Exploring drug targets and their effect on drug approval

Sandro Bruno & Sharadwata Ganguli
SIADS 591 & 592

Notebooks link
Intermediate Data link

Table of Contents

https://unsplash.com/photos/2fyeLhUeYpg

Safety features improved the accuracy of drug approval classifier models.

Executive Summary

Drug Development

Requires considerable time & investment

Late developmental stage failures common

Only efficacy is prioritized in preclinical research hunting for new targets

This project will focus on impact of efficacy as well as safety of drug targets on drug approval

13.4% of all genes under investigation are predicted to make it on the market

Motivation

Drug development is a ***resource*** and ***time*** intensive venture requiring ***decades*** and ***billions*** of dollars in investment. Most drug failures happen at late stage – often due to adverse safety findings or lack of efficacy.

This project consists of ***Phase1*** and ***Phase2***. In Phase1, ***relationship*** between ***efficacy*** and ***safety parameters*** with ***drug approval*** was explored. In Phase 2, ***prediction models*** were evaluated to ***predict drug approval*** based on efficacy and safety data.

Data Source

5 different data sources (1 efficacy, 1 ground truth label and 3 safety) were used – having ***different formats*** (JSON, txt, tsv, gmt), ***sizes*** (0.17 MB to 21GB) ***and access methods*** (wget, direct download)***.*** The data collected covered 96% of all currently known human protein coding genes.

Data Manipulation

We have used ***google colab*** with ***google Drive*** as ***file system***. The individual datasets were both ***quantitatively*** as well as ***visually explored*** to identify the need for any cleaning/manipulation/new feature generation. Datasets were ***relatively clean***. Post manipulation to generate new features, ***intermediate data files*** (not the initial raw data) were saved for ***further processing***.

Data Analysis & Visualization

The ***intermediate files*** from the data manipulation stage were ***joined*** using the gene ***'SYMBOL'*** key. The relationship between the efficacy, safety and ground truth labels were explored using ***bar plots*** – ***distinct differential relationship*** were found between the efficacy and safety features and the ground truth labels.

Prediction of Drug Approval (Classification)

At Phase 2, we found target class data to be ***imbalanced***, 4 different approaches were applied to model the imbalance dataset (up/down-sampling, cost-sensitive training, adapt scoring metric).

Various machine learning models were compared, ***GradientBoosting*** were the best performing classifier showing generalizability on the validation set ***(0.89)***. ***Safety features*** found to ***improve*** the classification ***AUROC***. The top 3 performing classifiers were explored using model agnostic methods such as SHAP. These uncovered the ***inverted directionality*** of 2 out of 3 hypothesized ***safety features***.***13.4%*** of all genes under investigation are ***predicted to make it on the market***

Potential Next Steps

1) Explore if we are facing a data leakage problem
2) Expand our model on disease-target pairs
3) Identify and explore additional features,

Can inclusion of additional features (like safety) improve drug development success?

Motivation

Background

The ***development of a new drug***, from target (a gene or a protein) identification to FDA (Food and Drug Administration) approval, takes over ***10 years*** and may exceed the cost of ***$1.7 billion***[1]. Most of the costs are spent into ***failures***, which often occur in ***late-stage clinical development*** with a considerable waste of time and money[1].
Reasons for drug failures are mainly ***safety findings*** or ***lack of efficacy*** and its rational link to the disease they were intended for.

Following the growth of publicly available data, computational platforms linking potential drug targets to diseases have risen. However, most of them ***focus only on efficacy*** and not on ***other aspects like safety***.

Project details

This Project involves exploring the relationship between drug targets (genes/proteins which a drug would affect) and their impact on subsequent drug approval using the parameters of ‘efficacy’ and ‘safety’. This constitutes ***Phase1*** of the Project.
In ***Phase2***, various prediction models of drug approval using these drug target parameters will be explored.

Key questions to answer

- What is the ***relationship*** between ***efficacy and safety*** measures of a drug target and ***drug approval***?
- Explore ***prediction models of drug approval*** given efficacy and safety profile of a drug’s targets.

[1] Paul, S., Mytelka, D., Dunwiddie, C. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9, 203–214 (2010). https://doi.org/10.1038/nrd3078

5 datasets utilized - of varying size covering efficacy, safety and ground truth labels.

Data Sources

Dataset title	Description	Key variables	Size	Shape	File Format	Access Process
Efficacy – association	Contains data about the ***rational link*** to the disease .	SYMBOL, affected_pathway, animal_model, genetic_association, rna_expression, somatic_mutation, druggability_smallmolecule, druggability_other_modalities, druggability_antibody	20.7 GB	7.4M X 84	JSON	wget
Safety – protein network	Contains 11,759,454 ***interconnections*** between 19,354 unique genes.	protein1, protein2, combined score	516 MB	11.8M X 3	txt	wget
Safety – baseline	Contains ***RNA expression values*** for 110 different organs.	ensembl gene, 110 different organs	15 MB	43,663 X 111	tsv	wget
Safety – Onco genes	Contains 189 gene sets with 30,546 genes that were all up or ***downregulated*** in a cancer context.	entrezgene	0.17 MB	189 X 1	gmt	Download from website (free registration required)
Ground truth	Contains information about the ***status of known current or past explored targets*** (e.g., patented, discontinued, etc.)	TARGETID, TARGNAME, TARGTYPE, UNIPROID	0.5 MB	3,473 X 4	txt	wget

The title of each dataset is linked to the corresponding URL to download the file.

6 step data workflow used for data manipulation.

Data Manipulation (Phase 1)

Basic Approach

The ***workflow*** below was used to ***guide dataset manipulation***. Libraries used were pandas, numpy, matplotlib and seaborn. API called 'mygene' was used to convert the gene keys to gene symbols, to be used as ***JOIN keys***.

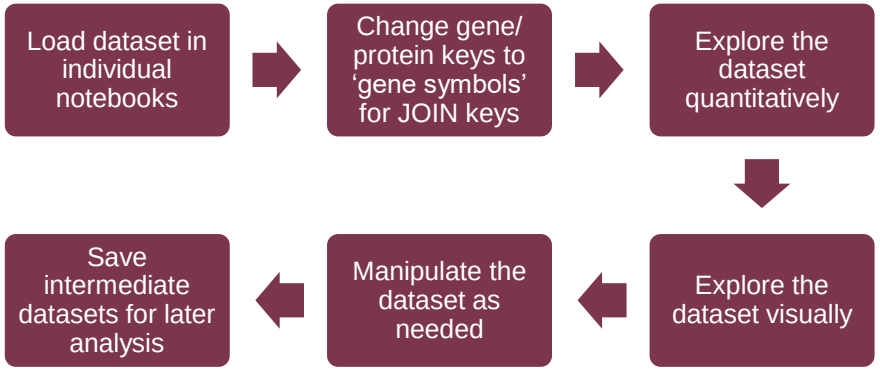


Figure 1: Data Manipulation workflow

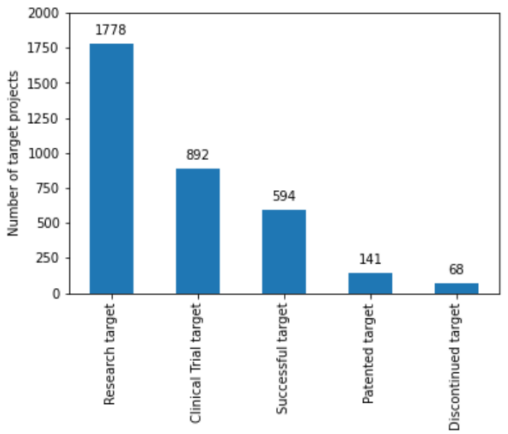
Efficacy – association file *(notebook link)*

This file contains the ***rational link*** of the gene target to the disease and was accessed using ***wget***. Due to the ***sheer size*** of the file, we had to load it as a ***Spark dataframe***. Via exploration, the dataset was found to have ***2M direct and 5.4M indirect associations*** between targets and disease. Redundant indirect associations were removed. To make the ***nested data*** usable, it was ***flattened***. There were ***no missing values*** as such. Gene 'SYMBOL' ID were used as JOIN keys.

Average association scores were computed for each target gene across disease. Additionally, ***tractability/druggability score*** for small molecules, antibodies and other modalities were retrieved. This was followed by combining the association and druggability scores into 1 file and then saving it as an ***intermediate dataset*** for further processing.

Ground truth label file *(notebook link)*

The file contains information about the known current or past explored targets – **Pre-/*Clinical*-/*Discontinued targets etc.***, and was accessed using ***wget***.



The file did not contain a traditional gene ID that could have been used as converter or direct key. Therefore, we derived the gene "SYMBOL" keys by several manipulation steps (showcased in the notebook). We found that only half of the entries were related to targets on the market (Clinical + successful targets). Finally. the datafile was saved as an ***intermediate dataset*** for the data integration step.

Figure 2: Targets counts across ground truth labels *(LNK)*

Section reference to refer to plots in the notebook provided as (LNK)

6 Tissue Specificity scores were evaluated – Gini score chosen.

Data Manipulation (Phase 1)

Safety – baseline file *(notebook link)*

This file contains ***RNA expression values for 110 different organs*** on which we determine ***tissue specificity scores*** - where 1 is a low specificity gene (high risk) and 0 is a gene of high specificity (low risk). The file was accessed using ***wget***. The file was quantitatively explored – ***no missing values*** were found. 'ensembl:gene' IDs were converted to gene 'SYMBOL' for JOIN keys. Data was ***left skewed*** – which we ***normalized*** by ***log transformation***.

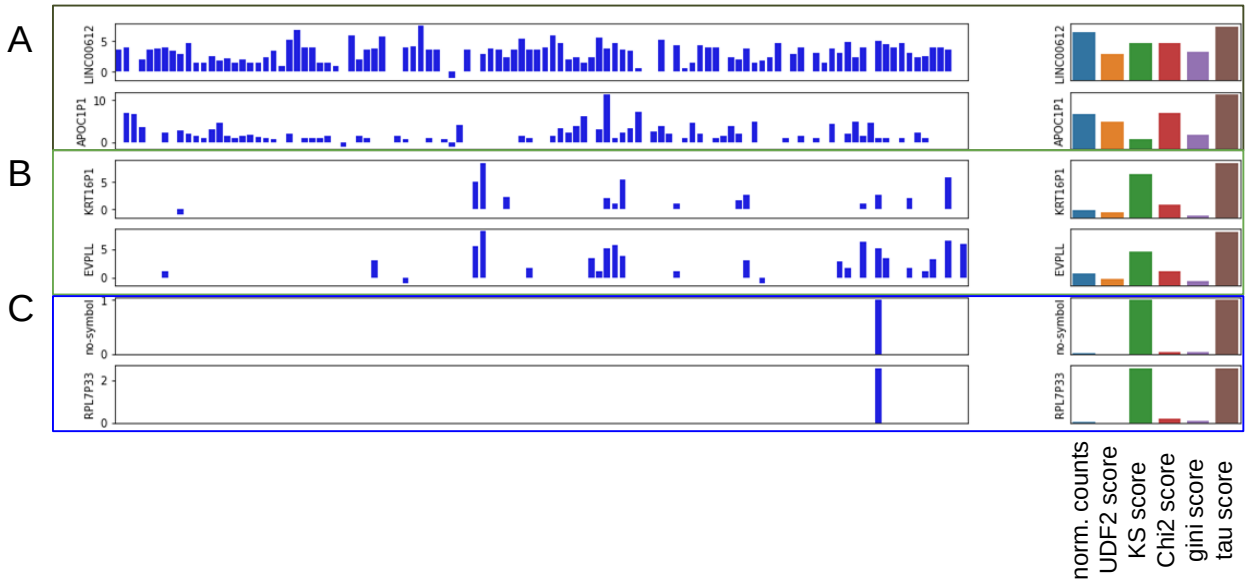


Figure 3: Random gene expression examples showcasing A) non- B) moderate- and C) tissue-specificity patterns. The barplot on the right indicates the scoring representation of the gene distribution on the left (*LNK*)

Finally, we developed ***6 Tissue Specificity scores*** – and found ***Gini score*** to match well with the distribution of the RNA expressions (refer Fig 3 below). The data frame containing the ***gene symbol*** and the ***Gini score*** were saved as an ***intermediate file***.

Safety – protein network file *(notebook link)*

This file contains 12M interconnections between 19K unique genes, accessed using ***wget***. To begin, 'ensembl:protein' IDs were converted to gene 'SYMBOL' for JOIN keys. Post this, ***degree centrality score*** was computed***,*** where 1 is a ***hub gene considered to be of high risk and*** 0 is a ***gene which has no interactions*** thus considered low risk. The ***gene symbol*** and ***degree centrality score*** were then saved as an ***intermediate file***

Safety – onco gene file *(notebook link)*

This file contains 189 gene sets with 30K genes that were all ***upregulated*** or ***downregulated in a cancer context***. The data is publicly available, requiring a login. We created an ***Onco data dictionary*** containing gene sets and associated genes. We then selected only the ***downregulated*** gene sets as they indicate a ***high risk*** for cancer. This halved the data dictionary. Subsequently the 'entrezID' keys were converted to gene symbols for JOIN KEYS. Finally, the data dictionary was converted to a dataframe with ***gene symbol*** and a ***Onco_score=1 column*** and saved as an ***intermediate file***.

Section reference to refer to plots in the notebook provided as (LNK)

Distinct relationship present between efficacy, safety features and ground truth labels.

Data Analysis & Visualization (Phase 1)

File combination and data manipulation *(notebook link)*

The 5 intermediate files created during the data manipulation stage were ***JOINED using pandas MERGE***. ***INNER JOIN*** was used for ***similarly sized data sets*** like efficacy - association, safety - baseline, safety - protein network. ***LEFT JOIN*** was used for joining the ***smaller data sets*** safety-onco gene and Ground truth with larger data sets. We have ***removed the 'druggability association scores'*** as there is ***chance of leakage of target status*** from these features. A TARGETSTATUS variable was generated which combines the ground truth labels into 'no-target', 'unknown-target', and 'target' categories. To ***explore the relationship between efficacy, safety features (numeric) and TARGETSTATUS (categorical)***, we converted the numeric features into categories of 'low' (less than mean of feature) and 'high'(mean of feature or higher).

Visualization of relationship between features and 'target' *(notebook link)*

When ***exploring the relationship*** between the ***ground truth labels*** (using the TARGETSTATUS variable) and the various ***efficacy and safety features*** using bar plots (shown alongside), we observe that the ***ground truth labels*** show a ***distinct relationship*** with the ***features*** – the ***labels are differentially distributed across a feature.*** Hence, these features may have effectiveness in predicting the ground truth labels.

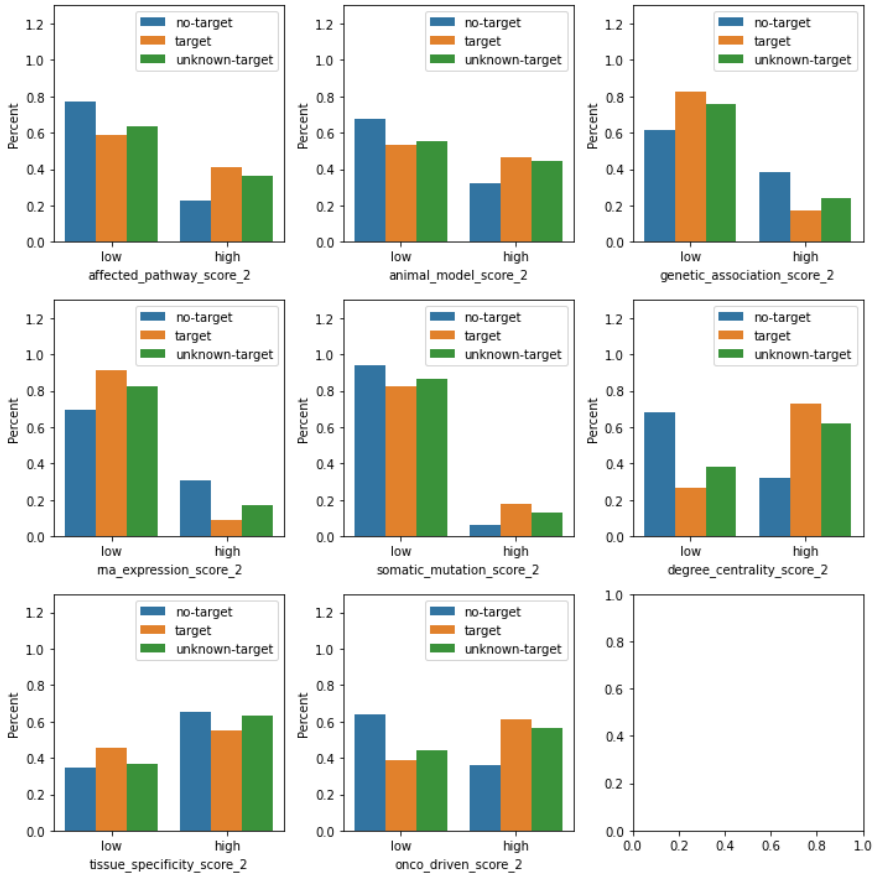


Figure 4: ground truth labels across features (*LNK*)

This comprehensive relationship across features and labels could not be found using any individual dataset alone.

Section reference to refer to plots in the notebook provided as (LNK)

Addition of safety features improved classifier prediction accuracy.

Prediction of Drug Approval (Classification) (Phase 2)

Building & Evaluating Classification models *(notebook link)*

Libraries used at Phase2 were sklearn, umap, hdbscan, seaborn and matplotlib while the Joined data from Phase1 was utilized.
As the classes were ***imbalanced*** (mostly 'no-target') – 4 different approaches were applied to model the imbalance dataset (up/down-sampling, cost-sensitive training, adapt scoring metric).

Next, various ***supervised classification models*** were compared on the AUROC metric (Fig.5). We found GradientBoosting as best performing classifier with great **generalizability** on the validation set (**0.89**).

Additionally, we found that the 3 best performing classifiers (GradientBoosting, RandomForest, LogisticRegression) above have **higher AUROC** for the combined feature set of **'safety + efficacy' vs only using 'efficacy' features** (refer to Fig 6 for roc curves for different Gradient Boosted Tree models). Thus, ***adding safety features*** *in addition to efficacy indeed* ***helped in improving*** *classifier prediction of drug approval–in line with our motivation*.

However, we are still trying to figure out if we are facing a data leaking problem because we did not expect such strong predictive abilities.

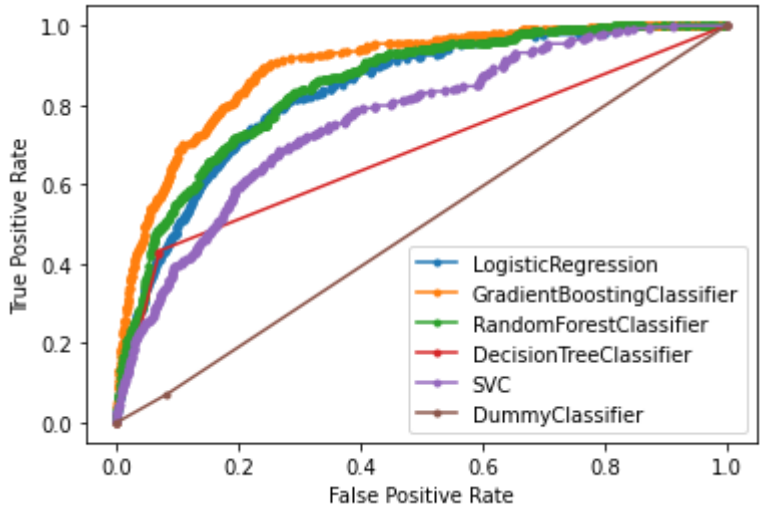


Figure 5: roc_curve for different classifiers (*LNK*)

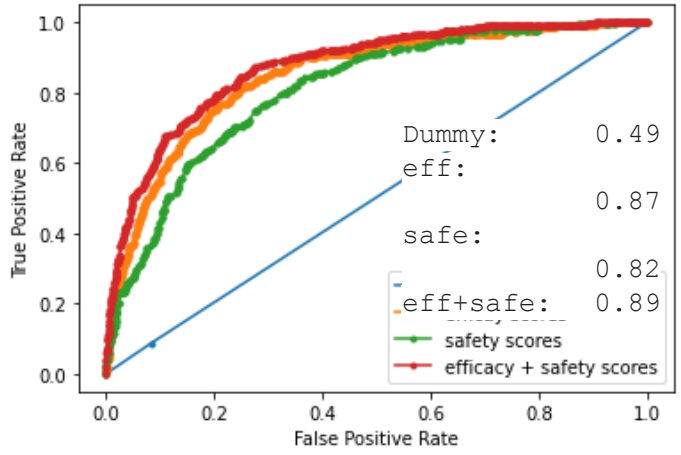


Figure 6: roc_curve for different Gradient Boosted Tree models (*LNK*)

Section reference to refer to plots in the notebook provided as (LNK)

Currently 191 genes are under investigation that are predicted to make it on the market.

Prediction of Drug Approval (Classification) (Phase 2)

Classification findings *(notebook link)*

We have extensively **explored the explainability** of the top 3 performing classifiers using **model-agnostic** methods such as SHAP, MDI and MDA. We could find that for all the three classifiers the **centrality scores** were always under the **top 2 most important features**. Interestingly, our hypothesis for the **centrality** and **onco driven score** as potential safety flag were disproved on all the 3 classifiers because these features showcased an **inverted predictability** (indicated by over represented dots on the right Fig.7). In addition, we would have **expected** that the **genetic associations** would have shown **higher predictability** because in the literature its highly postulated that a genetic association is doubling the chance of a target to make it on the market (Nelson et al 2015).

Prediction on unknown gene targets currently in preclinical stage

We have held back 1,429 potential gene targets which are currently in the preclinical stage with an 'unknown' label. Our best performing classifier suggested that **13.4%** of these unknown gene targets will eventually **make it on the market**. That is incredibly close to the recent observations of Wo et al 2019 who found a success rate overall diseases by around 13.8%. So this number seems to be realistically represented by our classifier.

Top5 of novel targets

SYMBOL	Likelihood
SOX2	0.97
CDH1	0.96
MYC	0.93
MMP9	0.92
CDKN1A	0.92

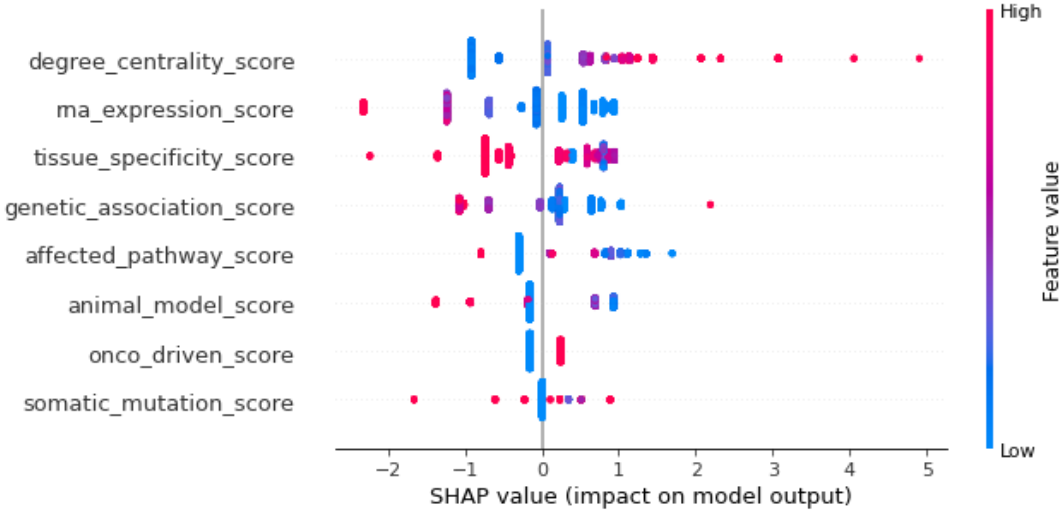


Figure 7: Feature Importance using SHAP for GradientBoosting Classifier (*LNK*)

Section reference to refer to plots in the notebook provided as (LNK)

Conclusion & Next Steps

Conclusion

Phase1:

Distinct and differential relationship present amongst the efficacy, safety features and ground truth labels. ***This relationship could not be found using any individual dataset alone.***

Phase2:

We found ***adding safety features increases the accuracy of classifiers*** – thus our ***motivation*** of adding safety to efficacy features while predicting drug approval was ***validated***. However, it seemed that 2 out of 3 self defined safety flags were showing rather target engagement instead of our hypothesized safety penalties.

Potential next steps :

1) Explore if we are facing a data leakage problem
2) Expand our model on disease-target pairs
3) Identify and explore additional features, which will enable identification of stronger relationships and thus further improve the Prediction Model

There is much work to do in improving efficiency of the drug development process, but it is a worthy goal as the societal benefits of efficient drug development are immense.

Statement of Work

Sandro Bruno

Sandro provided domain expertise in the field of drug development. He reviewed and rewrote the Project Report, co- contributed on the analysis and visualization plan. Additionally, he developed the scripts to manipulate Efficacy- association, ground truth labels, safety – protein network and the final integration of the files. He also worked on the predictive Phase 2 of the study and implemented the gini & tau scores.

Sharadwata Ganguli

Sharad focused on drafting and finalizing the Proposal and the Project Report. He co- contributed on the analysis and visualization plan. He also developed the scripts for manipulating the Safety-baseline, Safety-onco files as well as exploring the relationships between the final features. Additionally, he reviewed the integration scripts and predictive Phase 2 script .