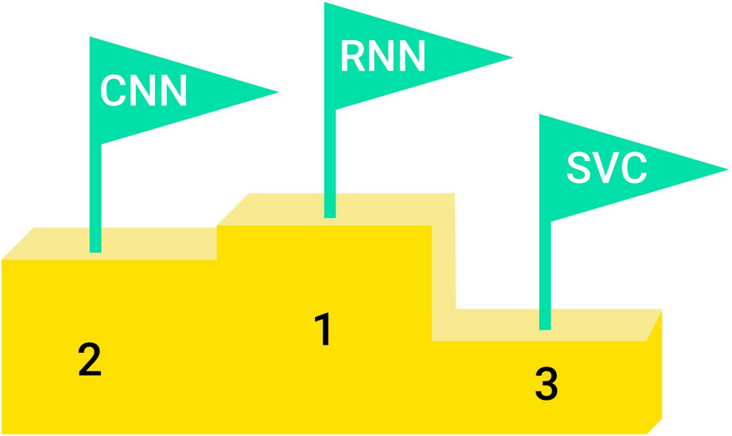
MODEL:

- Simple RNN
    - 5 hidden layers
    - Adam optimizer
- LSTM

RNN

Recurrent Neural
Network

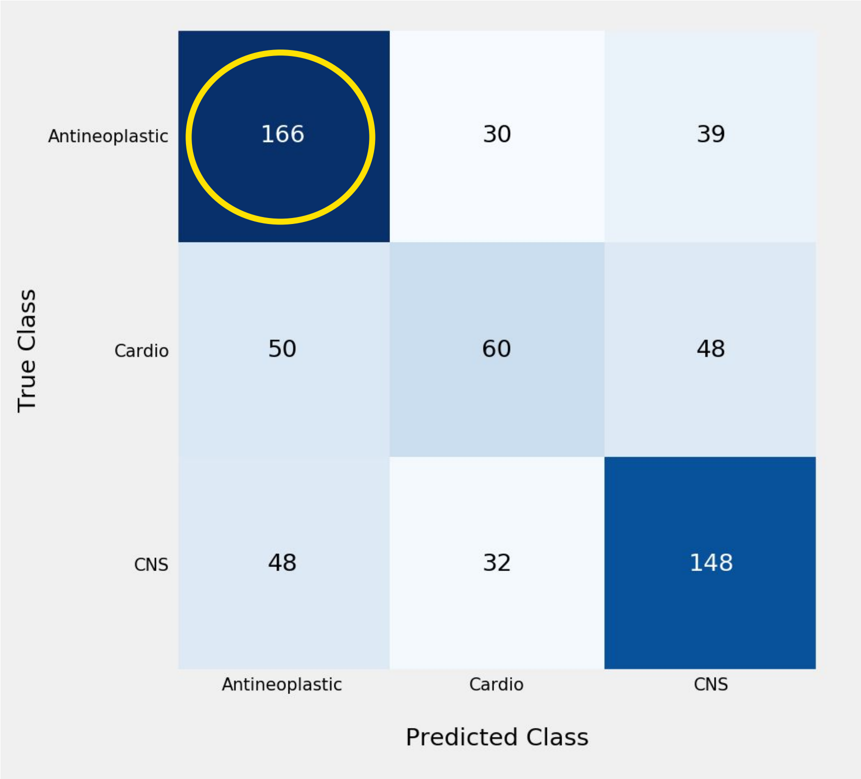




Model	Score
RNN	0.64
CNN	0.62
SVC	0.53
Baseline	**0.37**

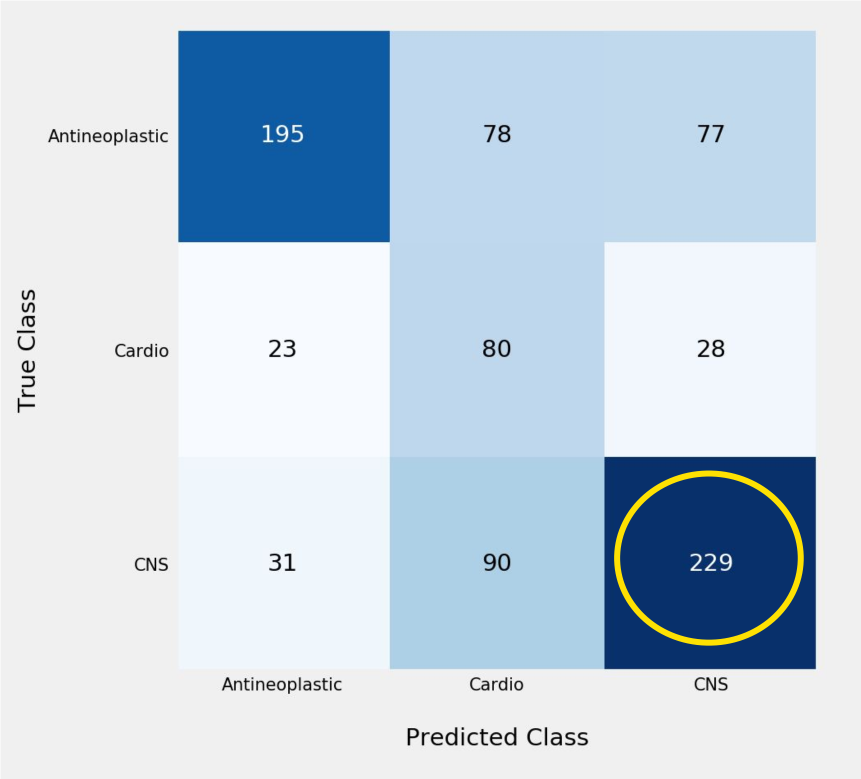
Both types of neural networks performed **relatively the same** in predicting drug classes. **This is really interesting** because running an RNN with text data is **computationally much lighter** than running a CNN with image data.

RNN



True Class \ Predicted Class	Antineoplastic	Cardio	CNS
Antineoplastic	166	30	39
Cardio	50	60	48
CNS	48	32	148

CNN



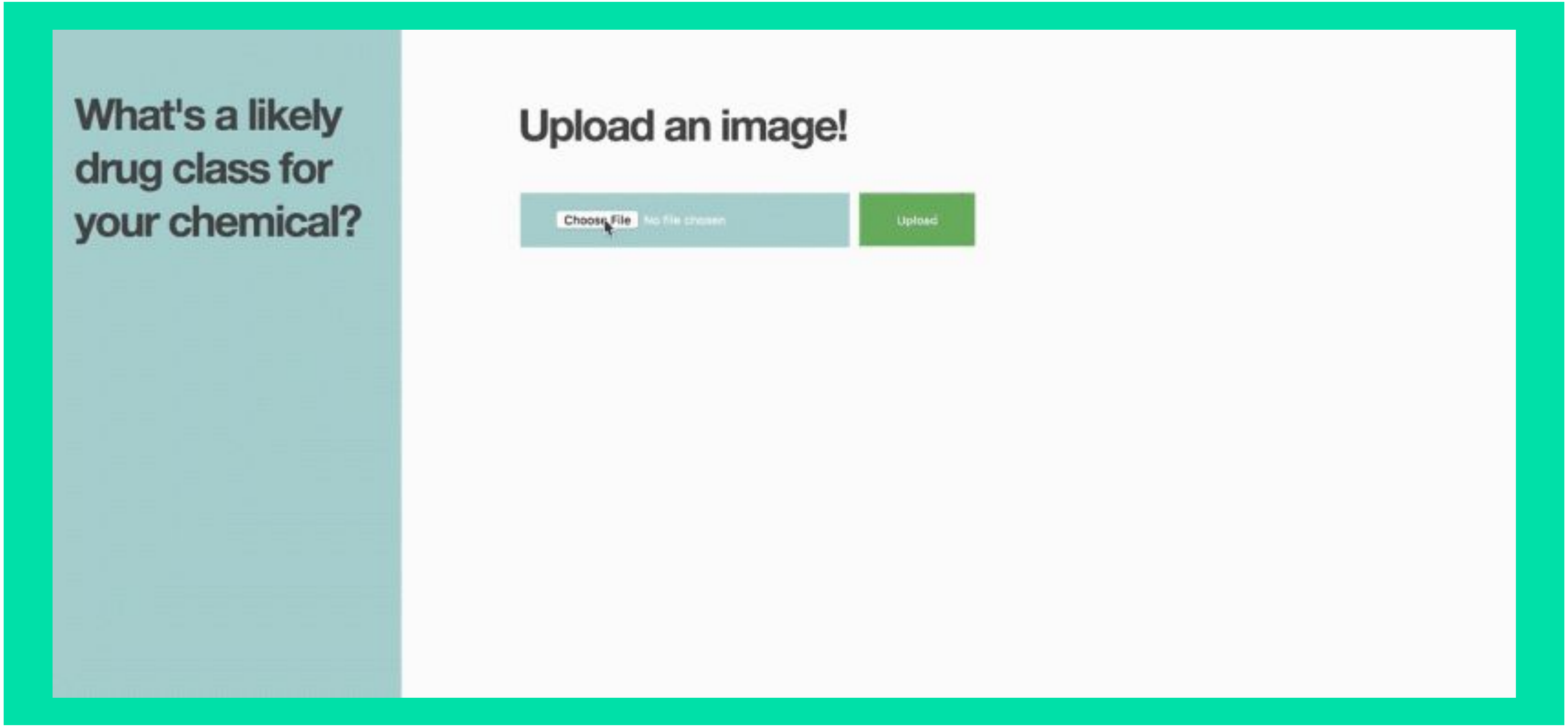
True Class \ Predicted Class	Antineoplastic	Cardio	CNS
Antineoplastic	195	78	77
Cardio	23	80	28
CNS	31	90	229

Models are good at predicting different types of drug classes.

Limitations

- Unable to tokenize SMILES to keep two-letter elements as one unit
- Lack of chemical expertise
- Unable to deploy app to Heroku

App Demo



What's a likely drug class for your chemical?

Upload an image!

Thank you!