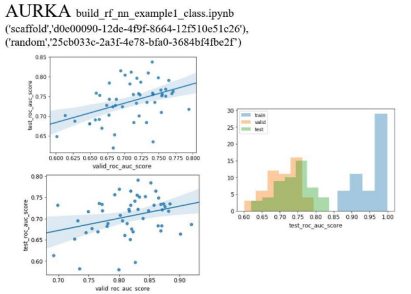
Data Mine Project: Evaluation of ATOM Capabilities

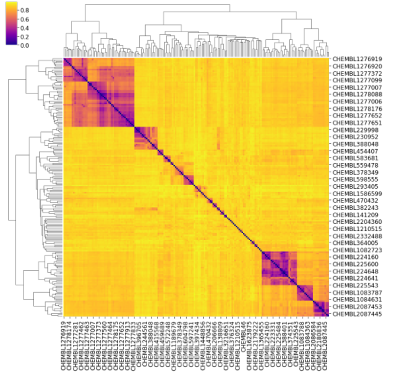
Dr. Jonathan Allen, Abigail Pati, Abigail Yu, Albert Zhang, Ali Hamidi, Anurag Sridhar, Camille Goenawan, Erika Meredith, Jason Qian, Journey Johnson, Kiernan Schuerman, Krystal Diaz, Patrick McCurry, Rosalie Wilfong, Sandokan Shahini, Seena Pourzand, Shan Lu, Sota Shishikura, Stephanie Close, Sylvia Liu, Terrence Ducksworth, Veer Pradhan, and Vidhi Singh

ATOM INTRODUCTION

- **ATOM**: Accelerating Therapeutics for Opportunities in Medicine
- Open public-private partnership for accelerating drug design using computation-driven drug design
- **Goals:**
  - Accelerate drug discovery process
  - Improve success rate in translation to patients
  - *Transforming drug discovery from slow, high-failure process into rapid, patient-centric model*



Example of data visualization from AURKA
Left: Example ROC_AUC scores from validation dataset compared to the test dataset for best and worst models.
Right: Histogram of the ROC_AUC test scores for all 3 datasets.



Data Visualization example from Fall 2020.
Heat map of Tanimoto distances

FUTURE GOALS

- Use the previous model training to create proper visualization and analysis tools
- Use created models to have a proper prediction pipeline that can run multiple molecules at one time and score them
- Run models through a virtual library to evaluate the created models against specified criteria
- *Impact:* With all of these tools, drug design and discovery will be significantly faster and cheaper

RESEARCH METHODOLOGY

Fall 2020:

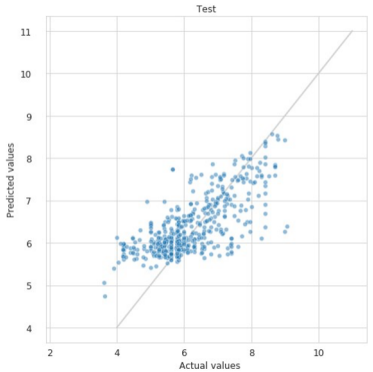
- Safety target background research and data extraction from:
  - Protein Database (PDB)
  - ExcapeDB
  - Drug Target Commons (DTC)
- Created and utilized datasets to generate graphs and compound counts
- Used data visualization to compare compound structure and determine structural diversity

Spring 2021:

- Construct machine learning models for various safety targets
- Generate predictive models to characterize interactions between various compounds with prospective targets
- Train models to be able to predict molecules which could be potential drug target candidate

CONCLUSIONS

- Models were generated for the protein safety targets AURKA, AURKB, HRH1, CHRM2, CHRM3 using neural network and random forest methodologies
- Heat maps of predicted vs actual values on validation and test splits demonstrated an upward trend indicative of model learning
- The best models created both using neural network and random forest showed test R^2 values ranging from 0.5 to 0.6 with training R^2 values topping off near 0.8
  - (Right) An example of data visualization conducted using the models provided.



Data Mine Project: Evaluation of ATOM Capabilities

A Deeper Dive Into Our Research

Dr. Jonathan Allen, Abigail Pati, Abigail Yu, Albert Zhang, Ali Hamidi, Anurag Sridhar, Camille Goenawan, Erika Meredith, Jason Qian, Journey Johnson, Kiernan Schuerman, Krystal Diaz, Patrick McCurry, Rosalie Wilfong, Sandokan Shahini, Seena Pourzand, Shan Lu, Sota Shishikura, Stephanie Close, Sylvia Liu, Terrence Ducksworth, Veer Pradhan, and Vidhi Singh

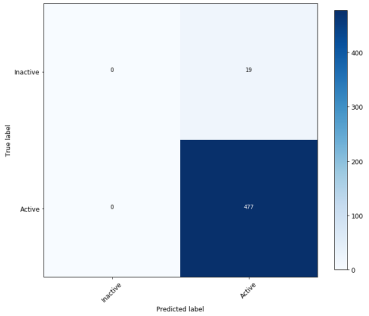
RESEARCH FOCUS

Aurora Kinase A and B

- These are molecules that attach phosphate groups to serine and threonine amino acids.
  - Play a regulatory role in cell division
- AURKA and B have a very structurally similar binding site.
- Our goal was to find a molecule that predominantly selects AURK A over AURKB, and AURK B over AURKA.

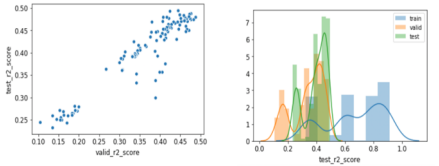
Antihistamines and muscarinic receptors (HRH1, CHRM2, CHRM3)

- Antihistamine drugs often causes undesirable side effects.
- Our goal was to design a drug molecule that is receptive for one histamine receptor and ignore the others that cause undesirable side effects.

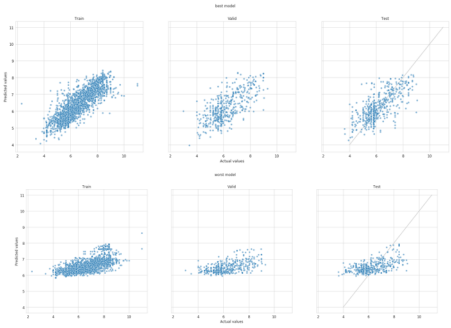


Data visualization example.
Confusion matrix

AURKB build_rf_nn_example.ipynb NN and RF
'scaffold','802f68ba-6b18-42fc-a9bc-e57ed614e784'



Assessment of model performance for AURKB. AURKB has slightly different graphs which allows us to compare between the R^2 scores of the datasets used.



Data visualization example. Distribution of R^2 scores between the 3 datasets (train/validation/test) for the best (top) and worst (bottom) models. Target used: AURKA

Acknowledgement: We would like to thank Dr. Jonathan Allen from Lawrence Livermore National Lab for his mentorship throughout this project and for providing us with the data and notebooks we used to produce our models.

RESEARCH METHODS

split_dataset_example.ipynb

- Utilizing datasets pertaining to the molecule of interest, this notebook was used to split the datasets 2 different ways.
  - Scaffold split: splits the dataset based on 2D structural framework of molecules
  - Random split: randomly splits the dataset
- Within the splits, we are creating 3 different datasets: training, validation, and testing.
- The data visualization output from this notebook is two 2D UMAPs which compares the training and the testing datasets from the scaffold split and the random split.
  - Allows us to visualize the similarities

build_rf_nn_example1.ipynb

- Using one split dataset (random or scaffold) to create and train models
  - Random forest
  - Graph convolutional neural networks
- Specifying hyperparameters for the models
  - rf_max_estimators
  - rf_max_depth
  - rf_max_features
  - Layer_sizes
  - Dropouts
- Received performance tables and graphs comparing R^2 scores, and predictions for the best and worst models separated by train/valid/test

build_rf_nn_example1_class.ipynb

- Using both datasets to generate random forest and neural network models
- Specifying hyperparameters for our models
- Displays a performance table and graphs comparing best and worst models, ROC_AUC scores, a confusion matrix, and a display of the amount of active and inactive molecules.

split_dataset_example_with_binary_classes

- Reads in original dataset (before the splits) and plots active and inactive molecules based on their standard_value, or pIC50 values.

DATA VISUALIZATION AND RESULTS -

Taking a Deeper Dive Into the Performance Table

The build_rf_nn_example1.ipynb and build_rf_nn_example1_class.ipynb both produce performance tables which contain results from the produced models.

The results that we focused on were:

- Model_type: identifies which model the results are from, either random forest or neural network
- Model_uuid: specifies the unique identifier for the specific model
- rf_max_estimators: the maximum numbers of trees in the forest
- rf_max_depth: maximum levels in each decision tree
- rf_max_features: Identifies the maximum number of features Random Forest models are allowed to try in each individual tree. Generally, increasing the features improves the performance of the model at each node.
- valid_roc_auc_score: classification score for models using the validation dataset. Measure of model performance at distinguishing between classes.
- valid_r2_score: regression score for models using the validation dataset; a statistical measure of how close the data is to the fitted regression line.

Data Visualization

- Confusion matrix
  - A table that is used to describe the performance of a classification model. Identifies the predicted active and inactive models and the actual active and inactive models
- Area Under the Curve (AUC) and Receiver Operating Characteristics Curve (ROC)
  - Performance measurement for classification models that numerates how capable models are at distinguishing between classes. The higher the number, the better it performs.
  - Can be used to generate confusion matrices.
- Active and Inactive molecules
  - Various plots and graphs that identify the number of active and inactive molecules.