Predicting Drug Sensitivity in Lung Adenocarcinoma Using Machine Learning

Jordan Dautelle, Hang Tran, David Yoon, Alli Warren

Arizona State University

LSC 585 Capstone II in Biological Data Science

Dr. Ken Sweat

May 2, 2025

I. Introduction

Lung cancer is the most lethal cancer in the world, with about 1.8 million deaths annually (World Health Organization, 2023). Non-small cell lung cancer (NSCLC) represents about 85% of all lung cancers, and Lung Adenocarcinoma (LUAD) is the most common. Despite the novel targeted therapies and immune checkpoint inhibitors, patients with LUAD tend to respond differently to drugs. This also involves developing resistance over time, resulting in a poor prognosis and poor success in treatment.

Traditional pharmacogenomic investigations have largely examined the contribution of somatic mutations to drug resistance and sensitivity. This approach has been successful to associate key cancer drivers with corresponding therapies—e.g., EGFR mutations with response to tyrosine kinase inhibitors (Engelman & Jänne, 2005)—but only accounts for a fraction of the molecular processes that govern drug effect. More recent research has demonstrated that gene expression profiling provides complementary information to mutation data, an even more dynamic picture of drug response and heterogeneity of tumor biology (Barretina et al., 2012; Geeleher et al., 2014).

Combining several types of data, including gene expression and somatic mutation profiles, allows us to learn and predict how patients will respond to drugs. By using large genomic datasets and complex machine learning models, researchers can identify intricate relationships between molecular features and their drug sensitivity. This strategy aligns with the goals of precision oncology, in which treatments are more tailored to the unique composition of each tumor instead of using the same approach for everyone.

In this study, we examine if it's possible to predict the sensitivity of LUAD cancer cell lines to drugs based on machine learning models leveraging genetic data. We consider two categories of models: regression models, which predict the natural log of half-maximal inhibitory concentration (IC50) values which is a standard value to determine drug sensitivity, and classification models, which categorize samples as sensitive or resistant. We then examine the impact of incorporating somatic mutation data into the models and ascertain whether inclusion increases predictive performance or interpretability.

We also expect to identify the most informative features that lead to increased model performance. These include gene-level expression signatures, drug-specific characteristics, and pathway-level annotations. The outcomes of this study have the potential to improve therapeutic classification and guide more precise treatment strategies for LUAD patients.

II. Methods

2.1 Dataset Sources

To study how LUAD cell lines respond to drugs, we used publicly available datasets from the Genomics of Drug Sensitivity in Cancer (GDSC) project. We downloaded these datasets through their bulk download portal at https://www.cancerrxgene.org/downloads/bulk_download. GDSC is a very comprehensive resource for studying how cancer cells respond to drugs, combining drug response data with the molecular features of human cancer cell lines. The data was created by exposing a panel of about 1,000 human cancer cell lines to hundreds of anti-cancer drugs and measuring their responses by high-throughput screening assays (Iorio et al., 2016).

Our project relied on five datasets:

- *GDSC2_fitted_dose_response_27Oct23*: This file contains drug sensitivity measures for a number of cancer cell lines. We were interested in the LN_IC50 values, the natural logarithm of the concentration to inhibit cell growth by 50%. It is a measure of how much of a drug is needed to reduce cell survival by half and is our target variable for both regression and classification models.
- *Cell_Lines_Details*: This information comprises extensive metadata for each cell line, including tissue type (e.g., LUAD), growth conditions, medium, and other phenotypic features. It was used to filter out LUAD-specific samples and add biological context for each cell line.
- *Cell_line_RMA_proc_basalExp*: A matrix of RMA-normalized gene expression values for thousands of genes across cell lines. There is one row for each gene and one column for each COSMIC ID (a unique code for each cell line). This dataset supplied the baseline gene expression data necessary for model training.
- *mutations_all_20230202*: This dataset catalogs binary mutation statuses for known cancer driver genes. Although these mutations are not linked via COSMIC ID, we used additional mapping strategies to integrate them with expression data and cell line metadata using model ID.
- *screened_compounds_rel_8.5*: Information on all the compounds screened by GDSC experiments, including drug names and target pathways. We used this dataset to add drug information to our model in the form of categories.

We chose these datasets because they are complete, compatible, and relevant to our research goals. Together, they allowed us to combine drug response data with genomic characteristics (such as expression and mutation data) and pharmacological characteristics to build predictive machine learning models for LUAD drug sensitivity.

2.2 Data Merging Process

Python version 3 was used and the first task was to create a combined dataset to filter for Lung Adenocarcinoma (LUAD) cell lines. From the *Cell_Lines_Details* dataset, samples were selected where the tissue type label matched LUAD. This would guarantee further analysis and model training would be on an important subtype of non-small cell lung cancer.

The gene expression dataset (*Cell_line_RMA_proc_basalExp*) has RMA-normalized values and is in wide format with COSMIC IDs as column headers. To match the other datasets, the matrix was transposed such that each row was a unique LUAD cell line, identified by COSMIC ID. This made it easier to merge with the drug response dataset (*GDSC2_fitted_dose_response_27Oct23*) using COSMIC ID as the key.

The drug-level data from the *screened_compounds_rel_8*.5 file were merged based on the DRUG_ID field. This included information on the drugs, such as their names, targets and target pathways.

The mutations dataset (*mutations_all_20230202*) required additional preprocessing due to its different indexing structure. Only those mutations that were labeled as cancer drivers were kept. The data was then transformed into a simple format, where all LUAD cell lines were given a value of 1 when a mutation took place in a gene, or 0 if it didn't. Since the mutation dataset was indexed by model_id, a final mapping step was done to link these IDs to the rest of the dataset via Sanger Model ID.

The combined dataset contained drug response values (LN_IC50), gene expression profiles, drug information, and binary mutation indicators for LUAD cell lines. This provided a complete multi-omic resource for machine learning analysis.

2.3 Preprocessing and Feature Engineering

Prior to model training, there were several preprocessing steps undertaken to ensure consistency and performance optimization throughout the merged dataset. Although gene expression and mutation datasets were relatively complete, we ensured to remove any feature with a high proportion of missing values to uphold data quality. Numerical features with high rates of missing values would be removed from modeling to avoid unreliable imputations. Fortunately, no features had more than 50% of missing values.

Categorical features, such as cell growth characteristics and tissue properties, along with drug-related properties, were transformed through one-hot encoding to be compatible with the XGBoost algorithm. One-hot encoding is a method that converts each category into its own binary column, where a value of 1 indicates the presence of that category and 0 indicates

absence. This transformation introduced a few hundred additional binary columns, particularly from the drug compound metadata.

To standardize the numerical data, we scaled all the continuous features with StandardScaler from the scikit-learn library (Pedregosa et al., 2011). StandardScaler transforms each feature so that it has a mean of 0 and a standard deviation of 1. This puts the features on the same scale and avoids features with larger ranges from dominating the analysis. Although XGBoost does not require feature scaling, we applied standardization when conducting Principal Component Analysis (PCA) to ensure appropriate dimensionality reduction. Also, to reduce computing costs, we converted all floating-point data into the float32 format. This conversion saved significant memory without sacrificing model accuracy.

In selected pipelines, we applied PCA for reducing the dimensionality of the dataset. As a benchmark, we applied PCA to retain the leading 100 principal components, which captured the majority of the variance in the original features. This was particularly helpful in determining if feature space reduction improved runtime and performance, especially when dealing with high-dimensional genomic data without reducing predictive accuracy.

III. Exploratory Data Analysis

3.1 Histogram of Natural Logarithm of IC50 Values

To explore the distribution of drug response values across LUAD cell lines, we first examined the natural logarithm of IC50 (LN_IC50), representing the concentration of a drug that reduces cell viability by 50%. The log transformation is a common preprocessing step to normalize IC50 values and reduce skewness caused by extreme outliers in pharmacogenomic data.

Figure 1

The histogram of log-transformed IC50 values distribution

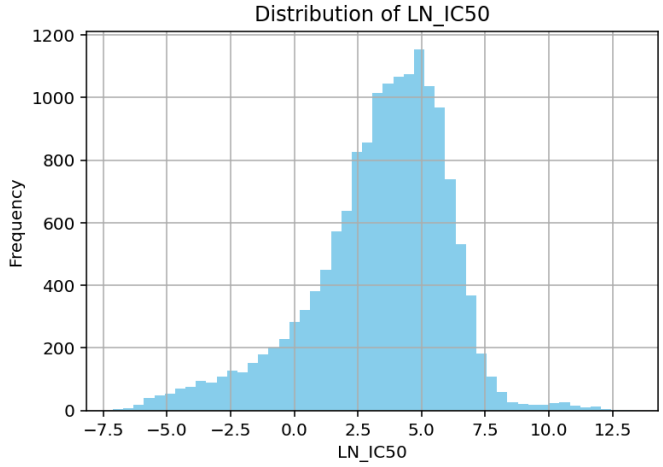


Figure 2

Q-Q plot of log-transformed IC50 values to test for normality

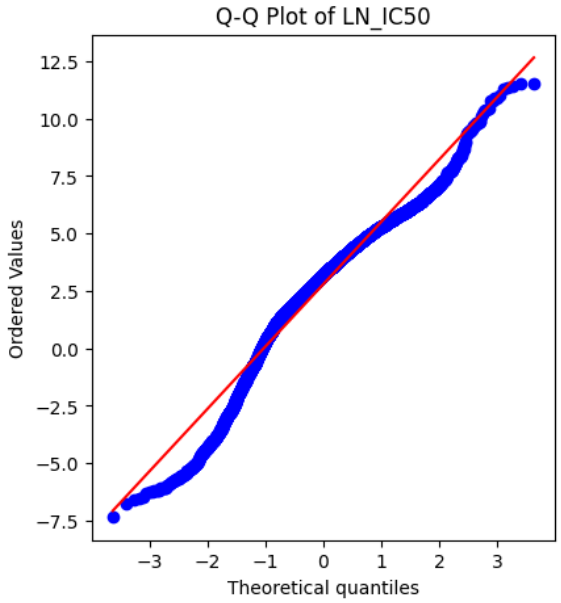


Figure 1 shows that the LN_IC50 values follow a single-peaked distribution that is slightly left-skewed and centered around 4. This shows that most drug–cell line pairs have a moderate degree of sensitivity, with fewer cases of very high sensitivity (very low LN_IC50) or very high resistance (very high LN_IC50). The Q-Q plot also displays deviations from perfect normality, especially in the tails. The Shapiro-Wilk test gave a statistic of 0.9554, which although near 1, signifies a slight deviation from normality in the data. This distribution contains several outlier values, which may indicate biological variation and are investigated further when considering mutational effects on drug response captured by IC50 values.

The findings thereby justify the use of log transformed IC50 values to stabilize variance and reduce skewness. In addition, this supports our use of tree-based machine learning models such as XGBoost that are resilient to non-normally distributed data and do not assume linearity or normally distributed inputs.

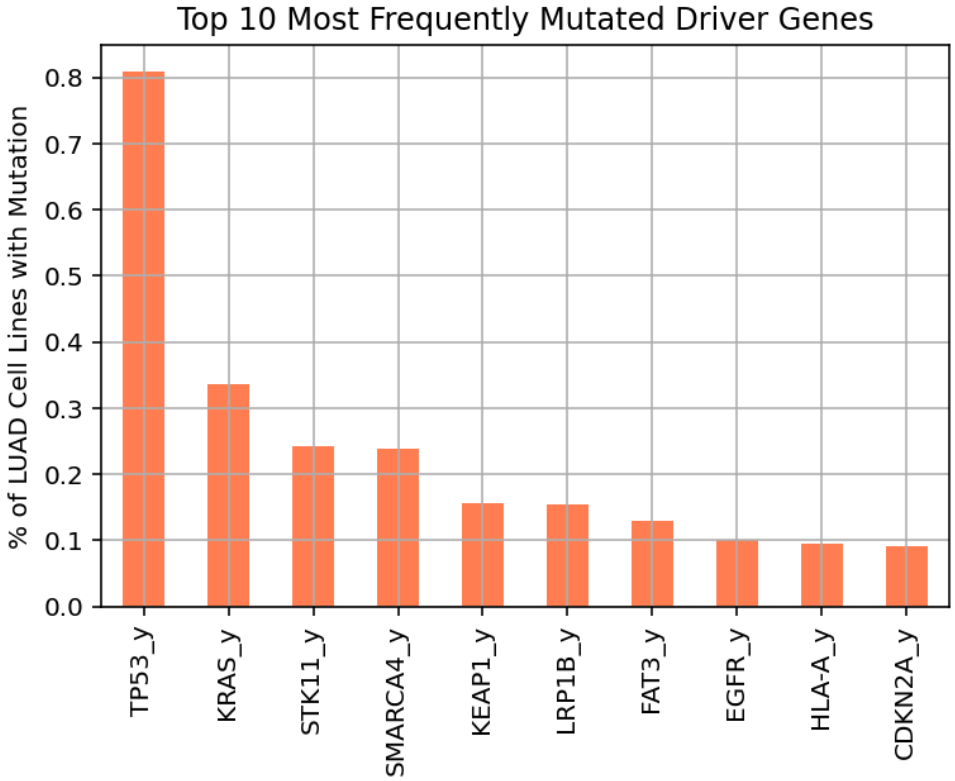
This distribution provides a biological rationale for splitting the continuous LN_IC50 values for our classification model. We set a threshold of LN_IC50 < 0 to define the "sensitive" class and LN_IC50 ≥ 0 to define the "resistant" class. Although conservative in this cutoff, capturing only the most drug-sensitive responses ensures high confidence in classification performance, and is consistent with previous machine learning studies on drug response (Menden et al., 2013).

This binarization also assists in reducing class imbalance more in a controlled manner, since it diminishes noise from borderline or unclear cases. In general, this exploratory visualization of

the natural log of IC50 values informed the downstream modeling strategy and the choice of conducting both regression and classification analyses.

3.2 Frequencies of Driver Gene Mutations

Figure 3
Frequencies of Driver Gene Mutations in LUAD Cell Lines



To explore the mutation pattern in LUAD cell lines, we analyzed the frequency of different somatic mutations in our dataset. The bar graph above displays the top 10 most commonly mutated driver genes in LUAD samples. The most common one was TP53, which was mutated in about 81% of cell lines. This is consistent with the existing literature, since mutations in TP53 are a feature of many cancer types and are associated with genomic instability and poor prognosis in LUAD (Li et al., 2023).

KRAS mutations were the second most frequent, present in approximately 34% of LUAD samples. These mutations have been implicated in drug resistance and aggressive tumor biology, particularly in smokers (Jones et al., 2021). Similarly, STK11 and KEAP1 mutations—present in 15–25% of our samples each—are implicated in affecting the therapeutic response and metabolic

regulation. STK11 is linked to immune evasion and weakens the effectiveness of immune checkpoint inhibitors (Kwack et al., 2020). KEAP1 helps the body respond to oxidative stress, causing resistance to drugs by changing metabolism (Yu & Xiao, 2021). The reason the mutation frequencies in the bar plot add up to more than 100% is because a single LUAD cell line can have mutations in multiple genes.

The other frequent mutations identified were SMARCA4, LRP1B, FAT3, EGFR, HLA-A, and CDKN2A. These were also identified in previous research on LUAD genes. Notably, EGFR, though not so frequent in our data, is significant nonetheless since it helps in predicting patient response to targeted therapy, particularly in non-smoking LUAD patients (Pao et al., 2004). These mutation frequencies provide biological context for downstream machine learning analysis and support the use of mutation features into models of drug sensitivity prediction

3.3 Correlation Between Mutations and LN_IC50

To investigate the relationship between somatic mutations and drug sensitivity in LUAD, we conducted point-biserial correlation analysis between mutation status (mutated or wild-type) and natural log of IC50 values. This type of test is appropriate for checking relationships between a binary variable and a continuous variable.

Figure 4
The correlation between mutations and drug sensitivity

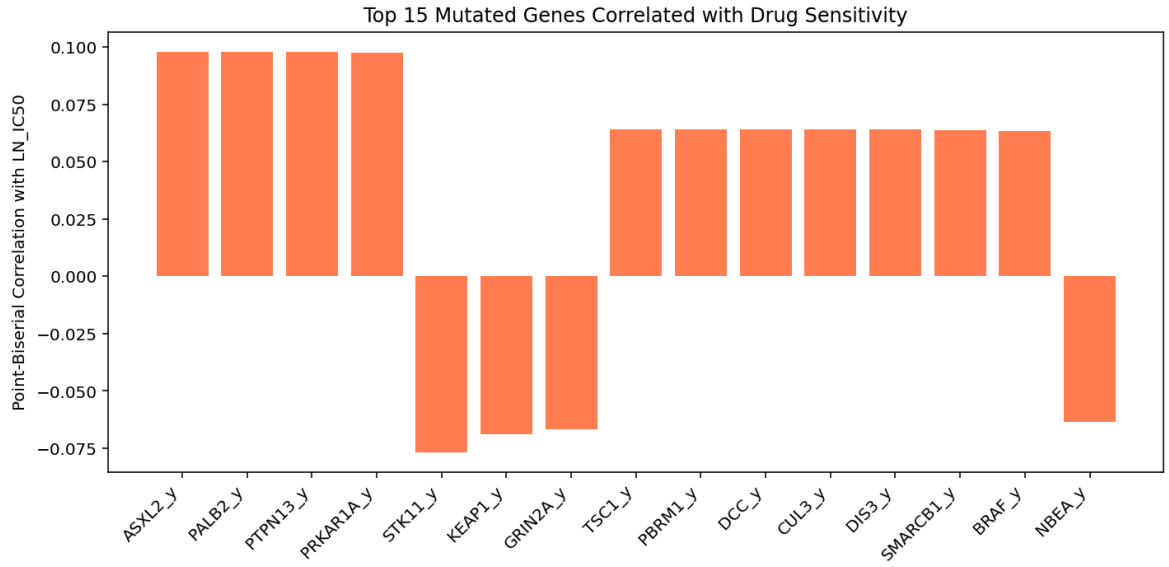


Figure 4 shows the 15 most correlated genes with LN_IC50. The positive correlation coefficients show that the variations of these genes are correlated with higher values of LN_IC50, meaning that there may be drug resistance. Some good examples are ASXL2, PALB2, PTPN13, and

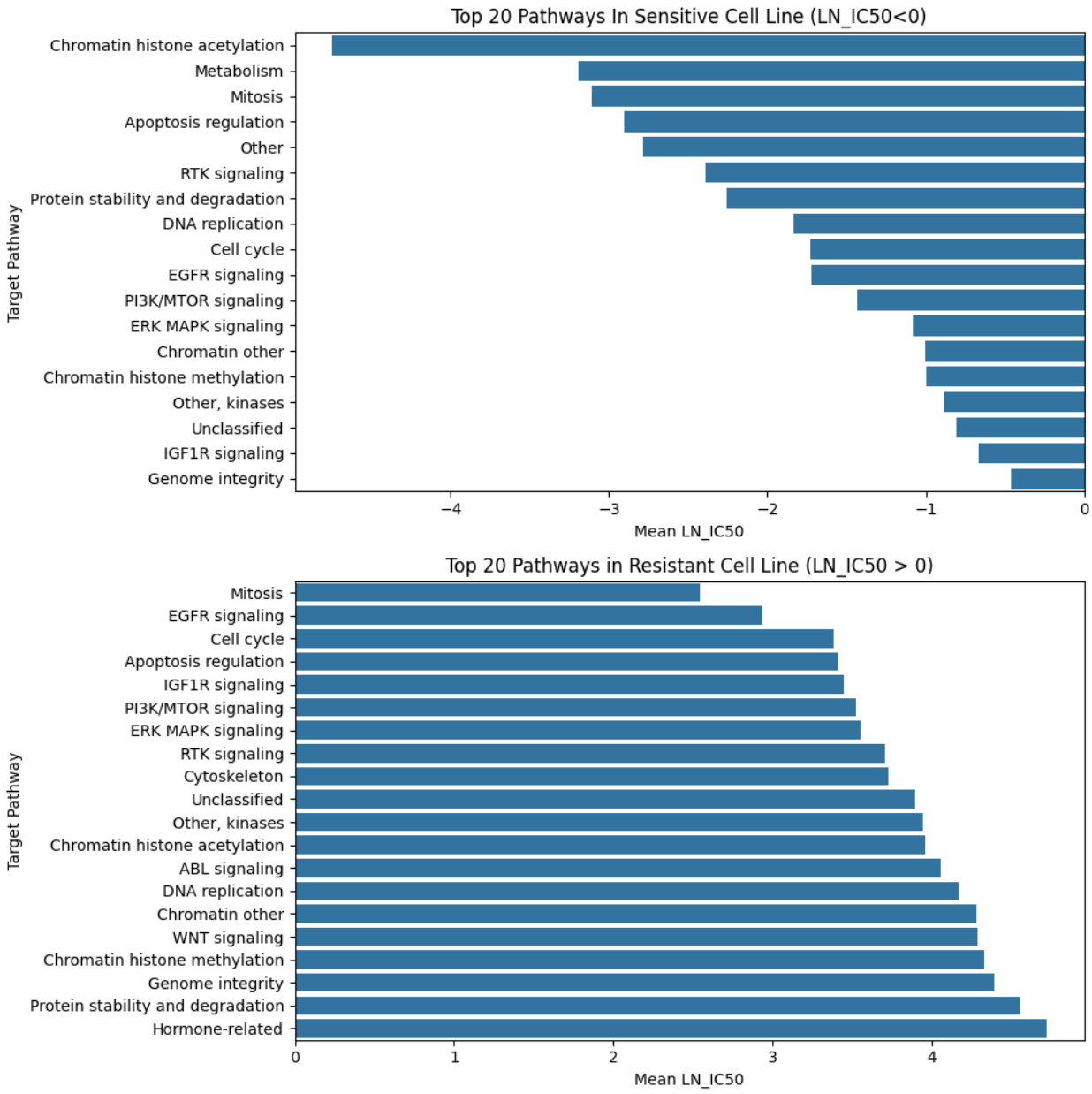
PRKAR1A, which all had correlation coefficients close to 0.10. These genes may be implicated in processes that render cells less susceptible to treatments.

In contrast, the negative correlation coefficients show that mutated cell lines are drug-sensitive. The most strongly negatively correlated genes are STK11, KEAP1, and GRIN2A, whereas the gene STK11 has the strongest negative correlation (~ -0.07). The results support earlier findings that suggest STK11 mutations make cells more sensitive in terms of metabolism, which may lead to better response to some treatments. This analysis identifies candidate biomarkers whose mutational status would guide treatment stratification or targeted therapy decision-making in LUAD.

3.4 Pathway-Level Enrichment

In order to investigate biological pathways related to drug sensitivity in LUAD, we computed the mean natural log of IC50 values for each drug pathway. We divided the findings into resistant cell lines (natural log of IC50 > 0) and sensitive cell lines (natural log of IC50 < 0). We displayed and compared the top 20 most enriched pathways in both sets.

Figure 5
The top 20 pathways in sensitive cell lines and top 20 pathways in resistant cell lines



In the sensitive group, the highest activity pathways were chromatin histone acetylation, metabolism, and mitosis, which had low average natural log of IC50 values. The implication here is that LUAD cell lines with higher activity in modifying chromatin structure and gene regulation are more sensitive to targeted therapy. Chromatin histone acetylation, specifically, allows for the activation of transcription and potentially enhances drug entry or sensitivity by making DNA more accessible (Steele et al., 2009).

In the resistant group, the pathways associated with the highest average LN_IC50 values (so the least responsive to drug treatment) were protein stability and degradation, hormone-related

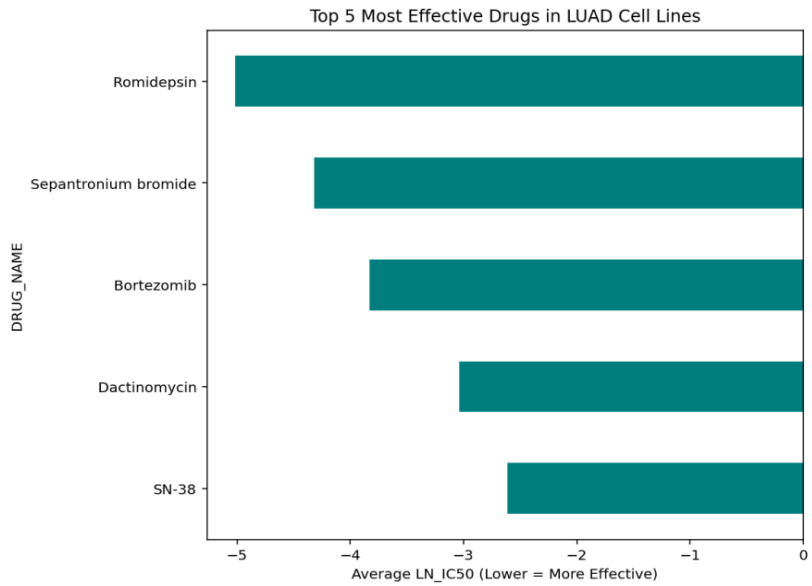
pathways, and genome integrity mechanisms. Such pathways can reflect ways that cancer cells survive drug stress, such as by upregulating protein turnover or genome repair.

Pathways such as mitosis, EGFR signaling, and cell cycle regulation are present in the resistant group as well but with lower LN_IC50 values. That means moderate sensitivity is present even in resistant cases. These findings suggest that drug resistance in LUAD is not straightforward. Some cancer-related pathways, including EGFR and PI3K/MTOR, could play a part in the drugs' mechanism of action, based on how they relate to other molecular features (Santoni-Rugiu et al., 2019). These results emphasize how diverse LUAD biology is and demonstrate that pathway analysis can help uncover treatment opportunities.

3.5 Drug-Level Efficacy Analysis

To complement our predictive modeling, we conducted an exploratory analysis to identify the most effective compounds in LUAD cell lines based on their average natural log of IC50 values. The lower the natural log of IC50, the more potent the drug because it means a smaller quantity is required to inhibit 50% of cell viability. Out of the 286 drugs screened with the GDSC2 dataset, Romidepsin was the most effective, followed by Sepantronium bromide, Bortezomib, Dactinomycin, and SN-38.

Figure 6
The top 5 most effective drugs in LUAD cell lines



Romidepsin is an FDA-approved histone deacetylase (HDAC) inhibitor (Grant et al., 2010). Romidepsin is used to treat some types of T-cell lymphomas. Though it is not normally used in

lung adenocarcinoma treatments, its low LN_IC50 values in LUAD cell lines indicate it may be useful. Romidepsin works by reactivating silenced tumor suppressor genes and altering transcription through epigenetic modification, mechanisms that are particularly promising given the epigenetic deregulation observed in many lung cancers (Falkenberg & Johnstone, 2014). Histone acetylation pathways were also found to be more active in drug-sensitive samples by our pathway enrichment analysis, and this adds to the evidence of why Romidepsin's action is important in LUAD. While experimental validation is needed, these findings suggest that Romidepsin may be a strong candidate for drug repurposing efforts in lung cancer research.

IV. Machine Learning Modeling

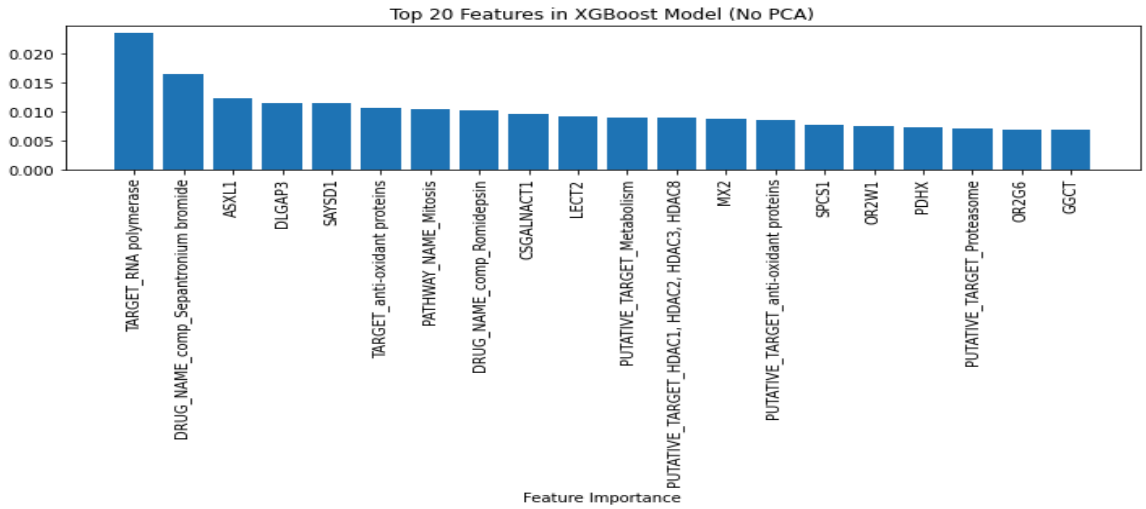
4.1 Regression Models

4.1.1 Regression Model Without PCA and Without Mutation Data

First, we trained an XGBoost regression model on LUAD cell line data without mutation features and without using Principal Component Analysis (PCA) for dimensionality reduction. We used this setup to keep it simple so that the importance of individual genes, pathways, and drug-related features can be interpreted easily using the model's internal feature importance metric.

Figure 7

The top 20 features in XGBoost regression model without mutation data and without PCA



The model identified a number of significant features influencing IC50 values, revealing potential biological drivers for variation in drug response. The most significant feature was the target RNA polymerase, suggesting that RNA polymerase-related targets play a significant role in determining the efficacy of drugs in LUAD cell lines. This is consistent with previous research demonstrating that the transcriptional machinery represents a key vulnerability in rapidly proliferating tumors (Ferreira et al., 2020; Saproo et al., 2023).

The second most important feature was Sepantronium bromide (YM155), a drug that induces cell death and autophagy in various types of cancer (Sasaki et al., 2015; Zhang et al., 2015). This indicates that the unique characteristics of drugs are among the dominant factors in response prediction. Other significant contributors were drug targets involving antioxidants, metabolism, and cell division, such as drug target antioxidant proteins, drug target metabolism process, and drug target mitosis pathway. These findings are relevant because oxidative stress and cell cycle regulation are frequently not functioning correctly in LUAD.

On the gene level, features like ASXL1, DLGAP3, and SAYSD1 were among the top 20 but have yet to be comprehensively studied within the scope of LUAD-specific drug response. The model also picked out drugs like Romidepsin (which we discussed previously), showing that certain compounds explain large parts of the natural log of IC50 value difference between cell lines.

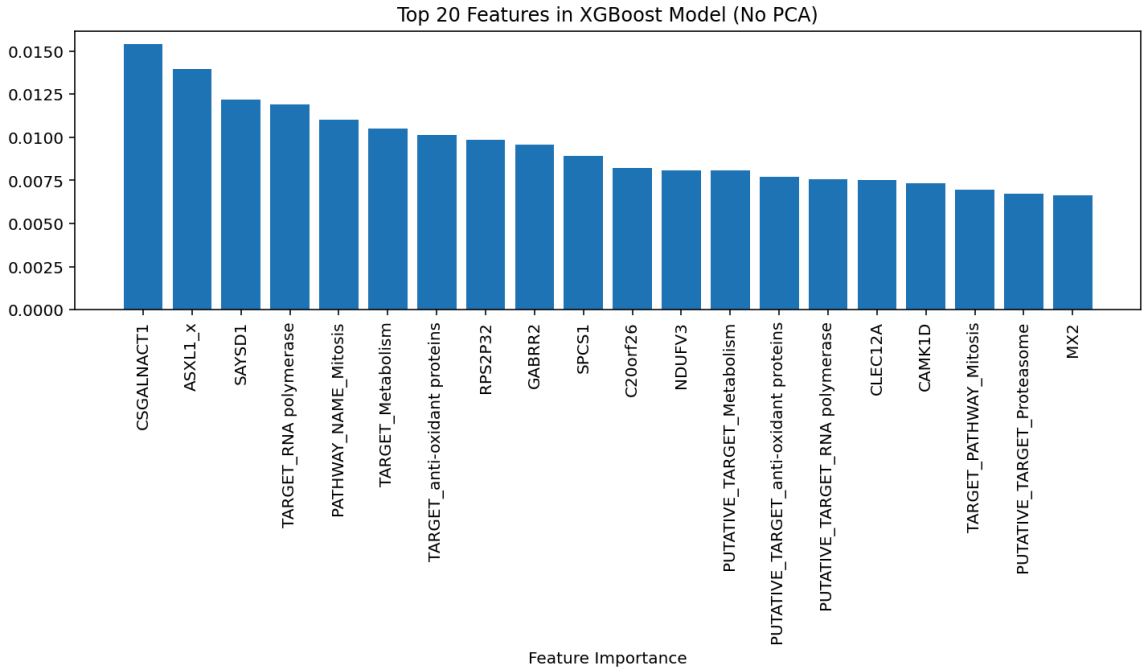
Notably, by not using PCA on this model, we preserved the native feature dimensions. This allowed us to make more precise conclusions about which individual genes or drugs affect sensitivity. This model was a valuable precursor to the addition of mutation data or applying dimensionality reduction for performance optimization.

4.1.2 Regression Model Without PCA and With Mutation Data

To see if adding mutation data improved model interpretability and performance, we retrained our XGBoost regression model using the full merged dataset which included binary indicators of somatic driver mutations. Just like we had done for our earlier model, PCA was not used to reduce dimensionality in order to retain all the single features of gene expression and mutation status.

Figure 8

The top 20 features in XGBoost regression model with mutation data and without PCA



Feature importance was also computed to identify top drug sensitivity predictors as measured by the natural log of IC50 values. Notably, the majority of the top 20 features remained expression-based, indicating the predominant role of gene expression in predictive power. However, one mutation feature emerged among the top important ones: ASXL1, indicating that ASXL1 mutational status contributes importantly to IC50 prediction in LUAD cell lines. For more information, ASXL1 is Additional Sex Combs Like 1 gene that plays a role in cancer survival since it regulates tumorigenesis and cancer progression (Jafarbeik-Iravani et al., 2024). This means that while the genetic expression profile is extremely informative, particular mutations can inform more about drug response mechanisms in certain pathways or subtypes.

Overall, the most important features of the model were still targets that correspond to RNA polymerase, drug-pathway descriptors (e.g., metabolism and mitosis), and antioxidant response elements. The addition of mutation data did not change performance substantially but enhanced the biological interpretability by allowing the effect of mutations to be explicitly measured.

4.1.3 Regression Model With PCA and Without Mutation Data

To evaluate model performance under dimensionality reduction, we trained an XGBoost regressor on a 100-component PCA-transformed version of the gene expression and drug feature matrix excluding mutation data. While PCA removes noise and reduces training time, it also

conceals the original feature identities, which makes biological interpretation difficult. The model performed well with an R^2 of 0.7943, an RMSE of 1.2418, and an MAE of 0.9345. R^2, the coefficient of determination, measures how much of the variance in the target variable is explained by the model. The nearer to 1, the more satisfactory the performance. RMSE (Root Mean Squared Error) predicts the average magnitude of prediction errors and gives heavier weight to big errors. MAE (Mean Absolute Error) describes the average size of the errors without considering outliers.

The results indicate that the majority of the variance in drug sensitivity (LN_IC50) can still be accounted for using a lower-dimensional representation of the data. Presenting multiple metrics allows us to capture both average prediction error as well as how good the model is at explaining variance.

4.1.4 Regression Model With PCA and With Mutation Data

To explore if the inclusion of mutational data could increase predictive power, we trained a final regression model on both gene expression and binary mutation features, and then applied dimensionality reduction using PCA (100 components). While this approach sacrifices interpretability—since single features can no longer be traced back to single biological variables—it is computationally rapid, performs better, and designed to work with high-dimensional data.

The model had an R^2 score of 0.7976, RMSE of 1.2316, and MAE of 0.9321, which is a marginal improvement over the model that did not have mutation data. This suggests that somatic mutations can provide a marginal extra hint when paired with gene expression data. However, since the performance increase is not substantial and it is harder to interpret the features, these models may be better suited for deployment or ranking rather than biological interpretation.

4.1.5 Lasso Regression Trial

We used Lasso regression for IC50 prediction to see if the model performed better. Nevertheless, the model did not perform as well as XGBoost, and its R^2 value was only 0.68. Because our dataset has many features and complex interactions, Lasso's simple assumptions and focus on a limited number of features most likely affected its performance. Therefore, we decided not to use it for final analysis.

This confirms why we chose XGBoost, which is better at capturing nonlinear relationships and interactions among genomic features.

4.2 Classification Models

4.2.1 Classification Model Without PCA and Without Mutation Data

To test how well we can perform binary classification, we trained an XGBoost classifier on all features except mutation data and without PCA for dimensionality reduction. The binary outcome was based on drug sensitivity: cell lines were marked as sensitive if their LN_IC50 was below 0 and resistant otherwise.

Figure 9
ROC curve of the XGBoost classification model without mutation data

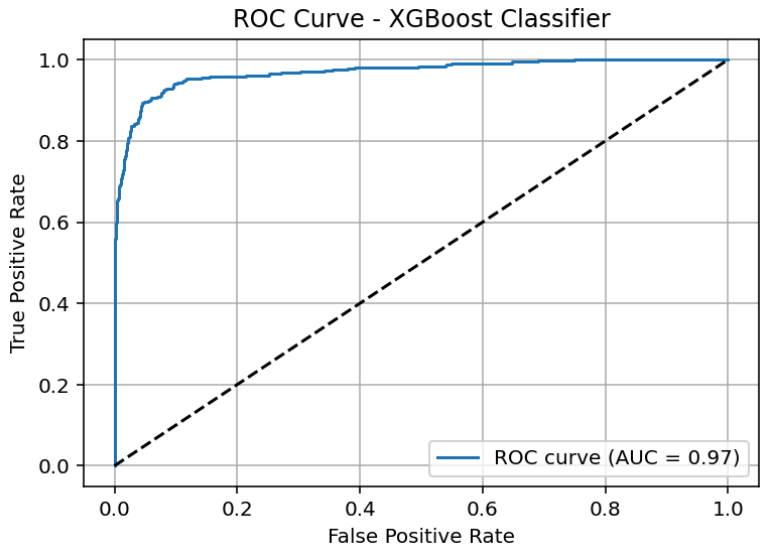
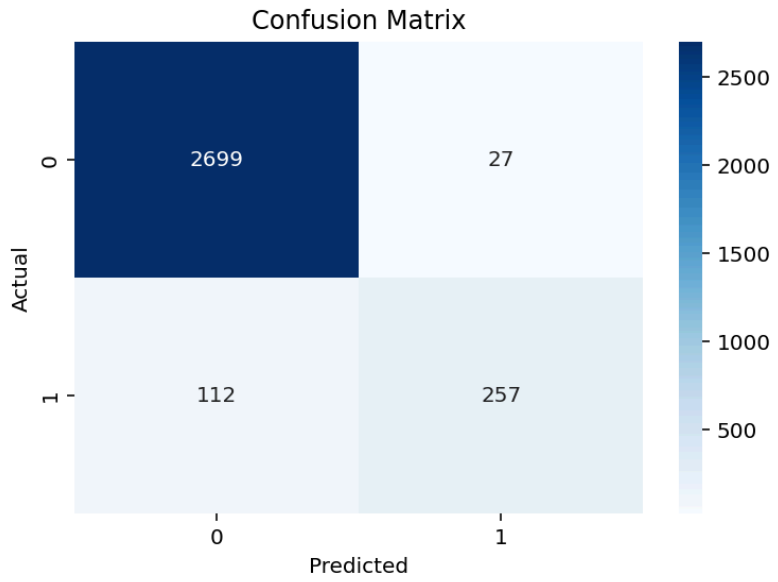


Figure 10
Confusion matrix of the XGBoost classifier without mutation data

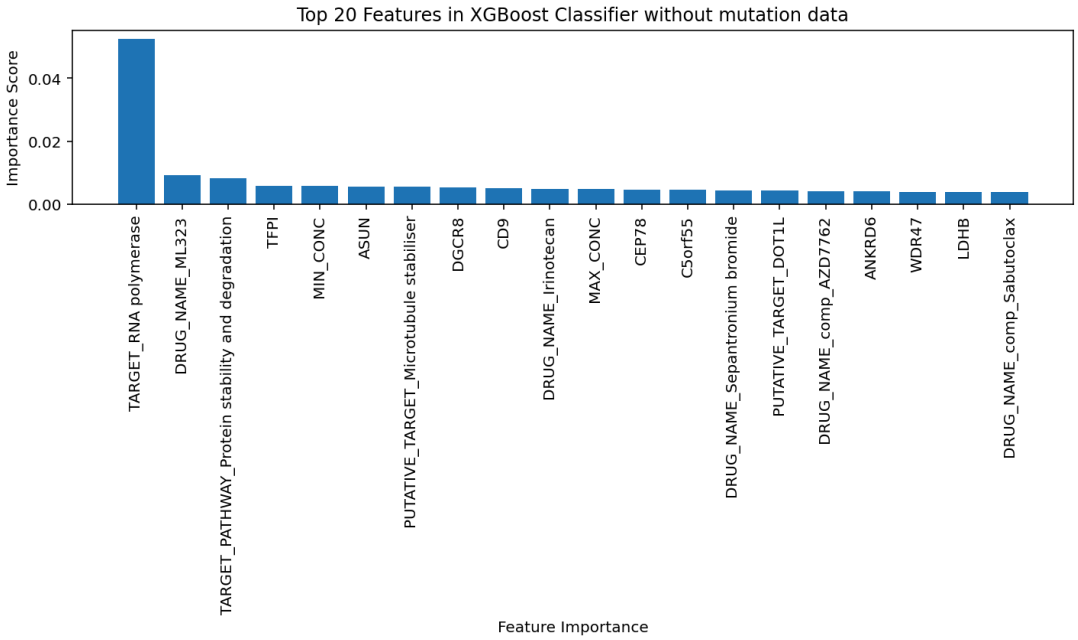


The classifier achieved high performance with an ROC AUC of 0.97, suggesting it has excellent ability to separate the classes. For more information, the ROC AUC is computed by finding the area under the Receiver Operating Characteristic (ROC) curve. The curve is obtained from the model's prediction scores. It gives a single value to evaluate classification performance (Wikipedia, 2025). From this, we were able to create a graphical plot illustrating the performance of our binary classification model.

Overall accuracy was 96%, meaning that 96% of all predictions, whether resistant or sensitive, were correct. Precision for the sensitive class (label = 1) was 0.90, recall was 0.70, and F1-score was 0.79. The majority class (resistant) was very accurately predicted, both in terms of precision and recall (0.96 and 0.99, respectively), as one can see from the confusion matrix. We can observe that the model correctly identified 2699 resistant samples and 257 sensitive samples, while making few mistakes. Indeed, 27 resistant samples were incorrectly predicted as sensitive (false positives) and 112 sensitive samples were missed and predicted as resistant (false negatives).

These results suggest that the model is great at identifying resistant cell lines, but not as good at identifying all sensitive cases. This is probably due to class imbalance.

Figure 11
The top 20 features in XGBoost classification model without mutation data and without PCA



The most important feature in the model was RNA polymerase, and it was much more significant than all the others. This means that LUAD cell lines treated with drugs that target RNA polymerase may be very different in their response. A few other important predictors were certain compounds like ML323 and information about pathways such as Protein stability and degradation. Features related to drug concentration (like minimum and maximum concentration) were also important, showing that the model is sensitive to pharmacological screening conditions.

These results demonstrate that drug information and gene expression—minus the incorporation of mutation data—can accurately classify drug sensitivity in LUAD. Certain molecular targets (e.g., RNA polymerase inhibitors) may drive the most distinct biological responses.

4.2.2 Classification Model Without PCA and With Mutation Data

In this classification model, we have integrated binary mutation indicators along with gene expression and drug information without using PCA. The model performed similarly to the one without mutation data, with a minimal difference in precision and recall: precision for class 1 (sensitive) rose from 0.90 to 0.91, and recall dropped minimally from 0.72 to 0.70. Overall ROC AUC was still very high at 0.9708, showing it's still capable of making distinctions well.

Figure 12
ROC curve of the XGBoost classification model with mutation data

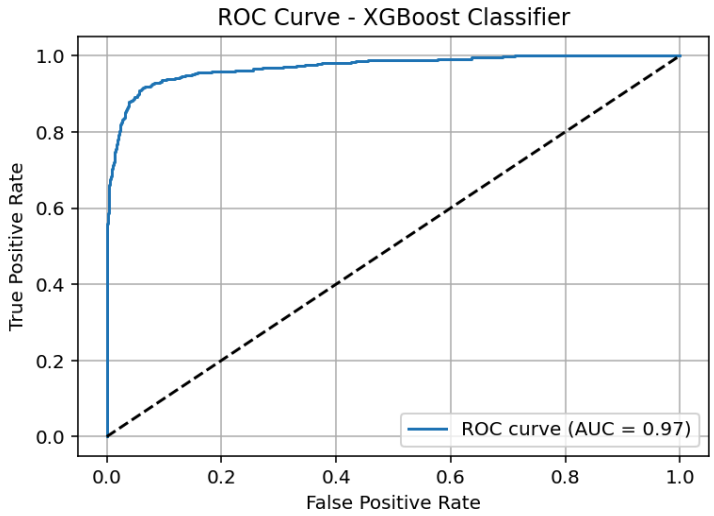
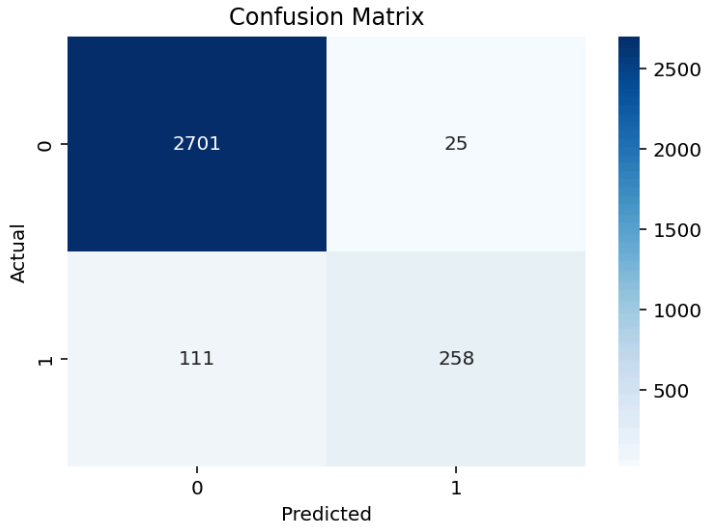
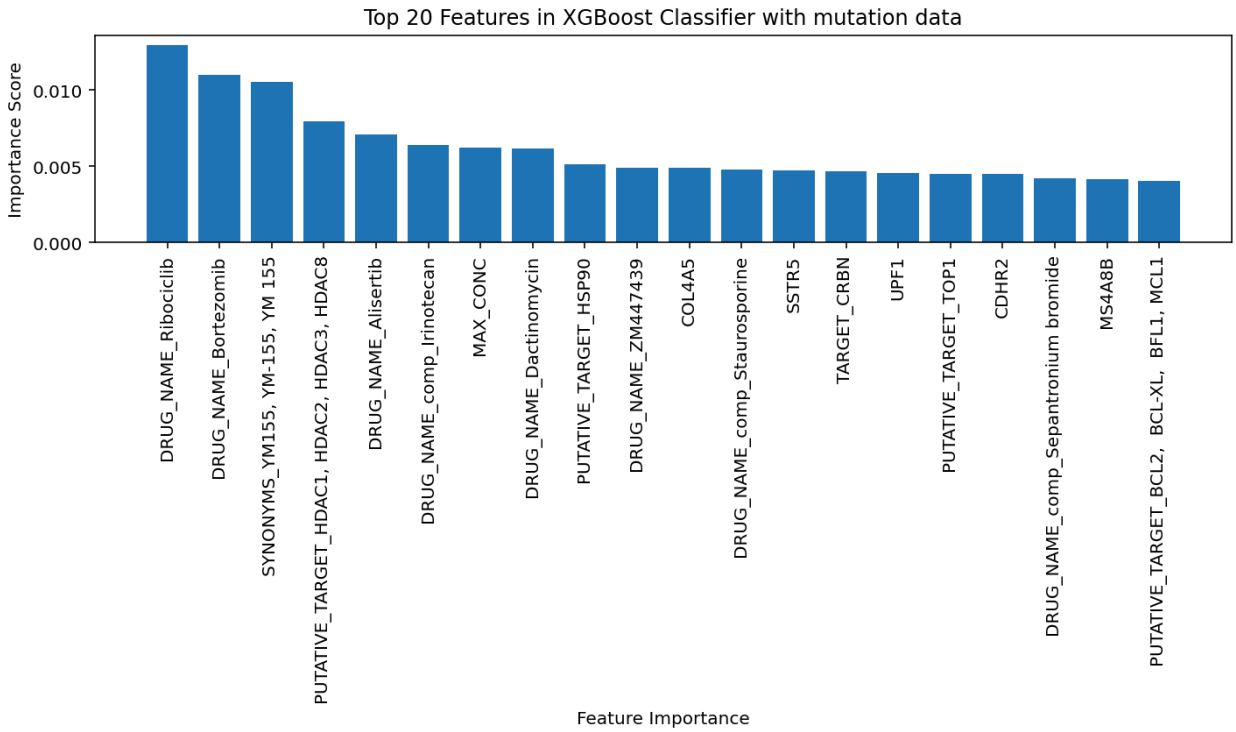


Figure 13
Confusion matrix of the XGBoost classifier with mutation data



The confusion matrix showed a nearly balanced number of false negatives and false positives, with 258 true positives and 111 false negatives.

Figure 14
The top 20 features in XGBoost classification model with mutation data and without PCA



Interestingly, adding mutation data introduced novel predictive features into the top 20 feature importance plot. One of these was a variable representing a group of proteins helping cancer cells survive: BCL-2, BCL-XL, BFL-1, and MCL-1. They are part of the BCL-2 family of proteins that tend to prevent cells from going through apoptosis, the normal process of programmed cell death. In LUAD, the cancer cells tend to overexpress these proteins, which enables them to survive treatments that would typically cause cell death. The strong predictive nature of this characteristic suggests that a cancer cell's sensitivity to a drug may depend on whether the drug can successfully target and turn off these "survival" proteins, making the cancer cells more likely to die if treated (Wesarg et al., 2007).

Similarly, other drug-specific and pathway-level features are drugs such as Ribociclib and Bortezomib, and the HDAC gene family (HDAC1/2/3/8) remained predominant, corroborating their promising predictive capability.

These findings indicate that mutation data may not significantly enhance performance, but it informs the model more by highlighting key mutations and target pathways. Therefore, we want to keep mutation data for analysis, particularly where clinical translation or therapeutic stratification may benefit from genomic interpretability.

4.2.3 Classification Model With PCA (With and Without Mutation Data)

We tried using PCA on the classification model both with and without mutation data to assess performance. However, PCA severely impaired model interpretability and yielded no performance improvement compared to the non-PCA model. In fact, classification accuracy and recall for the sensitive group (class 1) dropped, and the most discriminative genomic and drug-associated features could no longer be retrieved from the transformed components. Since our goal in the classification model was not only to predict but also to interpret drug response patterns in LUAD, we will omit classification models using PCA from the main results and discussion.

V. Results Summary

To assess drug sensitivity prediction in LUAD cell lines, we trained XGBoost regression and classification models. We evaluated models both with and without PCA, and with and without mutation data. The table below is a comparison of some of the key performance metrics.

Table 1

Summary of model performance metrics

Model Type	Mutation Data	PCA	R^2	RMSE	MAE	Accuracy	ROC AUC	Notes
Regression	No	No	–	–	–	–	–	Used for feature importance only; performance metrics not evaluated
Regression	Yes	No	–	–	–	–	–	Mutation feature ASXL1 ranked high
Regression	No	Yes	0.7943	1.2418	0.9345	–	–	Good balance of speed and performance
Regression	Yes	Yes	0.7976	1.2316	0.9321	–	–	Best regression performance
Classification	No	No	–	–	–	0.96	0.9701	RNA Polymerase = top feature
Classification	Yes	No	–	–	–	0.96	0.9708	BCL2/MCL1 genes enter top predictors
Classification	No/Yes (we did both)	Yes	–	–	–	↓	↓	Excluded due to drop in accuracy / interpretability

Lasso Regression	No	No	0.6771	1.5558	1.2373	–	–	Underperformed; not used

Interpretability vs Performance Trade-Offs:

- For regression, the PCA-based model using mutation data ($R^2 = 0.7976$) performed best in terms of predictive power, though non-PCA models yielded interpretable feature importance rankings.
- For classification, models without PCA outperformed their counterparts consistently. ROC AUC was consistently high (~0.97) with or without the presence of mutation features, and the classification accuracy was high (~96%).
- Mutation features contributed minimally to performance but gave additional biological insight. Driver genes such as ASXL1, BCL2, and MCL1 emerged as top features in models with mutation data.
- Models have a tendency to show that drug-related characteristics (like RNA Polymerase, Ribociclib, and HDAC targets) and biological processes (like Mitosis and Antioxidant proteins) are good predictors.

VI. Discussion

Our modeling findings highlight the predictive potential and biological interpretability of machine learning approaches to pharmacogenomics data for LUAD. Feature importance across models consistently identified biologically significant markers and drug-related features.

In the model without mutation data, ASXL1, DLGAP3, and SAYSD1 were the top genetic predictors of cells' drug sensitivity. ASXL1 (Additional Sex Combs Like 1) is a gene that is often mutated in blood cancers but how it works in solid cancers like LUAD is unknown and needs more research to identify its possible role in drug response by cells (Gelsi-Boyer et al., 2012). DLGAP3 (DLG Associated Protein 3) is mostly studied in brain tissues, yet not much is known in lung cancer (Wei et al., 2024). This makes it a potential new marker influencing drug sensitivity in LUAD. Likewise, SAYSD1 (SAYSVFN Domain Containing 1) is not well characterized in existing studies, and appearing as an important feature suggests a potential to detect new biological pathways associated with how patients respond to treatment in lung adenocarcinoma.

When including mutation data in the model, ASXL1, a chromatin modifier gene, reappeared as one of the top features in the regression model. While it had appeared in the expression-only model earlier, its re-appearance after adding mutation data suggests that certain driver mutations might meaningfully interact with gene expression in shaping drug response. The majority of the top features were drug or gene expression related, rather than mutation related. This suggests that gene expression is still the stronger signal in our predictions.

In the classification models, we observed very good performance (ROC AUC = 0.97) both with and without incorporating mutations. Interestingly, when we left out the mutation data, the most highly ranked feature was the target RNA Polymerase (just like with our regression model without mutation data), highlighting the impact of transcriptional control in drug sensitivity. The

other significant features of our classification models were drug-specific characteristics such as the compound Alisertib, a known Aurora kinase inhibitor (Pitts et al., 2016). There were also pathway-level annotations such as Mitosis and Protein stability and degradation. These pathways are known to be dysregulated in proliferative LUAD tumors (Eymin & Gazzeri, 2010). These findings indicate that LUAD sensitivity to drugs is determined by both the genetic composition and how the drug acts.

When the mutation data was added, the classifier's performance metrics did not change much, and none of the mutation-based features were among the top 20 most predictive features. The majority of the important prediction features were derived from gene expression or drug properties. Nevertheless, some mutation markers are still significantly important in understanding how drugs work in LUAD. Specifically, TP53, STK11, and KEAP1 mutations are well known to affect tumor growth, tumor adaptation to the environment, and resistance (Skoulidis et al., 2018). Our prior analysis of mutation frequencies also showed that these genes were among the most frequently mutated in LUAD cell lines. This highlights their clinical and biological relevance, though they were not top-scoring features in the model.

One of the primary key observations from our results is that using PCA for dimension reduction made the model marginally quicker but always harder to interpret and only had a minimal impact in terms of predictive power for regression. For classification, PCA lowered performance dramatically. This demonstrates a common trade-off in biomedical machine learning—dimension reduction may decrease noise but can also obscure important biological variation.

Lastly, the mutation correlation analysis revealed further information on how LUAD cells respond to drugs. The study showed that gene mutations like STK11, KEAP1, and GRIN2A had correlations with lower LN_IC50 values, suggesting increased drug sensitivity in mutated cell lines. Particularly, STK11 had the highest negative correlation (~ -0.07), which indicates its potential role in metabolic vulnerabilities that can be exploited therapeutically. In contrast, genes including ASXL2, PALB2, PTPN13, and PRKAR1A had positive correlations with LN_IC50. This implies that they potentially play a role in drug resistance in LUAD.

In summary, our results validate integrating multi-omic data for predicting drug sensitivity. They demonstrate that gene expression profiles are more important than mutation status alone for predicting treatment response in LUAD. This is useful knowledge for creating individualized treatment strategies in precision oncology.

VII. Limitations and Future Work

One of the main problems with this project is the size of the dataset. The study merely included LUAD-specific cell lines in the Genomics of Drug Sensitivity in Cancer (GDSC) database, even as the larger GDSC dataset includes over 800 cancer cell lines representing a range of cancer types. This limited scope might make our findings more challenging to apply to cancers outside of lung adenocarcinoma. Furthermore, while we did include gene expression and mutation data, we did not make use of other revealing data types, such as proteomics or epigenomics.

Another limitation is in terms of computational demands. The high dimensionality of the dataset, featuring over 17,000 numerical features, required memory optimization and was challenging in terms of hyperparameter tuning and model training. Although mutation features were incorporated successfully, they provided only marginal gains in model performance, suggesting their primary utility may be for biological interpretation rather than predictive ability.

Future directions involve extending the project to include more cancer types from the GDSC dataset or incorporating data from other research platforms. Validating the trained models on external datasets would assist in verifying their reliability. Methodologically, future work can also explore alternative machine learning methodologies apart from XGBoost, such as deep learning models, random forests, or graph neural networks. Additionally, incorporating drug synergy data or pathway-level embeddings might enhance predictive accuracy and biological insight.

VIII. Clinical Implications

The prediction models made in this project hold great potential to advance precision oncology for lung adenocarcinoma (LUAD). By quantitatively predicting the sensitivity of tumors to drugs according to genomic features—e.g., gene expression and mutations—these models could potentially allow clinicians to predict how each LUAD tumor would react to particular therapies, facilitating more precise and more personalized treatment approaches.

The XGBoost regression and classification models performed very well, with a ROC AUC around 0.97 and an R^2 around 0.79. This means that reliable predictions can be made even in high-dimensional feature spaces. The models and analyses showed important drug and gene-related characteristics, including TP53, KEAP1, and RNA polymerase targets, that align with our knowledge of LUAD biology and its clinical significance. These findings support the integration of machine learning into early drug screening, treatment selection, and the identification of new targets for treatment.

While our findings show strong predictive potential in LUAD cell lines, clinical translation will require validation in patient-derived samples and in vivo models, as cell lines may not fully capture the complexity of tumor microenvironments. However, this framework provides a valuable basis for integrating pharmacogenomics into clinical decision-making in lung cancer.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., . . . Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607. https://doi.org/10.1038/nature11003

Brooks, E. A., Galarza, S., Gencoglu, M. F., Cornelison, R. C., Munson, J. M., & Peyton, S. R. (2019). Applicability of drug response metrics for cancer studies using biomaterials. *Philosophical Transactions of the Royal Society B Biological Sciences*, *374*(1779), 20180226. https://doi.org/10.1098/rstb.2018.0226

Chen, H., Carrot-Zhang, J., Zhao, Y., Hu, H., Freeman, S. S., Yu, S., Ha, G., Taylor, A. M., Berger, A. C., Westlake, L., Zheng, Y., Zhang, J., Ramachandran, A., Zheng, Q., Pan, Y., Zheng, D., Zheng, S., Cheng, C., Kuang, M., . . . Meyerson, M. (2019). Genomic and immune profiling of pre-invasive lung adenocarcinoma. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13460-3

Engelman, J. A., & Jänne, P. A. (2005). Factors predicting response to EGFR tyrosine kinase inhibitors. *Seminars in Respiratory and Critical Care Medicine*, *26*(03), 314–322. https://doi.org/10.1055/s-2005-871990

Eymin, B., & Gazzeri, S. (2010). Role of cell cycle regulators in lung carcinogenesis. *Cell Adhesion & Migration*, *4*(1), 114–123. https://doi.org/10.4161/cam.4.1.10977

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Falkenberg, K. J., & Johnstone, R. W. (2014). Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nature Reviews Drug Discovery*, *13*(9), 673–691. https://doi.org/10.1038/nrd4360

Ferreira, R., Schneekloth, J. S., Panov, K. I., Hannan, K. M., & Hannan, R. D. (2020). Targeting the RNA polymerase I transcription for cancer therapy comes of age. *Cells*, *9*(2), 266. https://doi.org/10.3390/cells9020266

Geeleher, P., Cox, N. J., & Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, *15*(3). https://doi.org/10.1186/gb-2014-15-3-r47

Gelsi-Boyer, V., Brecqueville, M., Devillier, R., Murati, A., Mozziconacci, M., & Birnbaum, D. (2012). Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. *Journal of Hematology & Oncology*, *5*(1). https://doi.org/10.1186/1756-8722-5-12

Grant, C., Rahman, F., Piekarz, R., Peer, C., Frye, R., Robey, R. W., Gardner, E. R., Figg, W. D., & Bates, S. E. (2010). Romidepsin: a new therapy for cutaneous T-cell lymphoma and a potential therapy for solid tumors. *Expert Review of Anticancer Therapy*, *10*(7), 997–1008. https://doi.org/10.1586/era.10.88

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., Van Dyk, E., Chang, H., De Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., . . . Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, *166*(3), 740–754. https://doi.org/10.1016/j.cell.2016.06.017

Jafarbeik-Iravani, N., Kolahdozan, S., & Esmaeili, R. (2024). The role of ASXL1 mutations and ASXL1 CircRNAs in cancer. Biomarkers, 29(1), 1–6. https://doi.org/10.1080/1354750X.2024.2304187

Jiang, Y., Chen, M., Xiong, Z., & Qin, Y. (2024). Predicting anti-cancer drug sensitivity through WRE-XGBoost algorithm with weighted feature selection. *Genes & Diseases*, *12*(2), 101275. https://doi.org/10.1016/j.gendis.2024.101275

Jones, G. D., Caso, R., Tan, K. S., Mastrogiacomo, B., Sanchez-Vega, F., Liu, Y., Connolly, J. G., Murciano-Goroff, Y. R., Bott, M. J., Adusumilli, P. S., Molena, D., Rocco, G., Rusch, V. W., Sihag, S., Misale, S., Yaeger, R., Drilon, A., Arbour, K. C., Riely, G. J., . . . Isbell, J. M. (2021). KRAS G12C Mutation Is Associated with Increased Risk of Recurrence in Surgically Resected Lung Adenocarcinoma. *Clinical Cancer Research*, *27*(9), 2604–2612. https://doi.org/10.1158/1078-0432.ccr-20-4772

Kadi, N. E., Wang, L., Davis, A., Korkaya, H., Cooke, A., Vadnala, V., Brown, N. A., Betz, B. L., Cascalho, M., Kalemkerian, G. P., & Hassan, K. A. (2018). The EGFR T790M Mutation Is Acquired through AICDA-Mediated Deamination of 5-Methylcytosine following TKI Treatment in Lung Cancer. *Cancer Research*, *78*(24), 6728–6735. https://doi.org/10.1158/0008-5472.can-17-3370

Knijnenburg, T. A., Klau, G. W., Iorio, F., Garnett, M. J., McDermott, U., Shmulevich, I., & Wessels, L. F. A. (2016). Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep36812

Krall, E. B., Wang, B., Munoz, D. M., Ilic, N., Raghavan, S., Niederst, M. J., Yu, K., Ruddy, D. A., Aguirre, A. J., Kim, J. W., Redig, A. J., Gainor, J. F., Williams, J. A., Asara, J. M., Doench, J. G., Janne, P. A., Shaw, A. T., McDonald, R. E., III, Engelman, J. A., . . . Hahn, W. C. (2017). KEAP1 loss modulates sensitivity to kinase targeted therapy in lung cancer. *eLife*, *6*. https://doi.org/10.7554/elife.18970

Kwack, W. G., Shin, S. Y., & Lee, S. H. (2020). <p>Primary Resistance to Immune Checkpoint Blockade in an STK11/TP53/KRAS-Mutant Lung Adenocarcinoma with

High PD-L1 Expression</p> *OncoTargets and Therapy*, *Volume 13*, 8901–8905. https://doi.org/10.2147/ott.s272013

Li, S., Wang, W., Yu, H., Zhang, S., Bi, W., Sun, S., Hong, B., Fang, Z., & Chen, X. (2023). Characterization of genomic instability-related genes predicts survival and therapeutic response in lung adenocarcinoma. *BMC Cancer*, *23*(1). https://doi.org/10.1186/s12885-023-11580-0

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE*, *8*(4), e61318. https://doi.org/10.1371/journal.pone.0061318

Naidoo, J., Sima, C. S., Rodriguez, K., Busby, N., Nafa, K., Ladanyi, M., Riely, G. J., Kris, M. G., Arcila, M. E., & Yu, H. A. (2015). Epidermal growth factor receptor exon 20 insertions in advanced lung adenocarcinomas: Clinical outcomes and response to erlotinib. *Cancer*, *121*(18), 3212–3220. https://doi.org/10.1002/cncr.29493

Natu, A., Verma, T., Khade, B., Thorat, R., Gera, P., Dhara, S., & Gupta, S. (2024). Histone acetylation: a key determinant of acquired cisplatin resistance in cancer. *Clinical Epigenetics*, *16*(1). https://doi.org/10.1186/s13148-023-01615-5

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., Mardis, E., Kupfer, D., Wilson, R., Kris, M., & Varmus, H. (2004). EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences*, *101*(36), 13306–13311. https://doi.org/10.1073/pnas.0405220101

Pitts, T. M., Bradshaw-Pierce, E. L., Bagby, S. M., Hyatt, S. L., Selby, H. M., Spreafico, A., Tentler, J. J., McPhillips, K., Klauck, P. J., Capasso, A., Diamond, J. R., Davis, S. L., Tan, A. C., Arcaroli, J. J., Purkey, A., Messersmith, W. A., Ecsedy, J. A., & Eckhardt, S. G. (2016). Antitumor activity of the aurora a selective kinase inhibitor, alisertib, against preclinical models of colorectal cancer. *Oncotarget*, *7*(31), 50290–50301. https://doi.org/10.18632/oncotarget.10366

Santoni-Rugiu, E., Melchior, L. C., Urbanska, E. M., Jakobsen, J. N., De Stricker, K., Grauslund, M., & Sørensen, J. B. (2019). Intrinsic Resistance to EGFR-Tyrosine Kinase Inhibitors in EGFR-Mutant Non-Small Cell Lung Cancer: Differences and Similarities with Acquired Resistance. *Cancers*, *11*(7), 923. https://doi.org/10.3390/cancers11070923

Sasaki, R., Ito, S., Asahi, M., & Ishida, Y. (2015). YM155 suppresses cell proliferation and induces cell death in human adult T-cell leukemia/lymphoma cells. *Leukemia Research*, *39*(12), 1473–1479. https://doi.org/10.1016/j.leukres.2015.10.012

Skoulidis, F., Goldberg, M. E., Greenawalt, D. M., Hellmann, M. D., Awad, M. M., Gainor, J. F., Schrock, A. B., Hartmaier, R. J., Trabucco, S. E., Gay, L., Ali, S. M., Elvin, J. A., Singal, G., Ross, J. S., Fabrizio, D., Szabo, P. M., Chang, H., Sasson, A., Srinivasan, S., . . . Heymach, J. V. (2018). STK11/LKB1 mutations and PD-1 inhibitor resistance in

KRAS-Mutant lung adenocarcinoma. *Cancer Discovery*, *8*(7), 822–835. https://doi.org/10.1158/2159-8290.cd-18-0099

Steele, N., Finn, P., Brown, R., & Plumb, J. A. (2009). Combined inhibition of DNA methylation and histone acetylation enhances gene re-expression and drug sensitivity in vivo. *British Journal of Cancer*, *100*(5), 758–763. https://doi.org/10.1038/sj.bjc.6604932

Wei, J., Li, Y., Jiao, F., Wang, X., Zhou, H., Qiao, Y., Yuan, Z., Qian, C., Tian, Y., & Fang, Y. (2024). DLGAP3 suppresses malignant behaviors of glioma cells via inhibiting RGS12-mediated MAPK/ERK signaling. *Brain Research*, 149334. https://doi.org/10.1016/j.brainres.2024.149334

Wesarg, E., Hoffarth, S., Wiewrodt, R., Kröll, M., Biesterfeld, S., Huber, C., & Schuler, M. (2007). Targeting BCL‐2 family proteins to overcome drug resistance in non‐small cell lung cancer. *International Journal of Cancer*, *121*(11), 2387–2394. https://doi.org/10.1002/ijc.22977

What is XGBoost? (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/xgboost/

Wikipedia contributors. (2025, April 10). *Receiver operating characteristic*. Wikipedia. https://en.wikipedia.org/wiki/Receiver_operating_characteristic

World Health Organization: WHO & World Health Organization: WHO. (2023, June 26). *Lung cancer*. https://www.who.int/news-room/fact-sheets/detail/lung-cancer

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., & Garnett, M. J. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, *41*(D1), D955–D961. https://doi.org/10.1093/nar/gks1111

Yu, C., & Xiao, J. (2021). The Keap1‐Nrf2 System: A Mediator between Oxidative Stress and Aging. *Oxidative Medicine and Cellular Longevity*, *2021*(1). https://doi.org/10.1155/2021/6635460